UNIVERSITY OF NEW SOUTH WALES

Thesis/Project Report Sheet

Surname or Family name: BARLOW

First name: MICHAEL Other name/s: GLYNN

Abbreviation for degree as given in the University calendar: PH.D.

School: COMPUTER SCIENCE Faculty:

Title: PROSODIC ACOUSTIC CORRELATES OF SPEAKER CHARACTERISTICS

Abstract 350 words maximum: (PLEASE TYPE)

This thesis describes an investigation of the encoding of the speaker characteristics identity, sex, and dialect, in the prosodic acoustic parameters energy, fundamental frequency, voicing, and zero crossing rate, of speech. The acoustic parameters are extracted from a database of sentences, repeated by nineteen adult speakers of Australian English.

Speech analysis experiments are described using four different sentences. Discriminant analysis is applied to the examinations of identity and sex, while least-square-fit analysis is applied for dialect. The twenty-one measures of properties for each parameter are logically divided into two groups:- dynamic measures pertaining to the time varying properties of the parameters, and static measures pertaining to the time invariant properties of the parameters.

Results reveal that all three speaker characteristics may be determined significantly above chance based on the parameters extracted. Identity and dialect are shown to be more strongly encoded in the time varying properties of the parameters, while sex is more strongly encoded in the time invariant properties. Measures of the dynamic time warping-path are found to contain significant encodings of speaker characteristic information.

All four parameters are found to have encoded information pertinent to each of the three speaker characteristics and encoding is found to be utterance, speaker characteristic, and speaker dependent.

Perceptual experiments are described using a linear prediction analysis-resynthesis scheme which allows the independent manipulation of energy, fundamental frequency, voicing, and timing. Perception of identity is found to be significantly influenced by the prosodics, and results are both speaker and parameter dependent. Listeners are found to use both time varying and invariant parameter properties in judgements of identity. Perception of sex is found to be primarily a function of mean fundamental frequency with no significant effect for the other parameters. Judgements of dialect are generally consistent across listeners with shortening of utterance shifting perception towards cultivated, and lengthening of utterance shifting perception towards broad dialect. No consistent significant effect on dialect is found for the other parameters. A number of topics for further research are suggested.

Declaration relating to disposition of project report/thesis

I am fully aware of the policy of the University relating to the retention and use of higher degree project reports and theses, namely that the University retains the copies submitted for examination and is free to allow them to be consulted or borrowed. Subject to the provisions of the Copyright Act 1968, the University may issue a project report or thesis in whole or in part, in photostate or microfilm or other copying medium.

I also authorise the publication by University Microfilms of a 350 word abstract in Dissertation Abstracts International (applicable to doctorates only).

Signature	Witness	Date
		2/7/93

The University recognises that there may be exceptional circumstances requiring restrictions on copying or conditions on use. Requests for restriction for a period of up to 2 years must be made in writing to the Registrar. Requests for a longer period of restriction may be considered in exceptional circumstances if accompanied by a letter of support from the Supervisor or Head of School. Such requests must be submitted with the thesis/project report.

FOR OFFICE USE ONLY Date of completion of requirements for Award:

Registrar and Deputy Principal

THIS SHEET IS TO BE GLUED TO THE INSIDE FRONT COVER OF THE THESIS

PROSODIC ACOUSTIC CORRELATES OF SPEAKER CHARACTERISTICS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY COLLEGE

UNIVERSITY OF NEW SOUTH WALES

AUSTRALIAN DEFENCE FORCE ACADEMY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

By

Michael G. Barlow

June 16, 1991

Certificate of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgement is made in the text.

I also hereby declare that this thesis is written in accordance with the University's Policy with respect to the Use of Project Reports and Higher Degree Theses.

Michael Barlow

Preface

This thesis was undertaken in the Department of Computer Science, at University College, The University of New South Wales (ADFA). It was supported by a Commonwealth scholarship.

I would like to thank Michael Wagner for his supervision of this work. His patience, guidance, knowledge of a vast field, and tolerance of a wayward student is most gratefully acknowledged.

The members of the Department of Computer Science, ADFA, have been helpful in ways too numerous to elaborate. In particular I'd like to thank David Slater for providing both dialect classifications of speakers and the IPA font. Ed Lewis and David Hoffman provided invaluable assistance regarding the statistical analysis of results.

I am greatly indebted to the numerous friends and associates who sacrificed their time as speakers or listeners in my experiments.

Finally, and most importantly, to my friends and family who provided the support and encouragement that kept me going. Especially Maria; typist, sempai, companion, confidante, and partner-in-life.

Thank you all.

Abstract

This thesis describes an investigation of the encoding of the speaker characteristics identity, sex, and dialect, in the prosodic acoustic parameters energy, fundamental frequency, voicing, and zero crossing rate, of speech. The acoustic parameters are extracted from a database of sentences, repeated by nineteen adult speakers of Australian English.

Speech analysis experiments are described using four different sentences. Discriminant analysis is applied to the examinations of identity and sex, while least-squares-fit analysis is applied for dialect. The twenty-one measures of properties of each parameter are logically divided into two groups:- dynamic measures pertaining to the time varying properties of the parameters, and static measures pertaining to the time invariant properties of the parameters.

Results reveal that all three speaker characteristics may be determined significantly above chance based on the parameters extracted. Identity and dialect are shown to be more strongly encoded in the time varying properties of the parameters, while sex is more strongly encoded in the time invariant properties. Measures of the dynamic time warping-path are found to contain significant encodings of speaker characteristic information.

All four parameters are found to have encoded information pertinent to each of the three speaker characteristics and encoding is found to be utterance, speaker characteristic, and speaker dependent.

Perceptual experiments are described using a linear prediction analysis-resynthesis scheme which allows the independent manipulation of energy, fundamental frequency, voicing, and timing. Perception of identity is found to be significantly influenced by the prosodics, and results are both speaker and parameter dependent. Listeners are found to use both time varying and invariant parameter properties in judgements of identity. Perception of sex is found to be primarily a function of mean fundamental frequency with no significant effect for the other parameters. Judgements of dialect are generally consistent across listeners with shortening of utterance shifting perception towards cultivated, and lengthening of utterance shifting perception towards broad dialect. No consistent significant effect on dialect is found for the other parameters.

A number of topics for further research are suggested.

NOTE: SEE CORRECTIONS, ERRATA AND NEW CONCLUSION IN BACK POCKET

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Within a spoken utterance there is much information beside the 'textual message'. A large part of this extra information details facts about the speaker. Such data as the speaker's identity, sex, and dialect are implicitly encoded, and other characteristics such as the speaker's state of health and psychological state may also be encoded. *Speaker Characteristics* is the term used to describe the characteristics or qualities of a person which may be encoded in an utterance.

A detailed understanding of the relationship between speaker characteristics and the acoustic parameters of the speech waveform has many applications. Foremost of these are:

- More realistic speech synthesis systems, allowing the production of speech with any desired speaker characteristic combination (e.g., a synthetic voice that captures immediate attention for a warning system).
- Speaker recognition systems (e.g., security access systems).
- More robust speech recognition systems which can account for speaker characteristic variability (e.g., adjustment to compensate for headcold in a speaker).

With these and other objectives in mind, a large body of research has been published in the area of speaker characteristics and their correlation to the acoustic parameters of the speech waveform. The greater part of this research has involved static acoustic features, and has concentrated upon speaker identity, however studies have been made both of dynamic acoustic features and other speaker characteristics. Speaker recognition systems have been used in limited applications for well over a decade with error rates less than one percent ([Dod85]). However, the exact relationship of speaker characteristics to acoustic features is still not fully understood. In particular, this lack of knowledge occurs at a 'sentence level', at which many speaker characteristics appear to manifest themselves most strongly (see Chapter 2).

It is the objective of this work to address some of the inadequacies with regard to the understanding of the manifestation of speaker characteristics at a *prosodic* level. The approach taken is to use a combination of analysis and resynthesis techniques to investigate the contributions of both time varying and time invariant properties of various acoustic parameters to automatically distinguishable, *and* listener perceived, speaker characteristics.

This thesis is organised as follows: *Chapter 2* is a review of the literature related to speaker characteristics. Section 2.1 details perceptually motivated investigations, while section 2.2 describes analytical investigations of speaker characteristics. Section 2.3 details the long running and controversial area of speaker identification by visual inspection of speech spectrograms, and section 2.4 draws several conclusions and implications based on earlier sections.

Chapter 3 provides a bridge between the previous work as detailed in the literature review and the approach as taken in this thesis. Particular elements of the literature review, relevant to the current work, are highlighted, and an outline of the experimental approach is given.

Details of the speech database collected, including spoken material, choice and quantification of speaker characteristics, and recording, are described in *Chapter 4*.

Chapters 5, 6, and 7 pertain to the analytical investigation of the relationship between prosodics and speaker characteristics, and form the main body of the thesis. *Chapter 5* details the experimental method used in examination. *Chapter 6* presents the results of the experiments as broken down by form of analysis and speaker characteristic. *Chapter 7* provides a discussion of the method and results of the previous two chapters.

Chapters 8, 9, and 10 pertain to the perceptual examination of prosodics and speaker characteristics. *Chapter 8* describes the experimental method of the investigation. *Chapter 9* presents the results of the experiments, subdivided by speaker characteristic and form of analysis. *Chapter 10* is a discussion of the approach and results of the perceptually based scheme.

Finally, *Chapter 11* provides a conclusion to the work by highlighting the major experimental results and indicating several directions for further research.

Chapter 2

Literature Review

In this chapter a review of research upon the acoustic correlates of speaker characteristics in speech will be made.

The definition of what *are* speaker characteristics is a troublesome problem itself. For the purposes of this study speaker characteristics will be defined as: "*A physical, physiological, or mental characteristic of a speaker that may be determined from their speech.*" This broad definition includes such features of the speaker as their age, sex, mental health, physical health, identity, race, socio-economic background and emotional state. Within the area of research upon speaker characteristics certain aspects have received more attention than others which is probably due to the relative difficulties of obtaining data, potential applications of findings, and areas of expertise of the researchers. This review will primarily deal with five speaker characteristics which appear to have received the greatest amount of attention:

- speaker identity
- speaker "emotional state"
- speaker sex
- speaker age
- speaker race, dialect and accent

as well as considering other studies of less fully investigated and "more unusual" speaker characteristics.

The review is broken into four sections. The first section deals with research motivated by human auditory perception of speaker characteristics. The second section deals with purely analytic research upon speaker characteristics. The third section covers the controversial applied area of speaker identification via visual inspection of speech spectrograms. Finally, some implications and conclusions will be drawn based on the current state of research into speaker characteristics.

2.1 Perceptually Based Investigations

Many researchers have sought to investigate speaker characteristics via human auditory perception. Such studies involve the presentation of speech sounds to listeners who evaluate the speech as to its content of one or more speaker characteristics (e.g. two voices played to listeners who decide whether they are the same or not). The answer to two main questions are sought by such studies:

- How accurate are listeners at identifying the characteristic(s)?
- Which acoustic parameters do listeners utilise in making their judgements?

Several sub-questions arise from these two primary questions including: to what extent is accurate identification of characteristics speaker and listener dependent; is acoustic cue utilisation listener dependent; which acoustic features are given heavier weighting by listeners etc.

2.1.1 Identity

The auditory identification of speakers is a reality which is familiar in daily life (e.g., listening to friends over the telephone). Many researchers have sought to investigate this phenomenon and answer such questions as human accuracy, the effect upon identification of different linguistic contents, and durations, correlations to acoustic parameters, and the effects upon identification of mimicry or disguise.

Human Accuracy

An area of interest, both in terms of legal applications of forensic speaker identification [Tos79], and as a yardstick for gauging performance of automatic speaker recognition schemes, is that of human ability and accuracy when identifying speakers from their voices.

In 1954 Pollack, Pickett, and Sumby [PPS54] conducted experiments investigating the effects of several variables upon speaker identification rates. A homogeneous set of 16 male speakers recorded a set text. These recordings were then presented to a group of 7 listeners who were familiar with the speakers through daily contact and who made guesses as to the identity of the speaker. Mean identification rates of 92% for 4 possible speakers and 84% for 8 possible speakers were found, and identification performance dropped only slightly as the possible number of speakers was increased up to 16. Pollack et. al. high and low pass filtered the utterance at various frequencies and tested the effect upon speaker identification rates. They found that speaker identification rates were resistant to both forms of filtering and concluded: "This result suggests that the identification of a speaker's voice is not critically dependent upon the delicate balance of different frequency components in any single portion of the speech frequency spectrum." Pollack et. al. also tested speaker identification rates for whispered (no voicing) speech and found it to be considerably lower than that of normal speech; such that an utterance of whispered speech 3 times the duration of the normal speech was required to achieve the same identification rate.

Clarke and Becker [CB69] compared speaker identification performance based on direct listening, ratings on psychophysical and semantic scales, and measurement of properties of the speech wave. A series of four-alternate forced choice speaker identification trials were conducted with a population of 20 male speakers and 5 listeners. A 67% identification rate was found for the direct listening experiments; and a 51% identification rate based on the values of 5 psychophysical and semantic ratings for each utterance (using a Euclidean distance measure). In paired speaker discrimination experiments (2 utterance presented, decision as to whether from same speaker or not), discrimination rates of 90%, 83% and 68% were found for direct listening, spectral distance, and psychophysical and semantic rating, respectively.

In 1971 Rosenberg [Ros71] conducted auditory speaker identification experiments using the same speech data as that employed by Doddington in his investigation of automatic speaker identification [Dod71a]. Stimulus material, consisting of a sentence from one of 40 speakers, was presented to a group of 10 listeners in pairs. Listeners were required to decide if the two utterances were produced by the same or different speakers. Rosenberg reports that a mean error rate of 4.2% for false acceptance (speakers were actually different) and 4.2% for false rejection (speakers were actually the same) was obtained.

Legge, Grosmann and Pieper [LGP84] investigated learning and recognition of unfamiliar voices. A total of 477 listeners were used in the experiments, with a population of 46 white female speakers aged 18 to 35 years. Listeners were previously unfamiliar with the speakers and were given an initial training period in which they heard a read text from all speakers. Experiments involved the presentation of two voice samples to the listeners, one known, the other not, the experimenter nominated one of the utterances and the listeners were required to indicate whether the speaker was part of the training set or not. Recognition rates were significantly above chance, and Legge et. al. found that recognition rates improved when the target speaker set size was reduced from 20 to 5 (approximately 13%), when photographs of the speaker set were continuously present (approximately 10%) and when the duration of the speech sample was increased. Further, Legge et. al. found that there was no significant difference in recognition rates for experiments administered 15 minutes after initial training or experiments administered 10 days after training.

In a series of experiments Schmidt-Nielsen and Stern [SNS85, SNS86] investigated auditory identification of speakers both familiar and unfamiliar to listeners. Schmidt-Nielsen and Stern recorded 24 speakers (15 male, 6 female) while playing the Battleship game over both unprocessed and LPC processed (DoD LPC-10 algorithm) channels. In the first experiment [SNS85] listeners were co-workers of the speakers and were asked to identify the speakers from an expected speaker population of 40. Prior to the listening experiment, listeners were asked to rate each voice as to its familiarity and distinctiveness. An identification rate of 69% for LPC processed speech and 88% for unprocessed speech was obtained. Speaker familiarity ratings were significantly correlated with correct identifications. In the second experiment [SNS86] exactly the same speech material was used but listeners unfamiliar with the speakers were employed. During the familiarisation period the listeners were asked to rate each speaker as to their 'distinctiveness'. The speaker population was then broken up into 3 groups of 5;

being the most 'distinctive' set of male and female speakers and the least 'distinctive' set of males. Identification experiments were then carried out and the 'more distinctive' speakers were found to have a far higher identification rate than the 'less distinctive' speakers for an unprocessed channel (60% versus 30%). When the channel was LPC coded identification rates dropped to 30% for male and 40% for female 'distinctive' speaker while identification rates for 'less distinctive' speakers dropped little.

Effects of Linguistic Content and Utterance Duration

Several researchers have addressed the questions of the effects that the length and content of utterances have upon auditory identification of speaker. Such questions address the relative contributions of segmental (phonemic) level and suprasegmental (prosodic) level uniqueness of speakers, and the relative worth of different phonetic material.

In their 1954 investigation of auditory identification of familiar speakers Pollack, Pickett and Sumby [PPS54] found that identification performance improved as the duration of the utterance increased. Pollack et. al. found that an identification rate of 100% could be achieved for a speaker population of 8 if 2 minutes of speech was used to represent each speaker. Further, Pollack et. al. found that identification performance improvement as a function of duration could be described by the growth function $1 - e^{-n\tau}$ and attributed this improvement to the admittance of more statistical samplings of the speaker's speech repertoire. To investigate this theory multiple utterances of a short passage were played to listeners for speaker identification. No consistent difference was found to the identification rate for only one repetition of the passage.

Bricker and Pruzansky [BP66] investigated the effect of linguistic content and duration of an utterance upon identification of 10 familiar speakers by 16 listeners. Excerpted vowels, excerpted consonant-vowels, monosyllabic words, disyllabic words and sentences were recorded from each speaker and presented to the listeners for speaker identification. Identification rates ranged from 56% for vowel excerpts up to 98% for the sentences and were found to improve directly with the number of phonemes present in the utterance. Two vowels, /a,i/ were used in the vowel based identification experiments, and identification rates and specific misidentifications were found to vary between them.

In another investigation of recognition of familiar speakers Lariviere [Lar71] recorded 8 male speakers uttering 2 sentences and 4 isolated vowels. The vowels were voiced, whispered (no voicing) and low-pass filtered at 200Hz (source characteristics preserved). Identification rates were 97% for the sentences and 40%, 22% and 21% for the voiced, whispered and filtered vowels. Acoustic parameters of the isolated vowel utterances were extracted and it was found that F_0, F_2, and F_3 were equally good predictors of speaker confusion in the identification experiments. Lariviere stated that he believed that the contribution of the vocal source and vocal tract to speaker identification are equal and additive.

Cort and Murray [CM72] investigated the ability of children to identify each other using utterances of different phonetic complexity. The 20 children (10 male, 10 female) each recorded a paragraph, a sentence, and a sustained vowel. Identification scores were significantly above

chance and were found to increase as the utterance heard changed from sustained vowel, to sentence to paragraph. Multiple repetitions of the same utterance did not increase identification scores.

Psychological and Semantic Scaling and Judgements

Another area of investigation of auditory identification of speakers is the nature and number of dimensions that listeners use in determining speaker difference and how they relate to acoustic parameters.

In 1964 Voiers [Voi64] used a modified semantic-differential form containing 49 bipolar items to obtain descriptions of the utterances of 16 speakers from 32 listeners. Based on an analysis of variance Voiers found that 4 orthogonal factors labeled clarity, roughness, magnitude and animation accounted for 88% of the between speaker differences in rating (inter-speaker ratings). In a later investigation Voiers [Voi79] used 18 professional listeners to investigate a total of 550 voice descriptors. Eight orthogonal dimensions labeled animation, pitch, continuity, charisma, roughness, vocality, clickiness and stability were found to account for the systematic inter-speaker variance.

Matsumoto, Hiki, Sone and Nimura [MHSN73] attempted to relate the perceived personal quality of isolated vowels uttered by 8 speakers to acoustic parameters of the speech wave. Measures of speaker similarity/difference as perceived by listeners were correlated with acoustic parameters. It was found that mean F_0 played the most important role in the perception of personal quality followed by characteristics of the vocal tract and then other characteristics of the vocal source. In a very similar experiment Yokoyama and Inoue [YI84] used 5 voice quality terms to describe a sustained vowel from 13 speakers. Similarity judgements between speakers were then based on the voice quality description and after correlation with extracted acoustic parameters it was found that personal quality was mainly characterised by F_0 and the higher formant frequencies.

Disguise, Mimicry and Stress

The effect of various distortions of the voice due to disguise, mimicry or stress has received relatively little attention from researchers though it has obvious applications in such areas as forensic speaker identification. Rosenberg [Ros72] describes the methodology of an auditory speaker identification experiment in which 4 intensively trained mimics were used to imitate a set of "true" speakers. Unfortunately, no results were published.

Reich and Duke [RD79] examined the effect of various vocal disguises upon auditory speaker identification rates. A series of sentences were recorded by 40 male speakers in normal voice and 5 disguise conditions: old age, hoarse, hypernasal, slow rate, and free disguise of speaker's choice. Sentences were presented in pairs (one sample always undisguised) to two sets of listeners, one naive and one speech professionals, who were asked to judge if the two sentences were from the same speaker or not. Both groups of listeners discriminated talkers at approximately 92% for undisguised speech. For disguised speech discrimination rates ranged from 59% to 81% depending upon the disguise used. Nasal disguise was found to cause the greatest reduction

in discrimination rate, sophisticated listeners discriminated between speakers better than naive listeners for disguised speech, and high confidence decisions by the listeners did not yield a significantly higher discrimination rate.

In 1982 Hollien, Majewski and Doherty [HMD82] investigated auditory identification rates for normal, stressed and disguised speech. 10 adult male speakers recorded a text using normal voice, using a free disguise of their choice and while receiving random electric shocks to induce stress. Three groups of listeners were used to judge the identity of a speaker. The first group of listeners were familiar with the speakers, the second group were unfamiliar with the speakers and the third group were not only unfamiliar with the speakers but could not speak the language. Identification rates ranged from 98%, 98% and 79% for familiar listeners under normal, stress and disguised condition down to 27%, 27% and 18% for non-English speaking listeners under the same conditions. Stress was found to have no significant effect upon identification rates while disguise did. Performance dropped across the 3 listening groups of the order of 60% from the familiar to unfamiliar listeners and 70% from familiar listeners to Non-English speaking listeners. Unfamiliar English speaking listeners were split into 2 groups based on their identification performance and a significant difference was found between their identification scores.

Acoustic Alteration

Very few researchers have examined the interaction of acoustic parameters and speaker identification by examining the effect of acoustic parameter alteration upon identification rates.

Bricker and Pruzansky [BP66] tested the effect upon identification scores of presenting utterances in reverse order (playing backward) to listeners. Bricker and Pruzansky found that no matter whether the speech material ranged from vowels to sentences the identification rate was approximately 10% below that for the equivalent material played forward; though identification rates were still significantly above chance.

Van Lancker, Kreiman, Emmorey and Wickers [LKE85, LKW85] describe a series of novel experiments involving the rate alteration and backward playing of famous voices. In the first experiment utterances from 45 "famous" male entertainers, politicians, film, television and radio personalities were obtained such that the contextual message gave little or no indication as to the speaker's identity. These samples were then played to 94 subjects whose task it was to identify the speaker. For 2 seconds of forward speech with no other information an identification rate of 26.6% was obtained; for 2 seconds of forward speech with a list of 6 possible speakers an identification rate of 69.9% was obtained; and for 4 seconds of reversed (played backward) speech again with 6 possible candidates an identification rate of 57.5% was obtained. Identification rates were found to be speaker and listener dependent. The relatively good performance of the backward played speech led Van Lancker et. al. to state "such relatively successful performance ...suggested that voice recognition can be achieved from acoustic information limited to pitch, pitch range, rate, voice quality, and vowel quality, but without acoustic detail reflecting specific articulatory and phonetic patterns, and orderly temporal structure." In their second experiment Van Lancker et. al. checked the effect of rate alteration upon the identification rate for the

30 most recognised voices from the first experiment. Utterances had their rate increased and decreased by $\frac{1}{3}$ without altering F_0. Identification rates dropped by approximately 14% for the slowed speech and 11% for the hastened speech. Again identification rates were found to be speaker dependent and those speakers who experienced large drops in identification score due to rate alteration differed from those severely affected by backward presentation in the earlier experiment. Van Lancher et. al. concluded that listeners utilise a subset of the potential acoustic cues to identify a speaker and the subset is speaker dependent.

Takagi and Kuwabara [TK86] investigated the independent manipulation of formant characteristics and fundamental frequency upon speaker identification rates. Takagi and Kuwabara recorded 2 speakers uttering a nonsense word composed of the 5 Japanese vowels. Systematic alteration was then made to the frequency and bandwidths of the lower 3 formants together and all formants together. The resulting speech was then played back to 3 listeners who were familiar with the speaker. An 8% formant frequency shift led to a total loss of speaker recognition, and identity appeared more related to the frequency of the 3 lower formants than of the higher formants. For bandwidth alterations accurate speaker recognition was lost at scale of 3 times and $\frac{1}{5}$ of the original, with identity appearing more related to the bandwidths of the higher formants than the lower 3. For the F_0 investigation 5 speakers were recorded and mean F_0 was shifted between half and twice its original value. Speaker identity was lost at double the original F_0, but at half original F_0 identification rates of 50% were still found. The effects upon identification were found to be speaker dependent.

Childers and associates [CYW85, CWH87b, CWH87a, CWHY89] have described the development of sophisticated resynthesis system for altering the speech of one talker to that of another. Several factors were identified as making significant contributions to either the perception of identity or the quality of the transformed utterance. These include spectral compression/expansion (format shifting), F_0 scaling, the need to modify the energy termed based on new spectral information, and accurate pitch synchronous measurements of F_0 and voicing. Childers et. al. mention: "...Formal intensive listening tests..." however no details of these tests are presented; merely the results gained thereby.

Dommelen [Dom90] has reported on listener perception of the identity of 5 female speakers known to the 11 listeners in the set. Speakers uttered a 9 syllable nonsense sentence composed from the syllable "mama", and via manipulation the three features mean $F0$, F_0 contour, and speech rhythm (segmental duration) were examined for their contribution to speaker identity. Mean F_0 was found to be the most significant, followed by F_0 contour and segmental duration.

Recently Johnson [Joh90] examined the role of perceived speaker identity for F_0 normalisation of vowels. Creating purely synthetic word and phrase utterance he found that alterations in mean F_0 do cause alterations in listener perception of identity. There were significant differences between individual listener's perceptions, indicating the weighting assigned by different listeners to different acoustic cues.

2.1.2 Emotions and Stress

Several questions arise regarding human perception of emotional and stressed speech. These include how accurate listeners are at detecting vocal emotions, does the degree of detection depend upon the emotion, and what acoustic cues do listeners use in the detection of vocal emotions.

Dusenbury and Knower [DK38] performed an early test of perception of emotions via filming two speakers facially expressing 11 emotions while uttering the letters in the alphabet from "A" to "K". Judgements of emotion ranged from 62% for naive viewers up to 91% for speech students; showing that emotions can be identified by facial expressions and that relevant emotional data is contained within the speech wave for those capable of utilising it. Knower [Kno41] used the same data but included whispering. Utterances were played to listeners forward and backward. Whispering dropped identification rates by approximately 30% and playing in reverse dropped identification by approximately 45% showing the importance of prosodics in emotion identification by listening. Pfoff [Pfo54] also investigated the ability of listeners to identity emotions from 'content-free' speech (digits 1 to 8). Pfoff found a mean identification rate of 50% for 10 emotions and found that some emotions were identified with higher rates than others.

Fenster, Blake, and Goldstein [FBG77] examined several aspects of the vocal transmission of emotions. An adult and child speaker set recorded six emotions (anger, fear, sadness, contentment, happiness and love) which were played to adult and children listeners. It was found that adult and children listeners perceived negative emotions more accurately than positive emotions. Adult listeners were more accurate in the perception of emotions and there was no significant difference in the ability of adults and children to express emotions.

Nilsonne and Sundberg [ANS85] evaluated the ability of listeners to recognise depression in the voice of patients. Speech of depressive patients before and after treatment was recorded. Listeners heard a constant vowel sound upon which the F_0 contour of the patient's speech was synthesised. Listeners identified the "depressed" speech samples with 80% accuracy. Ross, Duffy, Cooker and Sargeant [RDCS73] examined the effect of low-pass filtering upon emotion recognition rates. 9 emotions were examined and the recognition rate was found to be emotion dependent. Recognition rate dropped slowly as filtering dropped to 300Hz low-pass, but at 150Hz low-pass recognition rates were little better than chance.

Brown, Strong and Rencher [BSR73] experimented with independent manipulation of mean F_0 and speaking rate, and its effect upon perception of the personality of the speaker. Increased F_0 was perceived as more benevolent, while decreased F_0 was perceived as less benevolent. Rate increased voices were perceived as less benevolent, and rate decreased as less competent.

Streeter, Apple, Draus and Coalatti [SAKG83] conducted a novel investigation of acoustic and perceptual indicators of stress. Tape recordings were made of the System Operator for Consolidated Edison speaking to their immediate superior in the hour leading up to the 1977 New York power blackout. The utterances of both speakers were presented to listeners who rated them as to stress. Acoustic analysis was performed and it was found that the pitch and amplitude of the superior's utterances increased markedly with situational stress while the System Operators pitch was found the decrease. Listener stereotypes of stress included

evaluated F_0 and amplitude levels, as well as their increased variability.

Ladd, Silverman, Tolkmitt, Bergmann, and Scherer [LST+85] examined the effect of manipulation of voice quality, intonation contour and F_0 range upon listener judgement of affect—evaluation of arousal and attitude. Ladd et. al. found that F_0 range and contour, and F_0 range and voice quality had independent effects upon the way utterances were judged.

Pittam et. al. [PGC88, PGC90] have examined the long-term spectrum characteristics of perceived emotions. Thirty speakers recorded three passages with the intention of evoking the emotions pleasure, arousal, and control. A listener set of 120 students rated all utterances on a 15 adjective scale correlated to the intended emotions. Analysis of listener responses showed a listener perceptions of the passages a being representative of the intended emotions. Long-term spectrum measures were then extracted and compared on the basis of the three effective dimensions. The long-term spectrum was found to be systematically related to the affective dimensions based on particular frequency bands of the spectrum, and there was found to be no significant sex or ethnic group effects.

2.1.3 Sex

It has long been known that mean fundamental frequency is a primary indicator of sex of a speaker [Wea24, HP69, HHP88], however other acoustic parameters may well signal speaker sex reliably.

In a series of experiments in the late 1960's Schwartz and Rine, and Ingemann [Sch68, SR68, Ing68] showed that listeners were able to identify speaker sex from isolated, whispered vowels and consonants. Identification rates were found to depend upon the individual phoneme and ranged as high as 93%. Analysis of the spectra by Schwartz and Rine showed a general shifting to higher frequencies of the formants for female speakers.

Coleman [Col71, Col76] investigated the relative contribution of F_0 and the lower three formants to perception of maleness and femaleness of the voice. In an initial experiment using 10 male and 10 female speakers uttering isolated vowels an artificial F_0 of 85Hz was synthesised for all utterances before presenting to listeners. An identification rate of 98% for male speakers and 79% for female speakers was obtained; and analysis of the lower three formant frequencies showed a significantly lower mean for male speakers. In later experiments Coleman played backward speech to listeners and correlated judgements of speaker sex with mean F_0 and the mean frequency of F_1, F_2, F_3. Judgements of sex were extremely highly correlated with F_0 and less highly correlated with the vocal tract resonance. Coleman then synthesised speech using the vocal tract resonance of females and the F_0 of males. In both cases approximately two-thirds of responses from listeners were that the speaker was a male. Coleman concluded that in natural speech F_0 was the primary cue to speaker sex.

More recently Lass et. al. have investigated the effect of certain variables upon speaker sex identifications. Lass et. al. [LHB+76] also investigated the relative contribution of F_0 and vocal tract resonance to speaker sex identification. Isolated sustained vowels were spoken in normal and whispered voice as well as low-pass filtered at 255Hz. Speaker sex identification rates were 96% for voiced speech, 91% for filtered speech and 75% for whispered speech; showing the

higher importance of F_0 over vocal tract resonance for speaker sex identifications.

Lass, Mertz, and Kimmel [LMK78] investigated speaker sex identification rates for backward played speech and speech compressed (rate increased) by 40%. Speech material was sentences from 10 male and 10 female speakers and was presented to 30 listeners. Identification rates did not drop for speech played backward and dropped only 2% to 3% for compressed speech. Using the same set of speakers Lass, Tecca, Mancuso, and Black [LTMB79] investigated the effect of phonetic complexity upon sex identification. From isolated vowels through to sentences a mean sex identification rate of 98% was obtained. Again using the same speakers as previously Lass, Almerino, Jordan, and Walsh [LAJ80] tested the effects of filtering upon sex identification rates. Unaltered, low-pass filtered at 255Hz and high pass filtered at 255Hz speech samples were presented to listeners for sex identification. Identification rates of 96%, 94% and 95% were found for the unfiltered, low-pass filtered, and high-pass filtered samples; showing that sex identification is not affected by this form of filtering.

In an interesting study of the speech of male-to-female transsexuals Spencer [Spe88] used 46 listeners to judge the gender and male/femaleness of the voices of 8 male-to-female transsexuals. The transsexuals and a control group of male and female speakers recorded a passage that was played to listeners. Listeners judged the sex of the control group with 100% accuracy. Results varied markedly between the individual transsexuals; from 100% identification as male, through a part-male-part-female spread of responses, to 100% female identification. Listener judgements were found to be highly correlated with mean F_0 with a sharp discontinuity at 160Hz. Not all sex perceptions could be explained by mean F_0 in which case Spencer cited vocal tract size as the likely cue.

Childers et. al. [CYW85, CWH87b, CWH87a, CWHY89] have described a series of sophisticated resynthesis experiments involving the alteration of one speaker's voice to that of another, as perceived by listeners. Among other experiments were the transformation of a male's sentence to that of a female's. Various parameters were uttered individually and in combination and it was found that mean F_0 was the most significant cue for gender perception.

Recently Johnson [Joh90] described a series of identity and sex identification experiments as the initial stage in an investigation of vowel normalisation. A series of synthetic utterances with varying F_0 were created and the perceived cross over point between the sexes was found to be between 140Hz and 150Hz.

2.1.4 Age

Several investigators have sought to examine listener judgements of speaker age based on the voice.

Shipp and Hollien [SH69] gathered speech data from 175 adult males equally divided into the 7 age decades between 20 and 90. Three groups of listeners were employed to estimate age based on young, middle aged or old; the decile band within which the age of the speaker fell; and the actual age of the speaker. Shipp and Hollien found that listeners were able to accurately estimate speaker age (correlation co-efficient of 0.88 between judged age and chronological age). Ryan and Capadano [RC78] conducted a similar experiment with both male and female

speakers. Listeners reliably judged the chronological age of speakers (female speakers more accurately), and it was found that certain listener ratings of personality, such as flexibility and reservedness, were correlated with perceived age.

Linville, Fisher, and Korabic [LF85a, LK86] evaluated the ability of young and elderly adult female listeners to estimate vocal age of adult females from a sustained vowel. 75 women speakers equally divided into the age groups 25-35, 45-55 and 70-80. Speakers were trained to produce the sustained vowel /æ/ at a steady F_0 within the 200Hz to 220Hz range; as well as whispered. Listeners were asked to judge which of the 3 possible age groups a presented 1 second segment of the vowel belonged to. Young listeners were more capable of accurate judgement of age (51% voice, 43% whispered) than elderly listeners (45% voiced, 38% whispered). Correlation of acoustic parameters with judged age showed that F_0 (both mean and standard deviation) was the most important cue used by listeners for voiced speech, and F_1 was the most important cue for whispered speech (lower F_0, F_1 perceived as older).

Recently Neiman and Applegate [NA90] have examined the accuracy of listener judgements of speaker age. Thirty sex speakers as divided into 3 male and 3 female speakers in the age ranges 20-25, 30-35,... 70-75 recorded a passage of which the first three sentences were played to a listening group of aged 20-25. Listeners judged which of the six age categories a speaker belonged to and a correlation of 80 was found between listener judgements and actual speaker age categories. Further it was found that young speaker age was judged more accurately than old speaker.

Other researchers have sought to investigate the perceptual dimensions that listeners use in making judgements of vocal age. Ryan and Burke [RB74] used trained listeners to estimate the vocal age of 80 male speakers. Speech samples, from those speakers whose judgement vocal age corresponded closely (standard deviation less than 6 years) to their chronological age, were presented to speech pathologists to rate on 10 voice characteristics. It was found that voice tremor, laryngeal tension, air loss, imprecise consonants, and slow rate of articulation were strong predictors of perceived vocal age. Hartman and Danhauer [HD76, Har79] performed similar experiments and found such voice features as pitch, rate of speech, quality, and articulation to be strong predictors of perceived vocal age.

Jaques and Rastatter [JR90] recently examined the perception of young and old speakers by groups of young and old listeners. Speakers produced sustained vowels which were then processed and played to listeners. Listeners heard F_0 and resonance, F_0 alone, and resonance alone and were required to judge young or old speaker. Identification rates ranged from 40% up to 80% and were found to be dependent upon a number of factors. F_0 was found to be a significant perceptual cue.

2.1.5 'Race', Dialect and Accent

The area of speaker 'race' and dialect is extremely diverse and attempting to cover it entirety is beyond the scope of this work. In fact the whole question of 'race', its definition, and interaction with socio-economic environment is a complicated issue. Therefore, whenever the term 'race' is used the term is that chosen by the original authors of the papers and used in the context defined

or assumed by those authors. Various questions, however, arise as to the acoustic correlates of perceived dialect, the degree of correspondence between different listeners' perception of dialect and the linguistic level at which dialect differences manifest themselves.

In a series of experiments Abrams [Abr73] examined listener perception of 'race' based on different degrees of phonetic complexity; vowels through to sentences. With a speaker set of 36 black and white Americans, and 60 listeners Abrams found that listeners judged skin colour (race) correctly 62% of the time. Identification rates were found to be highly variable based both on individual listener and speaker group; while phonetic complexity had no regular effect upon identification scores. Further, it was found that response bias varied directly with the language community of the speaker. Thus speakers of Standard English were usually classified as White, while speakers of Non-Standard English were classified as Black.

Brennan, Ryan and Dawson [BRD75] investigated the degree of correspondence between 72 naive listeners' rating of accentedness in speech samples from 8 Spanish-English bilinguals. Brennan et. al. found there was a significant agreement between listeners on the perceived degree of accentedness in the test utterances and concluded that nonlinguistically trained listeners are capable of making accurate and consistent judgements of the degree of accentedness.

Lass et. al. [LTMB79] and Flege [Fle84] have both examined the effects of phonetic complexity upon listener's identification of accented speech. Lass et. al. recorded utterances of various phonetic complexities from 10 black American and 10 white American speakers. Identification rates were found to range form 55% (little better than chance) for isolated vowels up to 78% for sentence length utterances. Flege investigated the detection of French accent in American English by listeners at a segmental through to a phrase level. Correct detection rates ranged from 70% for isolated /t/ up to 90% for phrases.

Lass et. al. [LMK78, LAJ80] have looked at the effects upon 'race' identification of manipulation of the acoustics of the speech wave. Using sentences from 10 black American and 10 white American speakers Lass et. al. investigated the effects upon listener race identification rate of time compression and backward playing of the speech. Identification rates were 70% for forward speech and approximately 62% for backward speech and 65% for time compressed speech. Lass et. al. concluded that temporal clues play a role in speaker 'race' identification. Lass et. al. also examined the effect of filtering upon perceived speaker 'race' for the same set of speakers. Low-pass filtering at 255Hz was found to have a greater detrimental effect upon correct 'race' identification than did high-pass filtering at 255Hz.

Moon, Leeper, and Mencel [MHAL84, MML88] have investigated the speaker 'race' identification of North American adults and children. For both adult and children speaker sets utterances of varying phonetic complexity from sustained vowels through to sentences were recorded voiced, whispered and with an electrolarynx synthesised F_0 . It was found that identification scores rose as phonetic complexity did and identification scores dropped from voiced, to whispered to electrolarynx. Identification scores were well above chance given sufficient phonetic complexity and suprasegmental features were used by the listeners in decision making. Identification scores were examined on the basis of speaker 'race' for both speaker sets. Indian males and Non-Indian females were markedly better identified as to 'race' than Non-Indian

males, and Indian females had the lowest correct 'race' identification rate (below chance).

2.1.6 Other

Several other traits of a speaker such as physical health, physical size, and socio-economic background may also be detected by listeners.

Andrews, Cox Smith [ACS77] investigated the perceived characteristics of speakers after consuming alcohol and compared them with perceptions of non-intoxicated speech. Speech of subjects after consuming alcohol was rated by listeners as less efficient, reasonable, self confident, artistic, and theatrical, and more untrained than the equivalent utterances without prior consumption of alcohol.

Lass, Beverly, Nicosia, and Simpson [LBNS78] examined listener ability to judge the height and mass of a speaker. Both male and female speakers and listeners were used. Mean judged male speaker mass was overestimated by 1.50kg while mean judged female speaker mass was underestimated by 1.65kg. Mean judged male speaker height was underestimated by 6mm while mean judged female speaker height was underestimated by 32mm. Lass et. al. found that listeners were able to accurately judge the approximate height and mass of speakers.

Reich [Rei81] examined the ability of listeners to detect the presence of vocal disguise in the voice of male speakers. Both naive and sophisticated (speech scientists) listeners heard vocal stimuli from the speaker set in either normal or disguised voice and were asked to determine whether the utterance was disguised or not. Speakers chose their own form of disguise, the only criteria being that it disguise their identity as much as possible; and were not told to make the disguise undetectable. Naive listeners detected vocal disguise with 89.4% accuracy, while sophisticated listeners detected disguise with 92.6% accuracy.

2.2 Analytically Based Investigations

The alternate to the perceptual approach to speaker characteristic examination is a purely analytic approach. Characteristic of such studies is the analysis of speech data containing well-defined pre-measured speaker characteristics (e.g., speech samples from speakers of various known ages).

The two major questions to which answers are sought in such studies are:

- Which acoustic parameters show manifestations of the speaker characteristic(s)?
- What level of accuracy can be obtained for automatic identification of the speaker characteristic(s)?

Again, several sub-questions arise. These include: what degree of manifestation is there in the individual characteristics; to what degree are these speaker dependent etc.

2.2.1 Speaker Identity

By far the largest area of research into speaker characteristics is that of Automatic Speaker Recognition. Research in this area is principally motivated by practical applications such as access control systems, forensic speaker identification, and telephone banking.

Spectral Approach

In 1962 Smith [Smi62] described the first known automatic speaker recognition system. Output from a 35 channel filter-bank was analysed via multi-dimensional analysis of variance, and a smaller set of linear transformations of the original filter measurements was derived such that their variance ratio was optimal. These transformed parameters were then used to identify the speaker from a set of known speakers. Unfortunately, no more details are available.

In 1963 Pruzansky [Pru63] described a speaker identification system using the outputs of a 17 channel filter bank. Using 7 repetitions of 5 sentences from 10 speakers (7 male, 3 female) Pruzansky investigated the identification scores when different vectors were used. Selected words were excerpted from the sentences and representative vectors, based on the filter bank outputs, were formed. Vectors investigated were of 3 forms:- time-frequency-energy (one value per band per sample), time-energy (total energy per sample) and frequency-energy (mean energy in each of the bands across the entire word). Using matrix correlation as a measure of similarity speaker identification scores of 89%, 47% and 89% were achieved. These results tend to lead to the conclusion that the energy contour has little benefit in speaker recognition, but much of its poor performance may be attributed to the means of time alignment (moment of maximal energy matched).

In a follow-up investigation motivated by human performance in spectrogram reading Pruzansky and Mathews [PM64] tested recognition rates when various sized rectangular regions in the time-frequency domain were used. Using the same speech data as Pruzansky's previous work they defined various sized rectangular regions (different numbers of samples, and different numbers of filter bands) for which the mean energy was calculated. Recognition performance improved as larger time intervals were used and decreased as the number of frequency bands composing a region increased. Pruzansky and Mathews concluded that the energy in successive time intervals is relatively dependent.

Li and Hughes[LH74] investigated speaker identification based on correlation matrices of continuous speech spectra [LHH66]. Output from a 35 channel bandpass filter was used to generate correlation matrices for 8 speakers reading a text. Matrix differences, calculated via 3 different measures, were found to be consistently smaller in the intra-speaker as opposed to the inter-speaker case. When speaker identification and verification experiments were conducted for the 8 speakers from a total population of 30 error rates of 1% to 3% were found.

In 1974 Atal [Ata74] conducted a series of speaker recognition experiments using Linear Predictive Co-efficients (LPC)[1]. Atal recorded 10 female speakers pronouncing the all-voiced sentence: "May we all learn a yellow lion roar." 3 times on each of two days, the days being

[1]LPC analysis is based upon the model of the speech production process as a source-filter mechanism; in which the vocal tract is represented as an all-pole filter [MAHG76].

separated by 27 days. Each utterance was divided into 50 equal segments and 12 LPC predictor co-efficients were determined for each segment. From these predictor co-efficients other parameters, including cepstrum co-efficients and area function co-efficients were derived and each set was tested for its ability to discriminate the speakers. Using only one segment (approximately 50 msec) identification experiments were conducted for each of the parameter sets and cepstrum co-efficients were found to have the highest identification rate of 70%. When half a second of speech was used the identification rate increased to 98%. Similarly, the cepstrum co-efficients yielded a verification rate of 83% for 50 msec of speech, increasing to 98% for 1 second of speech. Comparing these results with those he obtained using pitch contours of the same data Atal stated: "...these results suggest that the spectral envelope information is more effective than the pitch contour information for automatic speaker recognition."

In a series of speaker identification experiments Sambur [Sam76] investigated the use of orthogonal LPC parameters derived via eigenvector analysis. 21 male speakers recorded the sentence "I was stunned by the beauty of the view." on 6 separate occasions. Orthogonal LPC parameters that showed little variance across the utterance were selected as representing the speaker, while those with high variance were selected as representing the linguistic content. The higher order co-efficients were found to be "stable", while the lower order varied greatly. Orthogonal LPC, orthogonal PARCOR, and orthogonal Log Area Ratio (LAR) co-efficients were all compared for their ability to identify the speaker with varying number of co-efficients identification scores of 96.8%, 99.2% and 98.9% were achieved for the LPC, PARCOR and LAR co-efficient sets respectively. Similar results were obtained for verification experiments.

In another set of experiments investigating orthogonal LPC co-efficients Shridar et. al. [SMB81] tested their utility for text-independent speaker recognition. Using a dynamic programming procedure for co-efficient selection Shridar et. al. found that the higher order co-efficients weren't necessarily the optimal co-efficients.

Furui [Fur81a] investigated cepstrum co-efficients for speaker verification. Time functions of the cepstrum co-efficients were used to represent the speaker. These were compared via Dynamic Time Warping (DTW) and error rates as small as 0.3% were reported.

Prosodics and Suprasegmentals

Many speaker recognition systems use prosodic and suprasegmental information as an adjunct to their main vectors of recognition which are based on spectral data [Luc69, MOJ77, NND89]. However few approaches primarily investigate prosodic or suprasegmental sources of information.

Doddington [Dod71b, Dod71a] investigated the use of time functions of the formant frequencies, fundamental frequency and speech energy for their ability to identify speakers. Using 40 speakers (8 designated customers and 32 imposters) Doddington found that proper time normalisation was an important factor in improving error rates. Non-linear time normalisation was performed by maximising the correlation between sample and reference second-formant profiles through a piece-wise linear transformation of time. Non-linear time normalisation improved error rates by a factor of 4 over simple end-point alignment. Average speaker verification

error rates after non-linear time normalisation were 5% for F_0; 4% for the formants; 4% for the energy; and a combined error rate of 1.5%.

In 1971 Levitt and Rabiner [LR71] attempted to quantify the various levels of variability in the production of F_0 contours. Three speakers repeated the two sentences "Larry and Bob are here." (with emphasis on "Bob") and "This is an olive." (emphasis on "olive") three times each. F_0 contours were then extracted and each was divided into "epoches", generally corresponding to voiced/unvoiced boundaries. Orthogonal polynomials were used to represent the F_0 contour within a time window. The co-efficients of the orthogonal polynomials themselves were then treated as contours and an orthogonal polynomial determining them was then calculated. Sources of variability in the F_0 contour were then compared and rank ordered. Inter-speaker variability was found to be significantly larger than intra-speaker variability. Also, intra-speaker variability was found to be greater during non-stressed than stressed sections.

In 1972 Atal [Ata72] described a series of speaker identification experiments using pitch contours. 10 female speakers made 6 recordings of the all voiced sentence: "May we all learn a yellow lion roar." The contours were all linearly normalised to a 2 second duration and represented by a 20 dimensional vector in the Karhunen-Loeve co-ordinate system. These vectors were then linearly transformed so as to optimise their intra-speaker to inter-speaker variance (F ratio [PM64]). Identification was based on the nearest-neighbour approach. Using a training set of 5 utterances and 1 for testing for each speaker, recognition rates of 93% to 98% were achieved. Atal then compared the system to several other well-known approaches to speaker recognition, namely:- minimum distance classifiers, cross correlation, moments of pitch period, which yielded recognition rates of 68%, 70%, and 78% respectively. These results showed the applicability of pitch contours to the recognition of speakers in general, the usefulness of the linear transformation of the feature space (97% versus 68%) and the significance of the moment to moment variations in pitch (moments of pitch period performance versus pitch contour performance).

In 1973 Lummis described another speaker recognition technique based on the use of prosodic features [Lum73]. The procedure was based on Doddington's earlier approach [Dod71b] where, however, gain was used for temporal alignment as opposed to Doddington's more computationally expensive F_2. Parameters analysed were pitch, gain, F_1, F_2, and F_3. Using the all voiced sentence "We were away a year ago." with a speaker set of 41 (8 customers, 32 casual imposters, and 1 twin brother of a customer) verification error rates of the order of 1% were achieved, and the contribution of the formant contours was found to be minimal.

Wasson and Donaldson [WD75] describe a speaker identification system based on orthonormal functions of the amplitude and zero crossing rate. For a mixed sex speaker population of 10 an identification rate of 96.6% was reported using the ubiquitous "We were away a year ago".

Johnson et. al. [JHH84] examined temporal measures of utterances of speakers for encoded speaker identity. Twenty male speakers read a passage under normal voice, disguise, and stressed conditions. Two vectors, a time-energy distribution and voiced/voiceless time contrast were extracted and examined via speaker discrimination experiments. Both vectors were found to discriminate speaker, energy-time better than voicing/voiceless, and combined yielded a

speaker discrimination rate of 55%.

Chen and Lin [CL87] have reported on the use of F_0 contours to identify speakers of Mandarin. Based on slope, mean, and duration F_0 measures of the four basic Mandarin tones a text-independent speaker identification rate of 99.2% was obtained for a speaker population of 11 males and it was shown that the period of one month between utterances did not alter recognition rates.

Barlow and Wagner [BW88] have examined the speaker discriminant ability of four prosodic parameters using the DTW paradigm. For a small population of five Australian males uttering four sentences the four parameters energy, voicing, F_0, and LPC error were all found to show significant encodings of speaker identity.

Phonetically Based

Another known source of speaker difference in is the uniqueness of segmental[2] production of each person. Both phonetic and acoustic definitions of segmental units have been used.

Glenn and Kleiner [GK68] were the first to investigate the uniqueness of segmental production as applied to speaker identification. In an investigation of the uniqueness of nasal phonation Glenn and Kleiner recorded 30 speakers (20 men, 10 women) reading two words lists containing 10 words each; each word containing a minimum of two nasals. Nasals were represented by 25 element power spectra in the range of 1khz to 3.5khz which were manually calculated. Vector similarity was measured by their correlation. Using the mean of 10 consonants as test input a speaker identification rate of 97% was achieved for a population of 10, and 93% for a population of 30.

In a similar investigation Su, Li, and Fu [SLF74] examined the identification of speakers via nasal co-articulation. Isolated utterances, containing nasal-vowel combinations, were phonated by two phonetically trained adult males and two adult females. Three front vowels/ i, e, æ/ and three back vowels /u, o, a/ were combined with the two nasals /m/ and /n/ for a total of 12 combinations. Each combination was repeated 3 times by each speaker. A 25 channel filter bank was used to extract the nasal spectrum in the range 250Hz to 3681Hz. Euclidean distance between mean vectors showed a far larger inter-speaker distance between [mV] combinations than [nV]. 10 further male speakers were recorded and identification experiments were conducted based on correlation matrix similarity. An identification rate of 100% was obtained.

Luck [Luc69] examined the cepstra of the first two vowels combined with the speaker's pitch and the length of the word "my" in the phrase "my code is -" for their ability to discriminate speakers. 34 adult male speakers were recorded. Four were designated customers and recorded the test phrase 25 times on 8 separate occasions. The other thirty were designated imposters and recorded the utterance 25 times on a single day. Spectral extraction was performed at the position of maximum amplitude in the word "my" and /o/. Speaker verification experiments were then conducted using a nearest neighbour criteria and error rates of 6% to 13% were obtained depending upon the inclusion of F_0 and the word length.

[2]Here defined as the production of individual speech sounds, whether these units are defined phonetically or acoustically.

In 1972 Hair and Rekeita [HR72] performed a series of experiments investigating the spectra of phonemes and their use in speaker verification. 40 speakers, each of whom had uttered 10 isolated words once on each of nine weeks were used as customers and imposters. Thirty element power spectra for each of 6 phonemes were used to characterise the speakers. Using a hypersphere decision rule (threshold distance) an error rate of 2% was found and Hair and Rekeita concluded that the technique was a reliable method for speaker verification.

Pfeifer [Pfe78] obtained more than 63 minutes of conversational speech from 20 speakers (10 male, 10 female) via an interview-like situation. The vowels /i, I, E, æ, a/ were manually detected and represented by 12 reflection co-efficients obtained at the stablest portion of the vowel. Five reference vectors (one for each of the vowels) were used for each speaker. Speaker identification experiments were then conducted using a weighted Euclidean distance. For straight vowel matching an identification rate of 39.2% was achieved. Interestingly, when only one reference vector was used to represent the 5 vowels (vowel space) the speaker identification rate increased markedly to 44.98%. When a sequential deferred decision process was used with only one reference vector for each speaker a speaker identification rate of 85% was obtained using speech data corresponding to approximately 30 seconds of speech. This approach has also been taken by Wood [Woo78] who expanded the system to automatic spotting of vowel-like sounds (vowels, nasals, and liquids) in running-speech.

Li [Li87] describes a series of experiments to separate speaker specific and phoneme specific features of the formants of vowels. Using both a conversational database from sixteen speakers, and a connected database from ten speakers, Li applied orthogonal analysis prior to speaker verification experiments for which error rates ranging from 21% down to 11% were reported.

Savic and Gupta [SG90] describe a speaker verification system based on HMM and a broad phonetic categorisation of utterances. The broad groupings of vowels, fricatives, nasals, plosives, and voiced are used and LPC parameters are extracted. Verification experiments with a speaker set of 43 are carried out separately for these categories, and in combination. Error rates were found to vary greatly between the different categories but all were found to have encoded speaker identity information.

Recently Rosenberg et. al. [RLS90] have reported on both an acoustic and phonetic transcription based approach to speaker verification using the Hidden Markov Model paradigm. With a 100 speaker database of isolated digit utterances, and using cepstral co-efficients as the vectors of recognition both approaches were found to have roughly equal speaker verification performance. For a single digit error rates of 7% to 8% were reported, dropping to 1% for 7 digits.

Feature Selection

The selection of "good" acoustic parameters for speaker recognition is a task facing all applications. In their 1964 investigation Pruzansky and Mathews [PM64] based feature selection on the basis of having a small variance between utterances from a given speaker, compared to the variation among utterances of different speakers. Thus, they defined a measure of this variance

ratio, known as the F ratio:

$$\text{F ratio} = \frac{\text{variance of talker means}}{\text{average within talker variance}}$$

to select the k best features based on the highest k F-ratios.

Wolf [Wol72] was the first investigator to examine in detail a large number of acoustic features for their speaker specificity. Using utterances from 21 adult American males speaking 6 sentences Wolf examined various parameters extracted at particular speech events (e.g., peak F_0 value in the syllable "not" of one of the 6 sentences). Wolf defined six properties that measured speech parameters should have for good speaker recognition:-

1. occur naturally and frequently
2. be easily measurable
3. vary widely among speakers, but be consistent within speakers
4. not change over time or be affected by health of speaker
5. not be unreasonably affected by background noises or transmission system
6. not be modified by conscious effort (unaffected by attempts to disguise).

Wolf investigated over 30 parameters on the basis of their F ratio, a measure of their inter-speaker variance to intra-speaker variance [PM64, DS71], being closely related to the above third property. High F ratio parameters were chiefly measures of F_0 (although these were often highly correlated with one another), though vowel and nasal formants also yielded high ratios. Using the best 9 parameters (based on their F ratio and measure of interdependence) Wolf achieved an identification error rate of 1.5%. Unfortunately Wolf's speech material was all recorded in one sitting, highly reducing the intra-speaker variability. Further, due to the nature of the measures investigated, it is first necessary to "phonetically" segment an input utterance before measures may be taken (a task which appears far more difficult than speaker recognition).

In 1975 Sambur [Sam75] investigated a set of 92 features extracted at particular speech events (including formant bandwidths and frequencies in vowels, durations of speech events, dynamics of formants etc). Using a probability of error criterion based upon a Gaussian error rate Sambur chose a best sub-set of features via a 'knock-out' strategy. Starting with the N features to be investigated N sets of $N-1$ features were generated (where each feature had been left out of 1 of the N sets). The set of $N-1$ features having the lowest probability of error was then kept (effectively eliminating one feature). The same process was then followed with this set of $N-1$ features (breaking into $N-1$ sets of $N-2$ features and selecting the best set of $N-2$ features). Using the same speech material as Wolf (with additional recordings to introduce intra-speaker variability over time) Sambur found the formant frequencies in vowels and nasals and the mean F_0 to be most important. In a speaker identification experiment using the best 5 features an identification error rate of 0.3% was obtained.

Goldstein [Gol76] used a similar method to obtain the best features for 10 American males based on the formant structure of 3 diphthongs, 4 tense vowels, and 3 retroflex sounds. 109 features were examined and those that yielded F ratios greater than 60 were further investigated (29 features). Based on Sambur's probability of error measure the 'best' k features were determined via an 'add-on' process. Of the N features the one yielding the lowest error estimate was taken. Then each of the remaining $N-1$ features were paired with the already selected feature. The pair yielding the lowest error estimate was then selected. This process was then repeated for triplets etc. until k features had been found. Goldstein found that minimum and maximum measures of the formants (particularly F_2) were particularly effective and attributed these differences to being more a matter of speaker habits than of vocal tract anatomy. Using only the two best features an identification error rate of 15% was obtained.

Cheung and Eisenstein [Che78] used a dynamic programming procedure to select the best k-subset of N features. For ten male speakers reading 4 of the Harvard Phonetically Balanced lists F_0, energy, PARCOR co-efficients, cepstral co-efficients, and normalised autocorrelation co-efficients were extracted and averaged over the voiced duration of the utterances. In an approach similar to Goldsetein's "add-on" technique the feature subset was built up from one feature. However, at any stage i, there are N subsets of i features, each subset having been started with a different one of the N features (as opposed to Goldstein's one subset). The best k subset is then selected from the N k-featured subsets. Speaker identification experiments were then performed using each of ten cepstral co-efficients, ten PARCOR co-efficients, the ten best features as determined by Sambur's "knock-out approach", and the ten best features as determined by the dynamic programming approach. The two sets of selected features performed significantly better than the PARCOR co-efficients and the cepstral co-efficients (no doubt due to the F_0 and energy information lacking in the straight spectral representation). Of the two selection process the features derived via dynamic programming yielded an error rate approximately 1% less than those derived via the "knock-out" technique.

Shridar et. al. [SMB81] used a dynamic programming procedure to select the k best set of 12 orthogonal LPC co-efficients. They found that on the criteria of $\sum$ Inter-speaker distance $-\sum$ intra-speaker distance that the best co-efficients were speaker specific and varied from one speaker to another.

Long Term Features

Two types of variability in the speech produced by a given speaker compound the problem of speaker recognition. One problem is that of intra-speaker variability; particularly over the long term (e.g., intervals of greater than 6 months). The second problem is that of the variability introduced by the linguistic content of the speech. Several researchers have addressed this problem by examining long-term features of the speech.

As early as 1963 Pruzansky [Pru63] tested a semi text-independent speaker identification system. Ten speakers recorded a series of sentences from which certain keywords were extracted. The mean energy across 17 different frequency bands for all words was then used as a reference and test vector for each speaker. Using a correlation similarity measure 100% identification was

obtained.

In 1972 Furui, Itakura, and Saito [FIS72] investigated the ability of the long-term power spectrum (LTS) to represent the characteristics of a speaker. Speech material, consisting of a fixed sentence of approximately 10 seconds length, was obtained from nine speakers over the period of a year and a half. Speaker identification and verification experiments were conducted and the effects of different distance measures, training interval, and interval between training and test utterances was investigated. It was found that recognition rates dropped significantly as time between a single training utterance and test utterances increased. With increased training sets spanning greater time, recognition performance improved markedly, although recognition scores still dropped as the time between training and test utterances increased. Furui et. al. concluded that a speaker's spectral pattern appeared stable for a period of between 2 to 3 days up to 3 months, but there-after variations were observable.

Li and Hughes [LH74] in their investigation of correlation matrices of continuous speech spectra showed that correlation matrices for speakers stabilised after approximately 30 seconds of speech. This lead them to conclude that a least 30 seconds of continuous speech was necessary to characterise the speech of a single speaker.

Both Atal [Ata74] and Sambur [Sam76] conducted text independent speaker identification experiments via LPC derived parameters. Atal used a random re-ordering of the original sentence and achieved an identification rate of 93% for 10 speakers. Sambur used a set of 6 different sentences (5 for training, 1 for test) and achieved an identification rate of 94% for a population of 21 speakers.

Hollien and Majewski [MH74, HM77] examined the use of Long-Term Spectra to identify a large number of speakers. Fifty American male speakers and 50 Polish male speakers (speaking in Polish) were recorded reading a text. Long-Term Spectra for each was calculated over 30 seconds of speech using the output of a 22 channel filter bank. Mean intra-speaker distances were calculated for both groups (American and Polish) and the Poles were found to have a significantly smaller distance than the American speakers. Speaker identification experiments were then conducted for each speaker group separately and identification rates of 96% for the Polish speakers, and 94% for the American speakers were found.

Markel, Oshida, and Gray [MOJ77] took a somewhat different approach to Furui et. al. [FIS72]. Markel et. al. obtained approximately 15 minutes of speech in a single sitting from each of 4 speakers via an interview situation. F_0, gain, and reflection co-efficients were extracted for the voiced frames over the entire duration of each speaker's utterance. Calculating the mean for each of these parameters over periods of up to 70 seconds the standard deviation across the entire utterance was analysed. The standard deviation of F_0 and the reflection co-efficients dropped rapidly showing the advantage of long-term averaging of these parameters. Inter-speaker to intra-speaker variance ratios [PM64, Wol72] were then calculated, again for means calculated over varying lengths of time. Variance ratios increased significantly as the averaging interval increased. Using a transformed parameter set of 3 and calculating mean references from 70 seconds of speech Markel et. al. were able to identify the speakers with 100% accuracy.

In 1978 Furui [Fur78] reported on the effects of long-term spectral variability on speaker

recognition. Using speech collected over a period of up to 5 years Furui showed that speaker recognition performance decreased markedly as the time interval increased. Furui found it desirable to collect reference samples over a long time period and weight the distance calculations based on the long-term variability of the individual parameters.

In a further investigation of the effects of averaged parameters Markel and Davis [MD78, MD79] used a database of over 36 hours of linguistically unconstrained extemporaneous speech. This data was obtained from 17 speakers (11 male, 6 female) in 10 separate sessions over a 3 month period and was band-limited to frequencies of telephone speech. In an extension of Markel et. al.'s earlier work [MOJ77] Markel and Davis showed that for F_0 and reflection co-efficients the standard deviation decreased and variance increased as the interval over which mean averaging was done increased. Interestingly, the reflection co-efficients with high variance ratios differed between the male and female speakers. Speaker identification experiments testing both the effect of the duration of averaging and that of using training data from several sessions showed the reliance of identification scores upon both these factors. As both the averaging interval, and the number of sessions from which training data were extracted increase, speaker identification rates also increase such that; for training data from two sessions and averaging done over approximately 1 second of speech an identification rate of 52% was achieved, when these were increased to 5 sessions and approximately 50 seconds respectively an identification rate of 92% was achieved.

In their recent investigation Harmegnies and Landercy [HL88] examined the intra-speaker variability of the long-term speech spectrum. 10 French speaking males recorded an 18 second phonetically balanced text 10 times each, as well as reading a passage from a novel. LTS for various sections of the passages were extracted and their correlation compared on an individual speaker basis. It was found that the degree of correlation varied for different speakers and thus Harmegnies and Landercy concluded the degree of LTS variability is speaker-dependent.

Millar [Mil86, Mil88] has examined variations in the long term distributions of the F_0 and energy, and timing of the speech of 33 Australians. The speakers recorded a number of passages, each approximately one minute in length, over a period of three months. Differences between speakers were found for many features examined.

Gelfer et. al. [GMH89] have examined the impact of sample duration and time between example and reference utterances upon a text independent speaker identification system using long-term spectra. It was found that for reference utterance durations of 5 seconds or less that identification rates were highly variable and utterance dependent, while for 10 seconds and above this was not the case; indicating the need for more than 5 seconds of reference utterance to obtain true text independence. Further, non-contemporary utterances were found to greatly degrade verification performance, by a factor of up to 40%.

Large Scale and Practically Motivated Investigations

Most speaker recognition experiments have involved relatively few speakers due to the problem of data collection; and have been recorded under ideal laboratory conditions. These parameters do not reflect the situation of most applications of speaker recognition. An increasing number

of investigations are now addressing these inadequacies in two aspects. Many experimenters have examined telephone quality speech [Fur81a, HM77, INN78, NND89] an area of obvious application, while others have used increasingly large speech databases [MD79, HM77, Dod85].

In 1971 Das and Saleeby [DS71] conducted verification experiments upon a population of 118 males. Over 7000 utterances of the sentence: "Check this terminal please." were used from the 118 speakers. An average "misclassification" rate of 1%, with a 10% "no decision" was obtained.

Ichikawa, Nakajima and Nakata [INN78] conducted a series of speaker verification experiments over telephone lines. Through the application of selective frequency band analysis and self inverse filtering to normalise the effects of transmission distortion verification performance exceeded that of PARCOR co-efficients. With the addition of F_0 frequency verification rates exceeded 95%, and with a single type of telephone handset verification rates of just below 100% were achieved.

Markel and Davis [MD78, MD79] obtained over 36 hours of linguistically unconstrained extemporaneous speech from 17 speakers (11 male, 6 female) via a series of interviews over a 3 month period. Speech was band-limited to telephone frequencies and mean F_0 and reflection co-efficients were used to represent the speakers. A text-independent speaker identification rate of 98.05% was obtained. In a speaker verification experiment an equal error rate of 4.25% was obtained.

Furui [Fur81a] investigates another verification system based on dynamic acoustic features, using telephone quality speech. Time functions of cepstrum co-efficients were represented by orthogonal polynomials over short time intervals. The utterance was then represented by the time function of the first two polynomial co-efficients [LR71]. Using DTW for time alignment several verification experiments were carried out using different data sets in an attempt to investigate several aspects of the process. Error rates varied from approximately 0.3% up to 4% depending upon the data set (10 male, 40 male imposters; 10 female, 40 female imposters and 21 male, 55 imposters). The technique was found to be robust in terms of transmission system variability, and cepstrum co-efficients were found to be better than log-area ratio parameters. Further, verification experiments using only the cepstrum co-efficients were found to have an error rate approximately three times that when the combined cepstrum co-efficients and polynomial co-efficients were used.

In 1985 Doddington [Dod85] reviews the area of speaker recognition. Included in the review is a description of the Texas Instruments voice verification system used to control access to their computer centre, and which had been operating 24 hrs per day for over a decade. This system uses the output of a 14 channel filter bank with simplified DTW for time alignment. An utterance consisting of 4 words randomly selected from a candidate set of 16 is prompted for, and multiple utterances may be used in the verification procedure (operational mean of 1.6 utterances). A gross rejection rate of 0.9% has been measured, together with a casual imposter acceptance rate of 0.7%; showing the validity of speaker verification for *well chosen* commercial tasks. In concluding the review Doddington makes the following points. There has not been the proliferation of speaker recognition systems that was envisaged in the mid 70's

and this is mainly attributed to a lack of robustness, lack of business interest, the problems of a man/machine interface, and the lack of a true need. System performance is highly dependent upon the operating conditions and there is a great need for benchmarking databases so that serious system evaluation may be carried out. Finally, Doddington recommends that research be directed toward high quality recognition over telephone circuits.

Recently Noda has investigated the distinctiveness of a speaker's utterance in the parameter space and its use in forensic speaker recognition [Nod89]. The distinctiveness of an utterance is defined as the distance between its position in the N dimensional parameter space and the mean position for that utterance from all speakers. Using speech from 523 males (20 to 50 speakers from each of 15 different Japanese prefectures) uttered over telephone lines Noda breaks the speakers into 5 equal groups based upon their 'parametric distinctiveness'. Conducting identification experiments, (of the form: nearest speaker, and nearest 5% of speakers) using cepstral co-efficients Noda examined the recognition rates on a distinctiveness basis. He found that the 'more distinct' groups always had higher recognition rates, whether on single vowel, multiple vowel, single word, or multiple word identification and verification and that error rates more than doubled from the most distinct group to the least distinct. These results re-enforced the fact that speaker recognition accuracy is a speaker-dependent phenomenon and therefore the need for large benchmarking databases.

Naik, Netsch and Doddington [NND89] recently investigated speaker verification over phone lines on two speech databases. The first database consisted of 20 speakers recorded using 10 varieties of telephone handsets. The second database consisted of 100 speakers recorded over long distance telephone lines.

Disguised Speech

Several researchers have addressed the area of disguised and/or mimicked speech and its effect upon speaker recognition systems.

In his 1969 investigation of speaker verification using cepstral measurements Luck [Luc69] gave 3 members of his 'imposter' population the opportunity to mimic one of the customers. The 3 imposters were selected on the basis of how closely, as measured by the recognition system, their normal voices matched the customer's (closest match for entire imposter population, worst match, and 10'th best match). The imposters were informed as to how the recognition system worked, and were repeatedly played the customer's code phrase to practice mimicking. The best (closest matching) imposter increased his equal error rate from approximately 10% to 20%. The other mimickers did not increase their equal error rates.

Rekieta and Hair [RH72] report on the success of a professional mimic's attempt to pose as a valid speaker for their speaker verification system based on phoneme spectra. The professional mimic was given recordings of 6 valid speakers, allowed to become familiar with them and then to record their utterance directly after the customer's utterance. Spectra analysis showed an increase in similarity for some speakers and phonemes. However, when verification experiments were performed using features from 5 phonemes, the impersonator was unsuccessful in all mimic attempts.

Lummis and Rosenberg [Lum72] conducted a more exhaustive experiment testing the effect of mimics upon speaker verification performance. Several professional mimics were selected from a large group via audition. These mimics were given intensive training on 8 speakers from Doddington's 40 speaker population [Dod71a]. Doddington's speaker verification system [Dod71a] based on time contours of F_0, energy, and F_1, F_2, F_3 was used to test mimic performance. During training the system gave feedback to the mimics on how close their utterance was to that of the speaker being mimicked. Recordings of the best utterances from the 4 best mimics were then processed by the verification scheme. On an equal error rate criteria, 27% of the best utterances by the best mimics would be accepted. This acceptance rate compares with 1.2% for non-mimicking imposters.

In their 1977 investigation of Long-Term Spectra for speaker identification Hollien and Majewski [HM77] examined the effects of vocal distortion due to task induced stress and vocal disguise upon identification scores. Two speakers sets, one of 50 American males, the other of 50 Polish males were recorded reading a text. Speakers then read the text a second time and mild electric shocks were applied at random intervals to induce stress. Finally, the speakers read the text again and were encouraged to disguise their voice as much as possible but not to put on a foreign accent or to whisper. Speaker identification rates dropped from the order of 100% for normal voice, to 92% for stressed, and 20% for disguised. This led Hollien and Majewski to conclude that while LTS appeared viable for detecting speaker identity under stressed and normal voice, but that it was incapable of determining identity when the voice was disguised.

Dynamic

Very few researchers have compared the utility of parameters for speaker recognition based upon their inherent property of time-variability. The terms dynamic and static, as used by different investigators, have different meanings. However for this review dynamic is defined to pertain to the time variability of a parameter, while static is defined to the time invariant properties (e.g., mean).

Saito and Furui [SF78] describe a novel approach to examing encoded speaker identity in the dynamics of speech spectra. Nine male speakers uttering two isolated words were recorded over a period of 2 years and Log Area Ratio, and F_0 were extracted. Standard DTW recognition experiments were conducted and an error rate of 2% was reported. The warp paths calculated for these experiments were examined on an intra-speaker versus inter-speaker basis. It was found that for intra-speaker comparisons there appeared regions on the warp path that remained close to the diagonal joining the start and end of both vectors, while for inter-speaker comparisons there appeared no such regions. A measure was defined to count the number of warp transitions that had come from the diagonally previous position. Using this measure and both words a speaker recognition rate of 63% was achieved. When this measure was added to that of the standard DTW experiment the error rate was approximately half that without the warp path measure. Saito and Furui drew the conclusion that "... the rate of similarity in a specific region is useful as a supplementary measure for talker recognition, although it is insufficient to be used as an independent measure for talker recognition."

Furui [Fur81b] compared the performance of dynamic and static approaches to speaker recognition. Two Japanese words were repeated by 9 male speakers over a period of 7 years. Log area ratios and F_0 were extracted. The statistical approach used correlation matrices derived from the mean value of each parameter. The dynamic approach compared time functions of the parameters via dynamic time warping. The two approaches were compared under a number of conditions involving training set data and duration between training and test utterances. The statistical approach was generally found to perform better, but needed a larger, more representative training set than did the dynamic scheme. When both schemes were combined the error rate was less than half that of either one individually.

Soong and Rosenberg [SR88] compared instantaneous and transitional schemes for speaker recognition over phone lines. 10 speakers (5 male, 5 female) each recording 200 utterances of isolated digits (20 per digit) over a 2 month period. Both instantaneous and transitional Vector Quantisation codebooks were made for each speaker. Soong and Rosenberg found that the instantaneous approach generally yielded higher results, but the transitional based scheme was more resistant to channel noise. Finally, they found that instantaneous and transitional schemes complemented each other and appeared relatively uncorrelated.

Bernasconi [Ber90] has also compared instantaneous and transitional measures of cepstral coefficients for text-dependent speaker verification. For a speaker population of 22 males reading phonetically balanced sentences over a period of 4 months it was found that the transitional scheme was practically as good as the instantaneous based scheme, though combined there were no further improvements. Error rates below 0.1% were reported.

Probabilistic Approach

Many speaker recognition systems implicitly assume an underlying Gaussian probability distribution [Ata74] and use distance metrics based on this assumption. Schwartz, Roucos and Berouti [SRB82] were the first to investigate the actual use of Probability Distribution Functions (pdf's) for speaker identification. Probabilistic classifiers compute the conditional joint probability that the successive test vectors $\tilde{x}$ were produced by speaker i (successive vectors are assumed to be independent). The speaker with the highest probability estimate is then selected. Using speech consisting of a read text of over a minute's duration, from 21 male speakers, three different probabilistic classifiers together with a Mahalanobis distance [Ata74] were examined and compared. The pdf's tested were Gaussian (gpdf), Gaussian with the addition of clipping (gdpf+c), and a non-parametric pdf (npdf) based on a k-nearest neighbour scheme. Test and reference vectors consisted of Log Area Ration (LAR) co-efficients. Performance was investigated under a number of conditions including the number of LAR Parameters used, the effects of noise and the duration of training. The pdf based schemes yielded higher identification rates under all conditions than the Mahalanobis distance based scheme. Of the pdf's the non-parametric scheme generally performed the best.

In their 1983 follow-up paper [WKK+83] Wolf et. al. investigated the same four recognition schemes' (namely Mahalanobis distance, Gaussian pdf, Gaussian pdf with the addition of clipping, and a non-parametric pdf) performance using speech recorded over radio channels. Over

30 seconds of speech from each of 19 speakers was obtained. Each radio transmission was of approximately 2 seconds duration and the noise characteristics of the channels were highly variable. Further, it was reported that: "The speakers vary from being calm and talking "normally" to being very excited and yelling." thus introducing an extra complication of speaker state (no further information is given regarding this variance, nor is the exact source of the speech data detailed). Speaker identification experiments for the four schemes tested performance under conditions of variable training length, test and training data from same and different sessions, and the number of Log Area Ratio co-efficients used. Interestingly, the Gaussian pdf (with and without clipping) yielded the highest identification rates as opposed to the Non-parametric pdfs superior performance in the earlier experiment [SRB82]. Generally identification rates dropped by 20% to 30% from those achieved using laboratory speech.

Matsumoto [Mat89] describes a text-independent speaker identification scheme for a 10 speaker population. Speaker's utterances in the k phonetic subspaces are modelled by gaussian pdfs and joint probabilities are calculated for an active utterance. Using cepstrum co-efficients and F_0 an identification rate of 90% for 0.5 seconds, ranging up to 100% for 1.4 seconds of speech is achieved.

More recently Rose and Reynolds [RLS90] have reported on a gaussian based (iterative maximum likelihood) approach to text independent speaker identification using acoustic segmentation. For a 12 speaker database of conversational speech and the use of Mel frequency cepstra an identification rate of 89% was achieved with a 1 second duration test utterance.

Reviews and Evaluations

Few comparisons or evaluations of the different approaches to automatic speaker recognition have been made.

In 1976 Atal and Rosenberg produced two excellent reviews of speaker recognition and speaker verification [Ata76, Ros76]. Atal said of speaker recognition:- "...motivation for speaker recognition research came from a desire to isolate speaker-dependent parameters of speech from message dependent ones. So far little progress has been made in this respect. ...Almost every acoustic parameter derived from speech is speaker-dependent to some extent." Both Atal and Rosenberg emphasised practical, commercially viable motivated research. Rosenberg highlighted the need to test recognition techniques using large databases of speakers and described both the Texas Instruments Entry System (a system still working a decade later [Dod85]) and the Bell Labs Systems [Lum73, Ros76].

More recently, both O'Shaughnessy [O'S86] and Doddington [Dod85] have provided excellent overviews and introductions to the area of speaker recognition, though with the subsequent application of both Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) to speaker recognition there is need for a more current review.

Recent Developments

More recently several newer approaches, such as Hidden Markov Models or Artificial Neural Networks [Lip87] have been taken to the problem of automatic speaker recognition.

In a recent investigation of verification performance using speech recorded over phone lines Naik, Netsch, and Doddington [NND89] investigated, the performance of DTW and HMM speaker verification algorithms together with speaker discriminant modelling. Two databases were investigated, one with 20 speakers using 10 different varieties of telephone handsets, and the other with 100 speakers recorded over long distance telephone lines. At a fixed true speaker rejection rate of 1.5% using DTW the rate of imposter acceptances dropped from 30% to 23.6% upon the second database. Using the handset database it was shown that the HMM approach yielded an equal error rate of 2.3% as opposed to an error rate of 6.2% for DTW. Naik et. al. concluded that speaker discriminating modelling gives a marked improvement in verification results and that word level HMM modelling is superior to whole-phrase DTW.

In their recent paper Xu, Oglesby and Mason [XOM89] investigated the application of perceptually-based parameters for speaker identification. With a speaker set of 5 males and 5 females a series of VQ based speaker identification experiments were carried out using LPC derived cepstral co-efficients. Utterances were pseudo text-independent with any of the ten digits being used. LPC parameters were perceptually weighted based upon 3 elements of psychoacoustic knowledge: - critical band (Bark scale) integration, equal-loudness pre-emphasis, and intensity loudness transformation. Perceptually-weighted LPC (PLP) consistently achieved higher speaker identification accuracy than the standard LPC, and speaker-specific information was found in the higher order LPC and PLP co-efficients. These results show the benefits of combining perceptual knowledge with automatic speaker verification systems.

Several researchers such as Lovell and Tsoi [LT90], Templeton and Guillemin [TG90], Bennani et. al. [BSG90], and Oglesby and Mason [OM90] have begun investigating the application of Artificial Neural Networks to speaker recognition. To date such experiments have been limited in scope; Lovell and Tsoi using a single isolated word from two speakers, and Templeton and Guillemin examining eleven vowels in [h-d] context from nine speakers. Bennani et. al. report identification rates of 97% for a speaker population of 10 and sentence utterances; while Oglesby and Mason describe a feedforward ANN with 10 speaker and digit utterances with an error rate of 8%. However, given the wide attention that Artificial Neural Networks (ANNs) are receiving in many sectors of the scientific community there appears little doubt that ANNs will receive far more in depth examination with regard to speaker recognition.

2.2.2 Emotions and Stress

Analytic research upon acoustic correlates of speaker emotions has been carried out from an early date.

In 1935 Skinner [Ski35] investigated the parameters of a sustained /a/ vowel from 19 speakers for indicators of happiness or sadness. Speakers read excerpts from books and were played classical music intended to induce the feelings of happiness or sadness. At the end of the stimuli presentation speakers phonated a sustained /a/. Acoustic analysis revealed that F_0 increased appreciability for happiness while for sadness it remained approximately at the same level as a 'neutral' phonation. Similarly, intensity increased for happiness and decreased for sadness.

Fairbanks and Pronovost [FP39] and Fairbanks and Hoaglin [FH41] investigated the pitch and temporal characteristics of 6 male actors during the expression of the emotions contempt, anger, fear, grief, and indifference. Fairbanks and Pronovost found significant differences in mean F_0, (e.g. indifference 108Hz, fear 254Hz) F_0 range, rate of F_0 change, and other measures of F_0 dynamics between the expression of the different emotions. Fairbanks and Hoaglin investigated such temporal characteristics as total speaking time, rate of phonation, duration and number of pauses; and found that these also differed markedly between the different vocal expressions of emotion.

Hecker, Stevens, Von Bismarck and Williams [HSBW68] used a combined reading and adding task under time constraints to investigate the manifestation of task induced stress in utterances. Intensity and F_0 parameters were extracted. Approximately half of the speakers examined showed significant differences between control parameter levels and parameter levels during stress. No common change in parameters was found between speakers; though for a given speaker parameters tended to change in a specific direction under stress.

Williams and Stevens [WS72] examined the acoustic correlates of 'acted' and 'real-life' emotions. Three method actors were used to portray the emotions: fear, neutral, anger and sorrow. Mean F_0, F_0 dynamics, formant frequencies and rate of articulation all varied between the emotional states. 'Real-life' emotional data was obtained via recordings of the radio announcer commentary on the approach and subsequent destruction of the Zeppelin Hindenburg at Lakehurst, New Jersey, in 1937. Spectrographic analysis of the announcer's speech prior to, and after the disaster show marked differences in mean F_0 (increased 30Hz), F_0 range and irregularities in the F_0 contour. Williams and Stevens concluded: "The aspect of the speech signal that appears to provide the clearest indication of the emotional state of a talker is the contour of F_0 vs time."

Both Ekman, Freisen and Scherer [EFS76] and Streeter, Kraus, Geller, Olson and Apple [SKG+77] have examined the acoustic characteristics of voice during attempted deception. Speech samples in both studies were elicited through an interview situation in which the subject attempted to deceive the interviewer on certain subjects. Both studies found that there was a significant increase in mean F_0 for deceitful utterances.

Cosmides [Cos83] attempted to determine whether acoustic changes for different emotions were invariant across speakers. Cosmides measured various static and dynamic parameters of F_0 and amplitude, together with timing; for utterances from 11 speakers (5 female, 6 male) expressing 10 different emotions. Based on a scoring system for parameter clustering Cosmides found that the results supported a common 'acoustic configuration' to express emotions independent of speakers.

More recently Jimenez et. al. [JPL+85] have investigated the ability to differentiate between emotions based on acoustic parameters. Repetitions of a Spanish sentence conveying 5 different emotional states were acoustically analysed. F_0, temporal, and energy parameters were extracted and several were found to differ significantly between emotions. It was found that for each emotion at least 3 parameters differed significantly from neutral value; but no one acoustic parameter sufficed to differentiate between any 2 emotions.

2.2.3 Sex

Many investigators [Wea24, MT34, Cow36, HP69, Lin73, HT78, Sto81, HHP88] have examined the mean fundamental frequency of male and female voices. Mean ranges for adult male and female F_0 typically have been found to span 80 to 150Hz and 200-300Hz respectively; with female pitch generally being regarded as being approximately one octave higher than male pitch.

Peterson and Barney [PB52] examined formants of vowels extracted from [h-d] context. Speakers were adult male and female, children, and some English 2'nd language speakers. Significant shifts to higher frequencies were found for all vowels when comparing female speakers' formants with those of males.

Brend [Bre71] in an initial exploratory study found differences in F_0 contour patterns between adult American male and female speakers. No systematic investigation was performed but Brend shows several sentences where typical male and female intonation patterns differed. Brend observed that male speakers tended to avoid certain intonation patterns including the use of high pitch and final patterns that don't finish at the lowest pitch level.

Millar [Mil86, Mil88] examined the long term variability of acoustic features for encoding identity and sex information. A total of 33 speakers, 15 male, 18 female, were recorded reading 5 different passages over a 3 month duration. Distributions, means, and durations of F_0, energy, and voicing were examined. Significant sex differences were noted for voicing duration, energy distribution, and F_0 measures. Currently it remains unclear whether the voicing differences are an artifact of the extraction process or not [Mil91].

Titze [Tit89] examined the physiological and acoustic differences between male and female speakers. Using data obtained by other investigations Titze found a prime scale factor of 1.6 between male and female mean F_0 based upon the length of the vocal folds.

Price [Pri89] examined the mono-syllabic utterances of 4 male and 4 female speakers. Spectrally Price found that female speakers, regardless of voice quality, had lower high frequency energy in the middle of vowels than men and that there were temporal differences between the two groups.

In an investigation of the voice quality breathiness Klatt and Klatt [KK90] recorded speech from 10 female and 6 male speakers. A number of features of the speech were extracted including aspiration noise in the F_0 region and amplitude and bandwidth of formants, a number of which showed sex specific differences.

2.2.4 Age

Relatively few researchers have sought to objectively measure changes in the acoustics of the speech wave due to advancing age.

In 1966 Ptacek, Sander, Maloney and Jackson [PSMJ66] compared the vocal performance of younger adult speakers (under 40) with older speakers (over 65). Over 50 speakers of each sex were studied and elderly speakers showed reduced scores in maximum F_0 range, speech rate, maximum vowel intensity and maximum vowel duration as compared with the younger

speakers.

Wilcox and Horii [WH80] examined the vocal jitter[3] of young and elderly males for sustained vowels. Elderly males were found to have a significantly higher mean jitter than younger speakers.

Ramig and Ringel [RR83] investigated the effects of physiological aging upon various acoustic parameters. A group of 48 males were divided evenly amongst 3 age categories (25-35, 45-55, 65-75) and 2 levels of health (good, and poor) based upon physical testing. Significant differences in parameter values were not only found between age groups but also between speakers of good and poor health within the same age group. Ramig and Ringel concluded that physiological age more than chronological age accounted for acoustic parameter differences.

Linville and Fisher [LF85b] and Linville [Lin88] have performed similar investigates of young and elderly female voices. Results showed that F_0 stability (as measure by F_0 standard deviation and jitter) decreased with age, and that there was significant lowering of F_1 frequency with age, signalling changes in vocal tract anatomy.

Recently Rastatter and Jaques [RJ90] examined the formant structure of the speech of young and old speakers. Twenty young speakers, 10 male, 10 female aged 20-22; and 20 old speakers, 10 male, 10 female aged 72-76 were recorded phonating sustained vowels. Inspection of spectrogram was used to visually determine F_1 and F_2 frequency values. Shifts in formant frequencies were found between the age groups but direction and degree of shift were found to be speaker sex dependent.

2.2.5 'Race', Dialect and Accent

Speech differences due to 'race' or dialect span the acoustic level of speech from segmentals through prosodics and suprasegmentals; and extended beyond to syntax and semantics. It is impossible, within the scope of this work, to address these differences more than in passing; or without particular regard to English and its dialects (particularly Australian English).

Bernard [Ber67] examined the formant structure of Australian vowels with regards to the three dialects of Australian English: Cultivated, General, and Broad. Speech samples from approximately 170 different Australians who covered the dialect spectrum of Australian English were obtained. Vowels were extracted from the ubiquitous [h-d] context and it was found that there were significant differences in formant frequencies and formant transitions across the three dialects.

Other researchers such as Flege [FH84] have examined segmental differences between native and non-native language speakers. Differences in formant structure are apparent even for phonemes which sound superficially similar, and there appears to be an inverse relationship between the presence of a 'similar' phoneme in the speaker's first language to their ability to produce a new phoneme accurately.

Many researchers [Ada71, Eng71, Bus71, FBS84] have examined prosodic differences between native and non-native speakers of English dialects. Adams showed that non-native speakers of Australian English were variable in their positioning of stress and characteristically

[3] cycle to cycle perturbations in F_0, a measure of F_0 stability.

had a higher ratio of pause intervals to total duration. Bush contrasted British, Indian, and American English and found marked differences in segmental durations and durational ratios.

Wagner [Wag78] details the application of a learning technique to the examination of dialect differences. Six sentences from 5 Australian and 5 North American male speakers were automatically analysed and broken into a series of speech events. Static and dynamic measures of features, such as intensity, F_0, and formant frequencies, were extracted at these events and compared on the basis of dialect. Differences were found in dynamic range, energy difference between vowels and fricatives, nasal spectral balance, F_3 range for vowels, F_0 changes, and duration of stop-vowel transitions.

Barry et. al. [BHN89] describe a dialect normalisation process in order to aid vowel recognition. Fifty eight speakers from 4 regional dialects - North American, Scottish, North English, and South English - recorded four calibration sentences. Stressed vowels were then automatically detected and scored as to dialect based on the first three formants, and speakers were assigned a dialect based on the total dialect results for the four sentences. Of the 58 speakers 43 were classified correctly as to dialect and the vowel recognition performance increased significantly after mapping of the reference vectors based upon the dialect assignment results.

Ingram and Pittam [IP86, PI90] have conducted a series of experiments examining the accent change of native Vietnamese living in Australia. Two separate experiments; one with a group of 10 immigrant school children aged 6 - 12, and another with 2 pairs of male/female siblings were conducted. Subjects were recorded initially then up to a period of 10 to 18 months later. Vowel formant trajectories were examined in the school children experiment while connected speech processes were examined for the siblings. Substantial changes or shifts towards the formant structure of Australian English were found and scoring of connected speech events improved.

2.2.6 Other

Several researchers have examined the effect of drugs, primarily alcohol, upon the speech of subjects. Trojan and Kryspin-Exner [TKE68] found that the speech of subjects under the influence of alcohol underwent a general linguistic dissolution as signified by increased repetitions, substitutions and phonetic errors. Changes in prosodic parameters such as F_0 and timing were also observed but were not consistent across subjects. Sobell and Sobell [SS72] conducted a similar study upon 16 male alcoholics and found that disfluency increased under the influence of alcohol. In a different approach Beam, Grant, and Mecham examined the effects of long-term alcoholism upon communication ability. Of the 15 subjects studied 14 had at least one aspect of unacceptable voice quality such as severe hoarseness or pitch deviation. Sobell, Sobell, and Coleman [SSC82] found that amplitude dropped and total speech time increased for nonalcoholics under the influence of alcohol. Recently Klingholz, Penning and Liebhart [KPL88] have reported on investigations of the detection of low-level intoxication from the speech signal. Systematic variation in the distribution of F_0 and the signal-to-noise ratio was found under the influence of alcohol. However, Klingholz et. al. concluded that the technique was not currently viable due to cost and inter-speaker variability factors.

In an interesting examination of voice quality Kuwabara and Ohgushi [KO84] investigated

the acoustic characteristics of professional male announcers' utterances. Analysis of professional announcers' and laymens' utterances showed several differences. Announcers were found to have a highly time-varying dynamic pattern of pitch and formant frequencies, together with a wider dynamic range. Differences in mean formant values were also found for certain vowels and the spectral envelope showed a characteristic peak between 3kHz and 4kHz not present in laymens' speech.

2.3 Forensic Speaker Recognition

An interesting combination of both analytic (objective) and perceptual (subjective) approaches to speaker recognition occurs in the area of forensic speaker identification via visual comparison of speech spectrograms[4]. Forensic speaker identification based on speech spectrograms has been an area of controversy in the speech science community for over 20 years. The technique was originated during World War II by Gray and Kopp and was used, in conjunction with goniometry[5], by the United States military to plot movements of German divisions via identification of their radio operators [Tos79] (apparently German divisions often used the same radio operator for communications).

In 1962 Kersta [Ker62] presented a paper in Nature entitled "Voiceprint Identification". Kersta described a series of speaker identification experiments based on visual comparison of speech spectrograms. Spectrograms of 10 words frequently used in telephone conversations ("the, to, and, me, on, is, you, I, it, a") were generated for a population of male speakers. Together with standard spectrograms, contour spectrograms (named "voiceprints") were generated where amplitude levels were represented by contours and each inner progression from contour to contour marked a doubling in amplitude. These contour spectrograms superficially resemble fingerprints and Kersta states:"Closely analogous to fingerprint identification, which uses the unique features found in people's fingerprints, voiceprint identification uses the unique features found in their utterances." 8 female high school students were given 1 week training in spectrogram matching and worked in panels of 2 on a series of identification tasks. Two forms of identification tasks were conducted. In one, 4 spectrograms of discrete utterances of one of the 10 words for 5, 9, or 12 speakers were presented to the subjects. The subjects were then required to cluster the spectrograms on a by-speaker basis. Mean error rates for these tests using conventional bar spectrograms ranged from 0.35% to 1.0%, and from 0.37% to 1.5% for contour spectrograms. The second form of experiment involved the identification of an unknown speaker against a set of known speakers. For a given utterance the subjects were presented a reference spectrogram from each of 9 to 15 speakers and were required to match an unknown spectrogram with that of one belonging to one of the known speakers. Kersta reported that for the 9 known speaker case an error rate of 1% was obtained. Kersta concluded :-"It is my opinion, however, that identifiable uniqueness does exist in each voice, and that masking, disguising, or distorting the voice will not defeat identification if the speech is intelligible."

[4] A means of representing the spectral characteristics of an utterance. Frequency is plotted against time (time the abscissa) with amplitude represented by density/darkness.

[5] A method of determining the source of radio signals.

In the years following Kersta's claim several firms began marketing "voiceprint" machines and voiceprints were admitted as evidence by several courts in the USA [Ano68]. Little investigation of factors governing intra-speaker variability was carried out and many professionals in the speech area became concerned [BCJ+69, Ano68].

In 1968 Stevens, Williams, Carbonell and Woods [SWCW68] conducted experiments comparing speaker identification based on auditory and spectrographic (visual) presentations of speech. Six subjects (3 male, 3 female) were presented with sets of 8 known speakers and one unknown speaker drawn from a population of 24 males. Both closed[6] and open[7] experiments were conducted. For both representations of the utterances the subjects had constant access to the reference material (i.e., for auditory tasks subjects could repeatedly play back any of the reference utterances). The subjects were initially untrained but the author's report that mean errors stabilised after approximately 4 hours of experimentation and a mean error rate of 6% for auditory presentation and 21% for visual presentation was obtained in the closed experiments. For the open experiments auditory presentation identification rates were approximately 90% while visual presentation identification rates varied from 60% to 70%.

In 1970 Bolt, Cooper, David, Denes, Pickett and Stevens presented a report on the reliability of speaker identification by speech spectrograms for legal purposes. This report was initiated by the Technical Committee on Speech Communication of the Acoustical Society of America after growing concerns about the use of speaker identification by speech spectrogram for legal purposes, without adequate supporting scientific evidence. The authors raised several questions including :- (1) when spectrograms are alike does this mean "same speaker" or merely "same word" (2) Would irrelevant similarities in spectrograms mislead laypeople (e.g. juries) (3) how permanent are spectrographic patterns (4) how unique or distinctive are the spectrographic patterns of the individual and (5) can spectrograms be disguised or faked. The authors raised the point that the speech signal carries several sub-messages (e.g. identity, mode of speaking, mood) all of which affect the parameters of speech (and hence, possibly spectrograms) in a complex and not fully understood manner. In addressing Kersta's analogy to fingerprinting the authors state that the differences between the two techniques exceed the similarities. With regard to previous experimental results the authors attribute differences in results to dependence upon: experimental test procedure, experience and training of the observer, speaking conditions the samples were obtained under, and instrumentation; and further state that none of the experiments to date have matched actual forensic applications. The authors state: "We concluded that the available results are inadequate to establish the reliability of voice identification by spectrograms."

Endres, Bambach and Flösser [End71] reported on the effects of age, voice disguise and voice imitation upon voice spectrograms. The formant and fundamental frequency structure of several phonemes were investigated. For the investigation of the effects of age, speech material was obtained from 6 speakers (4 male, 2 female) over the period of 13 to 15 years. Formant frequencies decreased markedly, as did mean F_0, while the range of F_0 decreased and became more centralised. For disguised speech Endres et. al. found that F_1 remained fairly stable,

[6] The unknown speaker is definitely one of the known speakers.
[7] The unknown speaker may or may not be one of the known speakers.

but the other formants increased or decreased markedly in frequency and some could not be traced. Investigation of the voice of imitators showed that while they were capable of misleading listeners their formant structure differed from that of the speaker being imitated.

In 1972 Tosi, Oyer, Lashbrook, Pedrey, Nocole and Nash released results of a 2 year investigation of speaker identification by speech spectrogram [TOP+71, TON72, TOL+72]. A population of 250 male speakers, which were randomly selected from a population of 25,000, was used; with a total of 34,996 trials of identity being conducted by 29 examiners who had received a minimum of a month's training. All trials used a known set of 40 speakers and one unknown utterance. Experimental variables examined included closed and open trails, contemporary and non-contemporary spectrograms, six or nine clue words, and words spoken in isolation, fixed context and random context; all of which were considered relevant to forensic applications (through neither mimicking or vocal disguise were examined). Examiners were forced to make a positive decision for each trial. A mean error rate across all trial conditions of 6% false identifications and 13% false eliminations was achieved. Based on confidence ratings of each trial, if the examiner had not been forced to make a positive decision then 26% of the trials would have been "non-decision"; and an error rate of 2% false identification and 5% false elimination would have been found. Based on these results Tosi et. al. stated that the results confirmed Kersta's data [Ker62] and that the technique could yield a negligible error if: (1) examiners were properly trained in phonetics, spectrography and speech science in addition to completing a 2 year supervised apprenticeship, and (2) examiners made no-positive decision if they weren't absolutely certain.

Bolt et. al. [BCJ+73] addressed the results and statements of Tosi et. al. [TON72]. Bolt et. al. noted the doubling in error rates when the population of speakers was increased from 10 to 40, together with the increase in error for the change in experimental condition of contemporary to non-contemporary spectrograms, and words spoken in isolation to word in-context. Further, Bolt et. al. stated that they believed the error rates were artificially low due to the controlled laboratory conditions, and that the understanding of the relationship between voice characteristics and spectrographic features was still poor. Bolt et. al. stated: "...But for less-than-ideal conditions encountered in forensic situations, the indications are that the probability of error will increase substantially." and "...We wish only to point out that present methods for such use lack an adequate scientific basis for estimating reliability in many practical situations ...".

Black et. al. [BLN+73] replied to Bolt et. al. [BCJ+73] pointing out their lack of personal experience in spectrographic identification and stated: "It is our contention that opinions based on feelings other than in actual experience are of little value, irrespective of the scientific authority of those who produce such an opinion." Black et. al. claim that Bolt and associates disregarded crucial facts that interacted with the decision process when professional full-time examiners are employed; such as professional training and responsibility, the possibility to make no-decision, the number of speech samples used, and the amount of time allowed to perform each comparison. In addressing the possibility of increased error as conditions departed from those of the laboratory Black et. al. claimed that the percentage of no-decisions would increase,

not the percentage of errors.

Hazen [Haz73] investigated the effects upon spectrographic speaker identification from different contexts in spontaneous speech. Five keywords were used and identification tasks consisted of closed and open trials with a known population of 50 speakers and a single unknown speaker. Error rates as high as 83% were reported and Hazen concluded that given the conditions of the study accurate identification of speakers by visual comparisons of spectrograms is not possible.

Reich, Moll and Curtis [RMC76] examined the effects of vocal disguise upon spectrographic speaker identification. Speakers produced 2 sentences under 6 different voice conditions: (1) normal voice, (2) old-age disguise, (3) hoarse disguise, (4) hypernasal disguise, (5) slow-rate disguise and (6) disguise of the speaker's choice. Identification rate for undisguised voices was a low 56.67%, but this dropped as low as 30% for hoarse disguise and 21.67% for the disguise of the speaker's choice. Reich et. al. concluded: "These experimental data obviously contradict Kersta's (1962c) claim that spectrographic speaker identification is essentially unaffected by attempts at disguising one's voice."

In 1979 Tosi published his book entitled *Voice Identification: Theory and Legal Practice* [Tos79]. Tosi presents an overview of research results on voice identification and a good description of the actual forensic practice of voice identification, together with a history of the legal cases involving voice identification. Addressing the adverse results of voice identification obtained by Reich, Hazen and others under conditions of vocal disguise, poor transmission channel etc. Tosi states that a trained examiner would merely make no-decision:- "In such a case, different samples can lead only to a no-opinion decision or at the worst a false elimination, but obviously they cannot lead to a false identification unless the examiner is not properly trained."

In 1980 Koenig, a Special Agent in the Technical Services Division of the American Federal Bureau of Investigation (F.B.I.) reported on an American National Academy of Sciences study of voice identification [Koe80a, Koe80b]. The investigation was instigated at the request of the F.B.I. in 1976 and consisted of a committee of 8 independent experts. In 1979 the report, entitled *On the Theory and Practice of Voice identification* was released and addressed the following points.

- Some information upon the identity of an individual is obtainable through listening and observing speech spectrograms.
- Spectrograms are fundamentally different to fingerprints.
- Investigation of error performance to date have only examined relatively few conditions and combinations found in real-life; and do not constitute an adequate basis for determining its reliability or acceptability from a legal standpoint in forensic applications.

More recently Koenig [Koe86] has reported on the results of 2000 voice identification comparisons, using spectrographic voice identification techniques, conducted by the F.B.I. over the previous 15 years. Koenig reports that the service is provided to requesting law enforcement agencies and is for investigative purposes only, as no expert evidence will be given in court by

the examiner. All examiners have a minimum of 2 years experience in spectrographic identification, had completed over 100 comparison in actual cases, completed a course in spectrographic reading and passed a yearly hearing test. Koenig reports that a combination of auditory examination and visual inspection of spectrograms is used. Spectral pattern matching comparisons are made by comparing beginning, mean and end format frequencies, formant shaping, pitch, timing and the other parameters of each word. Auditory examination is performed by playing both samples of a word simultaneously and switching rapidly from one to the other while listening through headphones. Koenig reports that of the 2000 comparisons, 1304 were rendered no decision, there was 1 case of false identification and 2 cases of false elimination (data based on outcome of cases), with 378 eliminations and 318 identification. This data corresponds to a decision rate of 34.8% with a 0.31% false identification rate and a 0.53% false elimination rate.

Shipp, Doherty, and Hollien [SDH87] questioned several statements made by Koenig. In particular they questioned the exact means of spectrogram analysis, the qualification of the examiners, and the means by which the F.B.I. determined their error rate. Shipp et. al. raised the question of independent testing of the examiners, and considered that court case outcomes and replies from organisations that used the service could not be considered as totally accurate scientific data.

In their reply Koenig, Ritenmour, Kohus and Kelly [KDVRKK87] made the following points. The F.B.I. does not consider voice identification to be a positive means of identification, nor is it error free. Court decisions are not a perfect criteria for determining the accuracy of the scheme, but it is the best source of data available. Koenig et. al. concluded that voice identification is a valuable investigative aid.

2.4 Conclusions, and Implications of Literature Review

Based on this review of the literature several major observations may be made.

The area of research into "speaker characteristics" is extremely broad and a large number of papers have been published in the area. The degree of investigation of the different characteristics varies greatly, for example automatic speaker recognition has received more research effort than all other areas combined.

Speech in general, and speaker characteristics more specifically, is a multi-disciplinary research area, and researchers come from, amongst others, such diverse backgrounds as medicine, engineering, computing, psychology, linguistics and physics. Such differences of expertise have naturally led to different methods and forms of experimentation, both in terms of the speech data used and the types of "trials" conducted.

Together with these variabilities based on experimenters' expertise are other variations between investigations including:- the size of the investigation (e.g., the number of speakers used in a speaker identification experiment), source of speech data (e.g., acted emotional speech versus speech obtained from a real-life emotional situation), degree of control over phonetic content and other (non-investigated) speaker characteristics, type of trial conducted (e.g., speaker verification versus speaker identification), and acoustic parameters examined. All such

variations between investigations make unification and comparison of results difficult at best and often impossible.

Much research has been carried out into the acoustic correlates of speaker characteristics taking either the perceptual or analytic approach. The choice of which of these two methodologies to use in investigating a specific speaker characteristic again appears to depend upon multiple factors such as the aim of the research and its potential application (e.g., a speech synthesis application would be oriented toward human perception of the characteristic(s)), the speaker characteristic being investigated, background of the researcher(s), and speech data considerations. However, relatively few researchers (e.g., Lass, Linville, Pruzansky) have bridged these two approaches and sought to combine or compare results from the two methodologies. Such a combination would appear to be extremely useful as, amongst other things, it facilitates the direct comparison of human and machine speaker characteristic recognition performance, shows potential areas of improvement or further experimentation in either approach; and in general provides a more thorough and rigorous investigation of the speaker characteristic in question.

The determination of the acoustic correlates of speaker characteristics was the major objective of this review. In answer to the question: *Which acoustic parameters are correlated to speaker characteristics?* the short answer appears to be *all of them!* Again, of course, many factors influence this correlation, such as the speaker characteristic itself, but it appears safe to say that for any acoustic parameter one or more speaker characteristics may influence the value of that parameter, and indeed, for several acoustic parameters, all speaker characteristics may exert some influence over the value of that parameter.

The degree to which a speaker characteristic affects or alters an acoustic parameter appears to be dependent both on the acoustic parameter and on the speaker characteristic. Thus, for example, it appears that speaker sex has a major influence upon mean fundamental frequency, but far less influence upon segmental durations. Similarly, the segmental formant structure appears to vary more between different dialects of Australian English than between sober and intoxicated speakers of the same dialect (other speaker characteristics remaining fixed).

For all the speaker characteristics examined, prosodic and suprasegmental features appear to be consistent and strong indicators of the characteristic. Unfortunately, in many cases examiners have failed or been unable to to quantify differences in such parameters in more than a crude or very limited fashion. There is a need, therefore, to apply existing techniques to help quantify these observed differences, and to develop new techniques, based on the unique features of these parameters, to measure these parameters and their variance.

A further significant observation possible from the review is the dependency of findings, of the acoustic correlates of speaker characteristics, upon the speaker and listener sets used in the investigation. Many researchers have shown that listeners vary in their ability to identify speaker characteristics based on their experience, background, language skills, and innate ability. Of still further significance is the fact that different listeners appear to utilise different acoustic parameters, or at least to apply different weightings to the parameters in their perceptual processes. Further, direct analysis has shown that the manifestation of speaker characteristics

in the acoustics of speech is to some degree speaker dependent. Not only does the degree of manifestation in a given acoustic parameter differ between speakers, but also the parameters most affected and the type of change in a parameter have been found to differ between speakers. Again, unfortunately, there has been no in depth study of such factors, and little is known of their potential effect upon the results of an experiment.

The following chapter, Chapter 3, will describe the broad details of the experiments to be conducted, and link that approach taken back to the literature reviewed in this chapter.

Chapter 3

Motivation and Approach

This chapter identifies the key questions arising from the literature review which will be explored further in this thesis and provide the basic outlines for the methods to be used.

3.1 Motivation

Section 2.4 (page 39) of the literature review illustrates several areas within which continued research upon the acoustic correlates of speaker characteristics may be carried out.

The division of research into that taking an analytic ("objective") approach, and a perceptual ("subjective") approach, highlights one area deserving of further investigation. While most experimenters have "specialised" in one approach or the other very few, e.g., Lass, Linville, Pruzansky, Xu [LMK78, LF85a, Pru63, XOM89], have sought to utilise or combine the two methods.

An analytical approach generally is motivated by machine performance, such as recognition systems, and is most often performed when the speaker characteristics are simple to measure and quantify. The perceptual approach utilises human perception of speech to study speaker characteristics, an approach often used when the speaker characteristics are hard to objectively quantify, or when human perceptual performance is to be assessed, and finding application in such areas as speech synthesis. The two techniques taken together should complement each other well, resulting in a more thorough examination of the topic and possibly yielding new insights based upon the combination of the two schemes.

As pointed out at the conclusion of the literature review it appears that most, if not all, speaker characteristics are correlated to some extent with most, if not all, acoustic parameters. Such a result is not unexpected but leaves the question of which parameters to investigate. One division of parameters is into prosodics, such as F_0 contours, and segmentals such as steady state vowel formant frequencies. Perusal of the literature review shows that by far and away the most analytic effort has been devoted to segmentals while the balance is far more even for perceptually based experiments. This deficit in research upon prosodics has several causes; however the two major reasons appear to be:- 1) it is difficult to quantify, measure, and compare prosodic parameters; 2) most research upon speaker characteristics is into

speaker identity (having the most obvious and numerous applications) and as has been shown by several investigators [Ata74, Luc69, Sam75, Fur81b, SR88] segmental parameters generally yield higher correlation (recognition performance) than do prosodic parameters. However, if this lack of a scheme for quantifying and comparing prosodics could be overcome, further analytical study of the correlates of prosodic parameters to speaker characteristics could be performed, and the results complemented, and enhanced by a corresponding perceptually based series of experiments. Further, such a scheme could find application in the investigation of such speaker characteristics as speaker dialect or emotional state where several researchers, for example Williams [WS72], have noted the fact and significance of prosodic correlates to the speaker characteristic but have not quantified or measured the relationship. Finally, several researchers [Lar71, Luc69, Fur81b, SR88] have shown that prosodic parameters complement segmental parameters in terms of recognition systems; the performance of recognition systems based upon segmental parameters increases when prosodics are added. Hence, prosodic parameters appear a worthy topic of further experimentation.

The inter-relationships of experimental factors, such as acoustic parameters examined, speaker set, speaker characteristics examined, experimental method, and speech material, contributing to the results derived via experimentation is extremely complex and occurs at multiple "levels". It appears that it is nigh impossible to separate the effects of any one factor; the factors being interwoven and overlapping with respect to their effect upon the results. Though such knowledge is generally implicit and assumed in most experimentation, few experimenters have sought to address the issue of the relationships between these factors. Several researchers such as Bricker and Pruzansky [BP66] have shown the effect that different linguistic material has upon results, while others such as Van Lancker et. al. [LKE85, LKW85] have illustrated that the choice of both speaker and listener sets affect results, and others such as Hecker et. al. [HSBW68] have shown that the form of manifestation of a speaker characteristic (in this case stress) in acoustic parameters is speaker dependent. All these and other papers hint at the complexity and inter-dependent nature that the choice of such factors has upon the final results of the experiments. There is therefore a great need to achieve better understanding of these relationships. In particular the effects of linguistic material, speaker set, acoustic parameter set, and speaker characteristic set, and how they are inter-related is worthy of deeper exploration.

Based upon these points the research shall take the following form. The experiments will examine the correlation between prosodic acoustic parameters and speaker characteristics. Both purely analytical experiments and experiments based upon human perception will be conducted independently, though using the same database and examining the same fundamental issues. Procedures will be utilised or designed to enable the quantification, measurement, comparison and alteration of prosodic parameters, and the results will be analysed as to the effect of linguistic material, speaker set, and parameter set, as well as the larger issue of acoustic correlates of speaker characteristics.

3.2 Approach

Given the above motivation and objectives for the research a method or approach is required. It is clear that several issues regarding the experimental method must be addressed; namely:-

- Choice of Speaker Characteristics
- Choice of Prosodic Acoustic Parameters
- Schemes for Quantifying, Comparing, and Altering Prosodic Acoustic Parameters

Each of these shall now be dealt with in turn.

3.2.1 Choice of Speaker Characteristics

As was shown in the literature review section there have been a great variety of speaker characteristics examined, and it is beyond the scope of this work to attempt to cover all such speaker characteristics. It is therefore necessary to choose a subset of speaker characteristics which will be examined in the current work.

Three speaker characteristics, namely:-

- Identity
- Sex
- Dialect

were selected to be examined.

Several factors were responsible for the selection of these three speaker characteristics. As mentioned above only a limited number of speaker characteristics may be examined in any one work and it was deemed that three characteristics, while few, would allow more detailed examination of each of the characteristics in question.

Secondly, there are two closely related issues:- the degree of research already conducted into the speaker characteristic and how significant the characteristic is in terms of the applications stemming from any findings. Speaker identity, as shown by the literature review, is the most thoroughly researched characteristic of all, and has the most numerous commercial applications. Speaker sex has also been comprehensively analysed with regard to static acoustic measures such as mean F_0 [Wea24, Cow36, HHP88, Col76] or vowel formant frequencies [PB52] but few investigators [Bre71, LMK78, LTMB79] have sought to examine prosodic or suprasegmental indicators of speaker sex; hence making it worthy of further experimentation. There are numerous dialects throughout the world, consider only the dialects of English, and it is beyond the scope of this work to attempt to cover them. However, of particular interest to Australians and for speech applications within Australia are the three dialects of Australian English [Ber67].

Thirdly, the issue of measurability, control, and consistency of the speaker characteristics is an important consideration. It is highly desirable to be able to objectively quantify the speaker characteristic in question and further to have a large linguistically constrained test bed

of utterances representing the speaker characteristic. Such a task is a major undertaking in itself for characteristics such as emotional state or physical health, whereas for speaker identity and sex it is comparatively simple. Speaker dialect may be quantified by the judgement of a trained linguist, and while it is well known that the dialect of a single speaker may vary [Ber67], such variation appears to be a function of the speaker's environment, and hence controllable.

3.2.2 Choice of Prosodic Acoustic Parameters

As indicated above, prosodic and suprasegmental acoustic parameters appear to be strong and consistent indicators of speaker characteristics, and hence worthy of further detailed investigation. Due to the nature and form of investigation - both purely analytical and human perceptual - to be used to examine the acoustic parameters (detailed in the following sections) a small set of acoustic parameters was chosen.

Properties of the parameters used as criteria for selecting whether they should be selected were:- previous research or reasonable expectation indicated that the parameter is correlated to speaker characteristics, the parameter be extractable with relative simplicity and a high degree of accuracy, the parameter be time varying over the duration of a sentence, and that a number of measures of the properties of the parameter be derivable from the parameter itself. To this end, the parameters energy, fundamental frequency, zero crossing rate, and voicing; all of which are extractable on a frame by frame basis, but measurable over the duration of an utterance, were selected.

Parameter Contours

Throughout the following text, the time series of values representing the frame by frame value of an acoustic parameter over the duration of an utterance will be referred to as a contour. Hence, for example, the sequence of F_0 values for an utterance will be called the F_0 *contour* for that utterance. Figure 3.1 is a contour presentation of the acoustic parameters energy, fundamental frequency, zero crossing rate, and voicing for the second repetition of the sentence: "We were away a year ago." by speaker 18.

Energy

Log Mean Squared Energy (LMSE) values were extracted for the each 25ms frame of an utterance (N=400) using the formula:-

$$E = 10\log_{10}(1/N\sum_{i=1}^{N} x_i^2) \tag{3.1}$$

The sequence of such values for a particular utterance is defined as the *energy contour* for that utterance (see Figure 3.1 for a sample energy contour).

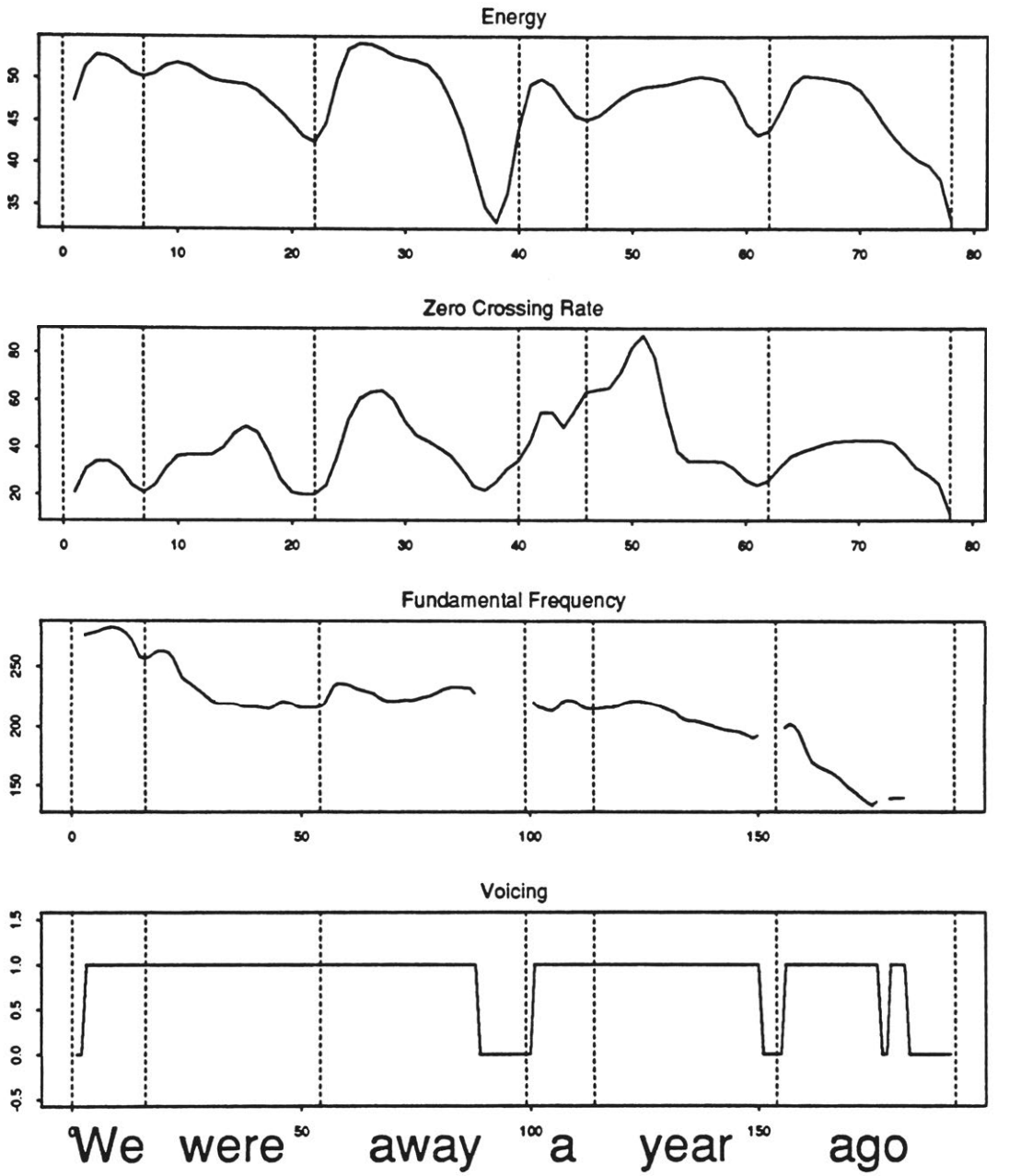


Figure 3.1: Energy, F_0, zero crossing rate and voicing contours for the sentence "We were away a year ago.", as uttered by speaker 18 on her second repetition.

Zero Crossing Rate

Zero crossing rate, or more simply zero crossings (ZC) values were extracted for each frame of an utterance using the formula:

$$ZC = \sum_{i=2}^{N} | sign(x_i) - sign(x_{i-1}) | \tag{3.2}$$

$$sign(k) = \begin{cases} 1, & k \geq 0 \\ 0, & k < 0 \end{cases} \tag{3.3}$$

The sequence of such values for a particular utterance is known as the *zero crossing contour* for that utterance (see Figure 3.1 for a sample zero crossing rate contour).

Fundamental Frequency

Fundamental frequency, F_0 (measured in Hertz) values for an utterance were extracted using 25 millisecond (400 sample) frames, with a 10 millisecond (160 sample) shift for greater accuracy, using Audlab's[1] time domain parallel pitch detector [GR69, SH87]. Male utterances were low-pass filtered at 300Hz, and female at 400Hz, before applying the pitch detector.

Several different representations of F_0 were examined in the analysis experiments (see Chapter 5 for details) however one representation was used as default for most experiments. In order to obtain a continuous F_0 contour, and to better separate the fundamental frequency from the voiced-unvoiced information, unvoiced frames were eliminated (voiced frames concatenated); leading to a shortened, continuous F_0 contour. Figure 3.2 shows an original F_0 contour, and an F_0 contour, as used in most experimentation, composed only of voiced frames.

Voicing

Voiced-Unvoiced (VUV) values were extracted for a frame size of 25ms at 10ms intervals based on the output of the time domain parallel pitch detector. Voiced frames were represented by the binary value 1, while unvoiced frames were represented by 0, leading to a square-wave *voicing contour* (see Figure 3.1 for a sample voicing contour).

3.2.3 Mechanisms for Quantifying, Comparing, and Altering Prosodic Acoustic Parameters

The previous discussion of experimental motivation [Section 3.1] made it clear that mechanisms were required to allow the quantification, comparison and alteration of prosodic acoustic parameters.

Quantifying and Comparing Prosodics

The analytic experiments require an objective mechanism to allow quantification and comparison of prosodic parameters. The Dynamic Time Warping (DTW) procedure has long been

[1] Copyright ©Edinburgh University 1987

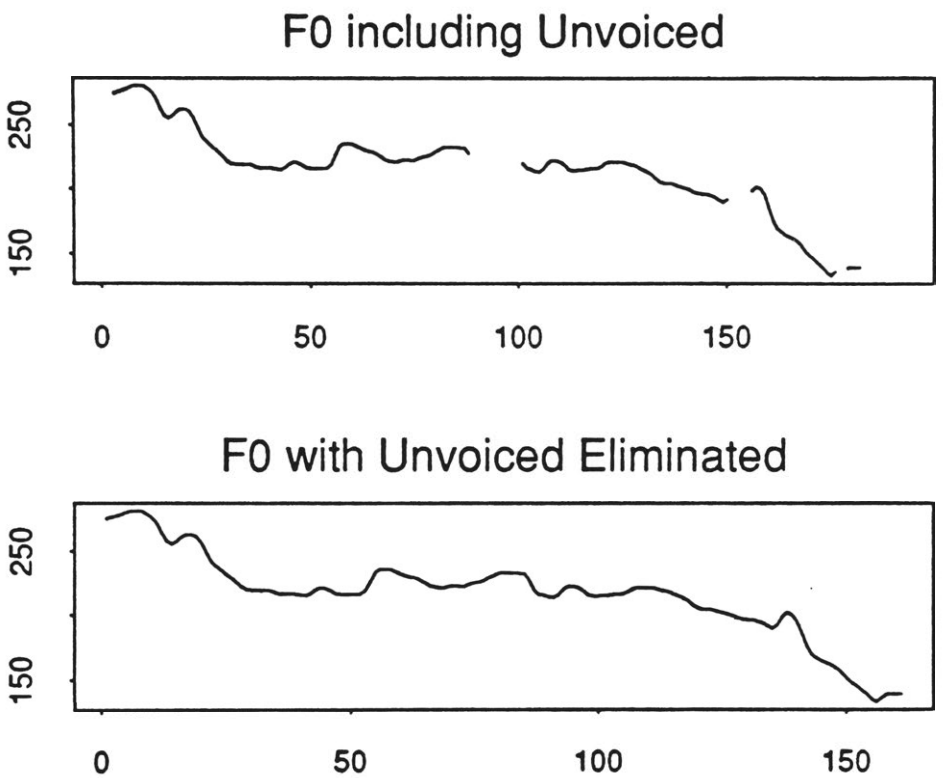


Figure 3.2: Sample F_0 contour showing the original contour (including unvoiced frames), and the contour as used in experimentation containing only voiced frames. Contour is speaker 18's second repetition of "We were away a year ago."

used in speaker characteristic analysis [Dod71b, Dod71a, Fur81b] to allow the comparison of two time series so as to minimise the calculated distance between the series (vectors). Section 5.4 will detail the process more thoroughly but intuitively DTW makes piece-wise timing adjustments, via the repetition of values from one or the other of the series, to the two original series to derive two new series of equal length that then may be simply compared. In effect DTW may be considered as "stretching" portions of each series, under certain constraints, so that both series are of the same length and most closely aligned.

While DTW has been heavily used to perform such comparisons few experimenters have sought to examine the DTW process itself, and in particular the calculated warp path (the traversal of the two contours to form the best match, see Section 5.4 for details) for further information regarding the *relative dynamics* of the two series being compared. Saito and Furui [SF78] conducted such a series of experiments where they sought to use information derived from the warp path calculation as an adjunct for their speaker recognition scheme. Using isolated words Saito and Furui found that there were regions of "high correspondence" (close to the diagonal joining starting and end points of both contours) upon the warp path for intra-speaker comparisons and that a measure could be derived based on these regions that led to enhanced recognition performance when combined with an extant recognition scheme (vector distance measures).

Saito and Furui's experiment was limited in scope but highlighted that DTW yields additional information to the calculated distance, in the form of the calculated warp path, which can be used to improve recognition performance. Hence, Saito and Furui showed that DTW provides a useful measure of the relative dynamics of two time series, in th form of the warp-path, and thus an 'approximate' means of measuring and comparing time varying parameters, beyond that already provided by the DTW distance. Their experiment was limited in terms of the acoustic parameters examined (PARCOR co-efficients), the utterances (isolated word as opposed to sentence), and in particular the type of measures based on the DTW warp path that were tested. It is therefore desirable to expand upon the scope of their experimentation in all three of these areas.

Altering Prosodic Parameters

Inherent in the process of using human perception to determine acoustic correlates of speaker characteristics is some method to link listener perceptions to values of the acoustic parameters. A technique frequently occurring in the literature [Lar71, LKE85, TK86, Kno41, LST+85, LHB+76, MHAL84, CWHY89], is that of altering the speech and comparing listener perceptions of altered and unaltered speech samples.

Such schemes vary in their complexity and refinement from relatively simple schemes involving filtering of the speech, typically to eliminate source or filter properties of the voice, [PPS54, Lar71], through methods such as rate alteration and reversal (playing backwards) [BP66, LKE85], to the more sophisticated methods involving individual alteration of acoustic parameters such as F_0 and formants [TK86, LST+85, CWHY89]. Very few experimenters have

taken this more sophisticated approach of the systematic alteration of parameters which, however, would appear to yield more detailed information. Further, most researchers have only examined a small subset of alterations and only applied these alterations to a limited group of parameters (where appropriate) to investigate a single speaker characteristic. Therefore, a refinement and more detailed application of this basic scheme, utilising judgements of altered speech, will be proposed as one of the methods employed in this thesis.

Chapter 4

Speech Data

In order to investigate the acoustic cues to speaker characteristics a suitable speech database was required. The important considerations in the design of such a database include:

- large speaker set
- minimal contextual effects
- multiple samplings of "each characteristic"
- controlled/measurable speaker characteristics
- presence of relevant speech events
- multiple characteristics
- large number of speakers commonly known to a set of listeners

The material selected for speakers to record consisted of a set of fifteen (15) isolated English sentences. Six (6) sentences were those used by Wolf [Wol72], a further eight (8) designed by Collins [Col77], and the all voiced sentence:- "We were away a year ago." [Dod71a] (see Appendix A for the sentence set). A fixed text was selected in order to minimise the effects upon acoustic parameters of textual variability. The set of fifteen sentences were selected due to the presence of many different speech events, the relative ease with which they could be segmented, and the fact that they are generally pronounced in a unique way [Sam75].

An initial speaker set of twenty one (21) adult Australian males and females was recorded. Utterances were then analysed by a linguist who nominated any speakers who weren't of one of the three Australian dialects [Ber67], and further, for each speaker assigned a scalar value representing the speakers position within the dialect spectrum. Two speakers were eliminated from the study on the basis of the linguist's advice (these two had spent a significant portion of their childhood outside of Australia), leaving a total of nineteen (19) speakers; twelve (12) male, and seven (7) female. These speakers were then randomly ordered and designated speakers 0 to 18 or speakers 'A' to 'S'; the two labelling systems being interchangeable.

The linguist was asked to assign each speaker a score based on their position in the Australian dialect spectrum [Ber67], a precise division into dialect groups not being desirable and also a far more difficult task. The range of possible scores was 0 to 10 with low scores reflecting a cultivated dialect, while high scores reflected a broad dialect. All scores were relative to other speakers in the data set, there was no absolute outside criteria used to assign a value on the spectrum. The linguist was provided with all recordings made by the speakers and via the process of repeated pairwise listening to utterances from the speakers ordered the speakers in a hierarchy of broadest to most cultivated dialect. This ranking system was then the chief means used to assign values for each speaker upon the spectrum. Figure 4.1 shows a plot of the speaker dialect scores.

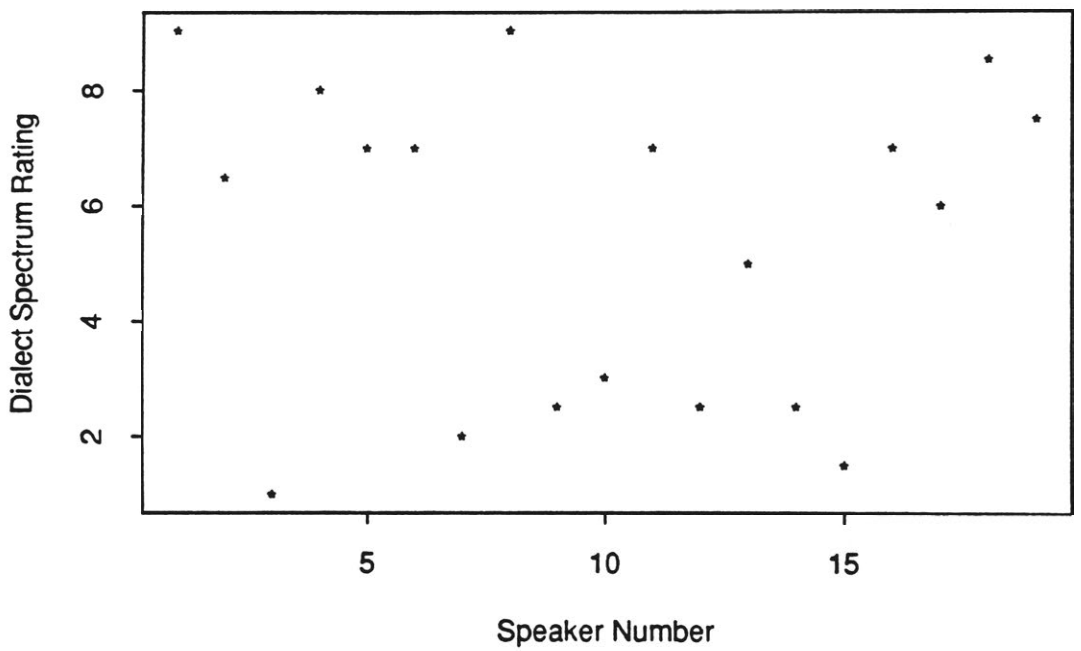


Figure 4.1: Speaker Dialect Ratings. All speakers were assigned a rating by a linguist on an Australian dialect scale ranging from 0 to 10; with 10 representing broader speech.

Speakers were asked to complete a form (particulars in Appendix B) detailing personal information of relevance to their language development. Information considered of relevance included age, occupation, years of formal education, sex, and places of residence of the speakers; together with similar details of the speakers' parents.

Speaker ages ranged from 20 to 48 (mean=31.5, σ=8.7), with female speaker ages ranging from 22 to 47 (mean=27.3, σ=9.7), and male speaker ages ranging from 20 to 48 (mean=33.2, σ=8.1).

Results of the survey showed that there was a wide range in the years of formal education between the speakers, ranging from 9 to 20. The mean was 16; approximately that of a bachelor's

degree. Similarly, there was a diversity of occupations with a high correlation between years of formal education and occupation.

All bar one of the speakers were born in Australia; and for greater than half the speakers both their parents were also born in Australia (the mean number of parents born in Australia per speaker was 1.4). For those speakers who's parents were not born in Australia the parent's place of birth was always within Europe.

Geographically the talkers were an extremely heterogeneous population; well representing the diversity of the Australian population. *All* states and territories of Australia had been the residence of one or more of the speakers for a significant period of time (not less than 3 years). On average each speaker had lived in more than four (4) different geographic locations (towns or cities) within Australia and in three (3) states and/or territories. Approximately 40% of speakers had "grown up" (spent the first sixteen years or more of their life) in a rural environment, while the rest had "grown up" in cities. A total of seven (7) speakers had spent a year or more in residence outside of Australia. For all bar one of these speakers this period of residence outside Australia was after passing the age of sixteen (16).

Speakers recorded the sentence set on five (5) separate occasions[1] over a period of no less than a week, with no two recordings from the same speaker being made on the same day. Speakers were asked to:- "Read the sentences naturally", or, "in your own voice", and to pause between each of the sentences. A different sentence order was used for each of the five sessions in an effort to further equalise the effects upon pronunciation of sentence position in the list.

Recordings were made in a soundproof film studio using an AKG D222 low impedance microphone. The recordings were then low pass filtered at 7.6kHz before 12-bit quantisation at a sampling rate of 16kHz.

The digitised recordings were then hand segmented to split them into their individual sentences. The hand segmentation process consisted of visual observation of the time series waveform, together with listening; and where necessary zero crossing, energy contours, and spectrograms were also examined.

[1] One speaker, nominated 18, only recorded two repetitions of each sentence due to unavailability. Her utterances were not analysed on an individual speaker basis but did contribute to the 'general population' analysis results.

Chapter 5

Analysis Method

This chapter describes the methods used to determine the prosodic acoustic correlates of speaker characteristics.

In brief, analysis experiments consisted of the extraction of acoustic parameters of all repetitions of a given sentence from all speakers. These parameter contours were then statistically and dynamically compared one to the other. Statistical analysis on the basis of each of the 3 speaker characteristics examined was then applied to the results of the prosodic speech parameter comparisons.

In Chapter 3, Section 3.2 a brief outline of the analysis method was given, but it remains to the current section to detail the process. Of particular importance are descriptions of the speech data used, preprocessing and *treatments* performed upon the parameter contours, an explanation of the DTW mechanism, details of the static and dynamic measures computed for each comparison, an overview of the file structures and experiment itself, details of the 'labelling' of speaker characteristics and finally the statistical methods used to analyse the computed measures on the basis of the afore said speaker characteristics.

5.1 Speech Data

In order to conduct the widest possible range of experiments given the speech material recorded and yet not exceed constraints of time and processing capability four (4) of the fifteen (15) sentences recorded were selected for analysis. These sentences were:-

1. "I cannot remember it."
2. "How do you know?"
3. "We are firm."
4. "We were away a year ago."

There were two major considerations in the choice of which four sentences to select. Firstly, pronunciation errors by speakers would lead to spurious results due to the strong influence

of linguistic content upon prosodic parameters [LR71]. Hence it was important that where possible sentences with no or few pronunciation errors, across all repetitions by all speakers, be selected. Secondly, it was desired that the individual sentences be disparate in various inherent properties; such as mean duration, mean voicing across entire sentence, and statement/question. Such a selection of sentences would be more representative of normal conversational speech than a single sentence chosen to yield optimal recognition performance. Further such a scheme allows the results to be compared upon the basis of sentence; and the differences examined with respect to these properties.

5.2 Treatments

The term *treatment* in this context is used to describe various pre-processing and transforms applied to the original contours before they are compared.

5.2.1 Smoothing

Two different smoothing filters were used in sequence to smooth transient data values that appeared inconsistent with those surrounding them. The filters used were median-5, and mean-3 [Hes83].

The formula for the median-5 filter is:

$$x_i' = \begin{cases} median(x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}) & i = 3, ..., N-2 \\ x_i & i = 1, 2, N-1, N \end{cases} \tag{5.1}$$

While the formula for the median-3 filter was:

$$x_i' = \begin{cases} 0.25x_{i-1} + 0.5x_i + 0.25x_{i+1} & i = 2, ..., N-1 \\ x_i & i = 1, N \end{cases} \tag{5.2}$$

As mentioned above the two filters were applied in sequence; median-5 then mean-3 to all raw contours. The median-5 filter has the property of eliminating either singular or paired outlying values, while the mean-3 filter 'averages' values in terms of their old value and those adjacent to it. In order that F_0 and voicing contours be treated 'appropriately' the median-5 filter was applied before voiced and unvoiced frames had been separated. Following this application the voicing contour was derived, and the F_0 contour generated as the mean-3 filtering of the voiced values, with either interpolation or elimination for unvoiced frames. The mean-3 filter was not applied to the voicing contour.

5.2.2 Normalisation

In order to accurately compare the dynamics (time varying properties) of two contours it is important that the scales of the two contours be the same. Consider, for instance, a comparison between an F_0 contour from a male speaker with a mean F_0 of 110 Hertz and an F_0 contour from a female speaker with mean F_0 of 230 Hertz. Without some form of normalisation to overcome

the non-intersecting ranges, application of the DTW process will tell us nothing significant of the relevant dynamics; being dominated by the disparate means; and returning only some measure of the differences in means.

In order to eliminate, or separate as thoroughly as possible, the static (time invariant) properties of a contour from the dynamic, so that solely dynamic contributions might be investigated the following pre-transform was derived:-

$$x_i' = \frac{x_i - \min(\tilde{x})}{\max(\tilde{x}) - \min(\tilde{x})} \tag{5.3}$$

having the property that the contour is linearly shifted into the range 0 to 1, with the old maximum now having the value 1 and the old minimum 0. This guarantees that comparisons between normalised contours occur within the same range of values and will be unaffected by the original range of the two contours.

The results achieved through experimental runs with normalised contours may then be contrasted and compared with experimental results for the same un-normalised original contours.

5.3 F_0 Representations

The question of what 'representation' to choose for F_0 contours was addressed by using four different contours, all to represent the same original F_0 contour. In the first two models F_0 was represented on a linear scale. The basic model, discussed in Chapter 3, ensures the separation of voicing from F_0 by concatenating all voiced frames and eliminated unvoiced frames. A variant on this, the second model, was to interpolate across unvoiced frames to ensure a continuous, less 'jumpy' F_0 contour.

Finally, Fujisaki [FH82, Fuj88] has proposed a log model of F_0 based on modelling lexical and syntactic constraints, and simplified here, to eliminate the constant F_{min} term, as:-

$$Log F_0(t) = \begin{cases} \log(F_0(t) - F_{min}) & F_0(t) \geq F_{min} + 1 \\ 0 & F_0(t) < F_{min} + 1 \end{cases} \tag{5.4}$$
$$t = 1..N$$

Such a representation has the property of accentuating differences for low values of F_0 while giving less weight to differences at the upper end of the F_0 scale. Based upon Fujisaki's work and observation of the F_0 contours of all speakers an F_{min} value of 60 Hertz for male speakers and 120Hz for female speakers was selected. Thus, applying the above formula to the two linear F_0 representations yields two *log* representations: Log-concatenated and Log-interpolated.

5.4 Dynamic Time Warping Mechanism

As stated in Section 3.2 dynamic time warping (DTW) is the chief technique used to compare the dynamic (time varying) qualities of the prosodic contours. Stated simply, DTW computes a warp-path, consisting of pairs of values from the two contours being compared, such that

the total distance between all pairs is minimised. Figure 5.2 is a graphic representation of the process. The warp path w_k joins the beginning and end of both contours, and obeys constraints of continuity and monotony.

Formally stated we have:

Calculate:

$$w_k = (i_k, j_k), \qquad k = 1..K$$

s.t.

$$w_1 = (1,1)$$
$$w_K = (M,N)$$
$$i_{k-1} \leq i_k \leq i_{k-1} + 1$$
$$j_{k-1} \leq j_k \leq j_{k-1} + 1$$
$$\mid i_k \tfrac{\min(M,N)}{\max(M,N)} - j_k \mid \leq \tfrac{\max(M,N)}{10}$$

Subject to

minimising:

$$D(i_k, j_k) = min \left\{ \begin{array}{c} D(i_k, j_k - 1) \\ D(i_k - 1, j_k - 1) \\ D(i_k - 1, j_k) \end{array} \right\} + d(i_k, j_k) \tag{5.5}$$

where

$$D(0,0) = 0$$
$$d(x,y) = |a_x - b_y|$$

with

$\tilde{a}$ a contour [vector] of length $M(1..M)$
$\tilde{b}$ a contour [vector] of length $N(1..N)$
$\tilde{w}$ the warp path of length K
$D(x,y)$ the total distance to (x,y)
$d(x,y)$ the distance between a_x and b_y

Included with the warp path constraints of continuity and monotony above is the restriction that the warp path must lay within the warp window. The warp window is the region about the diagonal between the start and end of the two contours being compared, $(1,1)$ and (M,N) (see figure 5.2). Based on an earlier investigation [BW88] the warp window width was assigned the value $(\max(M,N)/5)$ for all experiments. Finally, prior to application of the DTW process the contour of the maximum length was systematically assigned to be the abscissa, rather than any random process of assignment.

The above DTW mechanism is very simple, and more sophisticated schemes; particularly in regard to distance metrics and warp transitions; are commonly in use. However, simplicity was the objective, allowing the widest range of measures to be derived from the basic process. Hence, there are only three possible transitions for the warp path from point to point:- *diagonal, vertical,* and *horizontal*; and the distance measure between individual points on the two contours is simple; all allowing a wide range of more complex measures to be derived from a simple process.

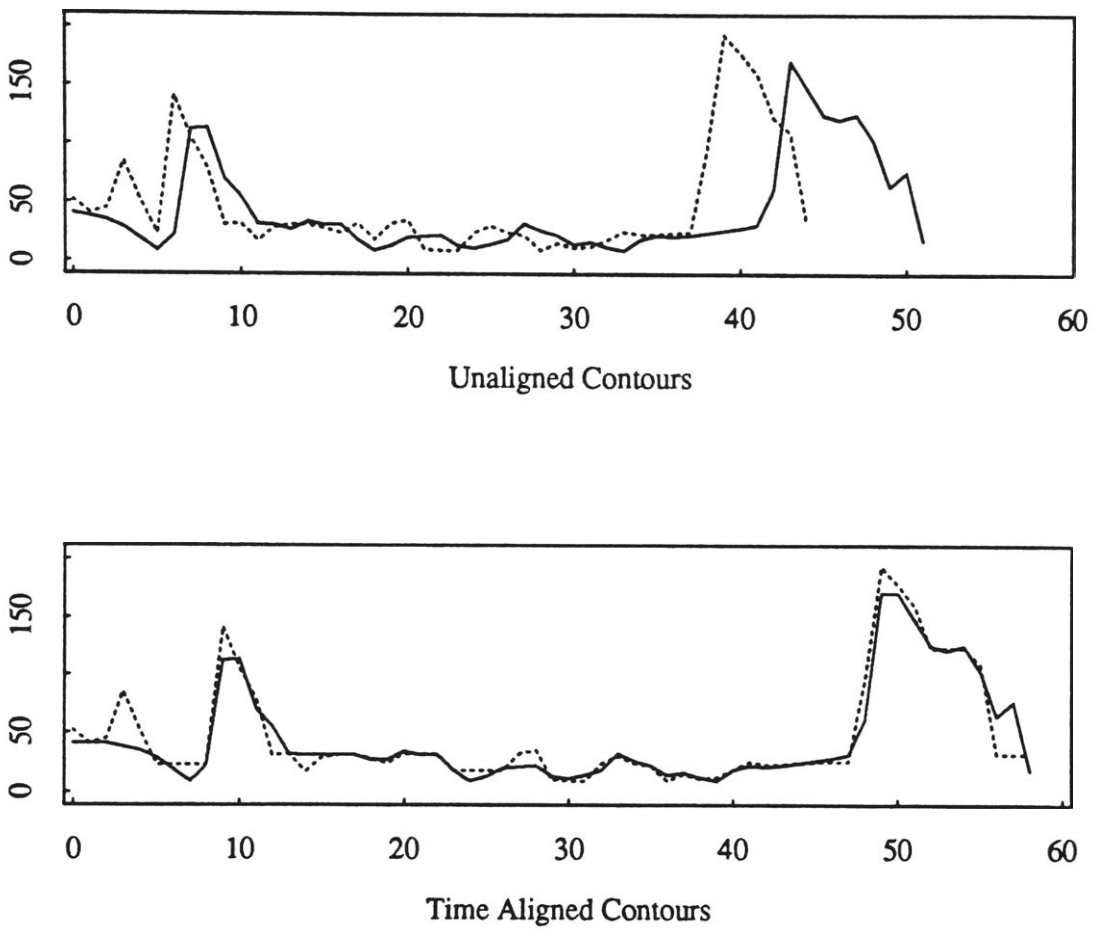


Figure 5.1: Time alignment of two Zero Crossing contours for the sentence "I cannot remember it.", via DTW. Initially the two contours are unaligned. However, after calculation of the path of best fit a far closer match is possible.

5.5 Measures

It is necessary to quantify or *measure* various properties of the prosodic speech parameters in order to relate them to the speaker characteristics being examined. Due to the complexity of the prosodic contours no single measure of the properties of that contour may be derived. Moreover, different properties of contours will undoubtedly be related to the different speaker characteristics to greater and lesser extents. It is therefore desirous to extract a melange of measures and examine each individually and in combination as to its ability to discriminate the speaker characteristic in question.

Measures of the time invariant properties of a parameter—static measures, may be extracted directly from a contour. The DTW process provides a rich and diverse series of measures relating to the relative dynamics—time varying properties, of two contours. Each of these two types of measures, static and dynamic, will now be dealt with in turn.

5.5.1 Static Measures

A suite of seven (7) measures of the static—time invariant, properties of the prosodic contours were extracted. These were:-

- Mean
- Standard Deviation
- Range
- Minimum
- Maximum
- Mean Absolute Rate of Change
- Length

Most of these measures are self explanatory however the final two need further elaboration. Length is simply a measure of the duration of an utterance. Hence for a given speaker characteristic-sentence pairing we would expect no difference between any of the results for the different prosodic parameters (Energy, voicing etc.), as the duration of the original utterance from which they are all derived is the same.

The Mean absolute rate of change of a contour may be considered as the 'speed' of the contour. The formula for its derivation is:

$$\text{MARC} = \frac{1}{N-1}\sum_{i=2}^{N} \mid x_i - x_{i-1} \mid \tag{5.6}$$

5.5.2 Dynamic Measures

As stated above the DTW process provides the means for the extraction of various measures of the time varying properties—relative dynamics, of two contours. There are two major sources of information provided by the DTW process. The DTW distance is the one commonly used by recognition systems that employ DTW for time alignment. However, the DTW warp path is also a potential source of information and is one that has received little investigation to date.

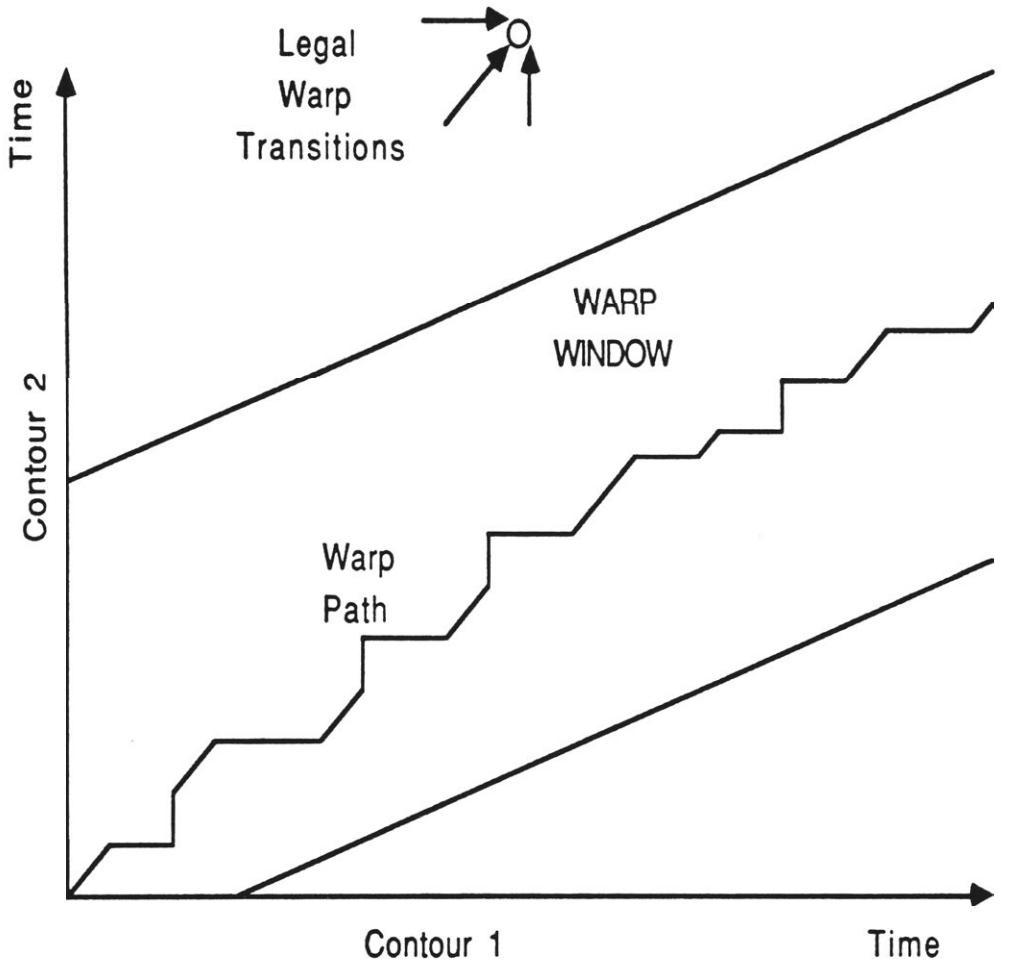


Figure 5.2: Dynamic Time Warping Schema. A best path match (Warp Path) is calculated between the two contours; subject to continuity constraints (Warp Window and Transition Types).

DTW Distance

The DTW distance is a direct product of the application of the DTW procedure to two contours. DTW distance is defined as the normalised sum of the distance between points on the two contours after they have been aligned in time. Referring to the earlier formulation of DTW (Equation 5.5) we see that:-

$$\text{DTW Dist.} = \frac{D(M,N)}{\max(M,N)} \tag{5.7}$$

$$= \frac{1}{\max(M,N)} \sum_{k=1}^{K} d(i_k, j_k) \tag{5.8}$$

Generally DTW distance is regarded as a good measure of the overall difference between two contours.

Two variants upon the basic DTW distance were also examined at certain stages within the experimental process and will be described here for completeness.

Firstly, one concern regarding the use of DTW is that spurious contour values may adversely affect its performance. As the original sentences were hand segmented it was felt that any human errors regarding the detection of phonation onset or termination would be detrimental to the performance of the DTW process. As such a simple variant on the above formula (Equation 5.8), labelled the Border DTW distance, was derived with the purpose of eliminating these possible poor matches that might occur due to the poor segmentation-

$$\text{Border DTW Dist.} = \frac{1}{\max(M,N)-4}\sum_{k=3}^{K-2} d(i_k, j_k) \tag{5.9}$$

In other words the final two matches from either end of the DTW warp path are eliminated. If the two contours were well aligned originally then this should have little effect upon the overall result; however if the process was being dominated by spurious leading or trailing values in one or the other of the contours then the above scheme should hopefully eliminate the this effect and allow the DTW process to calculate a more correct match.

Secondly, a scheme was considered that incorporated both properties of the warp path and of the DTW distance into a single measure. Simply, the warp path was calculated as previously, however after calculation the total distance was calculated as:

$$\begin{aligned} &\text{Weighted DTW Dist.} = \tfrac{1}{\max(M,N)}\textstyle\sum_{k=1}^{K} d'(i_k, j_k) \\ &\text{where:} \\ &\qquad d'(x,y) = (1 + W(x,y))d(x,y) \\ &\qquad W(x,y) = \tfrac{5}{2\max(M,N)}\left|x - y\tfrac{\min(M,N)}{\max(M,N)}\right| \\ &\qquad \Longrightarrow 0 \leq W(x,y) \leq 1 \end{aligned} \tag{5.10}$$

In effect, $W(x,y)$ is a weighting function based upon the vertical distance between the warp path and a theoretical optional diagonal between the start and end of both contours. Hence, individual distances are being "penalised", up to a maximum of doubling, based upon their distance from the theoretic optimal diagonal.

These two later measures, being derivates of other measures, will not, in general, form part of the general body or results but will be examined as to their individual performance against the standard DTW distance at a later stage.

Warp Path

As part of the dynamic time warping process a warp path is calculated. In essence this warp path is a recording of the *relative* dynamics or timing of the two contours. Therefore, it should be possible to extract meaningful measures from the warp path that could be analysed with regard to the speaker characteristics inherent in the two contours compared via DTW.

Interestingly, the warp path is itself a contour. Figure 5.3 shows an original warp path computed for the example DTW of figure 5.1 and a simple transformation to that warp path into a form which highlights salient features of the warp path. The transformation is accomplished by assigning a value of 1 for all vertical 'transitions' on the warp path, 0 for diagonal, and -1 for horizontal, which is similar to turning the original path on its side by 45°.

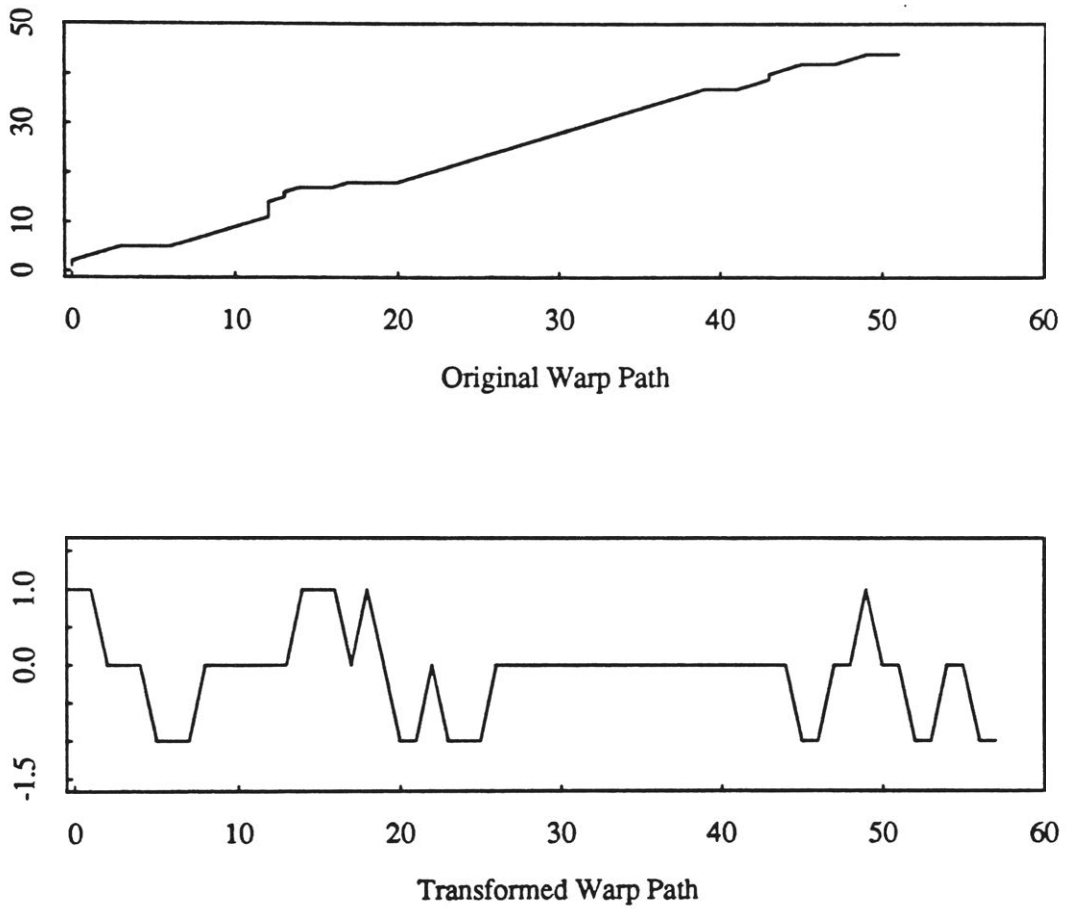


Figure 5.3: DTW Warp Path. Original warp path function, as computed between two Zero Crossing contours for the sentence "I cannot remember it", is shown together with a simple transformation of that warp path designed to highlight salient features.

Two terms will be used repeatedly to refer to qualities of the warp path so it is important that they be defined now. The first is the term *transition*. A transition is a movement from one point on the warp path to the following point. By definition there are $K-1$ transitions on the warp path and each transition is one of only three types:- vertical, diagonal, or horizontal. The definition of an *excursion* is then built upon that of a transition.

An excursion is a sequence of transitions all in the same direction (vertical, diagonal or horizontal) that is terminated by transitions in other directions, or the start or end of the warp path. Hence, for example in Figure 5.3, there are just under 60 transitions composing the warp path with 4 vertical excursions, 2 of which comprise single transitions only.

Based on examination and consideration of the warp path a total of eleven (11) measures of properties of the warp path were extracted. Nine measures related to the concept of transitions and excursions as sub-divided for the three possible types (horizontal, diagonal, vertical); while two measures considered the entirety of the warp path. Undoubtedly there is overlap with respect to to the scope of many of the measures yet each is sufficiently different to be worthy of investigation.

As stated, nine measures were extracted related to the concepts of transitions and excursions. These are:- *number of transitions, number of excursions,* and *length of maximum excursion*; as measured for each of the three possible types of transitions or excursions. All measures were normalised (divided by) the warp contour length K. Intuitively the number of transitions in each direction give an overall measure of goodness of alignment. If the number of diagonal transitions is high then intuitively its a good match. The number of excursions is a measure of the number of micro timing adjustments needing to be made to align the two contours. The maximum length excursion measures are a measure of the size (duration) of the best (diagonal) and worst (horizontal and vertical) regions of fit between the two contours being compared.

Of the other two measures one is a representation of the degree of non-optimal warping performed as measured by the ratio of the number of warp transitions to the maximum length of the two contours:

$$\text{Non-Optimal Warping} = \frac{K}{\max(M, N)} \tag{5.11}$$

The second measure is a quantification of the difference between the actual warp path line and a theoretically optimal warp path line passing diagonally from the start to the end of both contours. This measure is labelled Off-Diagonal-Warp-Distance (ODWD). Fundamentally, it is this measure that is applied as a weighting function to the Weighted DTW Distance of formula 5.10.

$$\text{ODWD} = \frac{1}{K} \sum_{k=1}^{K} \mid i_k - j_k \frac{\min(M, N)}{\max(M, N)} \mid \tag{5.12}$$

5.6 Experimental Steps

The process of experimentation involves several atomic units performed in sequence and in the aid of clarity an overview shall be given of this process. Figure 5.4 is a visual representation of the process. Utterances are digitised, and various extraction routines are run upon the digitised speech to derive the prosodic parameter contours. These contours are grouped according to parameter, sentence, and treatment; run through the static and dynamic (DTW) comparison routines to derive the raw experimental result files. Each experimental result file (sentences by parameters by treatments) comprises a number of records; one record for each comparison between two contours; and every contour pairing possibility is processed. Each record contains fields which are the twenty one (21) measures pertaining to that particular comparison.

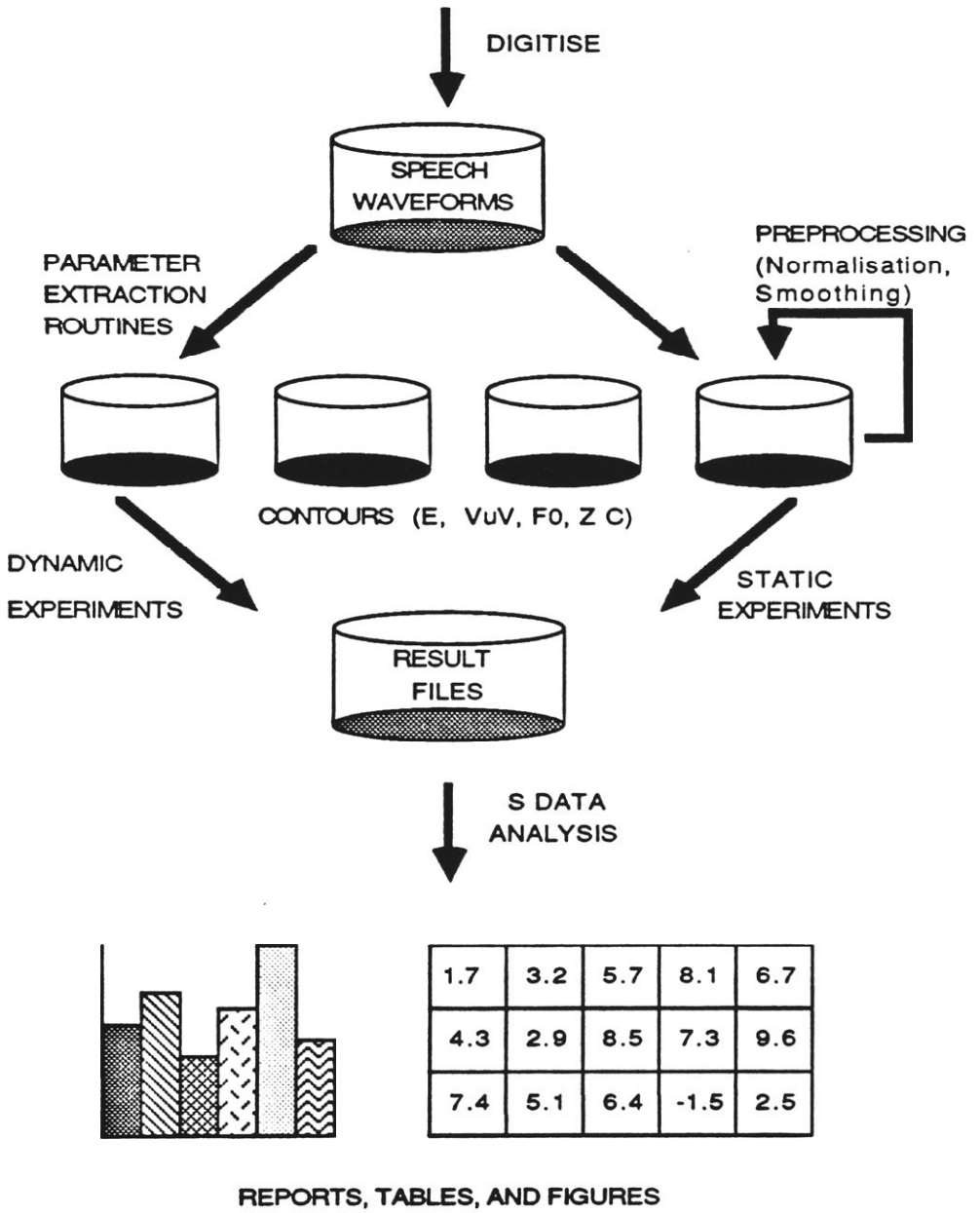


Figure 5.4: Schematic of Analysis Procedure. Spoken utterances are digitised and parameter extraction routines are run to extract the speech parameters E, F_0, Vuv, and Zc. Dynamic and static experiments are run comparing contour pairs and generating a number of result files. The statistical package S is then used to evaluate the results of the experiments.

These raw results are then 'loaded' into the S[1] statistical package [BCW88] where multivariate analysis is applied on the basis of speaker characteristic. Based on this analysis the various reports, figures, and tables showing the correlation of prosodic parameters to speaker characteristics are produced.

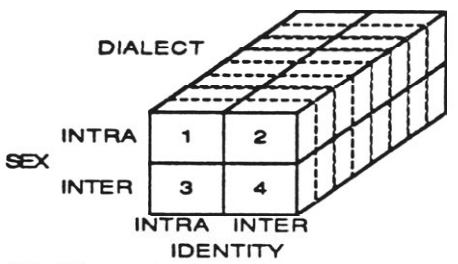
5.7 Speaker Characteristics: Labelling and Analysing

For the purpose of these experiments, each speaker; and therefore each contour derived from an utterance by that speaker, has three characteristics or quantities associated with them. These are the speaker identity, a number from 0 to 18; speaker sex, male or female; and speaker dialect score, a number from 0 to 10. Figure 5.5 presents a three dimensional representation of the above.

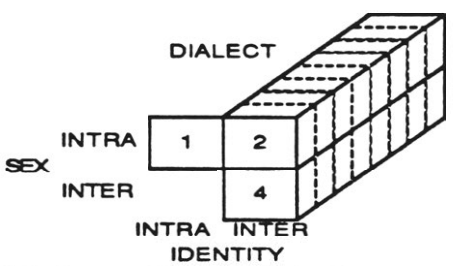
Inherent to the mechanism of DTW is that the properties of any single contour may not be measured in isolation but only relative to another contour. As all measurements then involve a paired comparison the question arises of how to quantify the 'speaker characteristics' of the comparison. For speaker sex, the simplest case, all comparisons can be broken down into two classes: - intra-sex and inter-sex comparisons. The inter-sex class contains measures derived from the comparison of contours derived from a male and a female speaker, while the intra-sex class contains both male-male and female-female measures. Similarly for speaker identity the measures may be split into two meta-classes:- intra-speaker and inter-speaker, with a possible subdivision into intra-speaker and inter-speaker classes for each of the individual speakers. For speaker dialect there is no obvious or simple division of comparisons into different classes. However, it is possible to consider the absolute difference between the dialect scores of the original speakers as a representation of dialect difference; yielding a number from 0 to 9. This quantity may then be used as a basis for statistical analysis with reference to speaker dialect. Figure 5.5 is a visual presentation of the above scheme.

Given the above classification of speaker characteristics for paired comparisons the question arises of how to analyse the results for each speaker characteristic with no or minimum effect or influence of the other two speaker characteristics. A simple, but effective scheme is to only consider those comparisons which are relevant to the speaker characteristic and which hold the other two characteristics fixed. If we consider Figure 5.5 it will help clarify the matter. For analysis of speaker identity we may eliminate the effect of speaker sex by only considering intra-sex comparisons; hence regions 1 and 2 of the figure. In effect this eliminates only the sub-class of inter-speaker comparisons between speakers of the opposite sex which, if included, would likely inflate the speaker discrimination rate, due to the excellent sex discrimination ability of such measures as mean F_0, rather than weaken the results in any way. For analysis of speaker sex we may eliminate the effect of speaker identity by considering only inter speaker comparisons; hence regions 2 and 4. In effect this ensures that the intra-sex comparisons will not be enhanced by including a number of intra-speaker comparisons. Finally, for speaker dialect analysis we may eliminate the effects of variance of sex and identity by considering only

[1] Copyright ©Bell Telephone Laboratories 1988



(A) Theoretic Characteristic Space



(B) Extant Characteristic Space

Figure 5.5: Conceptual (figure A) and actual (figure B) figures representing the 3-Dimensional speaker characteristic (identity, sex, dialect) space when two different contours, and hence two sets of speaker characteristics are compared.

inter-speaker but intra-sex comparisons; hence region 2. This subdivision of the data does not affect the statistical significance of the data due to the large number of comparisons (greater than 4,000) computed in each individual experiment.

5.8 Statistical Analysis

Based on the speaker characteristic the analysis using S [BCW88] took one of two major forms. For the characteristics identity and sex the resultant characteristic value after a comparison between two contours is a discrete value; either intra-characteristic or inter-characteristic; e.g., intra-speaker or inter-sex. For such data, discriminant analysis [HW71], was used to determine "recognition" rates for groups of measures. For individual measure analysis, Analysis of Variance (ANOVA) [HW71] was used to determine significance and an estimate ω^2 [HW71] of correlation to the speaker characteristic. It is worth noting that the two class (intra versus inter) discriminant analysis is 'far stricter' in terms of percentage rate returned and generally more demanding than equivalent decision processes for most verification or identification system implementations. A single overlap in distributions ensures that less than 100% discrimination is obtained, while a simple decision procedure such as k-nearest-neighbour would be 'unaffected' by such deviations.

On the other hand speaker dialect is a continuous linear spectrum from 0 to 10. Similarly the dialect-difference-score: the absolute difference between the two dialect scores, which is used when two contours are compared, is a continuum. Based on such data, least-squares-fit analysis [HW71] was used to quantify the correlation (R^2) between a suite of measures and the dialect-difference score.

Two items of data require mention regarding the use of the S package. When discrimination rates are given this value is calculated as the correlation between the derived vector of weights to be applied to the suite of measures, and the contrast between the two groups. Secondly S is limited in the amount of main memory that it may acquire; hence limiting the range of analysis possible. Several of the larger-scale analysis, e.g., all measures for all four parameters for all four sentences, actually had to be conducted piece-wise in several smaller runs, the results of which were then combined and rerun through the analysis function. At all times the piece-wise division of such experiments was performed so as to represent the most practical and logical division. For example, the speaker discrimination experiment when all twenty one measures of the four basic parameters from all four sentences were utilised would, if conducted as a single experiment, require a matrix of 4,000+ rows and 330+ columns which is beyond the bounds of the memory allocation capabilities of the version of S used. However if broken down into the four individual experiments, one for each of the sentences, the problem becomes manageable, and intuitively logical as a linked series of speaker recognition trials using each of four sentences in turn. Further, in all such cases where subdivision of the analysis was required it will be so noted in the text. Such an approach in no way invalidates the results, as in fact the presented discrimination or correlation value may be viewed as a lower limit which could reasonably be expected to be exceeded if a single all-encompassing analysis was conducted.

Chapter 6

Analysis Results

In this chapter will be presented the results of the various experiments whose methods were described in the previous chapter. The chapter is divided into a number of sections corresponding to major goals or questions that the analysis experiments seek to answer. Each section is further subdivided into three subsections corresponding to the three speaker characteristics—speaker identity, speaker sex, and speaker dialect, that are being examined.

Section 6.1 seeks to explore the basic question of discrimination or correlation performance given the four parameters and twenty one measures used. How well can speaker identity, sex, and dialect be discriminated on this basis? Discrimination and correlation results are found for both individual sentences and combinations of sentences, and the relationship between amount of speech material and discrimination rate is modelled using exponential growth functions.

Section 6.2 splits the twenty one measure set into two logical groups—dynamic and static—so as to compare their relative performance. As defined dynamic measures are extracted from the DTW process and quantify the time varying properties of a contour, while static measures quantify the time invariant properties such as mean and range. Comparison of the two sets shows the relative importance of the two approaches—is one markedly superior to the other, are they 'additive' in their contribution to total discrimination rate— and also further specifies the form of encoding (dynamic or static) of the speaker characteristics in the four parameters.

Section 6.3 examines discrimination and correlation rates when parameters are pre-normalised; i.e., linearly shifted into the range 0–1. In effect normalisation eliminates the static information of parameter range while maintaining the dynamics or 'shape' of a contour, thus allowing a more exacting examination of dynamics without an additional 'hidden' static factor. Further, normalisation may provide an approximation of system performance when properties of the transmission channel alter the static features of speech; e.g., additive noise on telephones or frequency shift in divers' speech due to breathing-gas mixture.

Section 6.4 compares each of the four basic parameters— F_0, voicing, energy, and zero crossing rate— as to their discrimination or correlation levels. The fundamental question to which an answer is sought is: What/which is/are the best parameter(s) to use to discriminate the characteristic?

Section 6.5 examines the four alternate representations of F_0 (Section 5.3) in order to determine which, if any, of the four versions is the best in terms of higher correlation and discrimination rates for the speaker characteristics.

Section 6.6 investigates properties of the warp path, calculated by the DTW scheme, for speaker characteristic encoding. The warp path calculated between two contours is a record of the relative dynamics of the two contours. Most DTW based schemes calculate the warp path implicitly without seeking to use it. Thus the dynamic measures are split into the DTW distance and measures of the warp path, and contrasted as to discrimination and correlation performance.

Section 6.7 contrasts the three varieties of the DTW distance examined as to discrimination and correlation performance. A simple fixed-end-point scheme was used as the basis of earlier experiments, however it is contrasted with a free-end-point scheme and a weighted distance measure using properties of the warp path.

Section 6.8 analyses and contrasts each of the twenty one measures as to their discrimination and correlation rates. Inherent to the process is an examination of their utility for recognition systems and a further refinement of where/how the characteristics ar encoded in the parameters.

Section 6.9 contrasts each of the four sentences used in experiments, based on properties of each sentence. The distribution of measure correlation values are examined in order to determine whether parallels may be drawn with discrimination and correlation rates for the sentences, and in an attempt to examine the question of: What makes a good sentence for speaker characteristic (e.g. identity) recognition systems?

Section 6.10 concludes the analysis by examining the results on the basis of the individual speakers who comprise the speaker set. This break-down allows the evaluation of the individual speaker variance from that of the total speaker population. 'Trouble' speakers, those with markedly lower discrimination or correlation scores, may be identified, and the applicability of the general population model to individual speakers may be evaluated.

It is worth noting certain salient points regarding the data, its analysis, presentation and significance. Both the discriminant and least-squares-fit analysis compute a weighted sum of the individual measures (columns) corresponding to each comparison so as to maximise division between the two classes (discriminant analysis) in the analysis, or the linear relationship between variables (least-squares-fit). Thus, for each comparison of two contours a single 'score', as a weighted sum of individual measures, is derived. The scalar value of this score— whether positive, negative, etc.— is not significant of itself, only in *relation* to the distribution of members of its own and the contrast class (discriminant analysis), or the determined line of best fit (least-squares-fit).

The most common figure, used for both identity and sex experiments, is a distribution plot. The distribution of intra-class and inter-class scores (see above) are plotted against each other. Obviously the particular score values, which may be linearly adjusted, are not significant, rather the overlap between the two distributions is.

Speaker dialect results are presented as a scatter plot. The score representing each comparison between contour pairings is plotted as a single point against the dialect difference between

the two speakers who originated the contours. A line of best fit, based on a smoothing of the data [BCW88] and showing the areas of maximum density, is plotted piecewise for steps of 0.5 in the dialect difference, the maximum resolution on that scale.

Finally, boxplots are used as a means of comparing multiple distributions. Each individual box represents a separate distribution. The horizontal line through each box shows the median of the distribution, while the upper and lower ends of the box are the upper and lower quartiles of the distribution. The "whiskers"—vertical lines, show distribution range, with extreme outliers being plotted individually.

Throughout the text a number of discrimination and correlation values are compared and contrasted. Where applicable a test of confidence interval between two proportions [WM72] was used to determine the significance of the differences. In many places in the text it is not explicitly stated that the difference is significant. However, any difference in speaker discrimination rates higher than 2.9% is guaranteed significant at the *5% level*, any difference in speaker sex discrimination rates greater than 2.1% is similarly significant, and any difference in correlation rates greater than .030 is significant at the *5% level*. These values are upper limits on differences before they become significant. In practice, smaller differences may also be significant in which case the significance will be explicitly stated.

6.1 'Discriminant Ability'

This section seeks to address the fundamental question of: "Given the speech material we have and the measures we are extracting, how well can the characteristic in question be discriminated?"

For the three speaker characteristics this was determined by applying discriminate analysis in the case of speaker identity and sex, or least-squares-fit analysis in the case of speaker dialect, to each of the four sentences individually. For each separate sentence the total twenty one (21) measures, seven static and fourteen dynamic, extracted for the four basic speech parameters—Energy, F_0, Voicing, and Zero Crossing Rate, were utilised.

The results for each of the four sentences were then combined in all permutations of 2,3, and 4 sentences in order to show the dependence of discrimination performance on the amount of speech material. Further, in an approach similar to Pollack et. al. [PPS54], growth functions, of the form $y = a(1 - e^{-bx})$ and $y = a(1 - e^{-b(x-c)}) + d$, are used to model this relationship. Residual errors are calculated to determine the 'strength' or goodness of the models together with the modelled asymptotic maximum discrimination rate for an infinite amount of speech data. The Non-linear regression procedure of the statistical package SAS [SAS85] was used to calculate all growth curves. In all cases the four methods: *gaussian, marquardt, gradient* and *dud* were applied and found to yield identical results results.

6.1.1 Speaker Identity

Table 6.1 and Figures 6.1, 6.2 present the results for the ability of the scheme to discriminate between speakers.

1	2	3	4	Mean	"Combined"
61.9	53.8	53.9	59.4	57.3	75.2

Table 6.1: Speaker identification rates for each of the four test sentences, a mean and a combined score. Discriminant analysis was applied for each sentence utilising all static and dynamic measures of the four acoustic parameters Energy, F_0, Voicing, and Zero Crossing Rate.

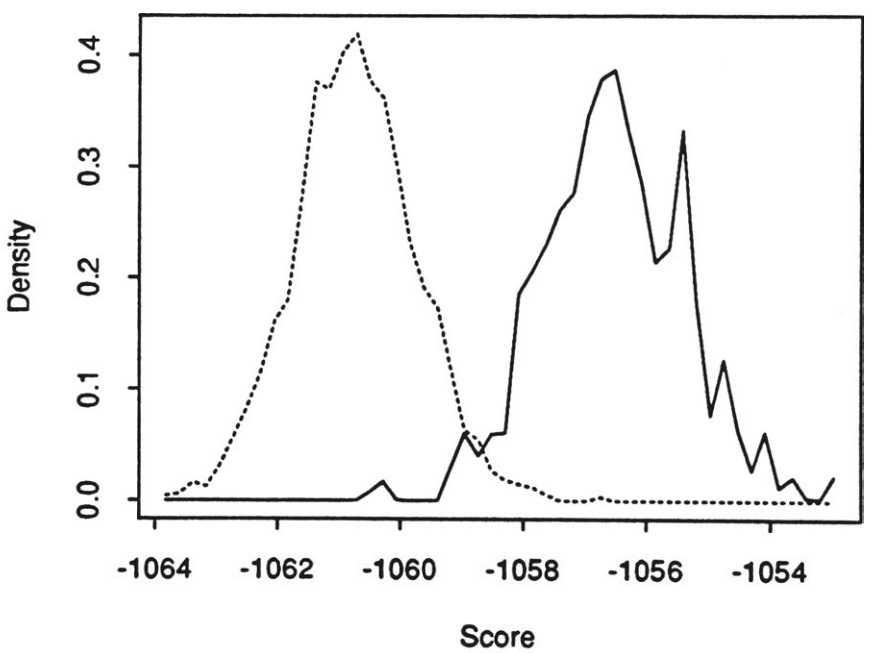


Figure 6.1: Speaker Identity Discriminate Plot - All four sentences combined. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figure represents the combination of the 21 measures for the 4 speech parameters E, F_0, Vuv, and Zc for all of the 4 sentences and corresponds to an identification rate of 75.2%

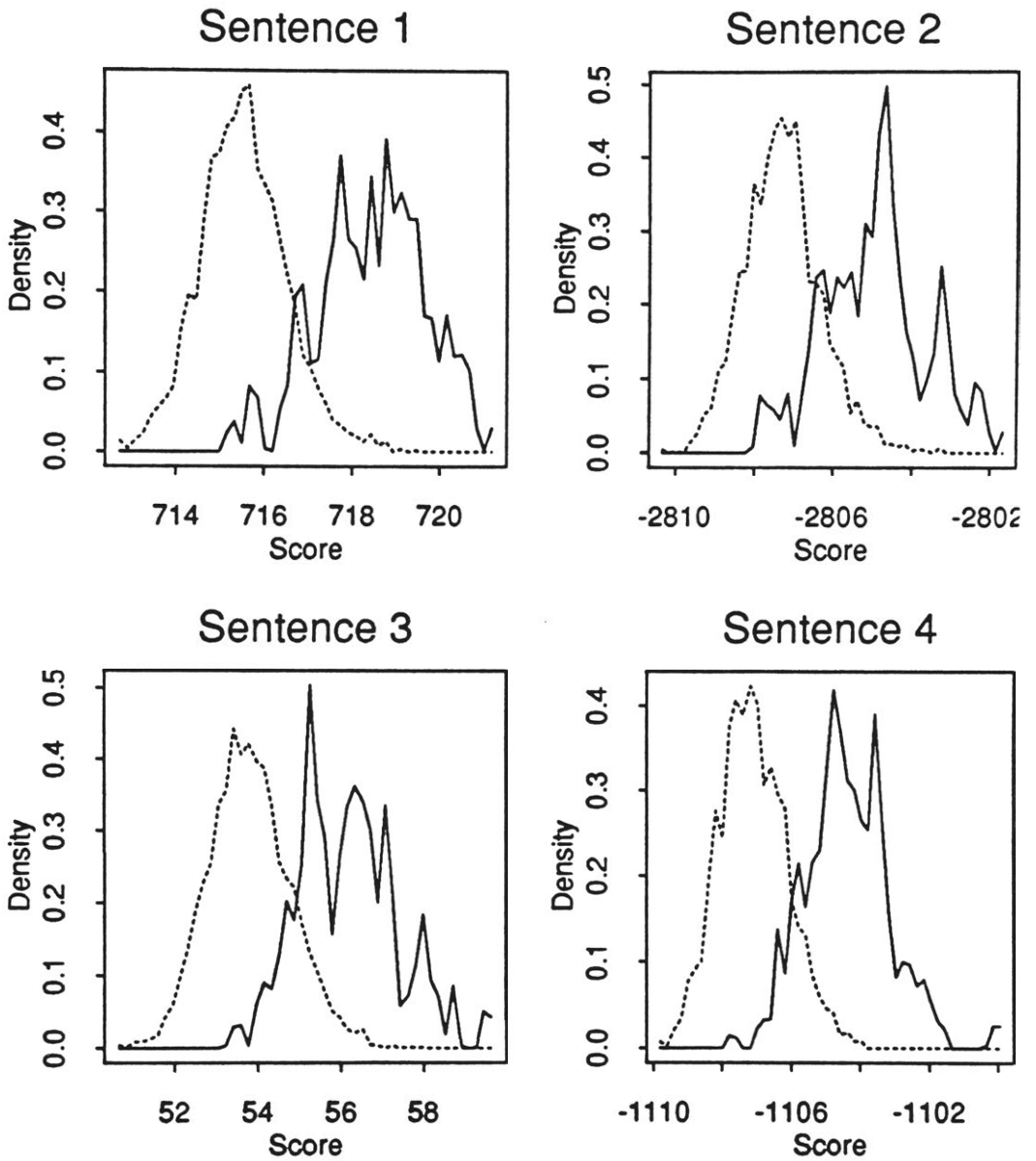


Figure 6.2: Speaker Identity Discriminate Plot - Results for each of the 4 sentences separately. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figure represents the combination of the 21 measures for the 4 speech parameters E, F_0, Vuv, and Zc.

It is clear from the table and Figure 6.2 that an identification rate well in excess of 50%, and over 60% in one case, is achieved for any particular sentence. These rates are markedly less than those of many current speaker recognition systems [Dod85, RLS90, Ber90]. However it must not be forgotten that no spectral parameters are being used, and further that this is a discriminant analysis trial, where a single overlap between inter-speaker and intra-speaker distributions yields a rate below 100%, far more strict and demanding than the a 'standard' speaker identification criterion such as, for example, picking the maximum gaussian classifier [O'S86]. In fact observation of Figure 6.1 would indicate that a speaker identification rate approaching 100% could be achieved with the implementation of a simple k-nearest-neighbour algorithm.

Comparison of Figure 6.1 with Figure 6.2 shows that a marked increase in discrimination rate is achieved when the four sentences are combined, corresponding to an increase from approximately 57% for any particular sentence to just over 75% for a combination of all four sentences. Further, sentences 1 and 4 yield significantly (*1% level*) higher discrimination rates than either of sentences 2 or 3.

Figures 6.3 and 6.4 present the results of the increase in speaker discrimination with additional speech material.

Two factors, number of sentences, and mean duration (expressed in seconds) of combined sentences are examined and their relationship to the discrimination rate modelled by a simple growth equation of the form: $y = a(1 - e^{-bx})$—Figure 6.3. A somewhat more complex formulation of the growth equation: $y = a(1 - e^{-b(x-c)} + d$, where c and d are chosen on an ad. hoc. basis to be the x, y values for the minimum length sentence, was also examined.

Based on all four plots the growth function appears to model discrimination rate as a function of speech material adequately, and show that there is a strong relationship there. It may be seen that discrimination rate improves markedly as more material is added. The four equations, with residual errors (square bracketed term) to shown closeness of fit, are:

$$\text{Discrim Rate} = 71.6(1 - e^{-1.525(\text{duration})}) \quad [109.3] \tag{6.1}$$

$$\text{Discrim Rate} = 72.3(1 - e^{-1.517(\#\text{sentences})}) \quad [125.0] \tag{6.2}$$

$$\text{Discrim Rate} = 53.8 + 23.5(1 - e^{-0.619(\text{duration}-0.8116)}) \quad [55.4] \tag{6.3}$$

$$\text{Discrim Rate} = 53.8 + 22.0(1 - e^{-0.945(\#\text{sentences}-1)}) \quad [146.6] \tag{6.4}$$

Comparing the residual error terms for all four formulations it may be seen that calculating discrimination rate as a function of the mean *duration* of the combined sentences gives a closer match than using the *number* of sentences.

In the two basic formulations the asymptotes as defined, 71.6 and 72.3, are less than already achieved discrimination rates with four, and some three sentence, combinations. Clearly then, these are less than adequate estimates of optimal discriminant performance given infinite speech material. With this in mind the asymptote defined by the formulation with the minimal residual, a value of 77.3% must also be regarded with some suspicion, though on the basis of the data a better estimate than any of the other 3. On the basis of these errors it would appear that amount of speech material, as measured in seconds or number of sentences, is inadequate to

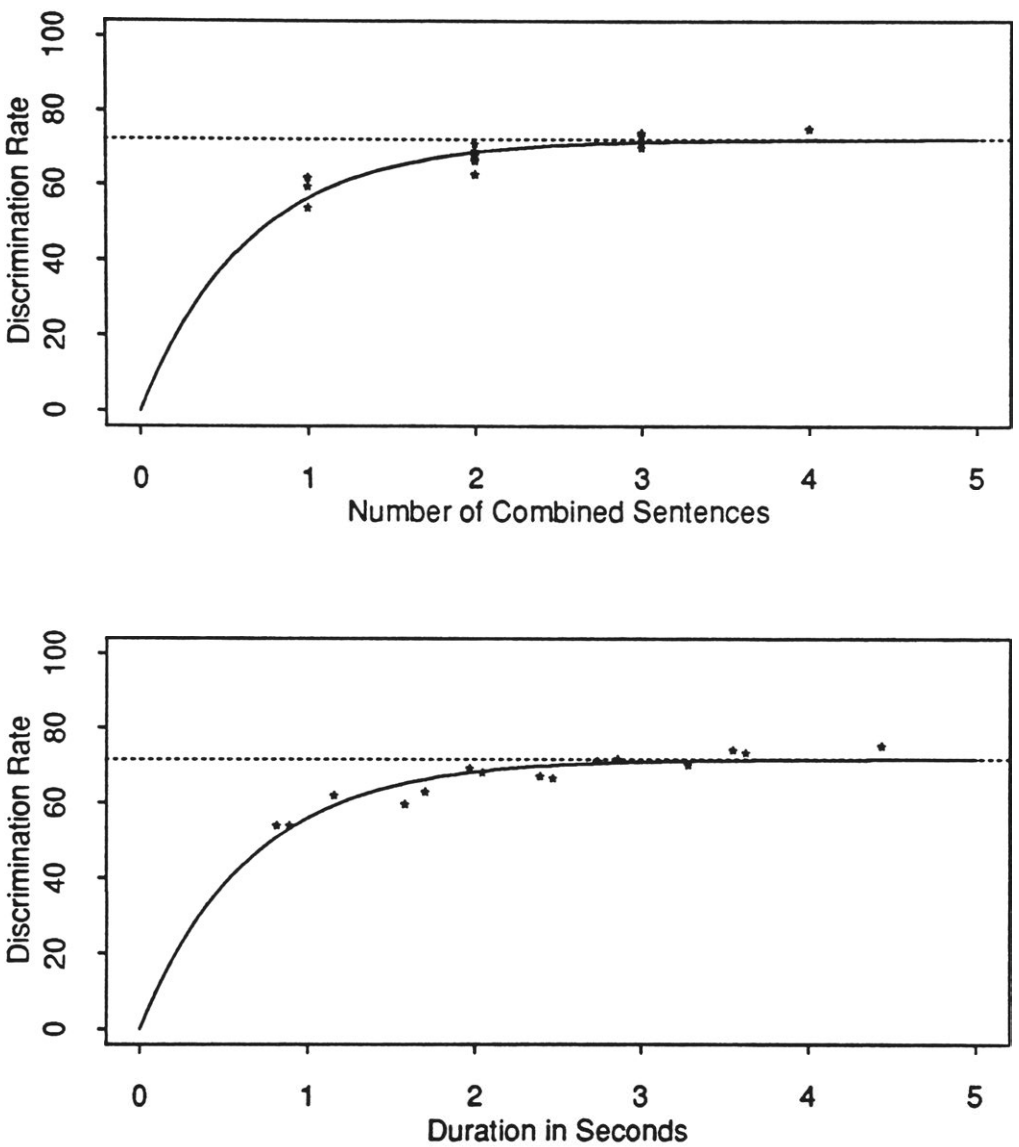


Figure 6.3: Speaker Identity Discriminate Plot - Plot of discrimination rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The first plot shows discrimination rate as a function of the number of sentences utilised, while the second as a function of the mean duration of the combined sentences. The curves are a least-squares fitted equation of the form: $y = a(1 - e^{-bx})$ with the broken line representing the asymptote a: in this case 72.3 and 71.6. For all experiments both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

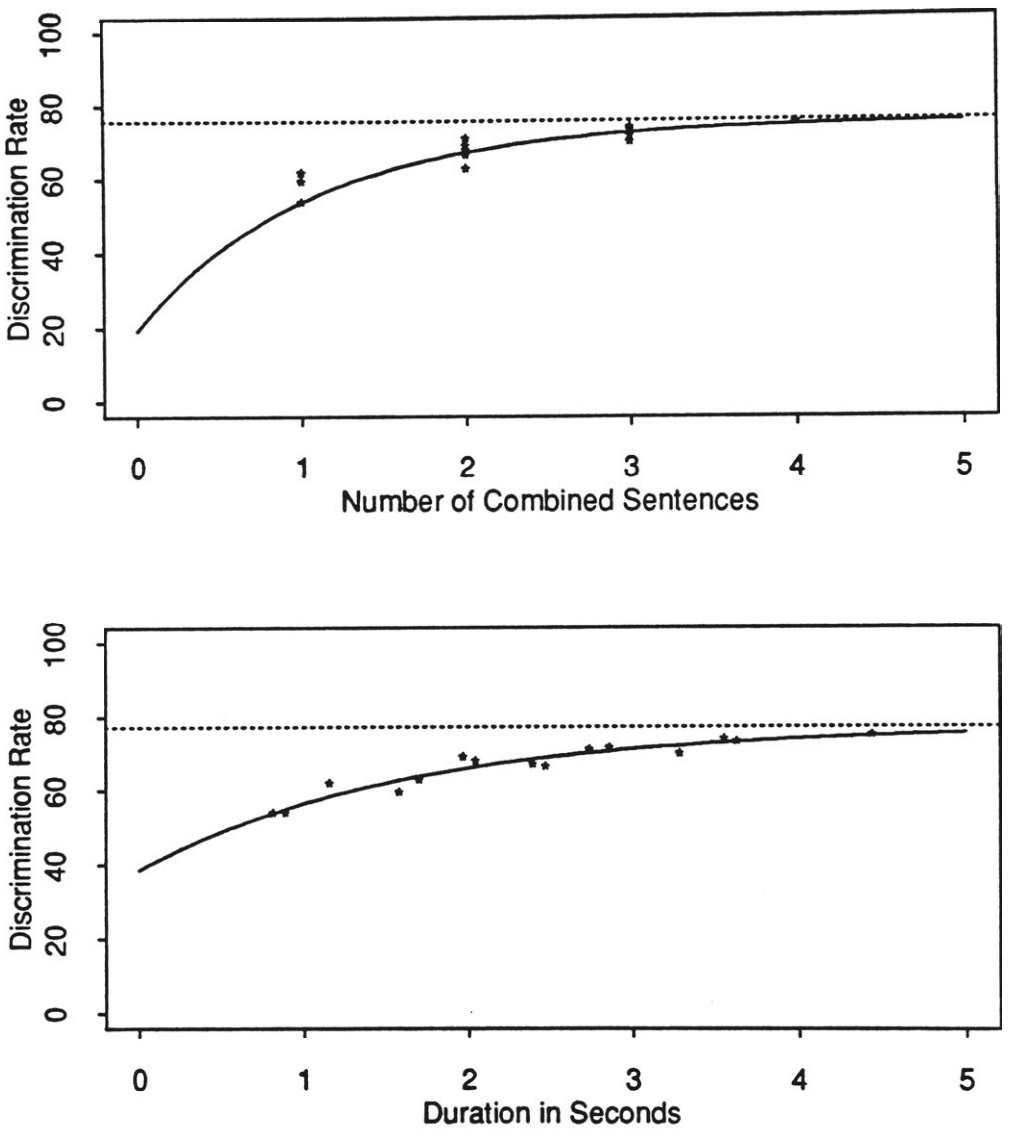


Figure 6.4: Speaker Identity Discriminate Plot - Plot of discrimination rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The first plot shows discrimination rate as a function of the number of sentences utilised, while the second as a function of the mean duration of the combined sentences. The curves are a least-squares fitted equation of the form: $y = a(1 - e^{-b(x-c)} + d)$ with the broken line representing the asymptote $a + d$: in this case 75.8 and 77.3. For all experiments both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

fully describe the discrimination rates obtained. Other factors, for example amount of voiced speech, or number of 'speech events' may need to be considered and incorporated into the formulation if truly accurate estimates of optimal performance are desired.

6.1.2 Speaker Sex

Table 6.2 and Figures 6.5, 6.6 present the results of our ability to discriminate speaker sex based upon the current scheme.

1	2	3	4	Mean	"Combined"
95.2	92.5	91.0	93.5	93.1	96.2

Table 6.2: Speaker Sex Discrimination Rates for each of the four test sentences, a mean and a combined score. Discriminant analysis was applied for each sentence utilising all static and dynamic measures of the four acoustic parameters Energy, F_0, Voicing, and Zero Crossing rate.

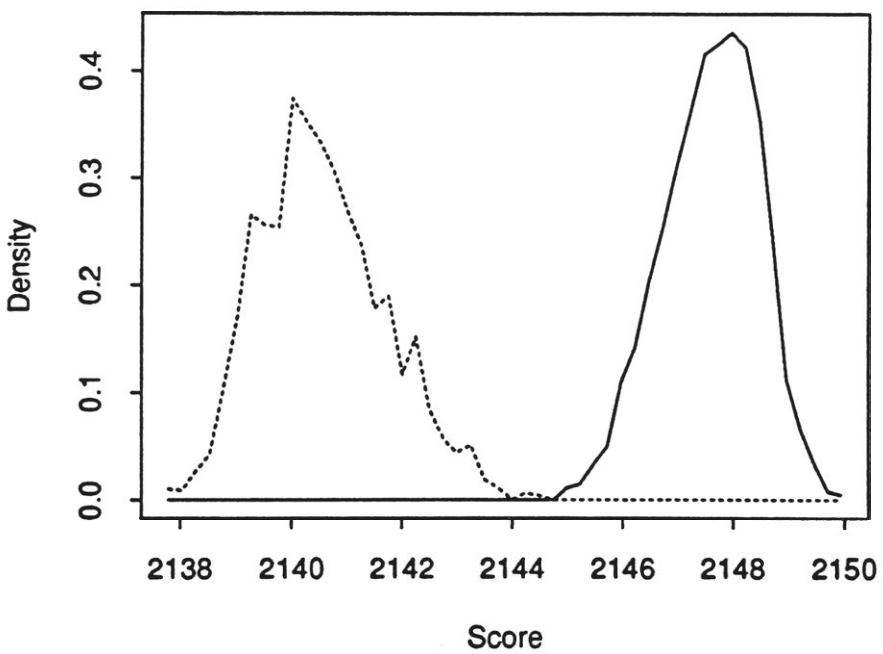


Figure 6.5: Speaker Sex Discriminate Plot - All four sentences combined. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figure represents the combination of the 21 measures for the 4 speech parameters E, F_0, Vuv, and Zc for all of the 4 sentences and corresponds to an identification rate of 96.2%

Table 6.2 and Figure 6.6 make it clear that for any given sentence a very high sex identification rate is achievable:- a mean rate of 93.1%. Such a rate is in no way surprising, given one of the four speech parameters utilised is F_0 and the well known [Wea24, HP69, Sto81, HHP88] difference in mean F_0 levels between male and female speakers. At the *1% level* sentence 1 is

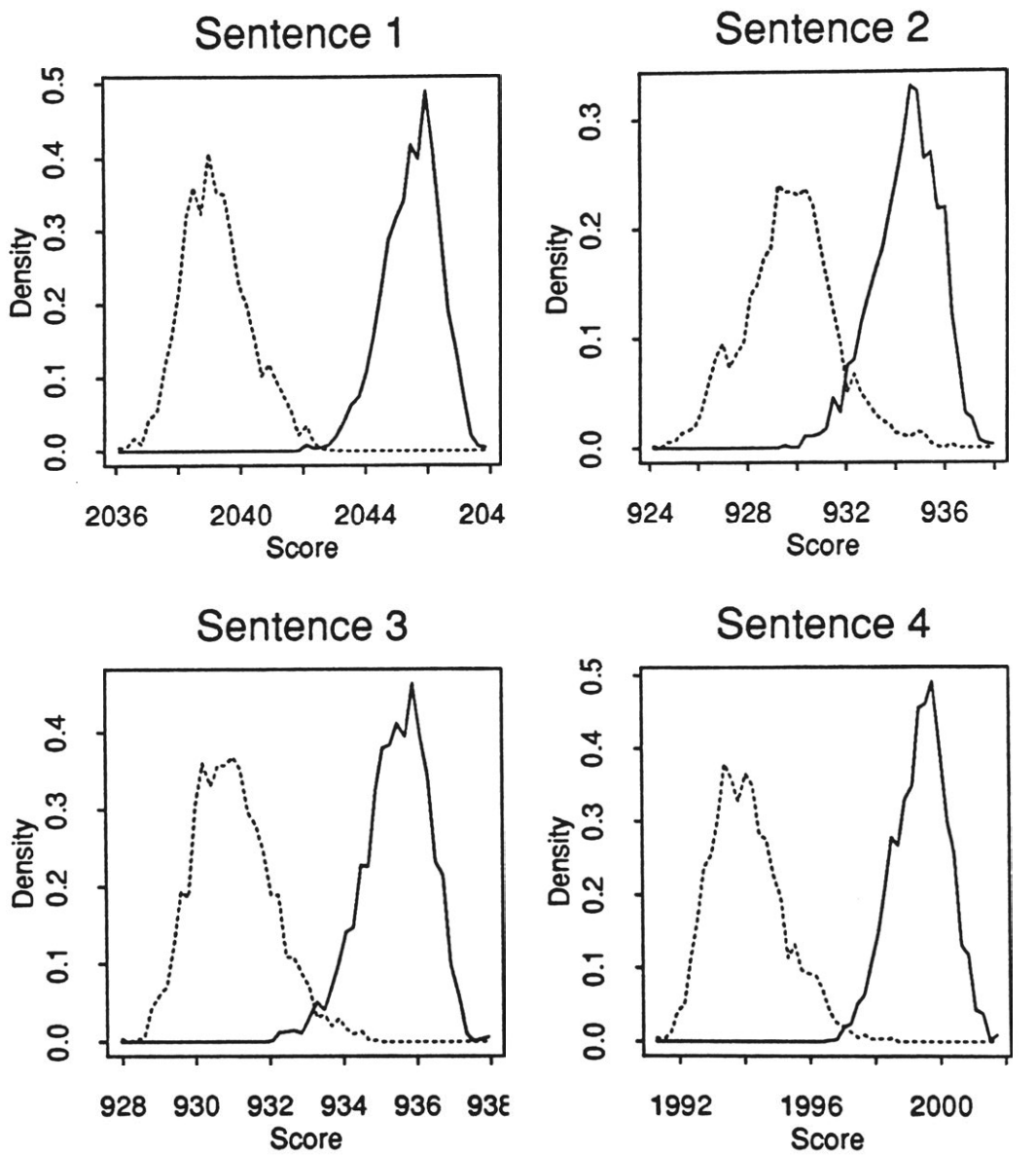


Figure 6.6: Speaker Sex Discriminate Plot - Results for each of the 4 sentences separately. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figure represents the combination of the 21 measures for the 4 speech parameters E, F_0, Vuv, and Zc.

significantly different to both 2 and 3, while 4 is significantly different to 3. Clearly there are differences between sentences as to sex discrimination levels.

Figures 6.7 and 6.8 present the results of the increase in speaker sex discrimination with additional speech material.

Two factors, number of sentences, and mean duration of combined sentences are examined, and their relationship to the discrimination rate modelled by a simple growth equation of the form: $y = a(1 - e^{-bx})$—Figure 6.7. A somewhat more complex formulation of the growth equation: $y = a(1 - e^{-b(x-c)} + d$, where c and d are chosen on an ad. hoc. basis to be the x, y values for the minimum length sentence, was also examined— Figure 6.8.

Based on all four plots the growth function appears to model discrimination rate as a function of speech material adequately, and show that a relationship between the two exists. Sex discrimination rate does appear to rise somewhat with additional speech, though only short samples of 0.8 of a second and more are adequate to discriminate sex at rates in excess of 90%. The four equations, with residual errors (square bracketed term) to shown closeness of fit, are:

$$\text{Discrim Rate} = 95.2(1 - e^{-3.992(\text{duration})}) \quad [16.5] \tag{6.5}$$

$$\text{Discrim Rate} = 95.3(1 - e^{-3.737(\#\text{sentences})}) \quad [19.6] \tag{6.6}$$

$$\text{Discrim Rate} = 92.5 + 4.0(1 - e^{-0.661(\text{duration}-0.8116)}) \quad [13.2] \tag{6.7}$$

$$\text{Discrim Rate} = 92.5 + 4.0(1 - e^{-0.836(\#\text{sentences}-1)}) \quad [17.9] \tag{6.8}$$

Comparing the residual error terms for all four formulations it may be seen that calculating discrimination rate as a function of the mean *duration* of the combined sentences gives a closer match than using the *number* of sentences.

In the two basic formulations the asymptotes as defined, 95.2 and 95.3, are less than several of the discrimination rates achieved in the experiments. Clearly, these are less than adequate estimates of optimal discrimination performance given infinite speech material. With this in mind the asymptote defined by the formulation with the minimal residual, a value of 96.5% must also be regarded with some suspicion, though on the basis of the data a better estimate than that derived by the two 'simpler' formulations. On the basis of these errors it would appear that neither the amount of speech material, as measured in seconds nor the number of sentences, is adequate to *fully* describe the discrimination rates obtained. Other factors, for example amount of voiced speech, or number of 'speech events' may need to be considered and incorporated into the formulation if truly accurate estimates of optimal performance are desired.

6.1.3 Speaker Dialect

As stated previously speaker dialect was investigated by correlating the difference between dialect scores of the speakers with the measures being examined. Figures 6.9 and 6.10, with Table 6.3 present the results for these analyses.

It can be seen from Figure 6.9 and Table 6.3 that a significant relationship does exist between the measures and the dialect-difference-score, implying that prosodic parameters do yield some

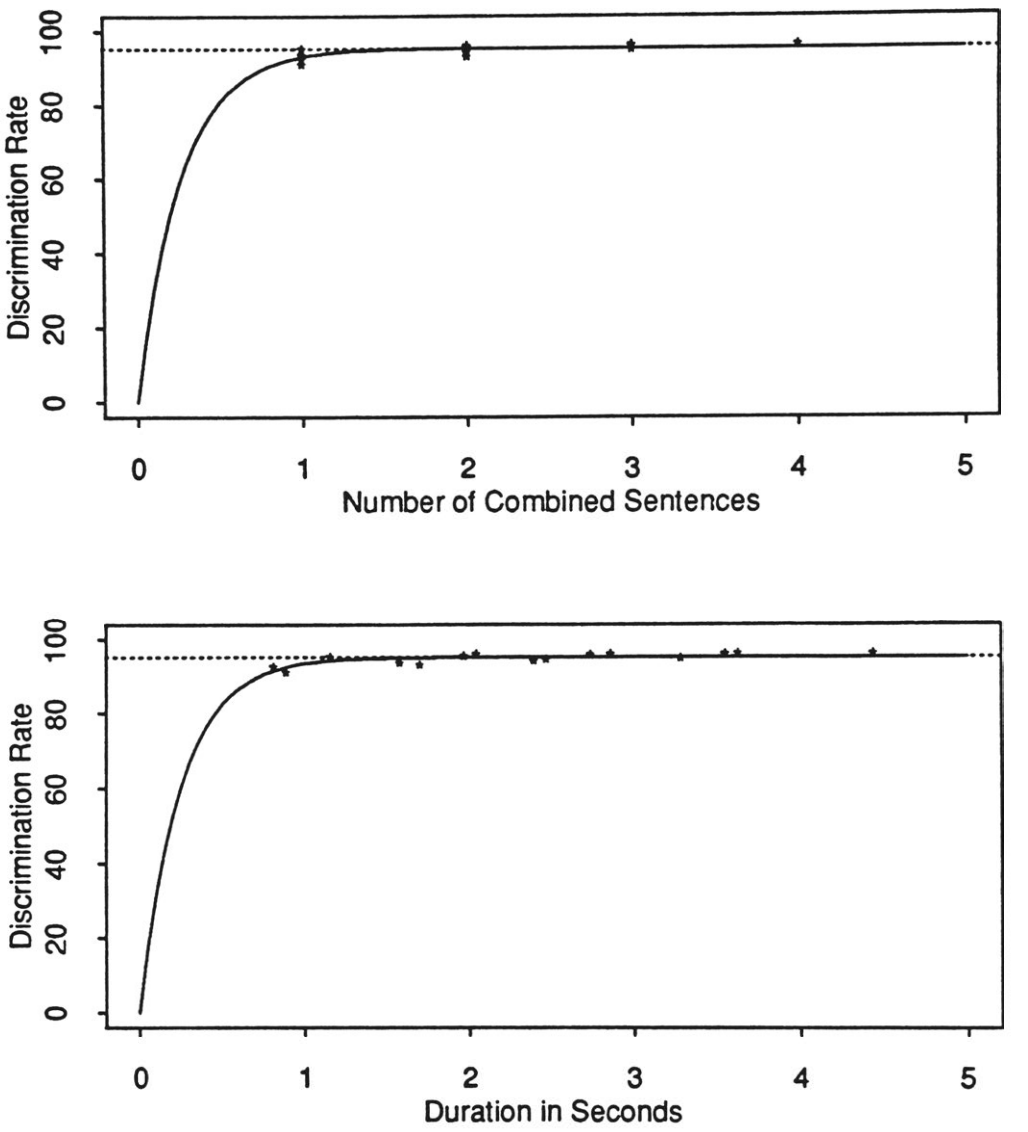


Figure 6.7: Speaker Sex Discriminate Plot - Plot of discrimination rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The first plot shows discrimination rate as a function of the number of sentences utilised, while the second as a function of the mean duration of the combined sentences. The curves are a least-squares fitted equation of the form: $y = a(1 - e^{-bx})$ with the broken line representing the asymptote a: in this case 95.3 and 95.2. For all experiments both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

1	2	3	4	Mean	"Combined"
.405	.408	.361	.349	.381	.584

Table 6.3: Speaker Dialect Correlation Scores. Least-squares-fit analysis is applied to each of the four test sentences in order to yield the highest correlation between the static and dynamic measures of the four acoustic parameters—Energy, F_0, Voicing, and Zero Crossing, with dialect difference values.

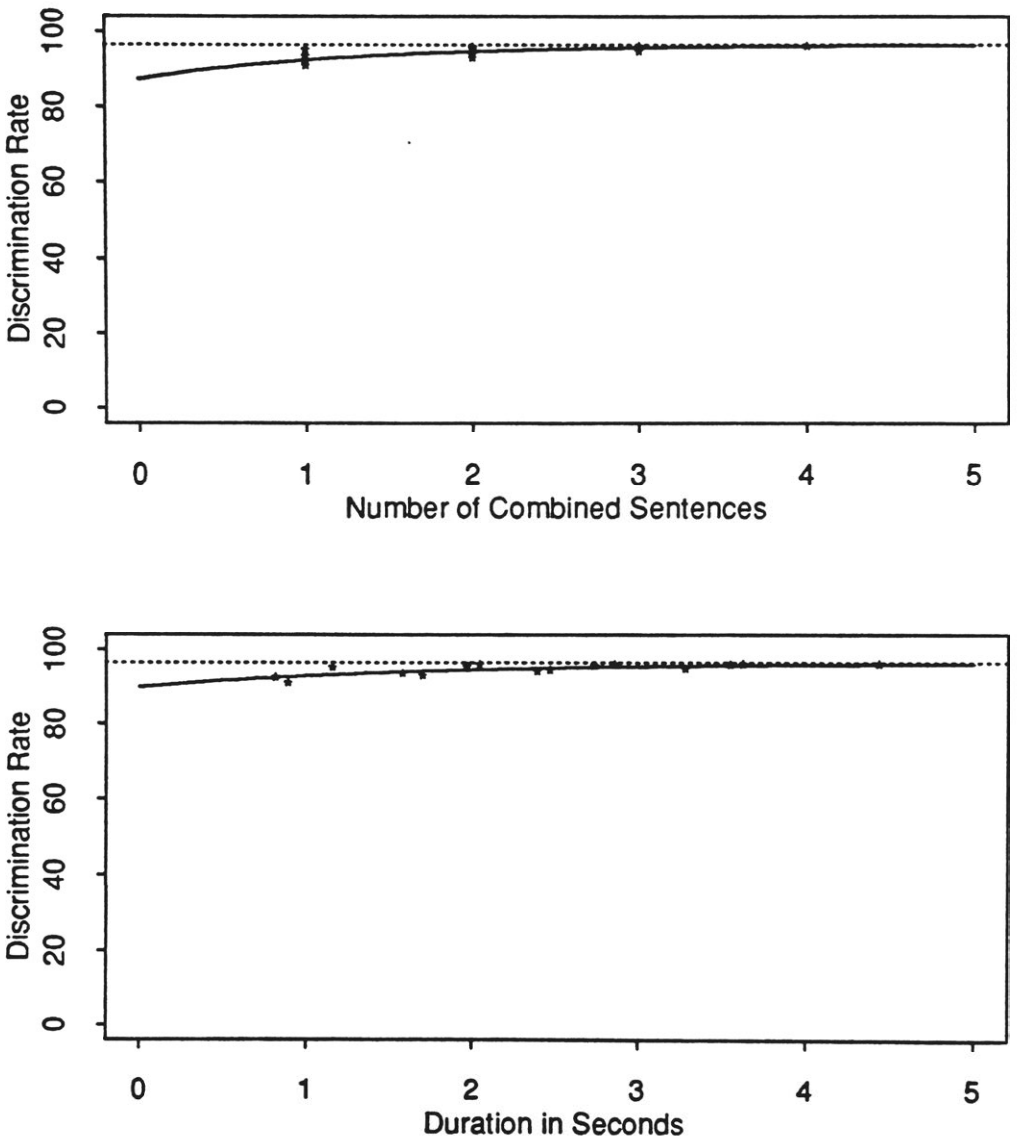


Figure 6.8: Speaker Sex Discriminate Plot - Plot of discrimination rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The first plot shows discrimination rate as a function of the number of sentences utilised, while the second as a function of the mean duration of the combined sentences. The curves are a least-squares fitted equation of the form: $y = a(1 - e^{-b(x-c)} + d)$ with the broken line representing the asymptote $a + d$: in this case both 96.5. For all experiments both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

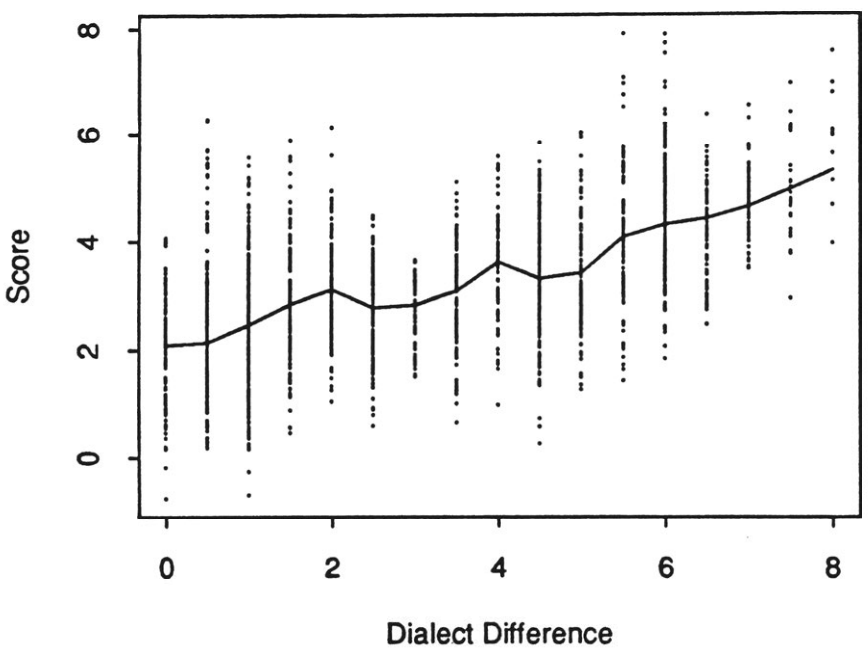


Figure 6.9: Speaker Dialect Least-Squares-Fit Scatter Plot - All four sentences combined. Figure represents the combination of the 21 measures for the 4 speech parameters: E, F_0, Vuv, and Zc for each of the 4 sentences and has a correlation rate of .584

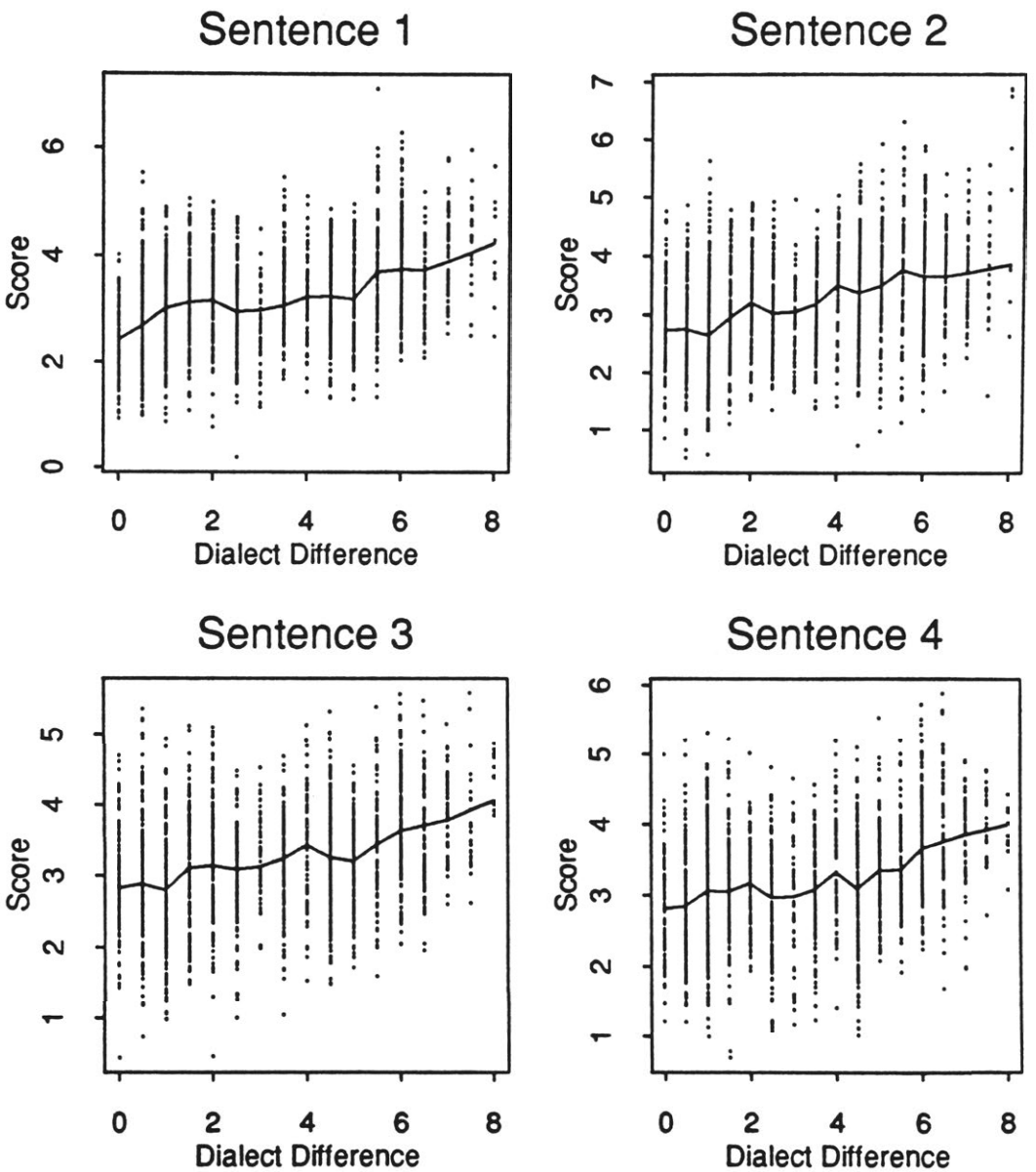


Figure 6.10: Speaker Dialect Least-Squares-Fit Scatter Plot - Results for each of the 4 sentences separately. Figure represents the combination of the 21 measures for the 4 speech parameters: E, F_0, Vuv, and Zc.

measure or quantification of the dialect of the speakers examined. Other researchers, such as Pittam and Ingram [PGC90] or Wagner [Wag78], have indicated that there appear to be prosodic differences in Australian dialect and this result confirms and furthers these previous findings.

It can be seen that the choice of sentence has a marked effect upon correlation values. Sentences 1 and 2 are significantly different (*1% level*) to both 3 and 4. Therefore sentences 1 and 2 may be regarded as being good choices and sentence 4 the poorest choice. Neither average sentence duration, nor degree of voicing within sentence, thought to be the two most likely explanations, can be related to the correlation value. It therefore appears that some other property of a sentence dictates the extent of dialect encoding within the prosodics of an utterance.

Examination of Figure 6.9 which presents the results when all 4 sentences are combined shows that utilising all 4 sentences yields markedly better results than if any single sentence is used. Numerically the difference is a correlation of .584 for the 4 combined sentences as opposed to a mean correlation value of .381 for a single sentence. Clearly a single sentence is insufficient to encapsulate all the possible dialect encoding within prosodics and the addition of further sentences improves ability to determine dialect.

Figures 6.11 and 6.12 present the results of the increase in dialect correlation value with additional speech material. As previously for speaker identity and sex, a growth formulation of the relationship between correlation rate and amount of speech material (expressed in number of sentences, and mean duration) is derived. Unfortunately no formulation of the relationship between adjusted duration and correlation value was obtainable as all four non-linear methods examined:- *gaussian, dud, gradient,* and *marquardt* did not converge.

Based on all four plots the growth function appears to model discrimination rate as a function of speech material adequately, and show that a relationship exists. Clearly correlation rates do increase markedly when additional speech material is used. The three equations, with residual errors (square bracketed term) to shown closeness of fit, are:

$$\text{Discrim Rate} = 0.537(1 - e^{-1.213(\text{duration})}) \quad [2.821] \tag{6.9}$$

$$\text{Discrim Rate} = 0.565(1 - e^{-1.061(\#\text{sentences})}) \quad [0.689] \tag{6.10}$$

$$\text{Discrim Rate} = 0.408 + 0.352(1 - e^{-0.273(\#\text{sentences}-1)}) \quad [1.630] \tag{6.11}$$

Comparing the residual error terms for all three formulations it appears that calculating correlation value as a function of the *number* of sentences gives a closer match than when using the mean *duration* of the combined sentences. This result contrasts with that of the previous speaker identity and sex where duration was found to be a more accurate factor in representing increased discrimination values. Section 6.8 shows that variations in total sentence duration appears to be a good measure for speaker dialect and hence the more sentences (rather than their actual length) the better the correlation value that could be obtained.

In the two basic formulations the asymptotes as defined, 0.537 and 0.565, are less than already achieved correlation rates with four, and some three sentence, combinations, clearly less than adequate estimates of optimal performance given infinite speech material. The other

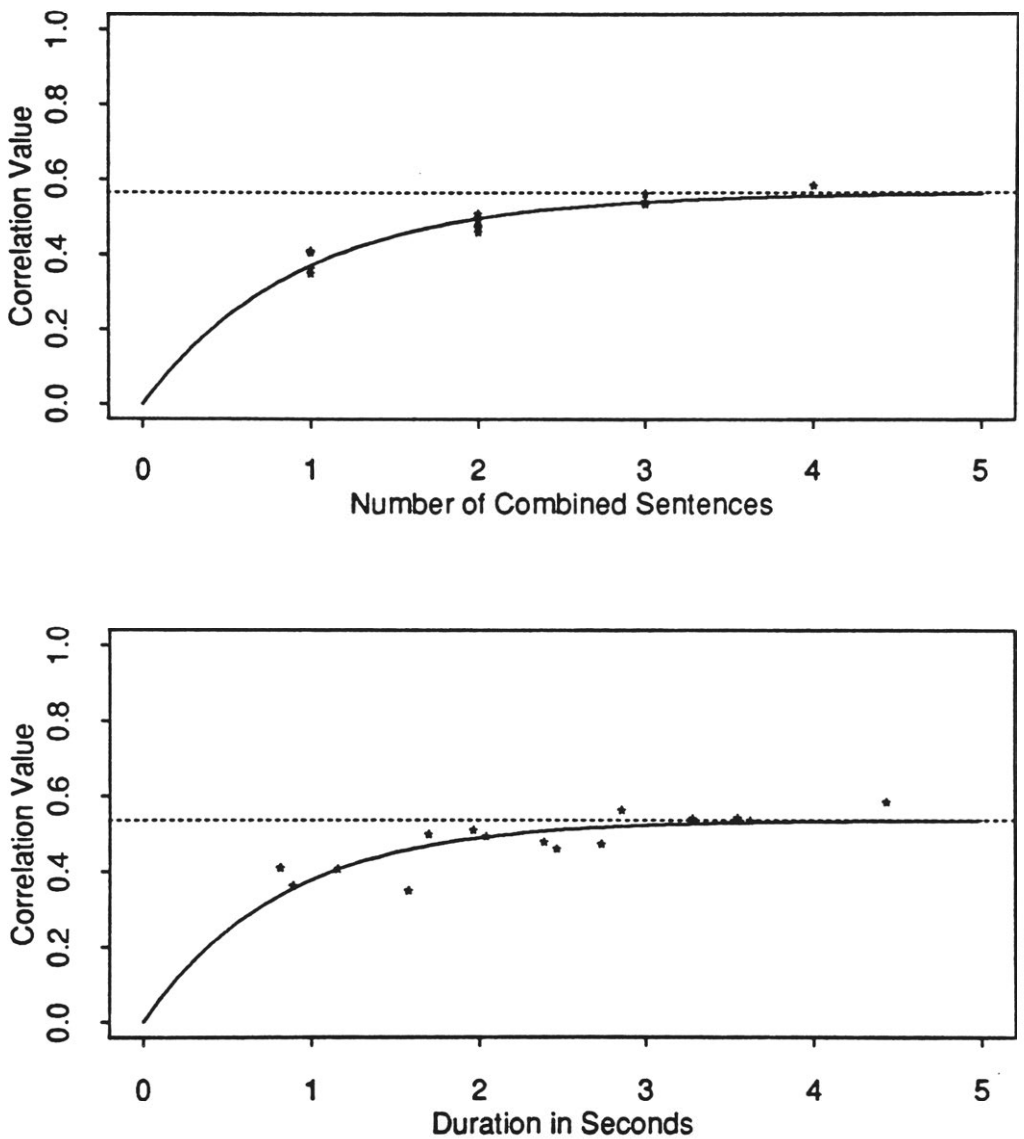


Figure 6.11: Speaker Dialect Correlation Plot - Plot of least-squares-fit correlation rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The first plot shows discrimination rate as a function of the number of sentences utilised, while the second as a function of the mean duration of the combined sentences. The curves are least-squares fitted equations of the form: $y = a(1 - e^{-bx})$ with the broken line representing the asymptote a: in this case .565 and .537. For all experiments both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

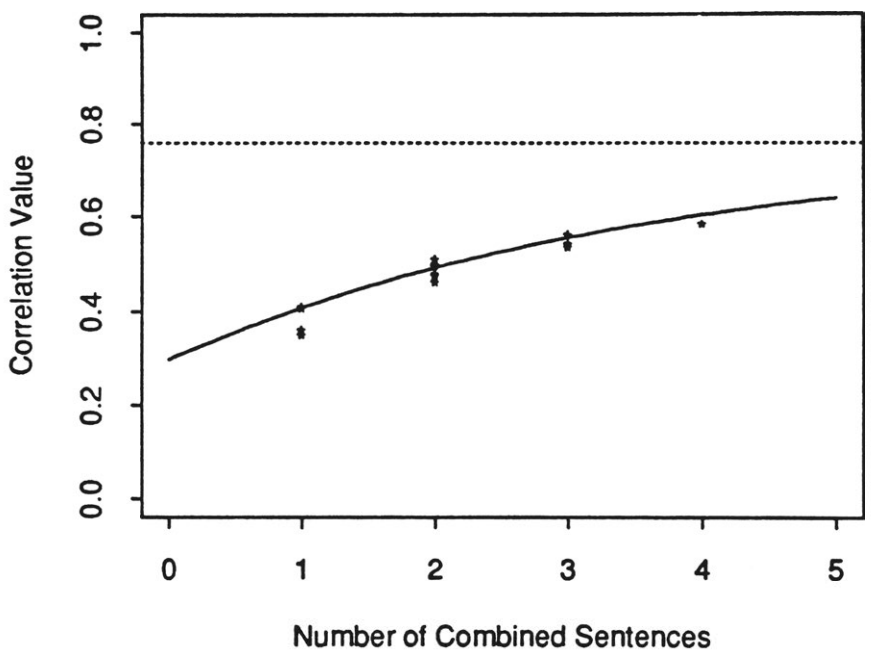


Figure 6.12: Speaker Dialect Correlation Plot - Plot of least-squares-fit correlation rate for all single sentences, and all combinations of 2, 3 and 4 sentences. The curve is a least-squares fitted equation of the form: $y = a(1 - e^{-b(x-c)} + d)$ with the broken line representing the asymptote $a + d$: 0.760. Both static and dynamic measures of the 4 speech parameters E, F_0, Vuv, and Zc were used.

estimate, 0.760 must also be regarded with some doubt due to the large residual error term associated with it; hence no truly informed estimate of optimal correlation value may be made. On the basis of these errors, and paralleling earlier results for identity and sex, it would appear that neither the amount of speech material, as measured in seconds nor the number of sentences, is adequate to *fully* describe the discrimination rates obtained. Other additional factors, as listed previously in the discussion of identity and sex results, may need to be considered to derive a truly accurate estimate of optimal performance.

6.2 Static versus Dynamic Measures

In Section 6.1 the basic ability to determine the speaker characteristics identity, sex, and dialect was examined. All measures for the speaker parameters Energy, F_0, Voicing, and Zero Crossing Rate were used. One division of the measures, carried through from the initial stages of experimentation, is that into dynamic—those measuring the time varying (dynamic) properties of a contour by the DTW process— and static—those measuring the time invariant (static) properties of a contour across its entire duration, such as mean and range.

A contrast of the results based upon these two classes of measures will highlight several salient features:- how significant are temporal related (dynamic) as opposed to static measures of the speech parameters examined in discriminating the speaker characteristics; and hence to what extent the characteristic is encoded in those properties of the speech parameters. Secondly do the two approaches complement each other [Fur81b] so that combined the result is superior to either alone, or is it sufficient to employ only one approach.

Analysis method was similar to that of Section 6.1, discriminate analysis was used for speaker identity and sex, and least-squares-fit for speaker dialect. Analysis was carried out on a sentence basis: in each case the static, or dynamic, measures for the four speech parameters— Energy, F_0, Voicing, and Zero Crossing Rate, were combined and examined. In order to yield a combined sentence result, the results from the four individual sentences were combined and run through the analysis scheme.

6.2.1 Speaker Identity

Table 6.4 and Figures 6.13 and 6.14 present the results of the speaker identity discriminate experiments utilising static or dynamic measures exclusively.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Static	41.4	39.2	40.0	44.9	41.4	54.8
Dynamic	57.8	50.4	51.0	56.9	54.0	74.6

Table 6.4: Speaker Identification rates contrasting Static and Dynamic Measures. Identification rates are given for all four sentences, the mean across the sentence and a 'combined' rate for all sentences together. In all cases the 4 speech parameters E, F_0 Vuv and Zc were employed.

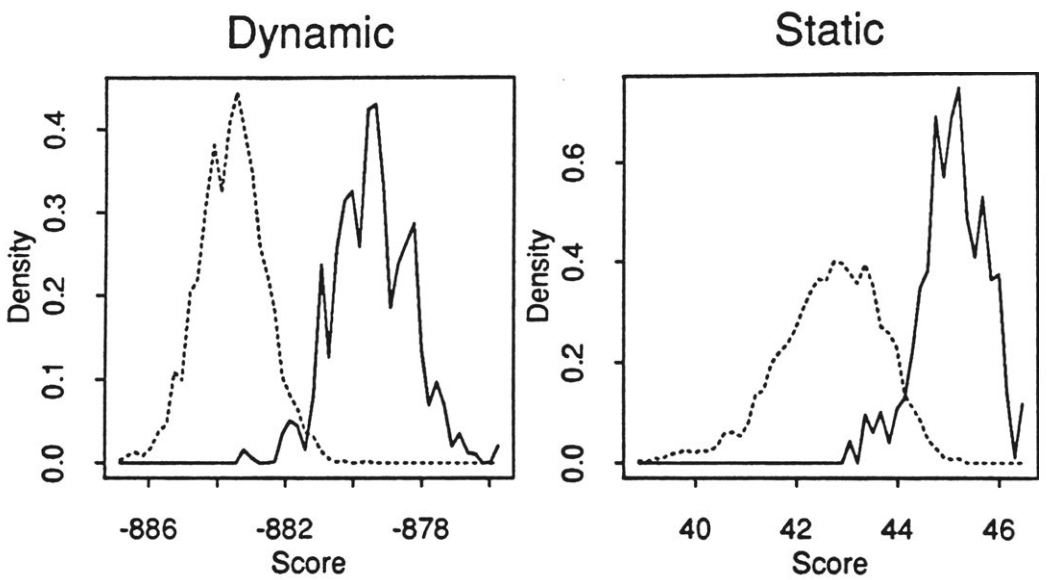


Figure 6.13: Speaker Identity Discriminate Plot - Dynamic versus Static Measures for all 4 sentences combined. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Each figure represents the combination of the relevant measures of the 4 speech parameters E, F_0, Vuv, and Zc for all of the 4 sentences and corresponds to a discriminate rate of 74.6% rate for dynamic measures and 54.8% for static measures.

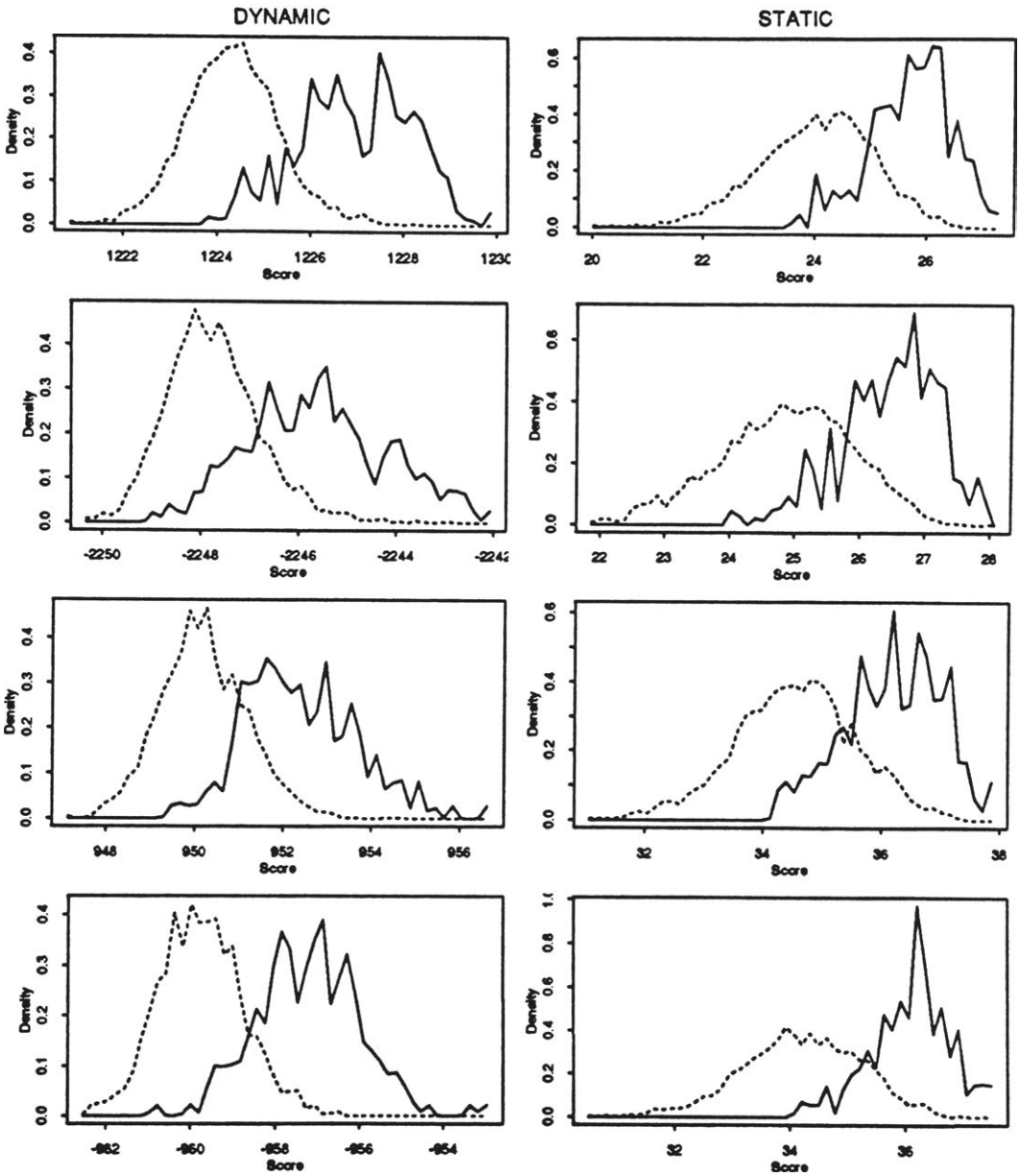


Figure 6.14: Speaker Identity Discriminate Plot - Dynamic versus Static Measures for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Results for each sentence were derived by combining the relevant (static or dynamic) measures of the 4 speech parameters E, F_0, Vuv, and Zc.

Examination of Table 6.4 and Figure 6.14 quickly reveals that dynamic measures taken alone give a far higher speaker discrimination rate than do static measures alone for all sentences (significant at 1% level). This result confirms those of Doddington [Dod71b] which showed the importance of the dynamics of F_0, and contrasts with Furui [Fur81b] who showed that for Log Area Ratios and F_0 parameters combined, static measures were better than dynamic.

For all four sentences the dynamic based discrimination rate is over 10% higher than the equivalent static based discrimination rate (*1% significance level*). Further, contrast of these results with those of Table 6.1 in which combined static and dynamic measures are used reveals that dynamic measures alone are little worse than combined static and dynamic measures, a mean difference of only 3.3%. Such a result implies a large overlap between static and dynamic measures but that combined they yield higher discrimination rates than either alone [Fur81b].

Figure 6.13 and Table 6.4 reveal further information regarding the identification rate for all sentences combined. It can be seen that the combined sentence discrimination rate of 74.6% for dynamic measures exceeds the static rate of 54.8% by just under 20%, a very significant (*1% level*) difference, and that in fact the discrimination rate for dynamic measures of sentence one or four taken alone exceeds that of static measures of the four combined sentences.

When regarding the increase in discrimination level from single sentence to the four combined sentences we see an increase of 20.6% from the mean for dynamic measures and 13.4% for static measures. Based on such figures it seems that not only do dynamic measures have a higher discrimination level, but that they are more 'additive', and more information may be extracted from them as more sentences are included in the experiment.

Finally, it is worth noting that the discrimination rate of 74.6% for dynamic measures for all four sentences is only 0.6% lower than the discrimination rate of 75.2% for the combined dynamic and static measures.

6.2.2 Speaker Sex

Table 6.5 and Figures 6.15 and 6.16 present the speaker sex discriminant results for analysis using exclusively static or dynamic measures of the 4 basic speech parameters: Energy, F_0, Voicing, and Zero Crossings.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Static	93.7	90.9	87.7	91.5	91.0	94.7
Dynamic	93.6	88.2	89.8	92.3	91.0	95.8

Table 6.5: Speaker Sex Discrimination rates contrasting Static and Dynamic Measures. Discrimination rates are given for all four sentences, the mean across the sentence and a 'combined' rate for all sentences together. In all cases the 4 speech parameters E, F_0, Vuv and Zc were employed.

Comparing the results of Table 6.5 there appears little to differentiate static or dynamic measures as to their ability to discriminate speaker sex utilising a single sentence, both having a mean of 91.0%, and there being individual variations for the four sentences.

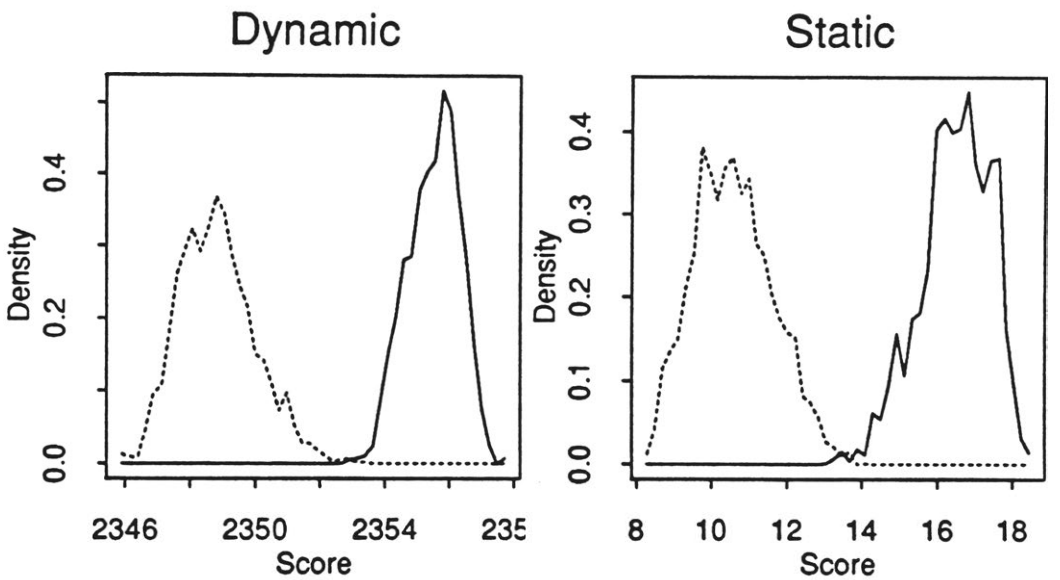


Figure 6.15: Speaker Sex Discriminate Plot - Dynamic versus Static Measures for all 4 sentences combined. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Each figure represents the combination of the relevant measures of the 4 speech parameters E, F_0, Vuv, and Zc for all of the 4 sentences and corresponds to a discriminate rate of 95.8% rate for dynamic measures and 94.7% for static measures.

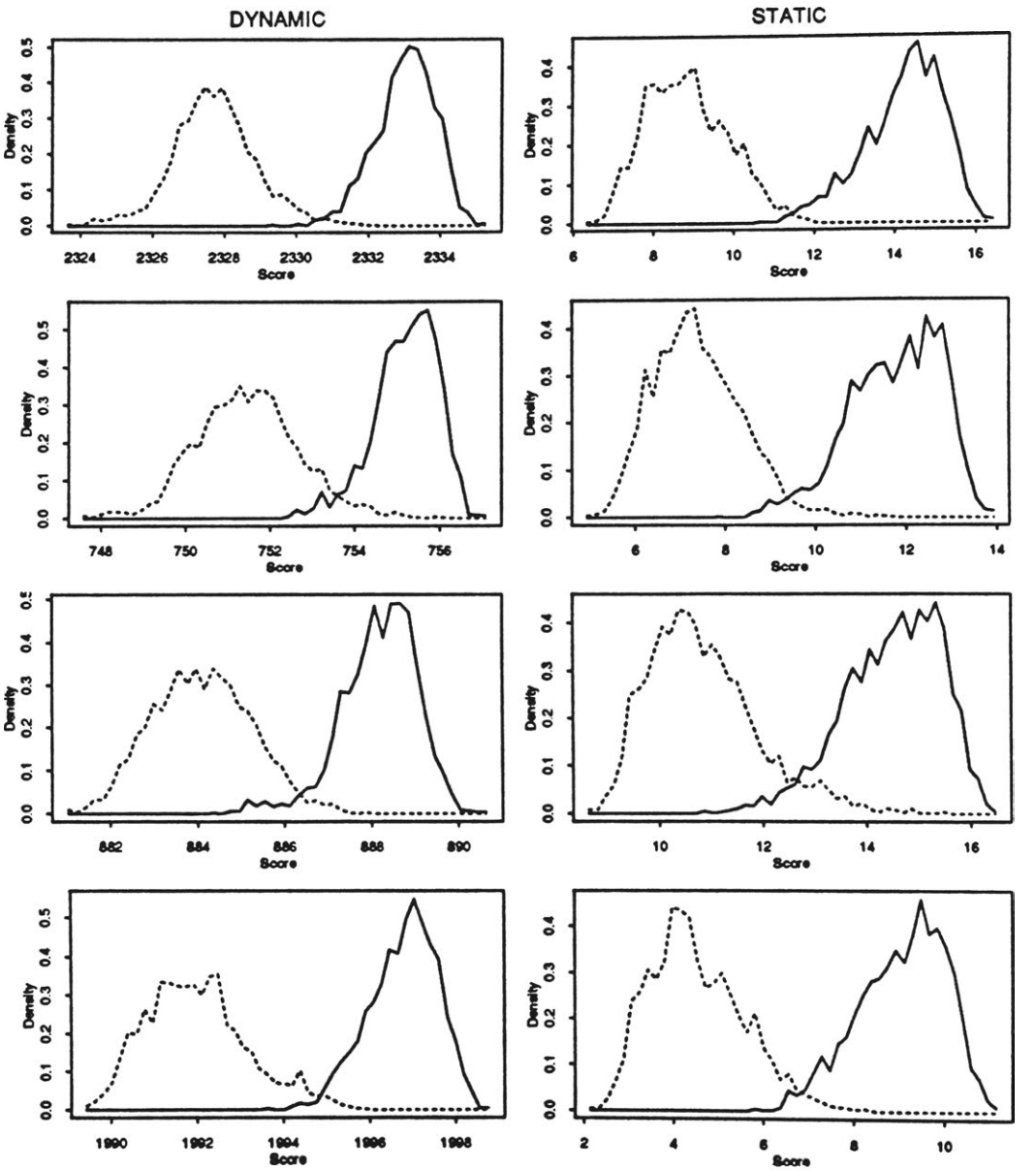


Figure 6.16: Speaker Sex Discriminate Plot - Dynamic versus Static Measures for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Results for each sentence were derived by combining the relevant (static or dynamic) measures of the 4 speech parameters E, F_0, Vuv, and Zc.

When the four sentences are combined it can be seen that the dynamic discrimination rate of 95.8% is markedly, 1.1%, higher than the static discrimination rate of 94.7%, though this is not significant at the 5% level. Further, the dynamic discrimination rate of 95.8% is only 0.4% less than the discrimination rate for combined static and dynamic measures, while the static measure discrimination rate is a non-trivial 1.5% less. Taken together these results would seem to imply that when multiple sentences are utilised that dynamic measures actually yield better discrimination for speaker sex and that static measures are largely redundant. This result of the superiority of dynamic over static appears attributable to the fact that primarily the dynamic approach is still measuring the differences in mean F_0 while additionally extracting possible further prosodic indicators of sex [Bre71, Mil88].

6.2.3 Speaker Dialect

Table 6.6 and Figures 6.17 and 6.18 are presentations of the results of the least-squares-fit analysis between dialect-difference-scores and static measure only or dynamic measure only experiments.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Static	.253	.263	.283	.253	.263	.452
Dynamic	.330	.322	.263	.291	.302	.563

Table 6.6: Speaker Dialect Correlation rates contrasting Static and Dynamic Measures. Correlation rates are given for all four sentences, the mean across the sentence and a 'combined' rate for all sentences together. In all cases the 4 speech parameters E, F_0, Vuv and Zc are employed.

From the table and Figure 6.18 it may be seen that for all sentences dynamic measures are significantly (*1% level*) more highly correlated with the dialect-difference-score than the static measures, a mean difference of 0.039. However, though higher than static measures, dynamic measures alone have a marked lower, 0.079, correlation value than when static and dynamic measures are combined.

When all four sentences are combined the difference between dynamic and static measures is enhanced, with the dynamic correlation value of 0.563 exceeding that of the static correlation, 0.452, by 0.111. Further, the difference between dynamic alone compared with dynamic and static combined is only 0.021, showing the overlap in encoding.

Clearly dynamic measures of the prosodic parameters appear to have more encoded dialect information than do static measures. This is not surprising as dialect is an acquired and to some extent alterable [Ber67] characteristic of a speaker, whereas many static measures, such as mean F_0, are highly related to physical characteristics of the speaker's anatomy.

6.3 Normalised vs. Non-Normalised Parameters

The previous Section 6.2 sought to determine somewhat of the significance of the dynamic versus static properties of the speech parameters being examined. However, much of the 'static'

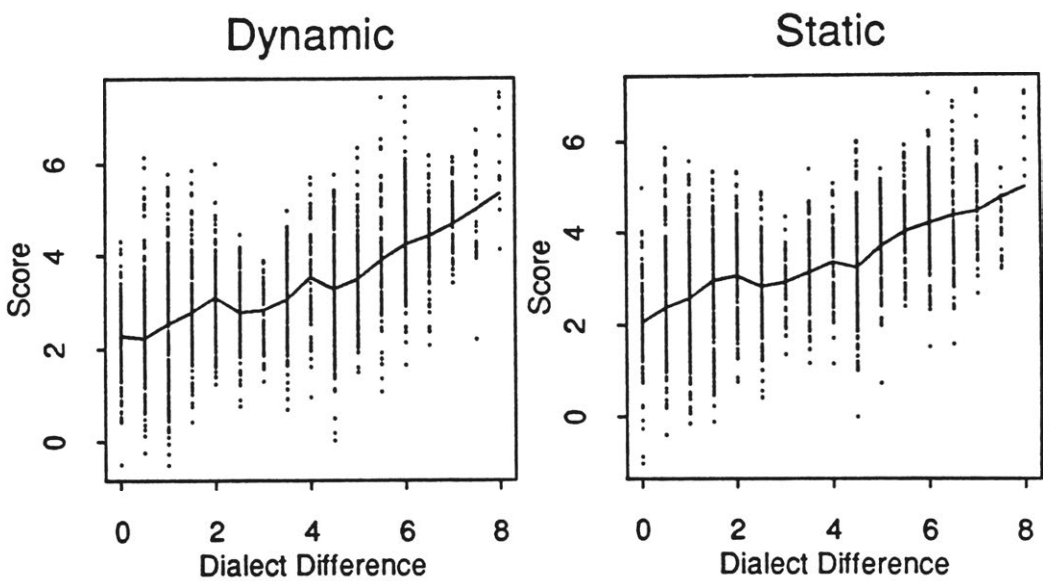


Figure 6.17: Speaker Dialect Least-Squares-Fit Scatter Plot - Dynamic versus Static Measures for all 4 sentences combined. Each figure represents the combination of relevant (static or dynamic) measures of the speech parameters E, F_0, Vuv, and Zc for all four sentences; and have a correlation values of .452 for the static plot and .563 for the dynamic plot.

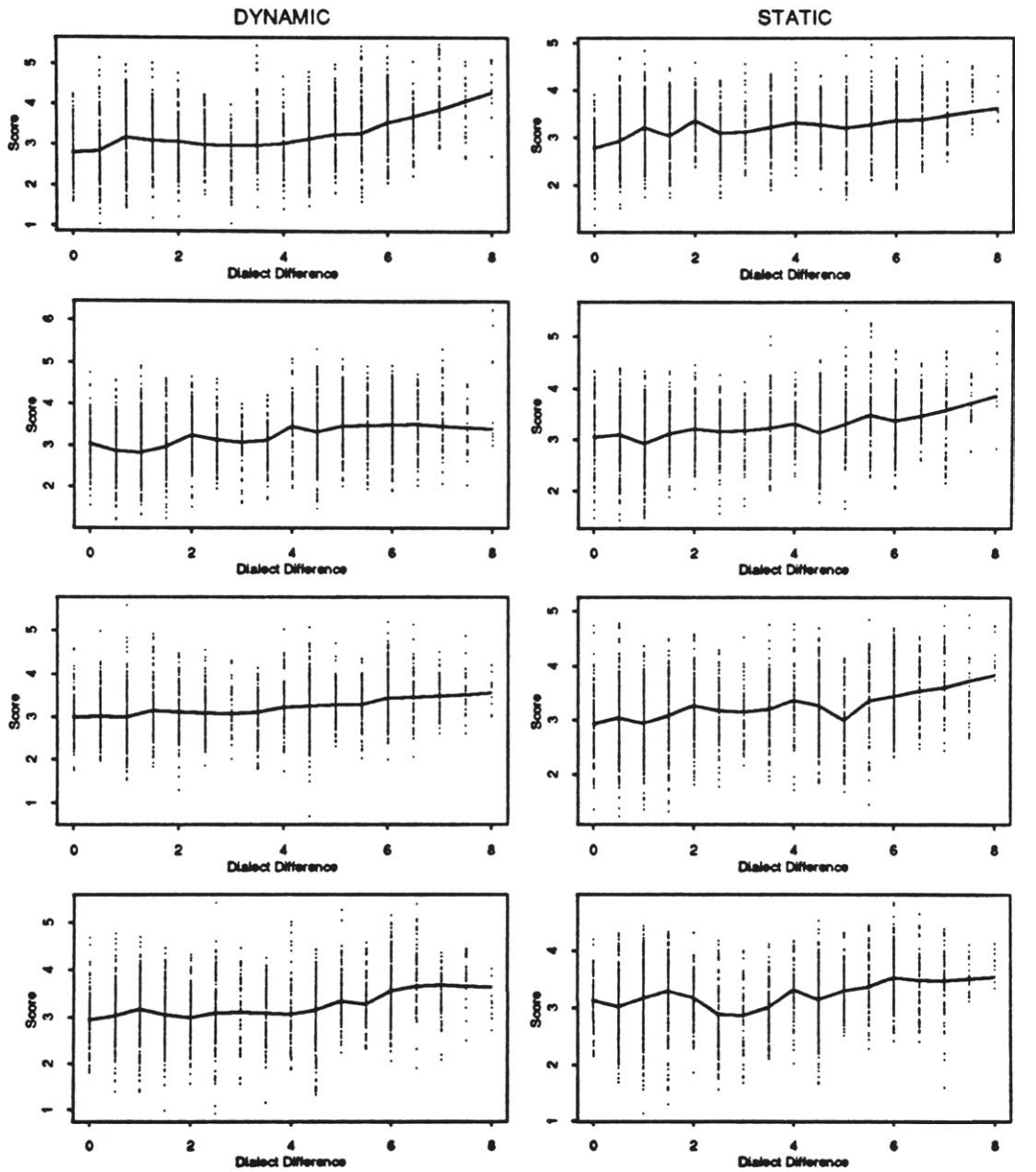


Figure 6.18: Speaker Dialect Least-Squares-Fit Scatter Plot - Dynamic versus Static Measures for each of the four sentences in turn. Results for each sentence are derived by combining the relevant (static or dynamic) measures of the speech parameters E, F_0, Vuv, and Zc.

information is still preserved in the contour during DTW experiments and no doubt affects the results (Section 6.2.2).

Normalisation is an attempt to address this issue by shifting all contours linearly into the range 0–1 prior to experimental runs (see Section 5.2.2). In effect the 'shape' of the contour is preserved but the static information of its range, and hence mean within that range, is eliminate and discarded. However no simple transformation may eliminate all static information entirely, and the static measures mean, standard deviation, 'speed', and duration may still carry encoded speaker characteristic information after normalisation, hence the static measures will still be examined.

No doubt such a transformation will adversely affect discrimination and correlation scores but the extent tells us much of the importance of the properties we have eliminated and those that remain.

Again, as for previous sections, the four basic speech parameters of the investigation, Energy, F_0, Voicing, and Zero Crossing Rate will be utilised in combination. The results will be analysed from three perspectives. The first, a straight contrast between discrimination or correlation rates for normalised and non-normalised parameters when all four sentences are combined. Secondly a contrast between normalised and non-normalised results for each of the 4 sentences in turn. Finally a two-way analysis where all 4 sentences are combined but examination is of normalised versus non-normalised as subdivided for static and dynamic measures.

6.3.1 Speaker Identity

Tables 6.7 and 6.8 and Figures 6.19, 6.20, and 6.21 represent the results of the analysis of the effects of normalisation upon speaker discrimination.

"Treatment" Type	*Sentence*					
	1	2	3	4	Mean	Combined
Non-Normalised	61.9	53.8	53.9	59.4	57.3	75.2
Normalised	52.4	44.1	41.2	51.4	47.3	70.5

Table 6.7: Speaker Identity Discrimination Rates contrasting normalised and non-normalised parameters on the basis of sentence. In all cases the 4 parameters E, F_0, Vuv, and Zc are employed.

"Treatment" Type	*Measure Type*		
	Static	Dynamic	Combined
Non-Normalised	54.8	74.6	75.2
Normalised	46.6	69.6	70.5

Table 6.8: Speaker Identity Discrimination Rates. Two way contrast of normalised and non-normalised parameters versus static and dynamic measures of the parameters. In all cases the 4 parameters E, F_0, Vuv, and Zc are employed.

It is greatly surprising that if all four sentences are utilised, Figure 6.19, there is only a drop of 4.7% in discriminant performance from non-normalised to normalised parameters. Clearly

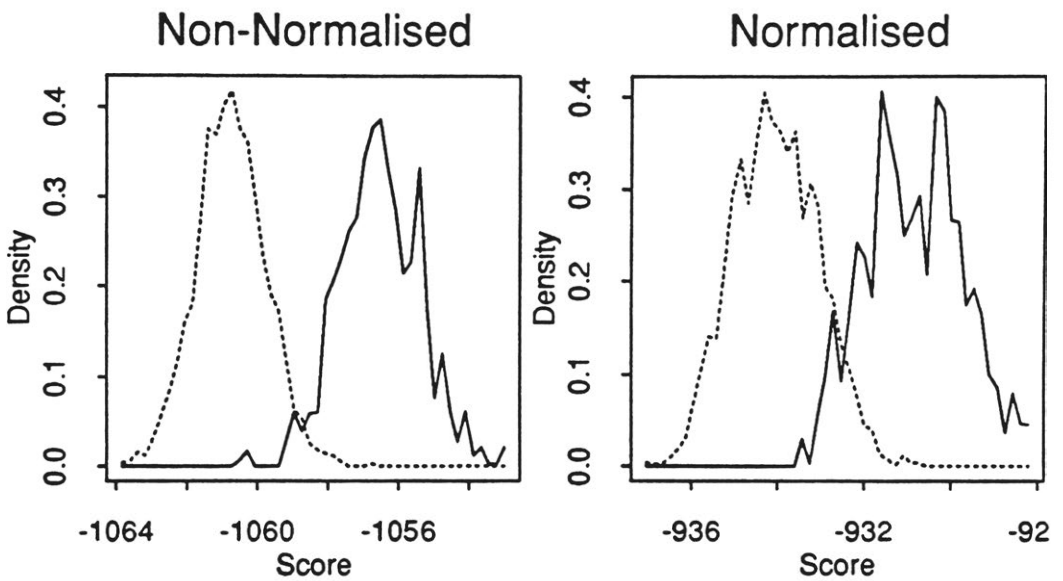


Figure 6.19: Speaker Identity Discriminate Plot - Normalised versus Non-Normalised parameters with all four sentences combined. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Results for each figure were derived by combining measures of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc across all 4 sentences.

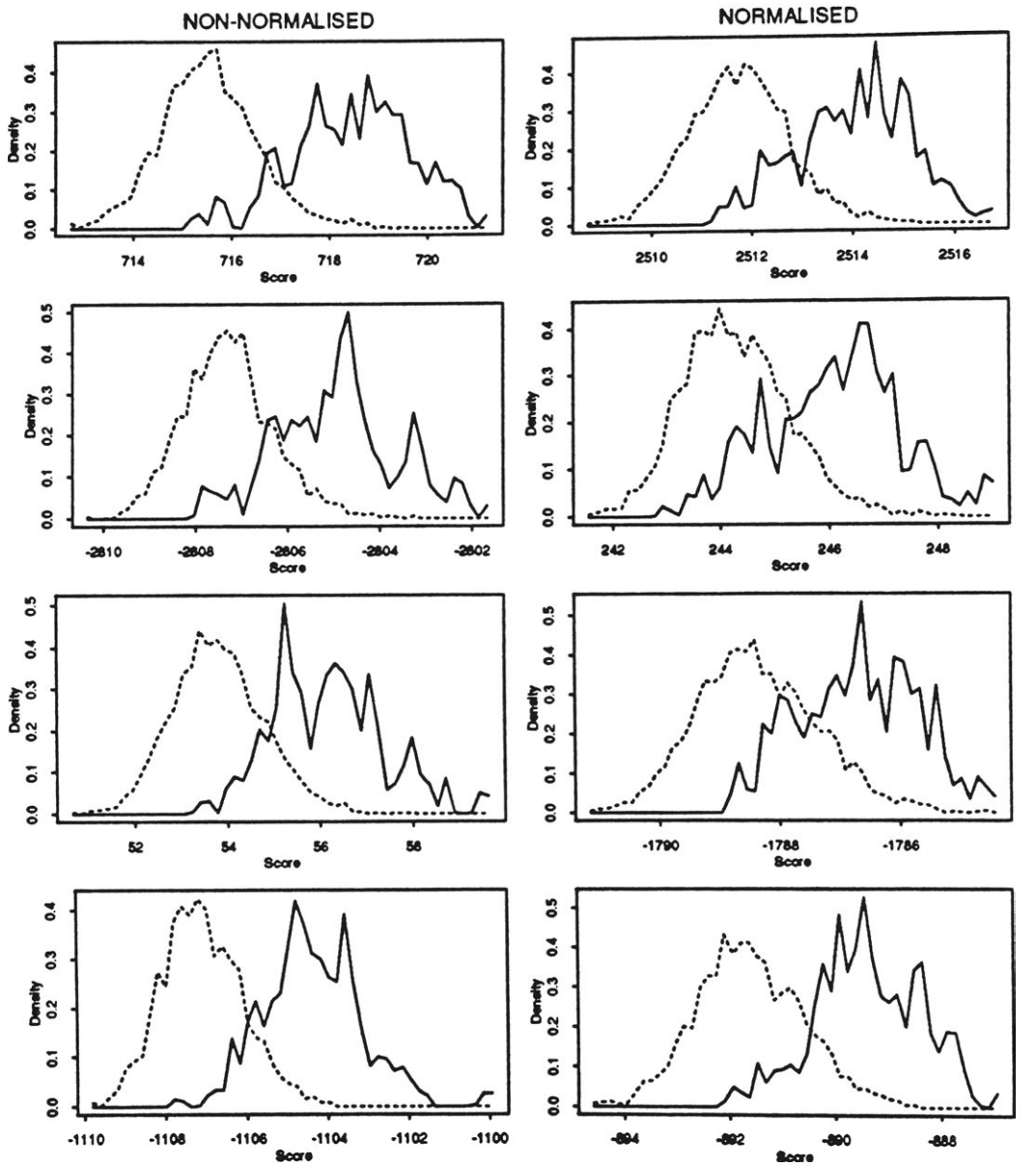


Figure 6.20: Speaker Identity Discriminate Plot - Normalised versus Non-Normalised parameters for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Results for each sentence were derived by combining measures of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc.

the elimination of static information such as range, and mean within range, has not greatly affected discrimination performance when the sentences are being utilised together, implying that there is a great degree of speaker identity encoding in the dynamics of the parameters.

Examination of the results on a single sentence basis, Figure 6.20 and Table 6.7 shows that normalisation drops identification rates an average of 10% which contrasts with the 4.7% when all sentences are combined. Clearly normalisation has more 'drastic' results upon discrimination scores when only a single sentence is used, however when sentences are combined these effects are lessened. It therefore appears that the addition of further speech material, in part at least, compensates for the reduction in discrimination rate due to normalisation. It remains to be determined whether, with sufficient speech material, this reduction may be eliminated or if normalisation imposes a permanent loss in speaker discrimination rate.

Figure 6.21 and Table 6.8 illustrate where the drop in discriminant performance due to normalisation is most marked. Not surprising is the drop in the static measures' discriminant ability from 54.8% to 46.6% a drop of 8.2%. However, of major significance is the difference between discrimination rates for dynamic measures when using non-normalised and normalised parameters. Surprisingly there is a drop of only 5% in performance, undoubtedly signifying that it is the dynamic or temporal properties of the contours that are being utilised as opposed to the static. Comparing discrimination rates for static measures of the non-normalised parameters with discrimination rates for dynamic measures of the normalised parameters it is clear that the later are significantly (*1% level*) superior to the former (54.6% versus 69.6%). On the basis of this result it appears that identity is more strongly encoded in the dynamic (time varying) properties, of the prosodic contours examined, than in the static (time invariant) properties.

6.3.2 Speaker Sex

Tables 6.9 and 6.10, and Figures 6.22, 6.23, and 6.24 represent the results of the application of normalisation upon speaker sex discrimination.

"Treatment" Type	*Sentence*					
	1	2	3	4	Mean	Combined
Non-Normalised	95.2	92.5	91.0	93.5	93.1	96.2
Normalised	55.9	37.9	58.6	62.9	53.8	77.7

Table 6.9: Speaker Sex Discrimination Rates contrasting normalised and non-normalised parameters on the basis of sentence. In all cases the 4 parameters E, F_0, Vuv, and Zc are employed.

"Treatment" Type	*Measure Type*		
	Static	Dynamic	Combined
Non-Normalised	94.7	95.8	96.2
Normalised	52.1	77.0	77.7

Table 6.10: Speaker Sex Discrimination Rates. Two way contrast of normalised and non-normalised parameters versus static and dynamic measures of the parameters. In all cases the 4 parameters E, F_0, Vuv, and Zc are employed.

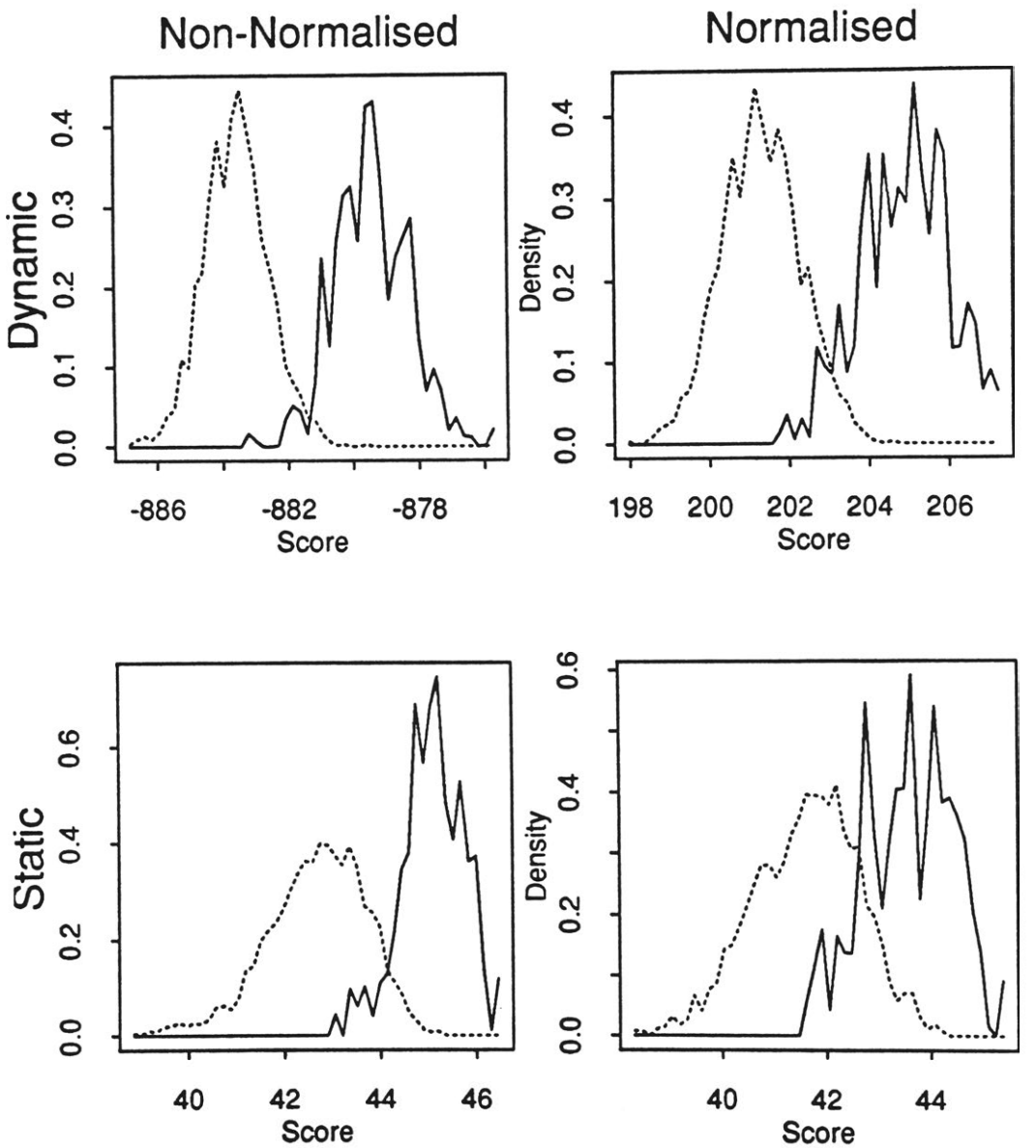


Figure 6.21: Speaker Identity Discriminate Plot - 2 Way breakdown of Normalised and Non-Normalised parameters together with static versus dynamic measures. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Results for each plot were derived by combining the relevant measures (static or dynamic) of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc across all 4 sentences.

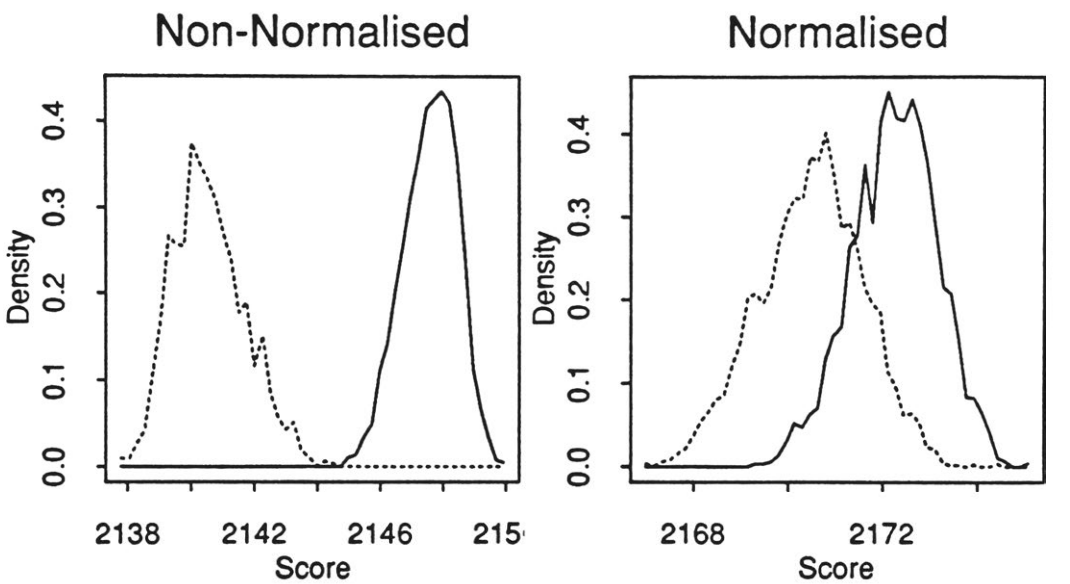


Figure 6.22: Speaker Sex Discriminate Plot - Normalised versus Non-Normalised parameters with all four sentences combined. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Results for each figure were derived by combining measures of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc across all 4 sentences.

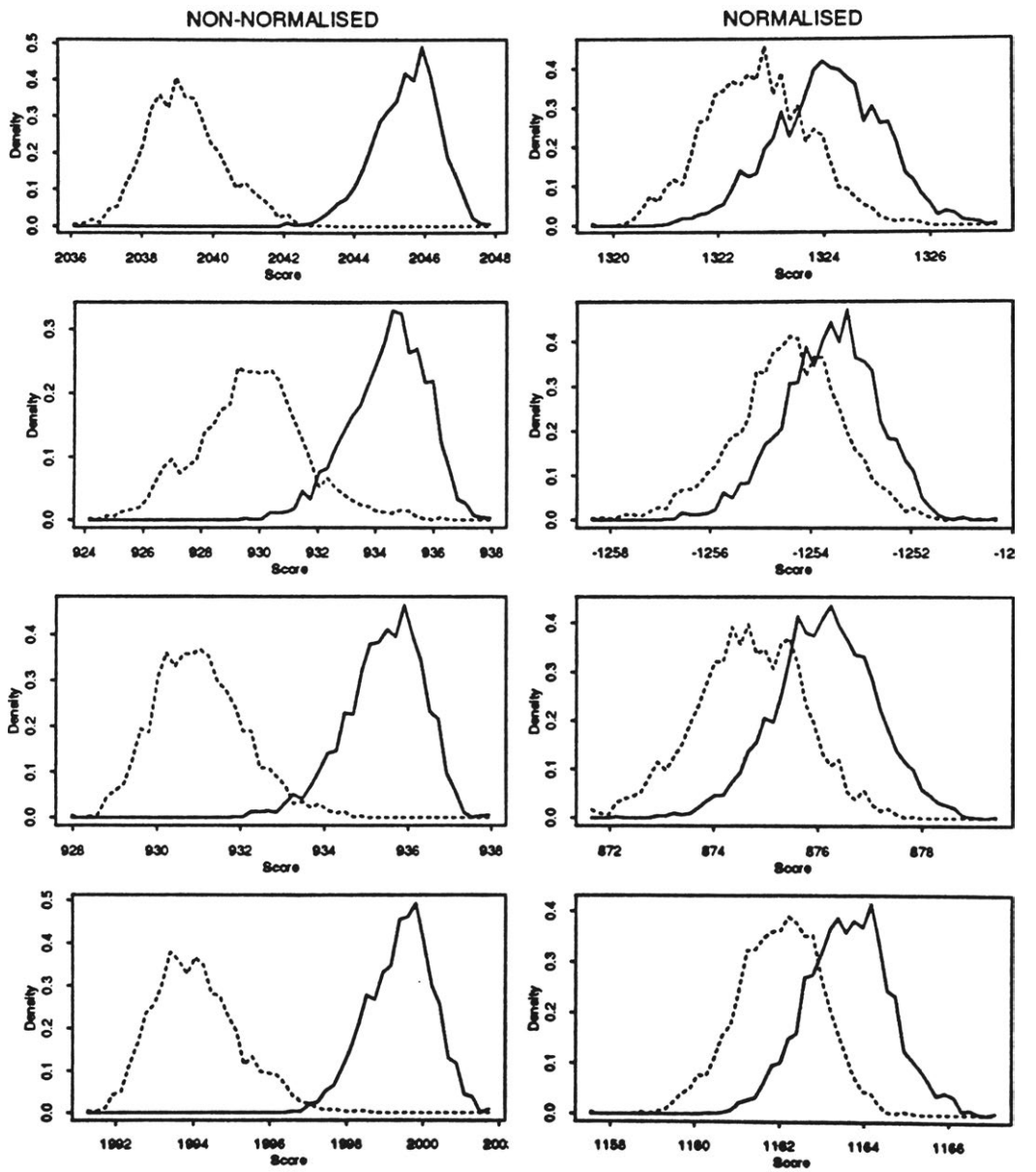


Figure 6.23: Speaker Sex Discriminate Plot - Normalised versus Non-Normalised parameters for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Results for each sentence were derived by combining measures of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc.

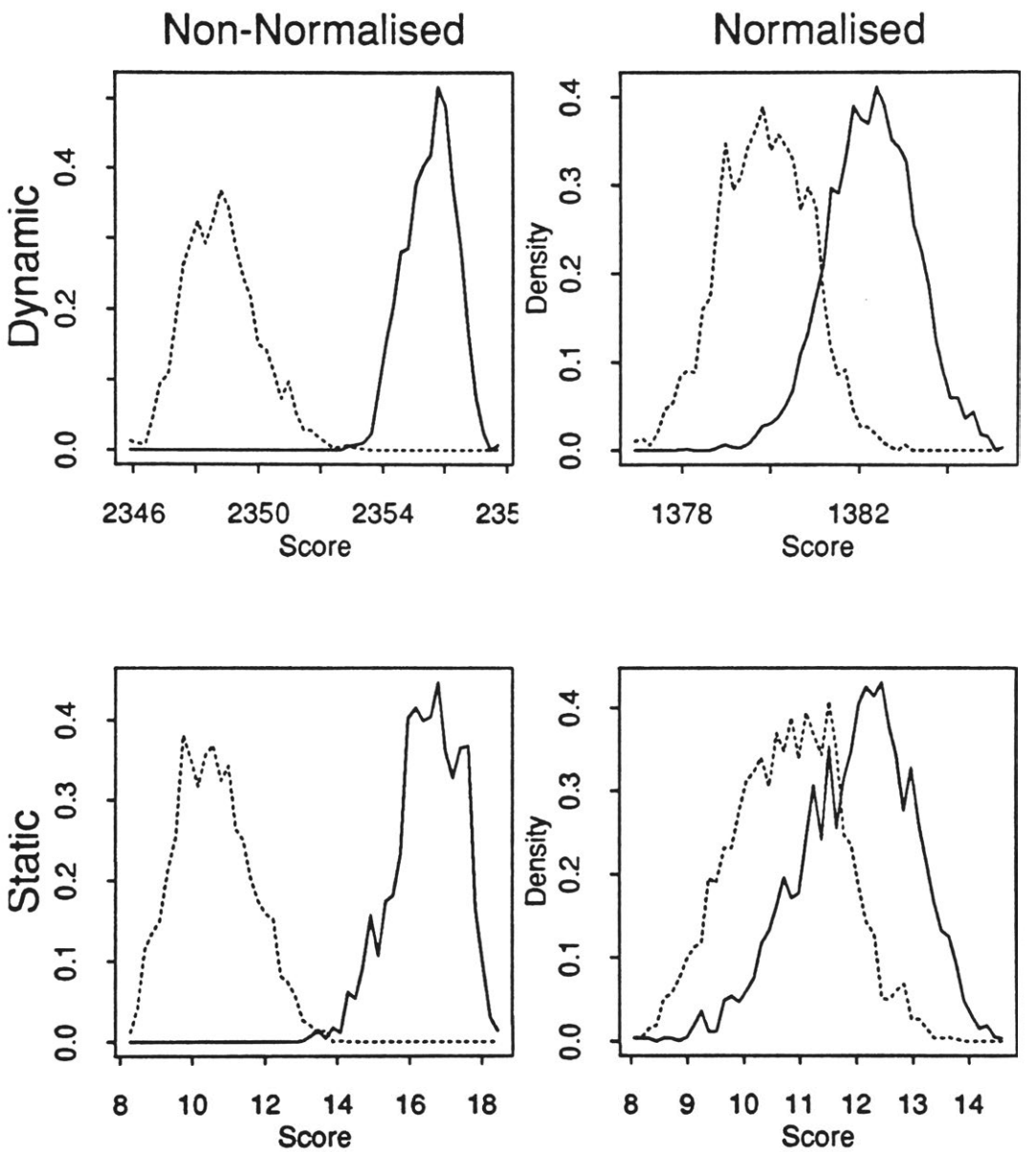


Figure 6.24: Speaker Sex Discriminate Plot - 2 Way breakdown of Normalised and Non-Normalised parameters together with static versus dynamic measures. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Results for each plot were derived by combining the relevant measures (static or dynamic) of the non-normalised or normalised speech parameters E, F_0, Vuv, and Zc across all 4 sentences.

Not surprisingly, as is evident from Figure 6.15, the ability to discriminate speaker sex is greatly affected by normalisation— a drop from 96.2% to 77.7% in the four-combined-sentence case. What is remarkable is the high discrimination rate of 77.7%, given that the information conveyed by the mean F_0 has effectively been eliminated. Clearly speaker sex is encoded in more than simply the mean F_0 of an utterance [PB52, Mil88], and in fact in the dynamics of prosodics.

An analysis upon a single sentence basis, as shown in Table 6.9, and Figure 6.23, reconfirms the drastic effect of the normalisation scheme upon sex discrimination. There is a marked difference between the sentences as to the effect of normalisation with the discrimination for sentence two being little better than chance, 37.9% as opposed to sentence four having the best discrimination of 62.9% (significant at the 1% level of confidence). Further, it is worth noting the difference between the mean discrimination rate for a single sentence of 53.8% and that of 77.7% for the four combined sentences, showing that it is necessary to utilise more speech material than a single sentence in order to determine speaker sex with a reasonably high (75%) degree of accuracy, if normalisation has occurred.

When the results are subdivided on the basis of static and dynamic measures, as shown in Table 6.10 and Figure 6.24, it is evident that static measures are the most strongly affected by the normalisation process, dropping from 94.7% to 52.1%, as opposed to dynamic measures' drop from 94.8% to 77.0%. Such a result further shows that after normalisation dynamic measures are far superior indicators of speaker sex when compared with static measures. However, comparing discrimination rates for static measures of non-normalised parameters and dynamic measures of normalised parameters it is clear that static measures of non-normalised parameters are significantly (*1% level*) better. Given this result it appears that speaker sex is more strongly encoded in the static (time invariant) properties of the prosodic parameters examined, rather than the dynamic (time varying) properties.

6.3.3 Speaker Dialect

Tables 6.11 and 6.12 and Figures 6.25, 6.26, and 6.27 represent the results of normalisation upon speaker dialect correlation rates.

"Treatment" Type	*Sentence*					
	1	2	3	4	Mean	Combined
Non-Normalised	.405	.408	.361	.349	.381	.584
Normalised	.357	.410	.280	.279	.332	.537

Table 6.11: Speaker Dialect Correlation Rates contrasting normalised and non-normalised parameters on the basis of sentence. In all cases the 4 speech parameters E, F_0, Vuv and Zc are employed.

It can be seen from Table 6.11 and Figure 6.26 that normalisation does, in general, reduce the correlation score for individual sentences, a mean reduction of 0.049.

Similarly, when all four sentences are combined, normalisation reduces the correlation rate, a drop of 0.047 from 0.584 to 0.537. However, for both single sentences and all four sentences

"Treatment" Type	*Measure Type*		
	Static	Dynamic	Combined
Non-Normalised	.452	.563	.584
Normalised	.270	.524	.537

Table 6.12: Speaker Dialect Correlation Rates. Two way contrast of normalised and non-normalised parameters versus static and dynamic measures of the parameters. In all cases the 4 speech parameters E, F_0, Vuv and Zc are employed.

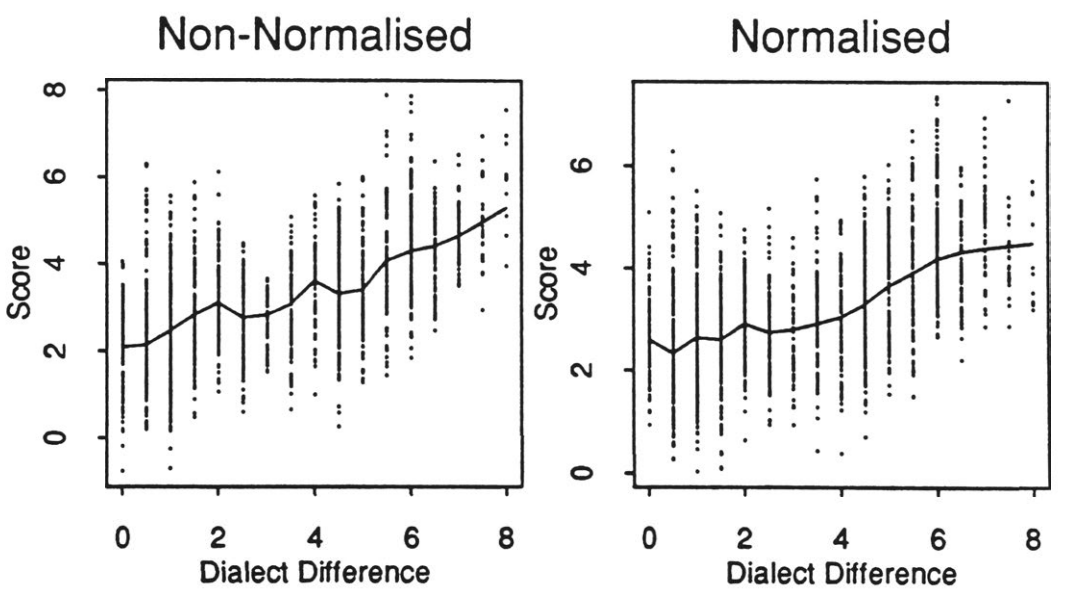


Figure 6.25: Speaker Dialect Least-Squares-Fit Scatter Plot - Normalised versus Non-Normalised parameters for all 4 sentences combined. Results for each figure are derived by combining all measures of the normalised or non-normalised parameters E, F_0, Vuv, and Zc across all 4 sentences.

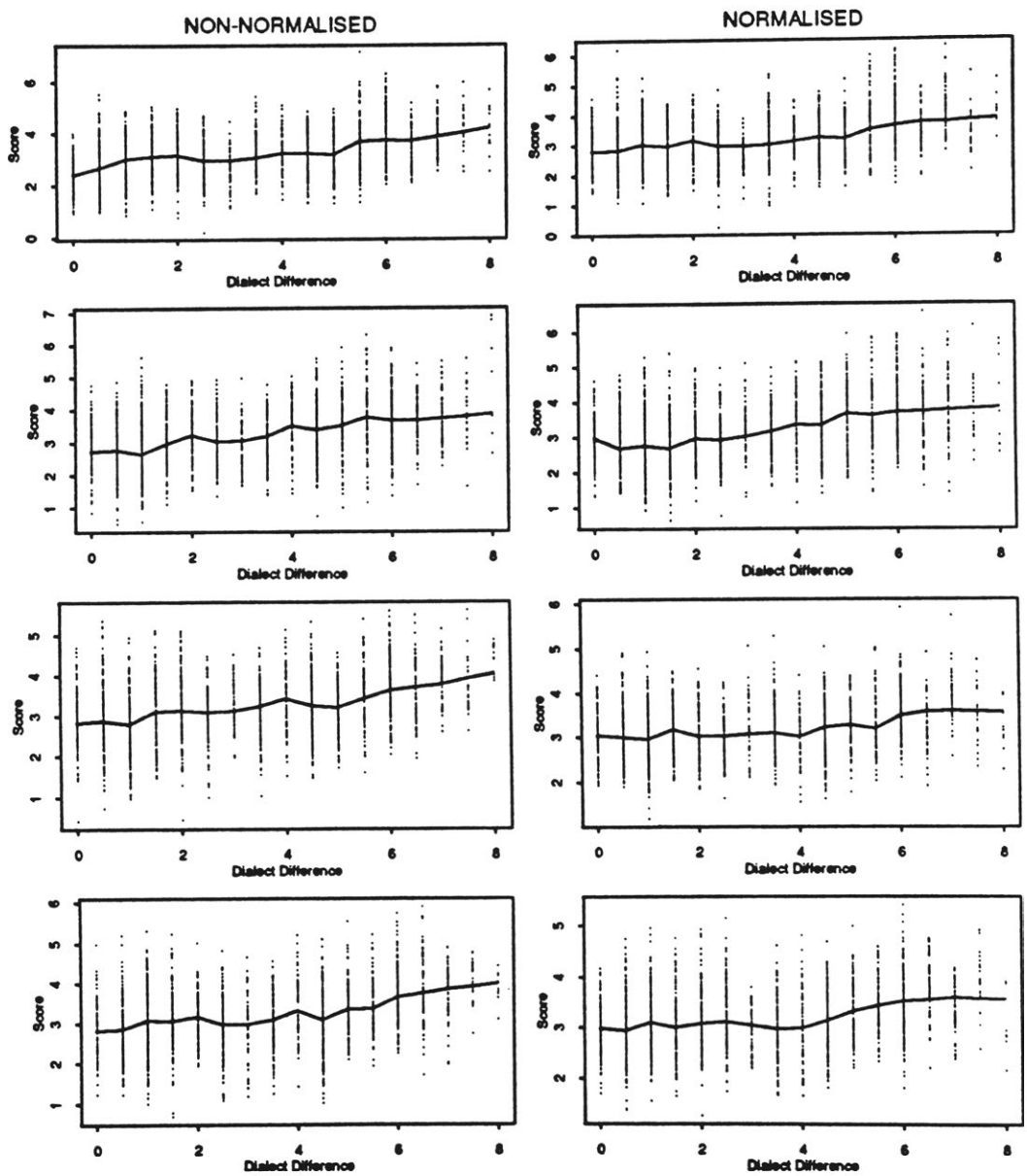


Figure 6.26: Speaker Dialect Least-Squares-Fit Scatter Plot - Normalised versus Non-Normalised parameters for the 4 sentences in turn. Results for each sentence are derived by combining all measures of the normalised or non-normalised parameters E, F_0, Vuv, and Zc.

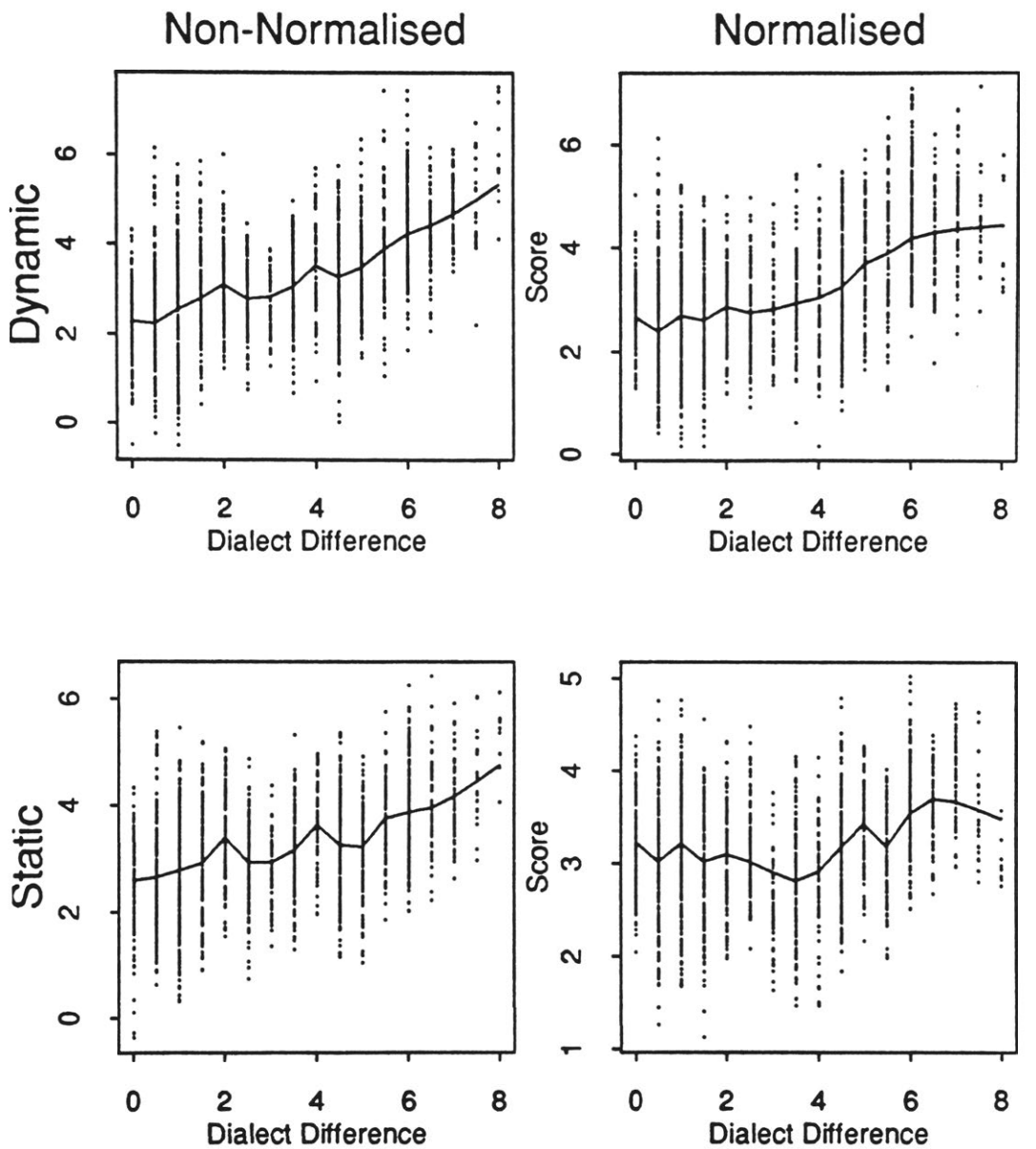


Figure 6.27: Speaker Dialect Least-Squares-Fit Scatter Plot - 2 Way breakdown of Normalised and Non-Normalised parameters together with static versus dynamic measures. Results for each figure are derived by combining the relevant measures (static or dynamic) of the normalised or non-normalised parameters E, F_0, Vuv, and Zc across all 4 sentences.

combined, the relatively minor effect of normalisation is somewhat unexpected, tending to indicate that it is the temporal measures that are largely being utilised. Examining Figure 6.27 and Table 6.12, which shows a breakdown for static and dynamic measures, further credence is lent to this theory. It can be seen that there is a relatively small drop for dynamic measures for un-normalised to normalised parameters, 0.039, and that for normalised measures the difference between correlation scores for dynamic measures, 0.524, and combined dynamic-static measures, 0.537, is minimal: 0.013.

Contrasting the correlation value for static measures of the non-normalised parameters, 0.452, with that for dynamic measures of the normalised parameters, 0.524, a significant (*1% level*) difference is noted. Given this result it appears that speaker dialect is more strongly encoded in the dynamic (time varying) properties of the prosodic parameters examined, than in the static (time invariant) properties.

6.4 Comparison of the 4 Basic Parameters

In all previous sections the four basic speech parameters under investigation, namely Energy, F_0, Voicing, and Zero Crossing Rate have been combined and their individual contributions submerged in order to examine other questions not directly related to the speech parameters. Many speaker recognition systems use F_0 [INN78, Fur81b, Mat89], or energy [Lum73, MOJ77] as additional vectors in their recognition system, while zero crossing and voicing appear to have received very little [WD75, JHH84, BW88] attention. However it is important to examine results on a parameter basis in order to further define encoding of speaker characteristics in speech and to aid in the choice of appropriate speech parameters for speaker characteristic 'recognition' systems.

For each speaker characteristic examined the results are analysed on a speech parameter basis, utilising the speech material of all four sentences. The results are also split on the basis of static and dynamic measures of the four speech parameters. Appendix D gives a more thorough breakdown of the results for all speech parameters, showing discrimination and correlation rates for each parameter for each sentence—using dynamic alone, static alone, and combined static and dynamic measures.

6.4.1 Speaker Identity

Figures 6.28 and 6.29 and Table 6.13 present the results of the speaker discriminant experiments as analysed on the basis of the four parameters Energy, F_0, Voicing and Zero Crossing Rate.

It is evident from Table 6.13 and Figure 6.28 that there is a definite ordering of parameters with regard to speaker discrimination levels, and the differences between all parameters are significant at the 1% level. F_0 yields the highest discrimination rate of 64.6%, followed by energy, zero crossings, and finally voicing with a discrimination rate of 47.6%. This result reconfirms the significance of F_0, as a parameter for speaker recognition, that many previous researchers [Dod71b, Ata72, Wol72, Fur81b] have previously shown. Based on these results it

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Energy	43.4	56.6	58.2
F_0	45.5	63.5	64.6
Voicing	34.0	47.0	47.6
Zero Crossing	39.1	50.9	51.5

Table 6.13: Speaker Identity Discrimination Rates for each of the four basic parameters. Energy, F_0, Voicing and Zero Crossing Rate are analysed separately as to their ability to discriminate speaker identity. Analysis is further split to dynamic and static measures of the parameters in question. Quoted rates are for all 4 sentences combined.

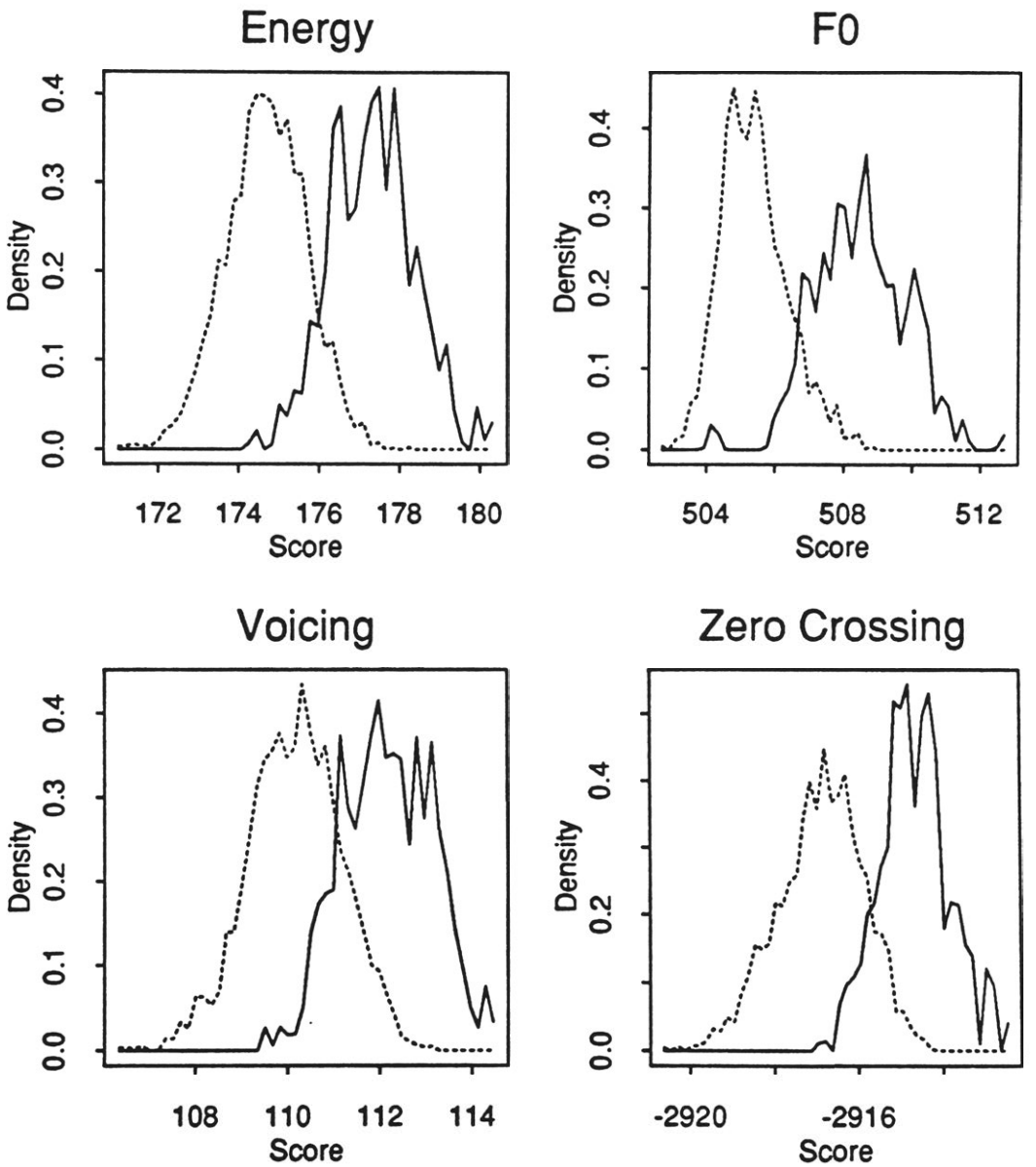


Figure 6.28: Speaker Identity Discriminate Plot - Contrasting the 4 speech parameters E, F_0, Vuv, and Zc utilising all 4 sentences. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the 21 measures of the appropriate speech parameter for all 4 sentences.

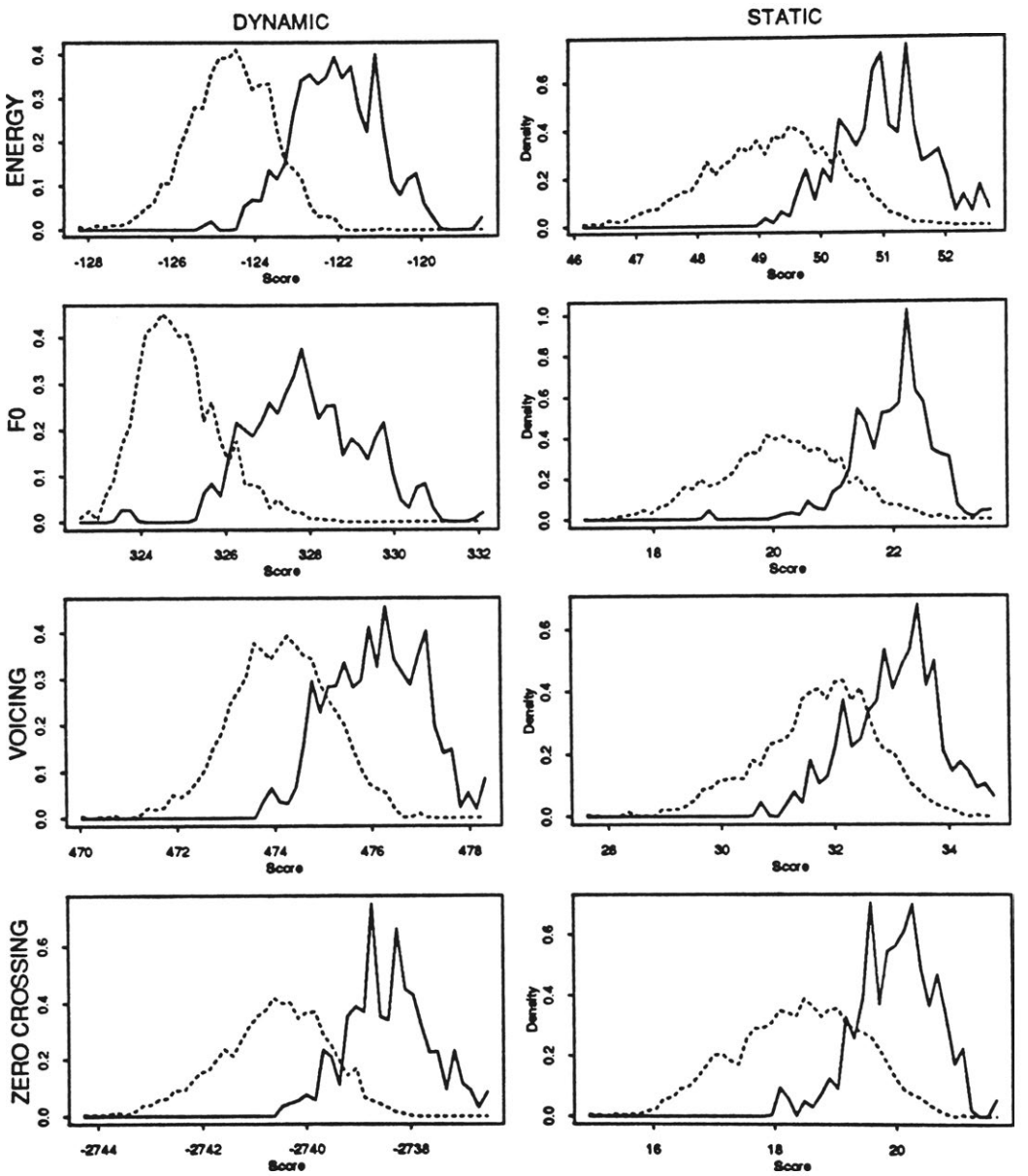


Figure 6.29: Speaker Identity Discriminate Plot - Contrasting the 4 speech parameters E, F_0, Vuv, and Zc subdivided as to static versus dynamic measures. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant (static or dynamic) measures of the appropriate speech parameter for all 4 sentences.

is evident that there are significant differences between the individual parameters with respect to their speaker discriminant abilities.

It is worth noting that while F_0 yields the highest discrimination rate of 64.6% this is still over 10% less than the discrimination rate when all four speech parameters are utilised. Clearly the encoding of speaker identity is distributed over the four parameters such that no single parameter is sufficient to encapsulate all the encoded speaker specific information.

When the parameters are further split with respect to static and dynamic measures, as shown in Figure 6.29, it is evident that dynamic measures of the parameter are always significantly (*1% level*) better at discriminating speaker than are static measures, the difference ranging from 11.8% up to 18%. Further, the dynamic alone discrimination rate is little different to the combined dynamic-static identification rate.

6.4.2 Speaker Sex

Figures 6.30 and 6.31 combined with Table 6.14 are presentations of the results of the sex discriminant experiments as analysed on the basis of the four speech parameters:- Energy, F_0, Voicing, and Zero Crossing Rate.

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Energy	47.5	54.8	55.9
F_0	95.0	95.1	95.5
Voicing	41.0	69.1	71.8
Zero Crossing	39.0	37.2	42.2

Table 6.14: Speaker Sex Discrimination Rates for each of the four basic parameters. Energy, F_0, Voicing and Zero Crossing Rate are analysed separately as to their ability to discriminate speaker sex. Analysis is further split to dynamic and static measures of the parameters in question. Quoted rates are for all 4 test sentences combined.

It is evident from Figure 6.30 and Table 6.14 that fundamental frequency yields a significantly (*1% level*) higher sex discrimination rate than any of the other three parameters, confirming this well known result [Wea24, HHP88].

However, it can also be seen that to lesser and greater degrees the other three parameters do provide some indication of speaker sex, with voicing the highest at a significant 71.8%. The interpretation of this voicing difference is unclear at this stage, it possibly being an artifact of the parameter extraction routine, though Millar [Mil88] has also reported a comparable result with a similar uncertainty as to its cause [Mil91].

When the results are further split based on static versus dynamic measures it can be noted that for all but zero crossing rate, which in fact has the lowest discriminant level, the dynamic measures of the parameters have higher discrimination rates than do the static measures.

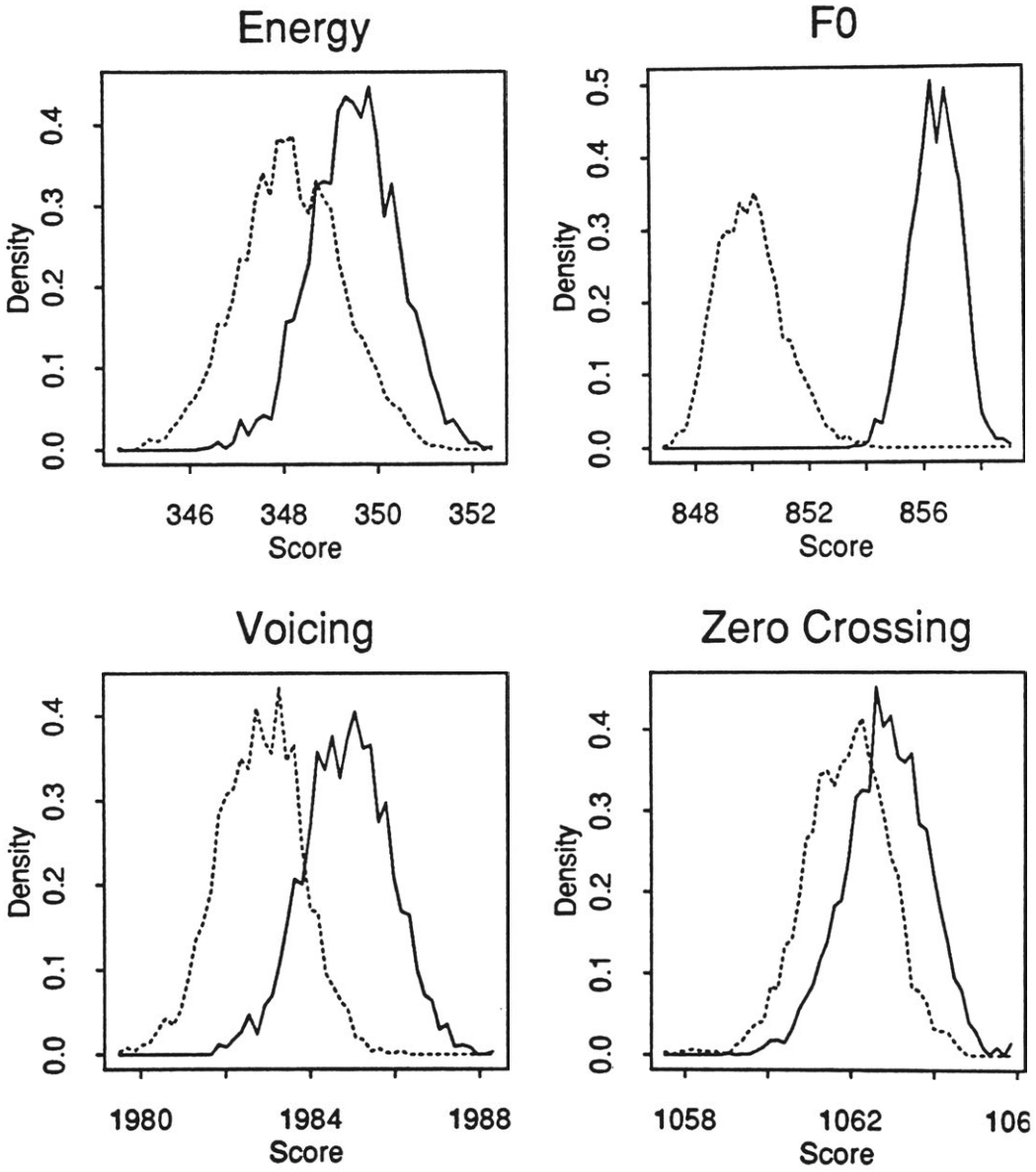


Figure 6.30: Speaker Sex Discriminate Plot - Contrasting the 4 speech parameters E, F_0, Vuv, and Zc utilising all 4 sentences. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the 21 measures of the appropriate speech parameter for all 4 sentences.

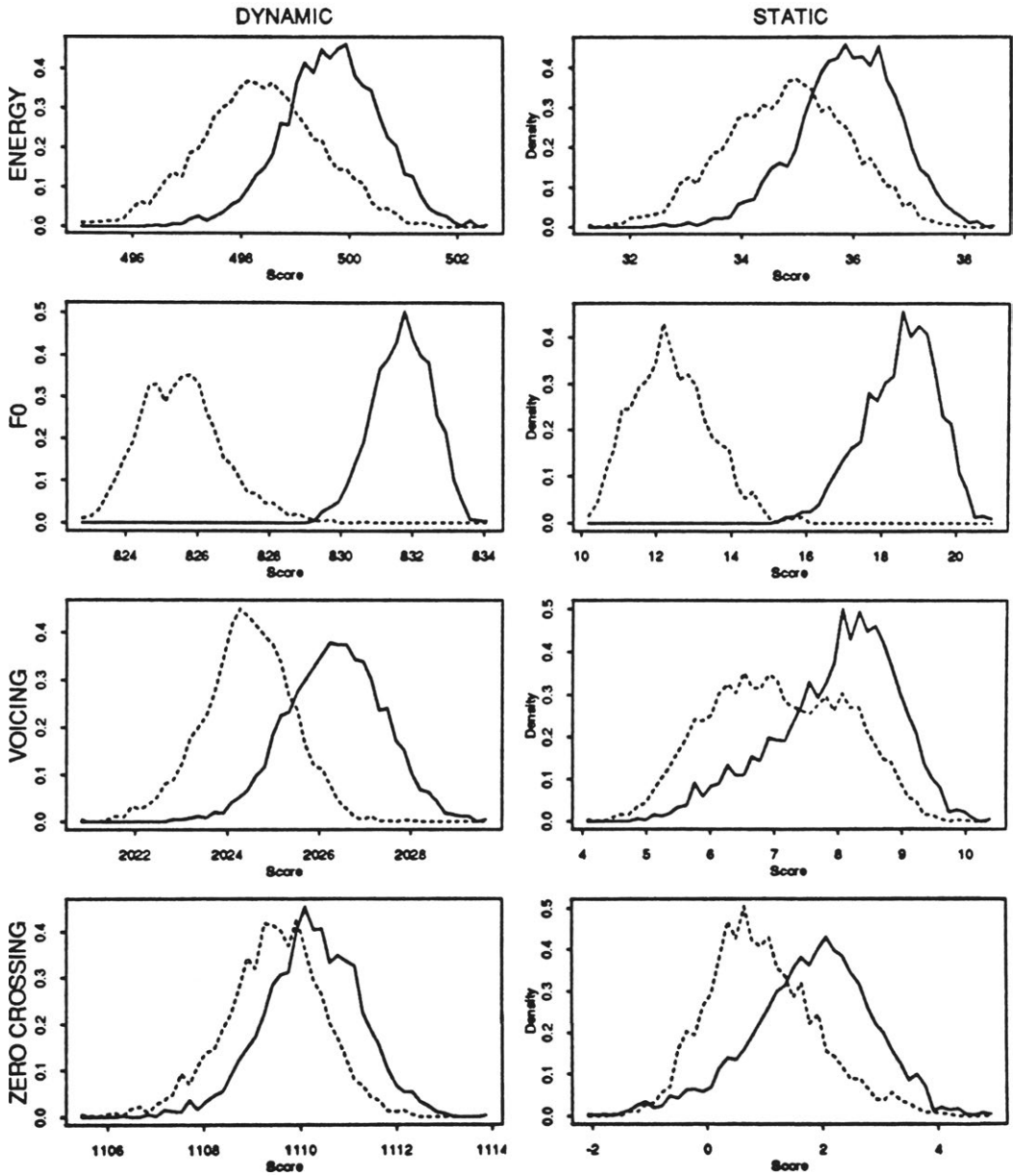


Figure 6.31: Speaker Sex Discriminate Plot - Contrasting the 4 speech parameters E, F_0, Vuv, and Zc subdivided as to static versus dynamic measures. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant (static or dynamic) measures of the appropriate speech parameter for all 4 sentences.

6.4.3 Speaker Dialect

Figures 6.32 and 6.33 combined with Table 6.15 present the results of the dialect-difference-score correlation experiments, analysed on the basis of the four speech parameters Energy, F_0, Voicing, and Zero Crossing Rates.

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Energy	.287	.350	.359
F_0	.322	.333	.364
Voicing	.216	.255	.261
Zero Crossing	.297	.268	.306

Table 6.15: Speaker Dialect Correlation Rates for each of the four basic parameters. Energy, F_0, Voicing and Zero Crossing Rate are analysed separately as to their correlation to speaker dialect. Analysis is further split into dynamic and static measures of the parameters in question.

It is evident from Figure 6.32 and Table 6.15 that of the four basic parameters examined F_0 and energy appear significantly (*1% level*) superior in terms of the correlation of their measures to the dialect-difference-score, with correlation scores of 0.364 and 0.359 respectively. Wagner [Wag78] has previously shown that for both F_0 and energy there appear dialect differences at particular speech events, and it now appears that these dialect differences exist across the duration of a sentence, not solely at particular speech events. Further, zero crossing rate has a significantly (*1% level*) different correlation rate to that of the lowest: voicing, with a value of 0.261. Such results shown that there are significant differences between the speech parameters with regard to dialect encoding.

When the results are split on the basis of static versus dynamic measures of the parameters it is found, as it was for speaker sex, that for all parameters, barring the zero crossing rate, the dynamic measures yield a markedly better correlation score than did the static measures.

It is also worth noting that the correlation rates for the two 'best' parameters alone: F_0 with 0.364 and energy with 0.359 are still markedly less than the combined parameter correlation value of 0.584, indicating the apparent importance of combining all possible speech parameters in an examination of speaker dialect, and that no single parameter has encoded all speaker dialect information.

6.5 Comparison of F_0 Representations

Up till this point the four parameters Energy, F_0, Voicing and Zero Crossing Rate have been investigated. In order to eliminate an extra dimension of complexity the linear-concatenated representation of F_0 was the only one of the 4 possible F_0 representations used. This representation had been selected as the default due to its simplicity and the fact that it lead to a stronger separation between F_0 and voicing parameters.

However, it is important to contrast the various possible F_0 representations to determine whether any particular F_0 representation is superior in its ability to discriminate the speaker characteristics.

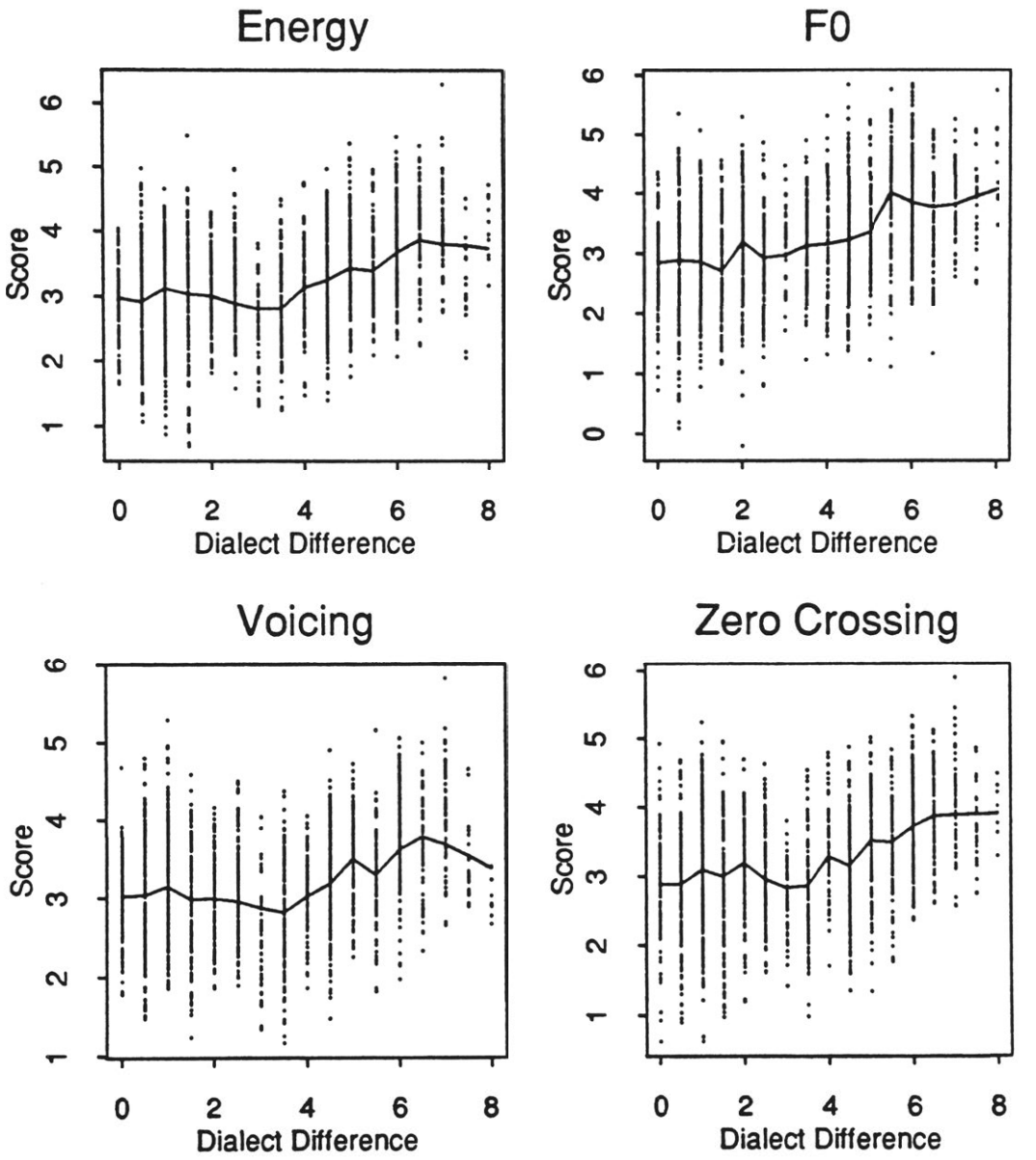


Figure 6.32: Speaker Dialect Least-Squares-Fit Scatter Plot - Contrast of the 4 speech parameters E, F_0, Vuv, Zc with all 4 sentences combined. Figures are derived via combining all 21 measures of the relevant speech parameter across all 4 sentences.

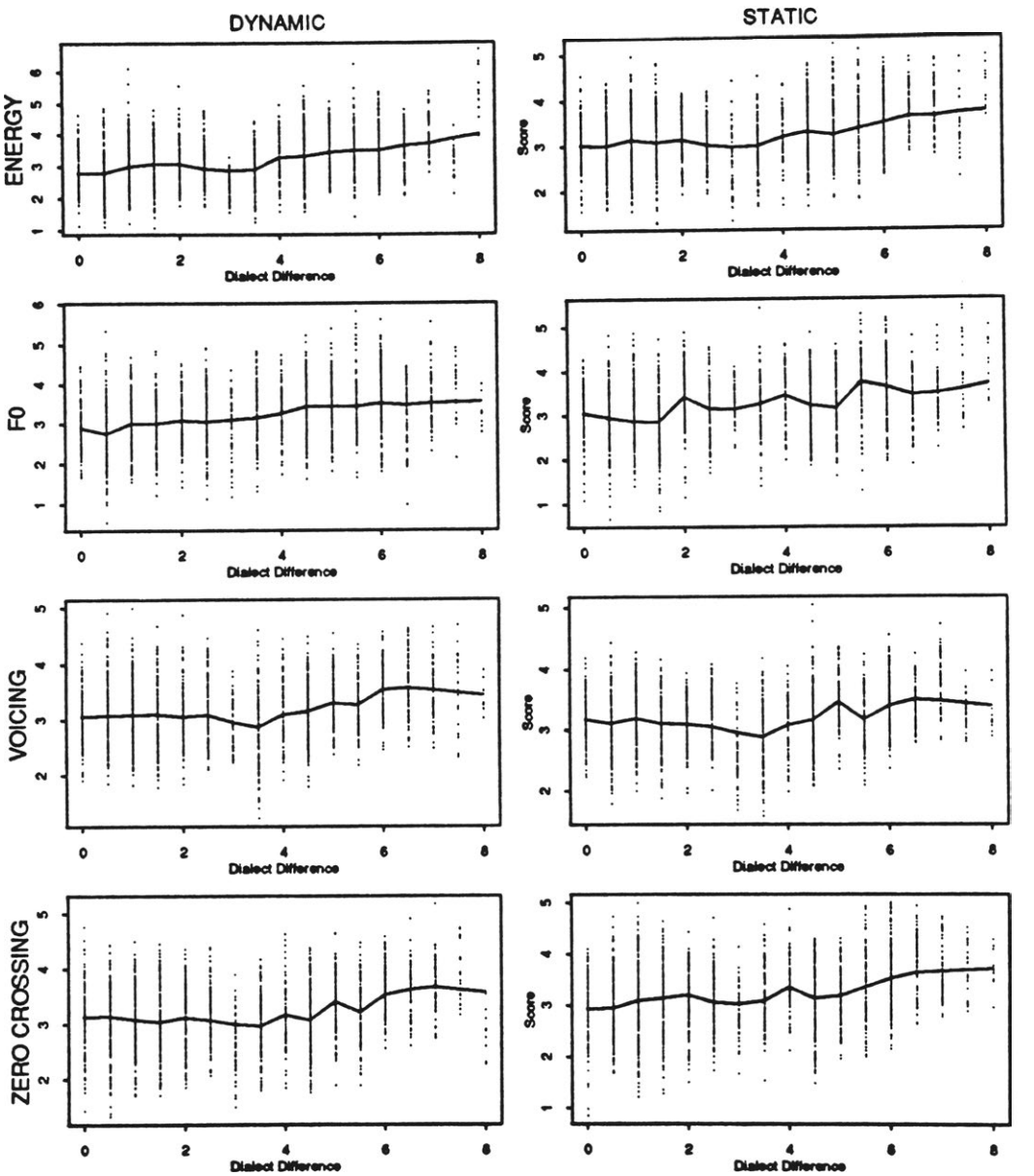


Figure 6.33: Speaker Dialect Least-Squares-Fit Scatter Plot - Contrast of the 4 speech parameters—E, F_0, Vuv, and Zc, subdivided by static versus dynamic measures, with all 4 sentences combined. Figures are derived via combining the relevant (static or dynamic) measures of the relevant speech parameter across all 4 sentences.

For each of the three speaker characteristics examined the results are analysed on the basis of the four representations of the F_0 parameter: linear concatenated, linear interpolated, log concatenated and log interpolated (see Section 5.3 for details), utilising the speech material of all four sentences. The results are also split on the basis of static and dynamic measures of the different F_0 representations. Appendix D gives a more thorough breakdown of the results for all speech parameters, including the four representations of F_0, showing discrimination and correlation rates for each parameter for each sentence—using dynamic alone, static alone, and combined static and dynamic measures.

6.5.1 Speaker Identity

Figures 6.34, and 6.35 together with Table 6.16 present the results of the analysis of the four F_0 representations as to their speaker discriminate performance.

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Concat. F_0	45.5	63.5	64.6
Interp. F_0	48.5	64.6	65.8
Log Concat. F_0	48.5	63.0	65.1
Log Interp. F_0	50.1	64.0	66.9

Table 6.16: Speaker Identity Discrimination Rates for each of the four different representations of F_0 that were considered. Analysis is further split into dynamic and static measures of the parameters in question.

It can be seen from Table 6.16 that the Log-Interpolated F_0 representation is superior of the 4 alternatives in terms of its ability to discriminate speaker, with a rate of 66.9%, an increase of 2.3% over the linear concatenated F_0 used as the default in all previous experiments. With such a difference in mind it becomes apparent that earlier quoted speaker discrimination, rates when all four speaker parameters were used, could reasonably be expected to be higher if the Log-Interpolated F_0 representation was substituted for the previously used linear concatenated F_0. Further, this result 'strengthens' the position of F_0 as the best of the four individual characteristics for speaker identification purposes.

Regarding the data two hierarchies of results may be seen. Interpolated F_0, either log or linear, yields better discriminant rates than the corresponding concatenated form. Secondly, Log F_0, either concatenated or interpolated, yields better discriminant rates than the corresponding Linear form. For all four versions of F_0, dynamic measures were markedly superior to static.

The superiority of interpolated over concatenated F_0 may be explained by examining the physical process of vocal cord vibration. For a change in F_0 the vocal cords must change their rate of vibration to the new 'target'. This cannot be done instantaneously but at a rate constrained by the biomechanical process. Interpolation may be viewed as a simple model of this process, as opposed to the instantaneous, 'jumpy', concatenated version. Fujisaki [FH82, Fuj88] considers a log version of F_0 on the basis that F_0 production under the constraints of syntactic

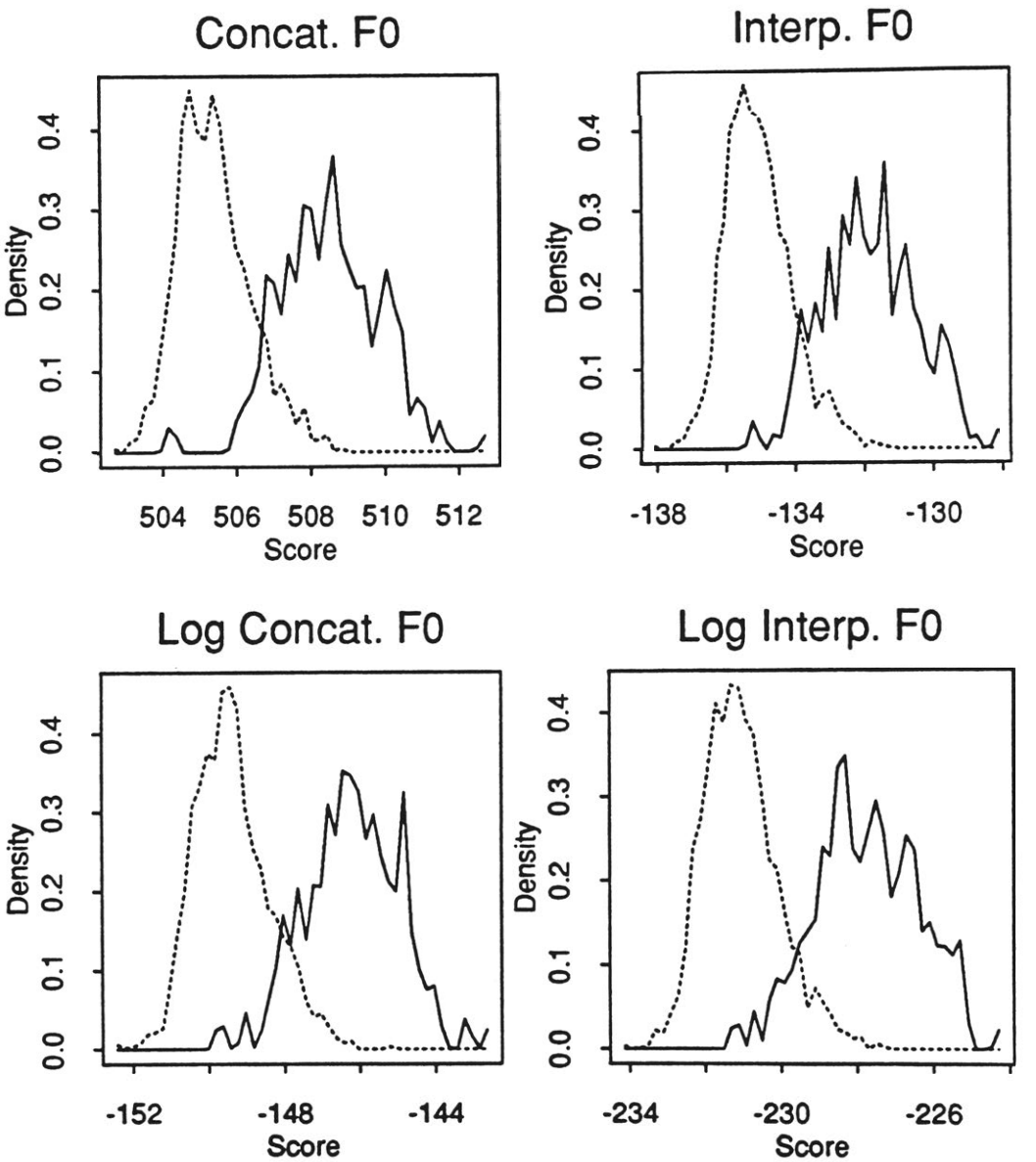


Figure 6.34: Speaker Identity Discriminate Plot - Contrasting the 4 representations of the F_0 parameter utilising all 4 sentences. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the 21 measures of the appropriate representation of F_0 for all 4 sentences.

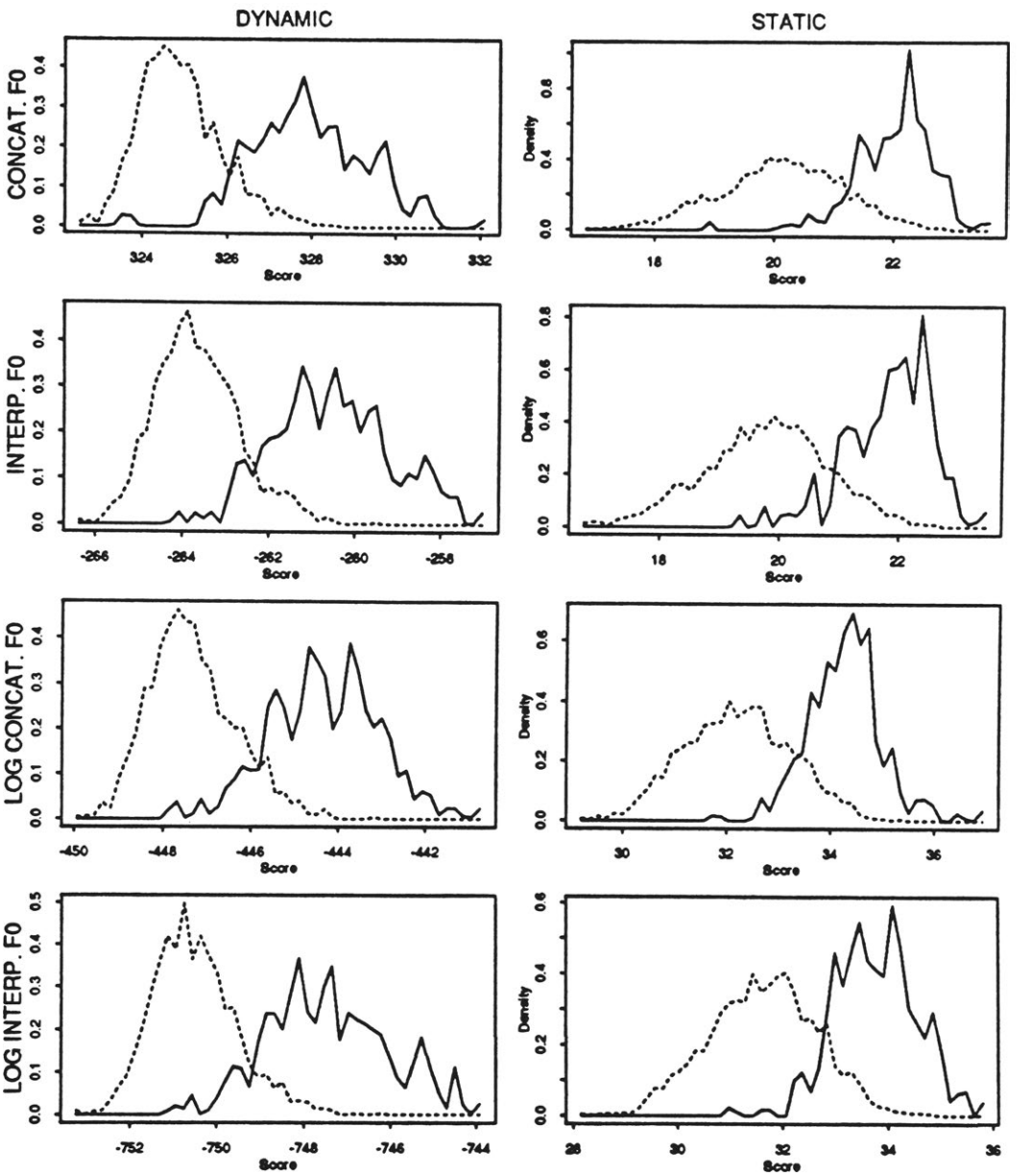


Figure 6.35: Speaker Identity Discriminate Plot - Contrasting the 4 representations of the F_0 parameter, subdivided as to static versus dynamic measures. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant (static or dynamic) measures of the appropriate representation of F_0 for all 4 sentences.

and lexical considerations may be more accurately modelled if on a log scale. On the basis of these results Fujisaki's model appears superior to one based on a linear F_0.

6.5.2 Speaker Sex

Figures 6.36 and 6.37 combined with Table 6.17 represent the results of the analysis of the four representations of F_0 as to their speaker sex discriminant ability.

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Concat. F_0	95.0	95.1	95.5
Interp. F_0	95.1	94.9	95.4
Log Concat. F_0	66.2	64.6	66.6
Log Interp. F_0	65.8	62.5	65.7

Table 6.17: Speaker Sex Discrimination Rates for each of the four different representations of F_0 that were considered. Concatenated F_0, Interpolated F_0, Log-Concatenated F_0, and Log-Interpolated F_0 are analysed separately as to their ability to discriminate speaker sex. Analysis is further split into dynamic and static measures of the parameters in question.

It is immediately clear from the results that Log representations of F_0 are significantly (*1% level*) inferior in terms of the speaker sex discrimination power. This is in no way surprising given that the formula for the derivation of Log F_0 involves the subtraction of a value F_{0min}, with a value of 60 Hertz for male speakers and 120 Hertz for female. Such a subtraction effectively eliminates much of the static information, such as the mean, that is so highly indicative of speaker sex. In fact, given such a scheme it is interesting to note that the sex discrimination rate is better than chance for the Log F_0 representations. Interestingly it appears that there is still static information pertaining to the speaker sex in both Log representations. Such a result may indicate that more appropriate values of F_{0min} could be selected for the two sexes.

Comparing results for the two linear representations there appears to be little difference between them in overall performance, a non-significant difference of 0.1%.

6.5.3 Speaker Dialect

Figures 6.38, all 6.39, together with Table 6.18 present the results of the analysis of the four versions of F_0 with regard to their fitted correlation to speaker dialect.

Two note-worthy results are discernible from the figures and table. Firstly, the log versions of F_0 have higher correlation scores than do the linear versions. Secondly, representations on which F_0 values were interpolated across unvoiced frames have higher correlation scores than do the corresponding concatenated versions. As for the speaker identity results it appears that an interpolated F_0 more closely models the physical reality of pitch production, and that Fujisaki's rationale [Fuj88] for log scale F_0 based on lexical and syntactic constraints is well founded.

Again, as for the speaker identity experiments, given the higher correlation rates for the Log-interpolated representation of F_0, it is reasonable to expect that earlier experiments in which the linear concatenated F_0 representation was used would have achieved higher correlation scores

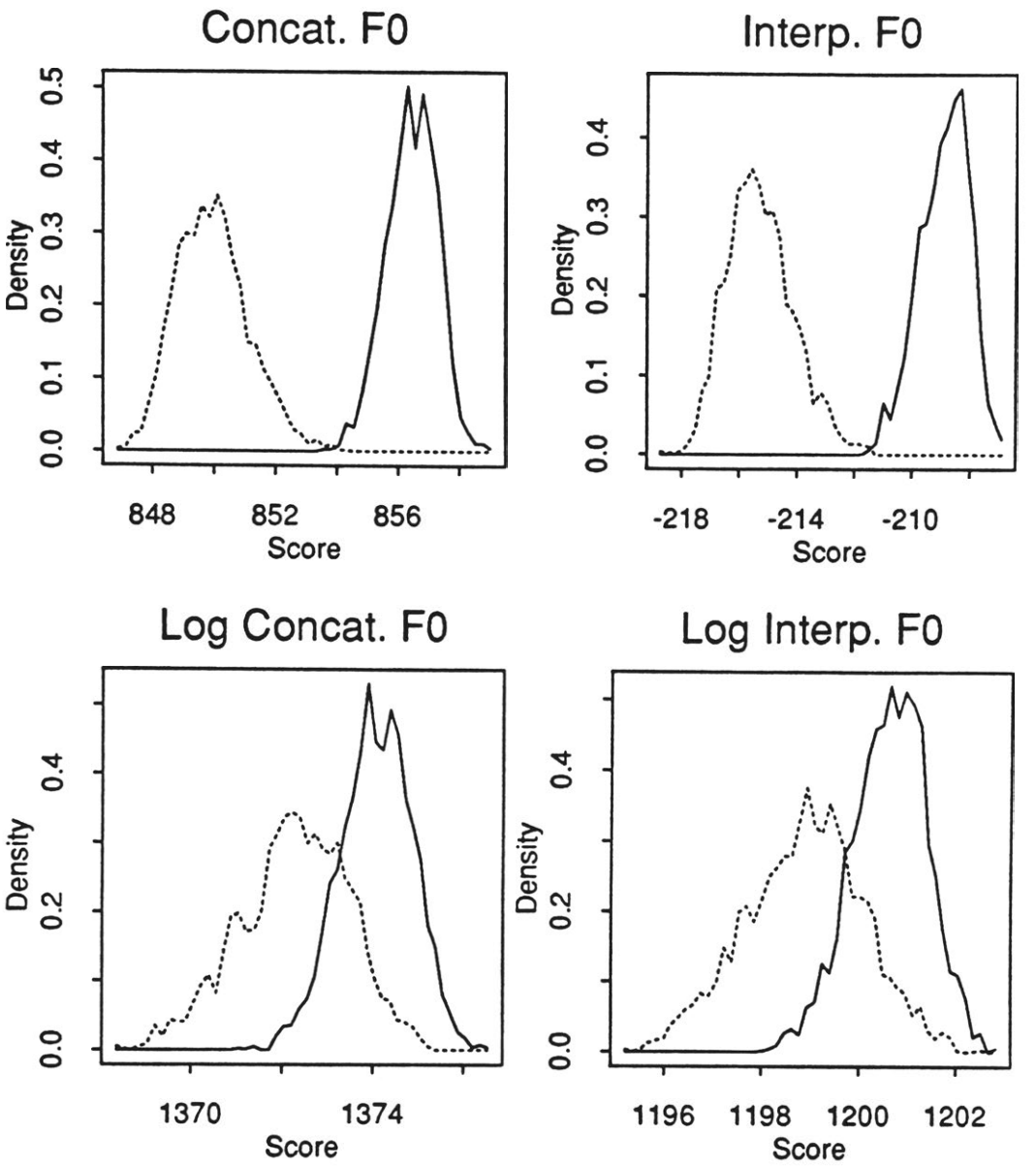


Figure 6.36: Speaker Sex Discriminate Plot - Contrasting the 4 representations of the F_0 parameter utilising all 4 sentences. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the 21 measures of the appropriate representation of F_0 for all 4 sentences.

Parameter	*Measure Type*		
	Static	Dynamic	Combined
Concat. F_0	.322	.333	.364
Interp. F_0	.351	..339	.378
Log Concat. F_0	.302	.333	.369
Log Interp. F_0	.319	.336	.396

Table 6.18: Speaker Dialect Correlation Rates for each of the four different representations of F_0 that were considered. Linear-Concatenated F_0, Linear-Interpolated F_0, Log-Interpolated F_0 and Log-Concatenated F_0 are analysed separately as to their correlation to speaker dialect. Analysis is further split into dynamic and static measures of the parameters in question.

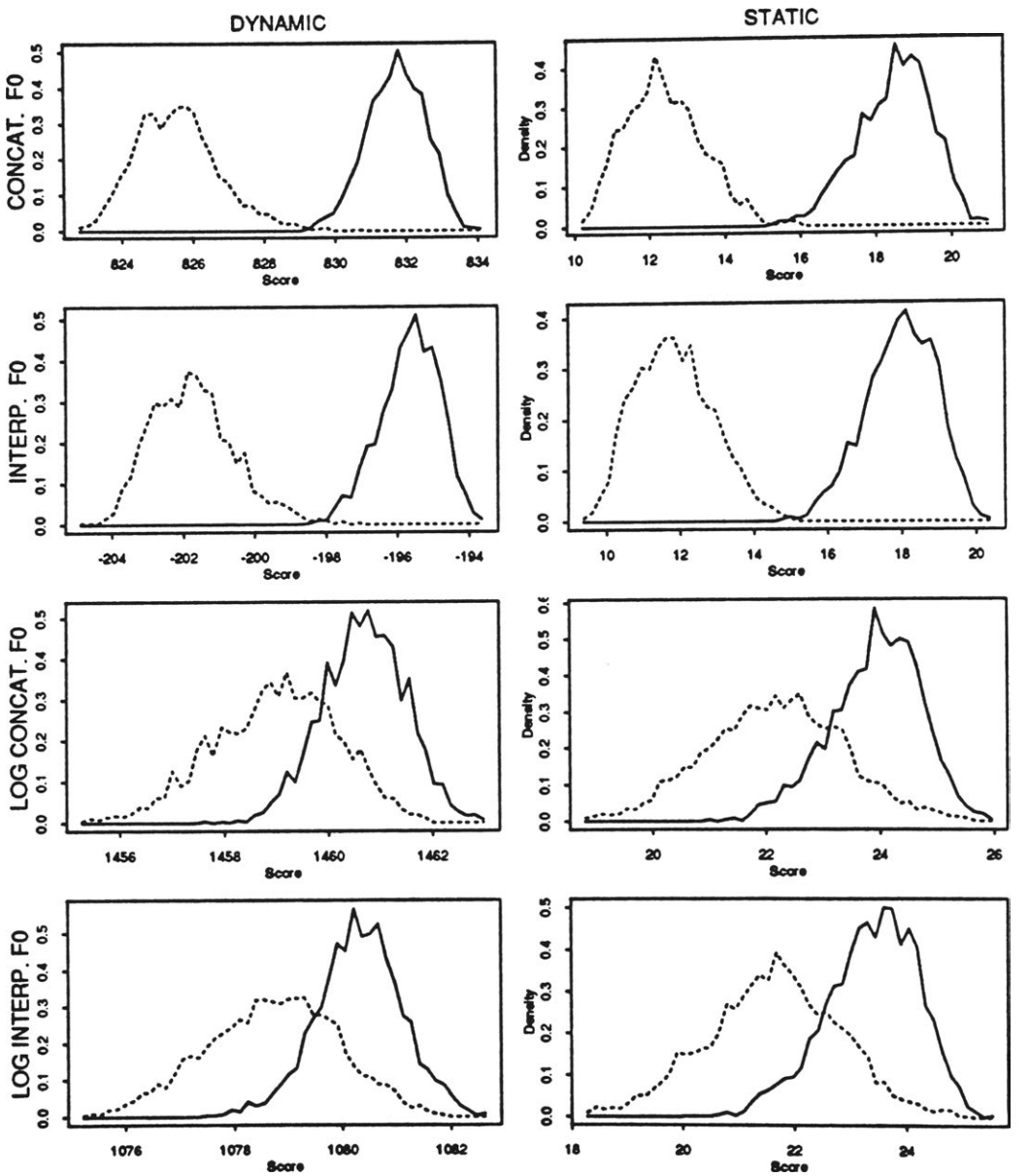


Figure 6.37: Speaker Sex Discriminate Plot - Contrasting the 4 representations of the F_0 parameter, subdivided as to static versus dynamic measures. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant (static or dynamic) measures of the appropriate representation of F_0 for all 4 sentences.

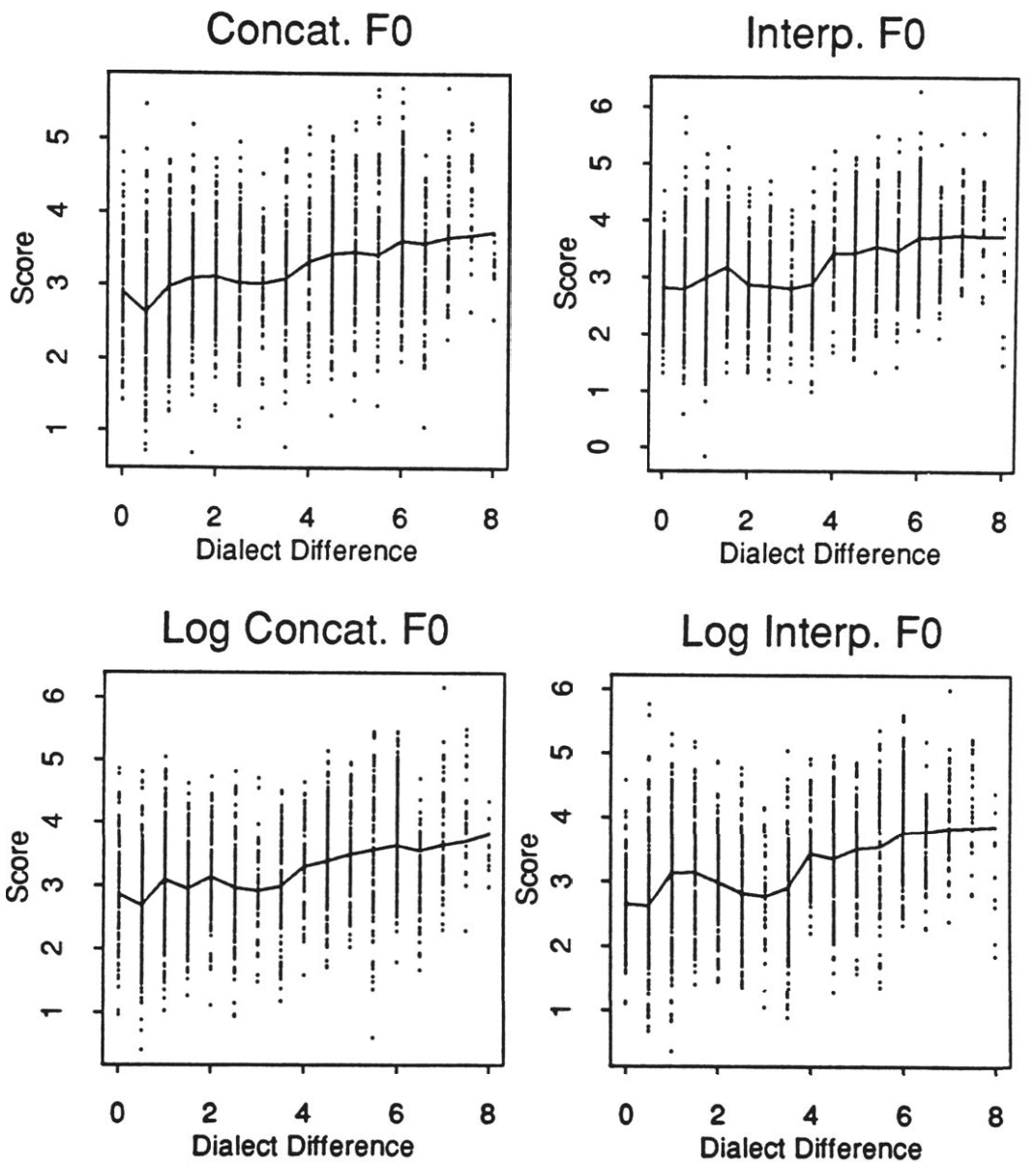


Figure 6.38: Speaker Dialect Least-Squares-Fit Scatter Plot - Contrasting the 4 representations of the F_0 parameter utilising all 4 sentences. Figures are derived by combining the 21 measures of the appropriate representation of F_0 for all 4 sentences.

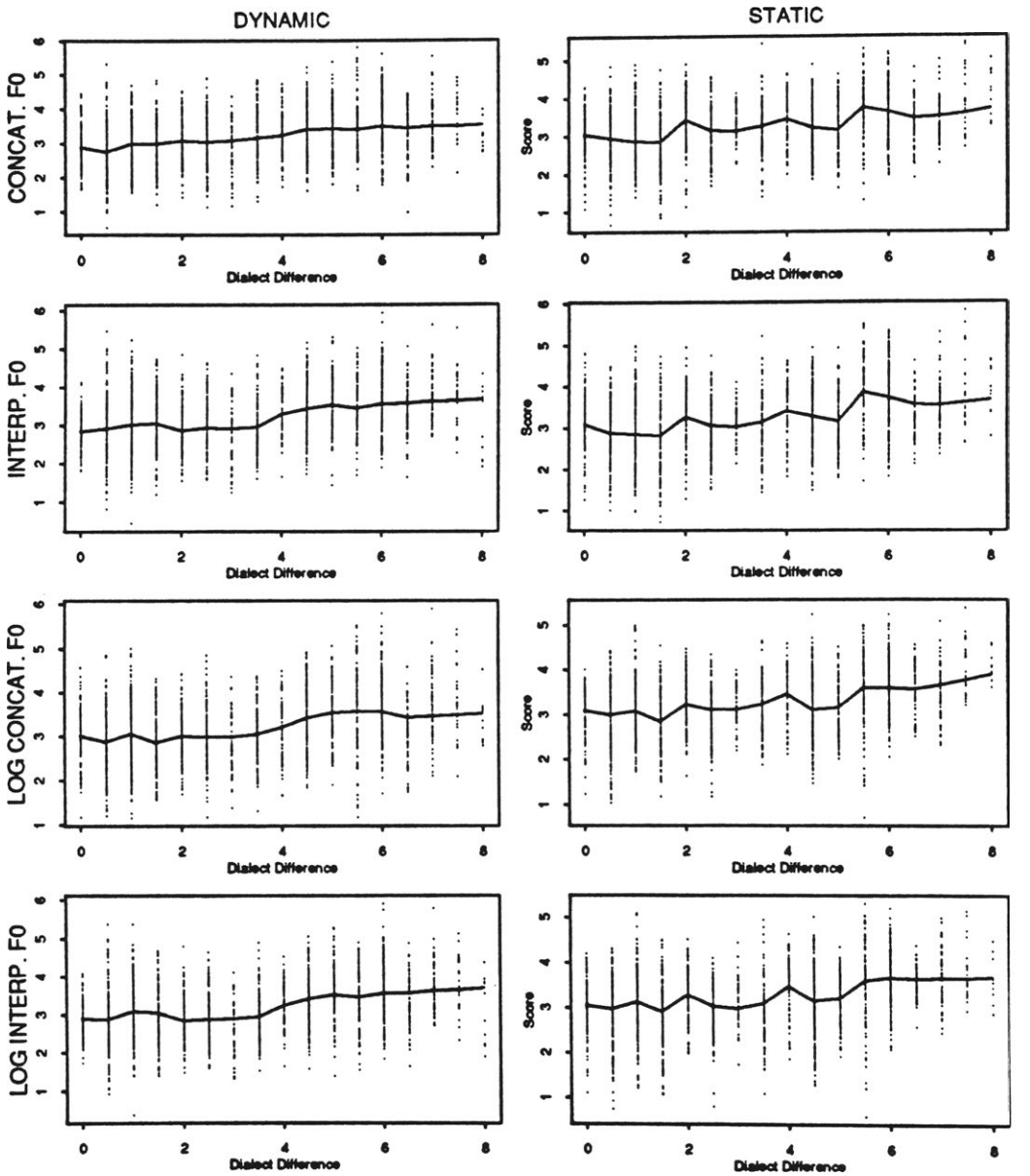


Figure 6.39: Speaker Dialect Least-Squares-Fit Scatter Plot - Contrasting the 4 representations of the F_0 parameter, subdivided as to static versus dynamic measures. Figures are derived by combining the relevant (static or dynamic) measures of the appropriate representation of F_0 for all 4 sentences.

if the Log-interpolated F_0 representation was used instead.

6.6 Contribution of Warp Path Measures

In the earlier Section 6.2 the results were divided and analysed upon the basis of static versus dynamic measures of the speech parameters. The dynamic measures, all of which are derived via the DTW process may be further split and analysed. Most schemes using DTW extract only the distance metric calculated as part of the process and ignore the possibility of relevant data from the warp path calculations. However Saito and Furui [SF78] described a speaker recognition experiment in which properties of the warp path were used. A number of measures based on properties of the warp path have thus been incorporated in the experiments.

A division of the dynamic measures into DTW distance and warp path measures will therefore illustrate whether there is useful information extractable from the calculated warp path and how such data rates against the standard of the DTW distance.

Analysis will be performed upon an individual sentence basis, and for all four sentences combined, at all times utilising the four speech parameters Energy, F_0, Voicing and Zero Crossing Rate.

6.6.1 Speaker Identity

Figures 6.40, and 6.41 together with Table 6.19 represent the results of the speaker identity discrimination experiments as analysed on the basis of comparing the DTW distance measure with the warp path measures.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
All Dynamic	57.8	50.4	51.0	56.9	54.0	74.6
DTW Dist	40.3	30.0	34.9	43.6	37.2	54.2
Warp Path	54.4	48.3	48.5	55.1	51.6	72.0

Table 6.19: Speaker Identity Discrimination Rates based on dynamic measures for all four test sentences and the four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate. Dynamic Measures are split into DTW Distance, and measures of properties of the warp path.

It is clear that on an individual sentence basis, or when all four sentences are combined, that the warp path measures out-perform the DTW distance measure by a factor of 10% and more, a significant difference at the 1% level of confidence. Such a result implies that there is more speaker specific information in the warp path than in the calculated DTW distance. This result contrasts with that of Saito and Furui [SF78] who found that warp path based recognition yielded inferior rates to that of DTW distance, and concluded that: "...the rate of similarity in a specific region" [warp path measure] "is a useful supplementary measure for talker recognition, although it is insufficient to be used as an independent measure for talker recognition." Clearly for prosodic parameters, and when a variety of properties of the warp path are examined the warp path significantly out-performs the DTW distance with respect to

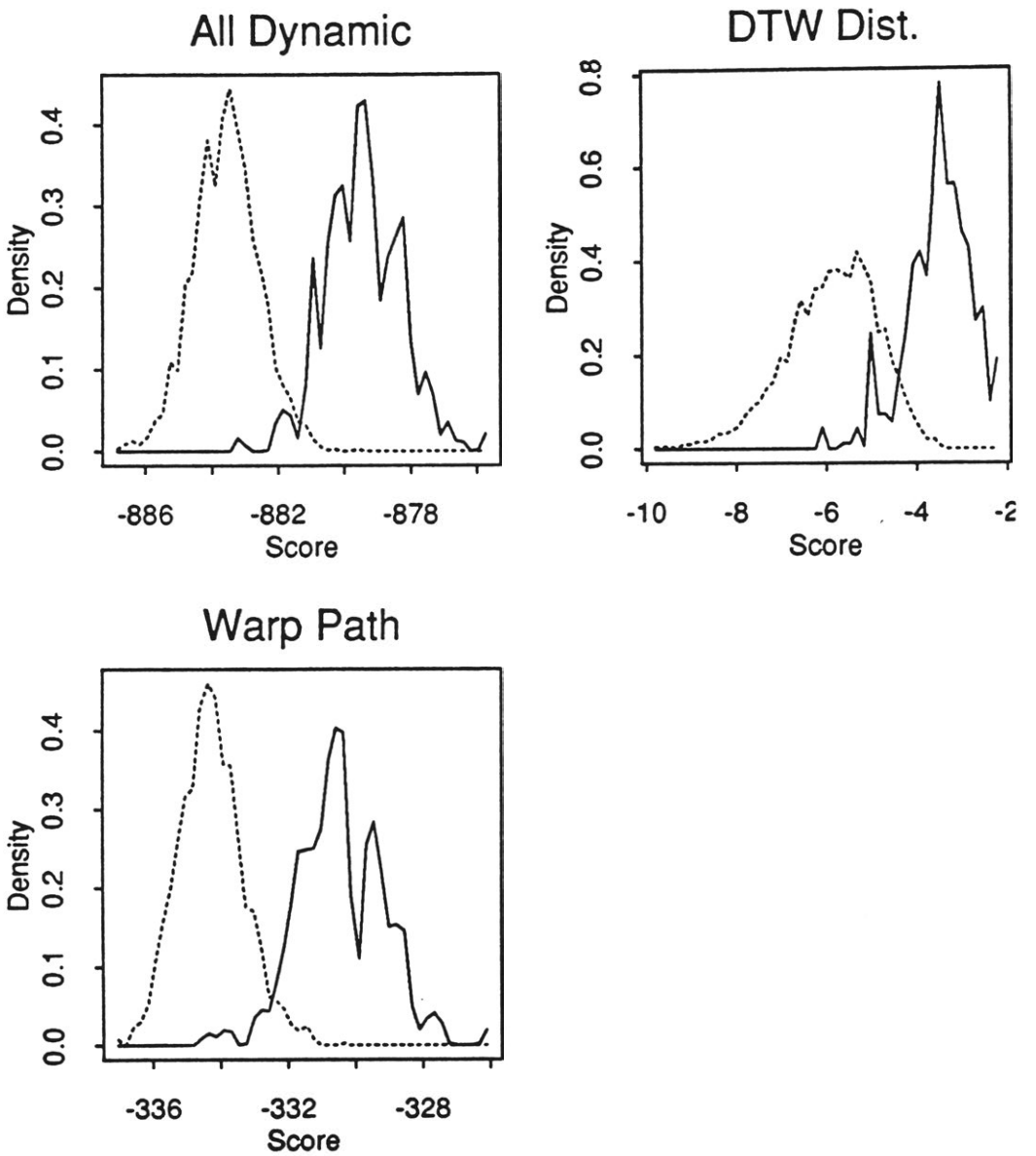


Figure 6.40: Speaker Identity Discriminate Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; utilising all 4 sentences. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for all 4 sentences.

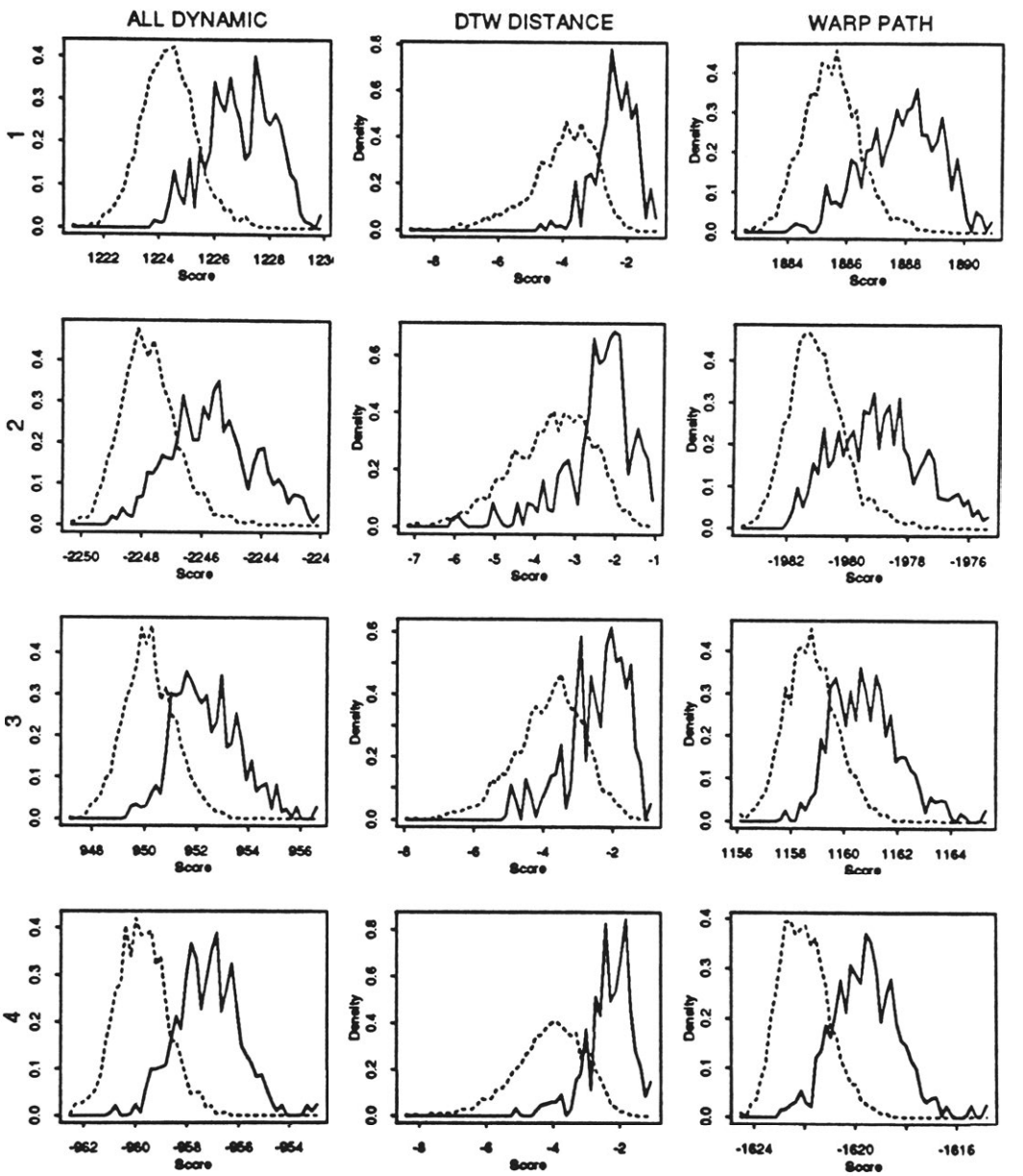


Figure 6.41: Speaker Identity Discriminate Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

the ability to discriminate speakers.

Upon examining the sentences individually it is evident that sentences 1 and 4 have the highest warp path based discriminant scores while sentences 2 and 3 have considerably lower scores. Both sentences 1 and 4 have a greater duration (more speech material), and a higher mean voicing duration (significance of F_0) either of which may explain their better performance.

Further, when the warp path measures are compared with the total dynamic measures there is little difference in discriminant ability. For example, when all sentences are combined, the discriminant rate drops from 74.6% for all dynamic measures to 72.0%, a difference of only 2.6%. In fact, the discrimination rate of 72.0% for warp path measures alone is little less than the entire discrimination score of 75.2% for *all* measures. Such a result implies that much of the relevant information is already encapsulated within the warp path measures and there is a high degree of overlap between them and the DTW distance measure.

6.6.2 Speaker Sex

Figures 6.42, and 6.43 together with Table 6.20 present the results of the speaker sex discrimination experiments as analysed on the basis of comparing the warp path measures with the DTW distance measure.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
All Dynamic	93.6	88.2	89.8	92.3	91.0	95.8
DTW Dist	91.0	86.3	85.7	89.6	88.2	92.8
Warp Path	84.8	72.2	77.9	86.1	80.3	91.7

Table 6.20: Speaker Sex Discrimination Rates based on dynamic measures for all four test sentences and the four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate. Dynamic Measures are split into DTW Distance, and measures of properties of the warp path.

It may be seen from both the figures and table that under all circumstances the DTW distance is superior to the warp path measures for determining speaker sex. This is not surprising as when applied to the two contours with vastly different means, for example the F_0 of a female and male speaker, it will measure little more than the difference in means between the two contours. Surprising is the discrimination rate for warp path parameters alone, 88.2% on average for a single sentence, indicating that whether on the basis of mean F_0 or 'true dynamics' of parameters, that the warp path measures provide a 'good' discrimination of speaker sex. Further the DTW distance measure is on average 3.1% below all dynamic measures in speaker discrimination level, indicating that the warp path parameters do make a contribution to sex discrimination independent of that made by the DTW distance.

6.6.3 Speaker Dialect

Table 6.21 and Figures 6.44, and 6.45 present the results of the speaker dialect experiments analysed with respect to comparing warp path measures with the DTW distance measure.

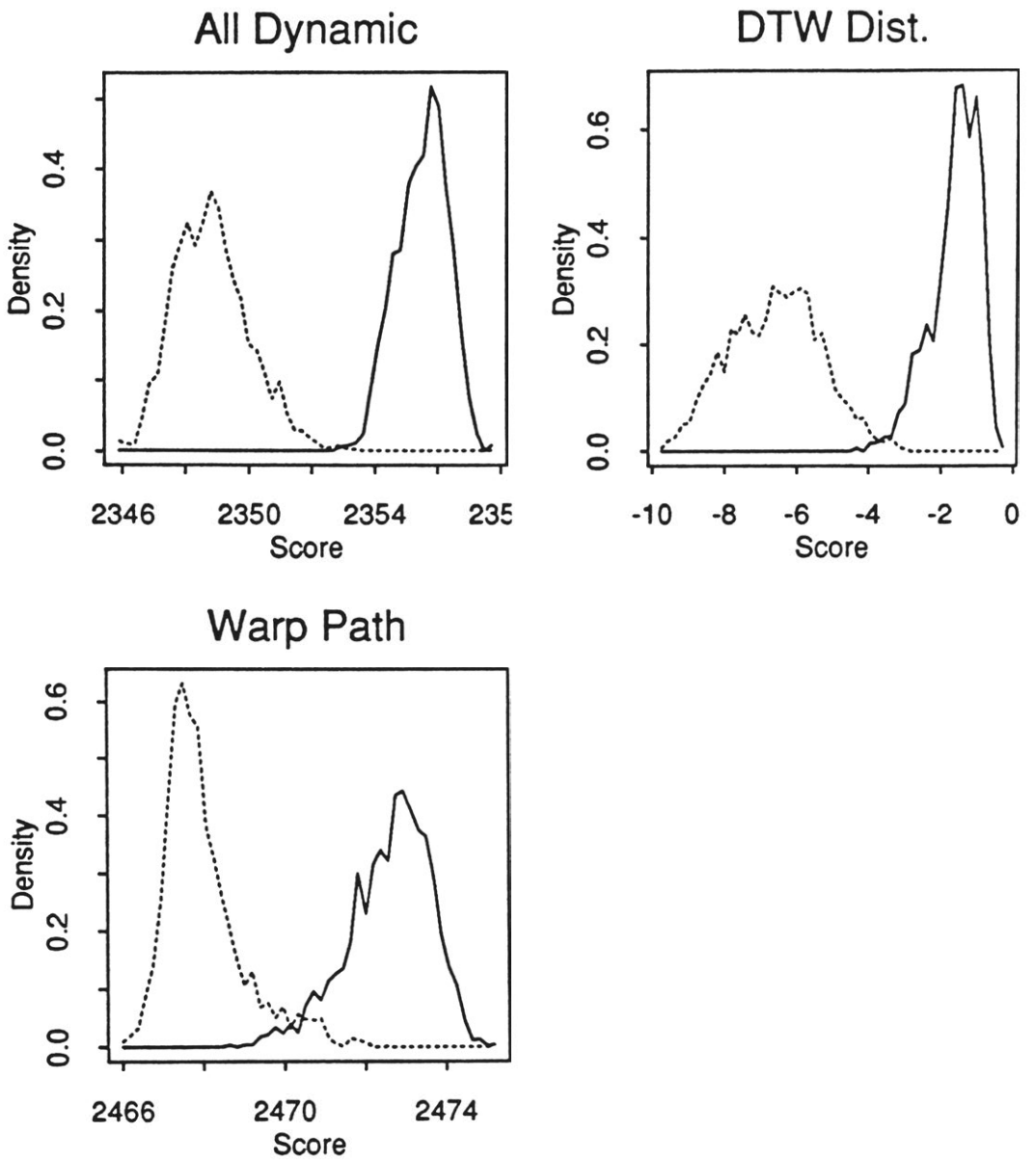


Figure 6.42: Speaker Sex Discriminate Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; utilising all 4 sentences. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for all 4 sentences.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
All Dynamic	.330	.322	.263	.291	.302	.563
DTW Dist	.184	.223	.151	.096	.164	.301
Warp Path	.311	.281	.223	.285	.275	.450

Table 6.21: Speaker Dialect Correlation Scores based on dynamic measures for all four test sentences and the four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate. Dynamic Measures are split into DTW Distance, and measures of properties of the warp path.

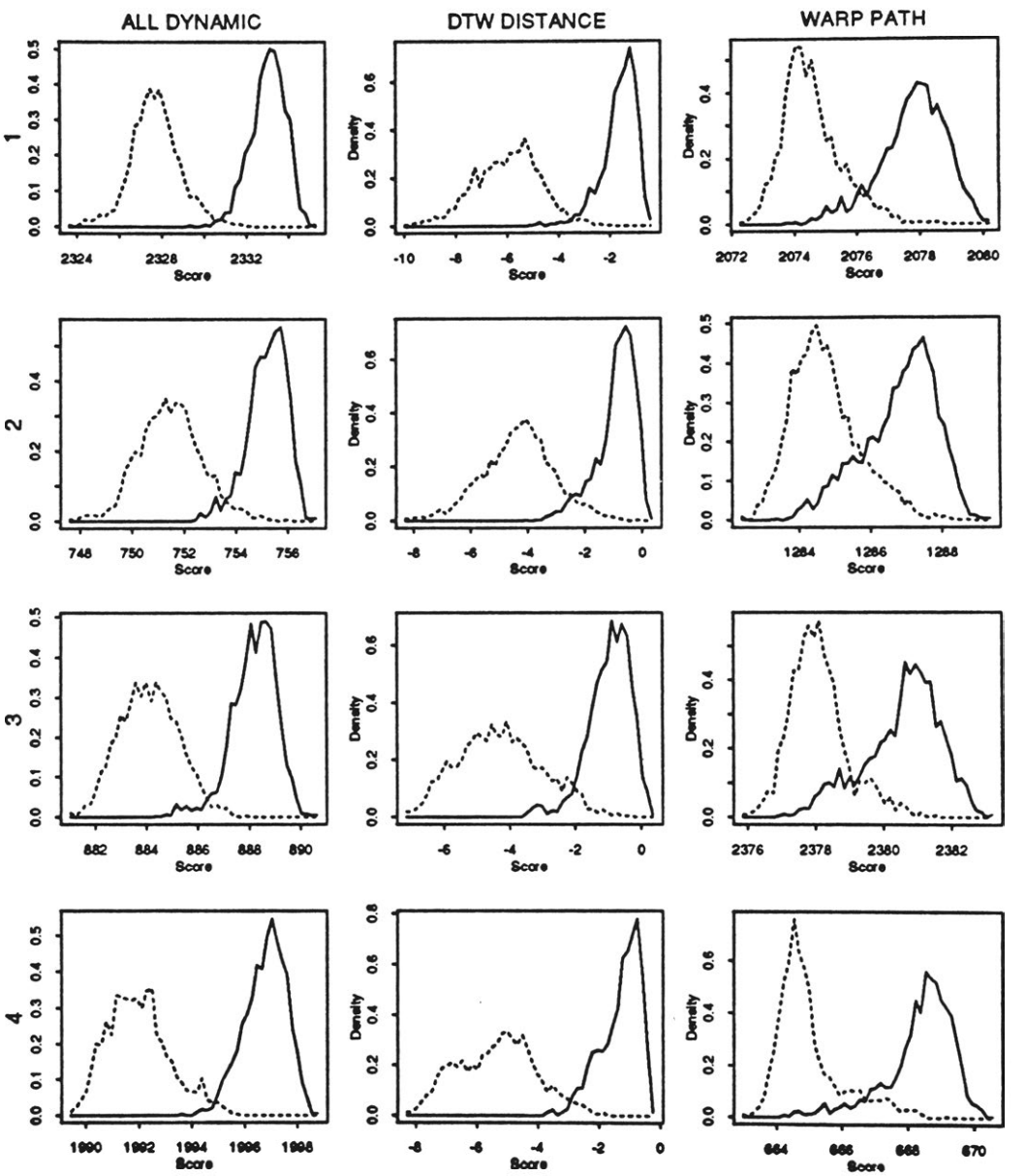


Figure 6.43: Speaker Sex Discriminate Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

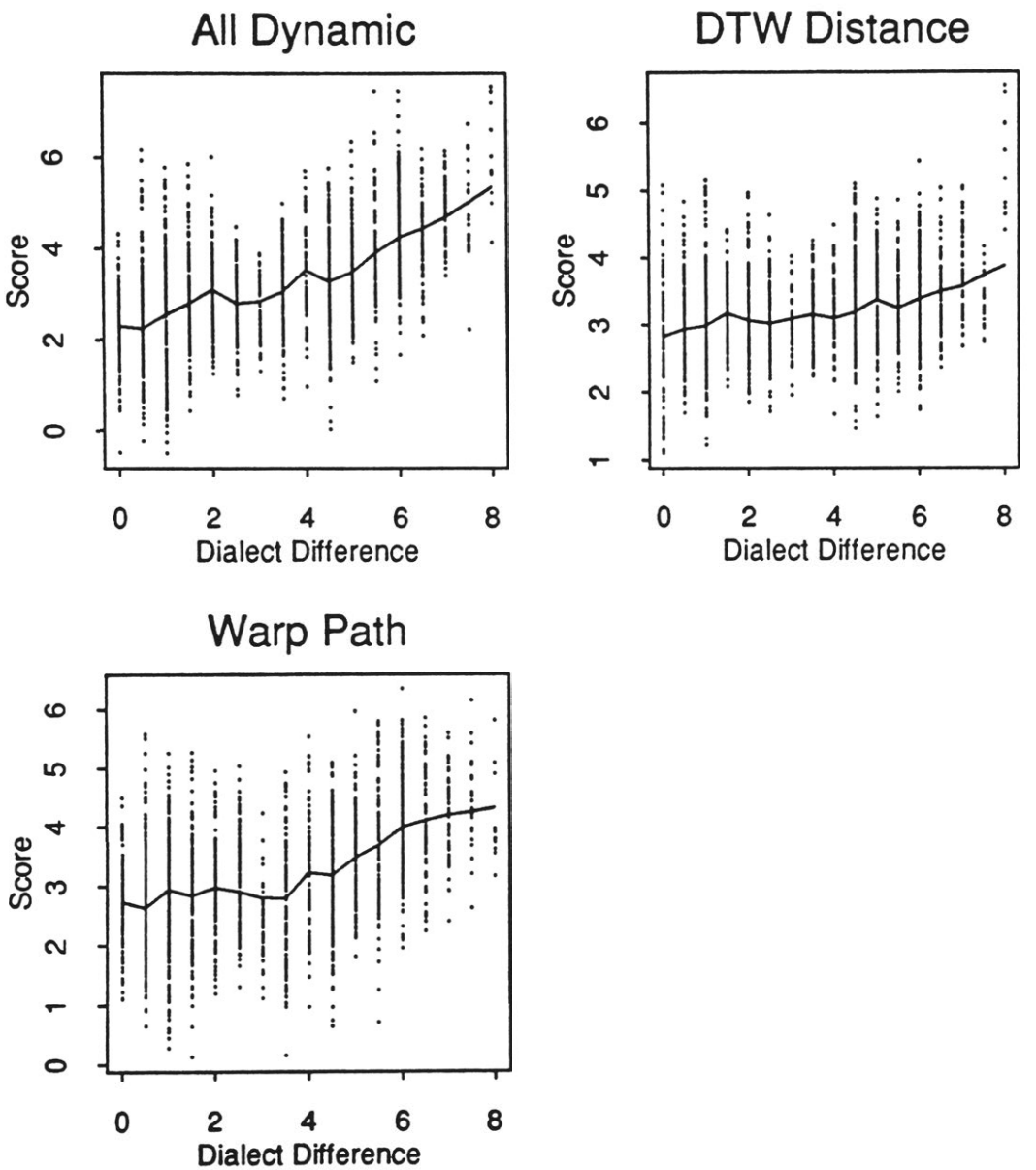


Figure 6.44: Speaker Dialect Least-Squares-Fit Scatter Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; utilising all 4 sentences. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for all 4 sentences.

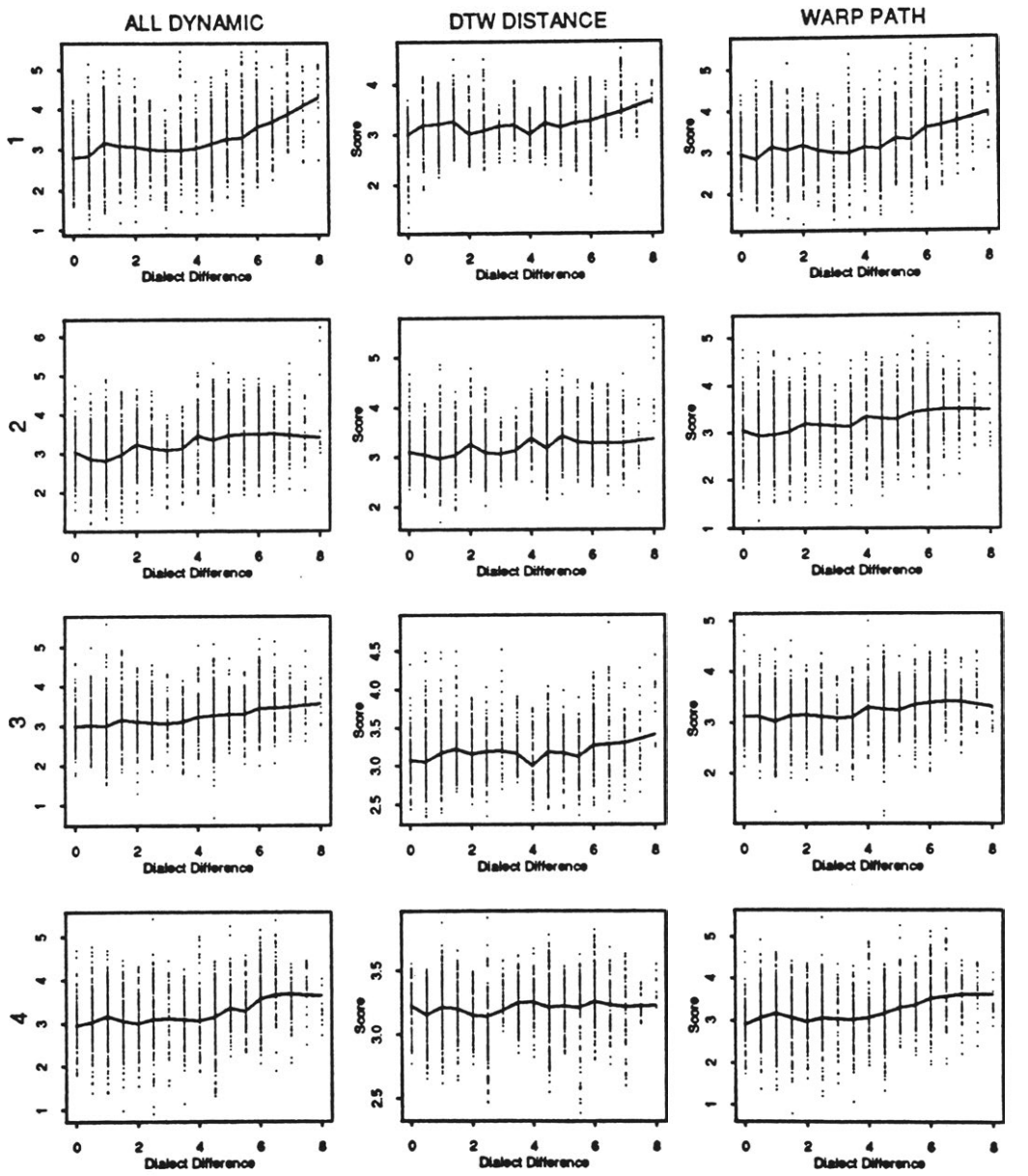


Figure 6.45: Speaker Dialect Least-Squares-Fit Scatter Plot - Breakdown of dynamic measures into DTW Distances versus Warp Path measures; for each of the 4 sentences in turn. Figures are derived by combining the relevant measures of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

It is apparent from the figures and tables that on a single sentence basis, the DTW distance measure has a low correlation with the dialect-difference-score, a mean value of 0.164. On the other hand the warp path parameters are markedly higher with a mean of 0.302 for individual sentences, a significant difference at the 1% level.

When the four sentences are combined the same relationship continues with the warp path measures having a significantly (*1% level*) higher correlation than the DTW distance measure.

Hence, it is apparent that the warp path measures are more strongly related to speaker dialect and contribute the greater portion of the correlation factor when all dynamic measures are considered together. However both approaches complement each other well and taken together yield higher correlation rates than either alone.

6.7 Examination of DTW-Distance Variant Measures

In Section 6.6 the DTW distance measure was contrasted with other measures derivable from the DTW process, and generally found to yield inferior discrimination or correlation rates.

It appears worth considering whether any variant upon the basic DTW distance scheme might yield better performance in terms of the ability to discriminate speaker characteristics. In Section 5.4 two variants on the DTW distance— referred to as the Weighted DTW distance, which sought to include warp path derived information in the distance, and Border DTW distance, which eliminated two leading and trailing values from the interval over which the distance is calculated ('freer' end-point conditions), were proposed. This section will contrast between these two variants and the DTW distance itself as to the discriminant or correlation scores for the three characteristics.

Analysis will be performed upon an individual sentence basis, where the four basic speech parameters: Energy, F_0, Voicing and Zero Crossing will be used in combination. The measures for each of the four parameters for all four sentences will also be combined and analysed to yield a result for when all four sentences are utilised together.

6.7.1 Speaker Identity

Figures 6.46 and 6.47 combined with Table 6.22 present the results of the analysis of the variant DTW distance measures with regard to their speaker discriminant ability.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Simple Distance	40.3	30.0	34.9	43.6	37.2	54.2
Weighted Distance	37.9	29.7	35.1	39.6	35.5	53.2
Border Distance	39.3	29.8	35.1	42.9	36.7	55.1

Table 6.22: Speaker Identity Discrimination Rates comparing the three quantifications of the DTW distance that were examined. The four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate were used across all 4 sentences.

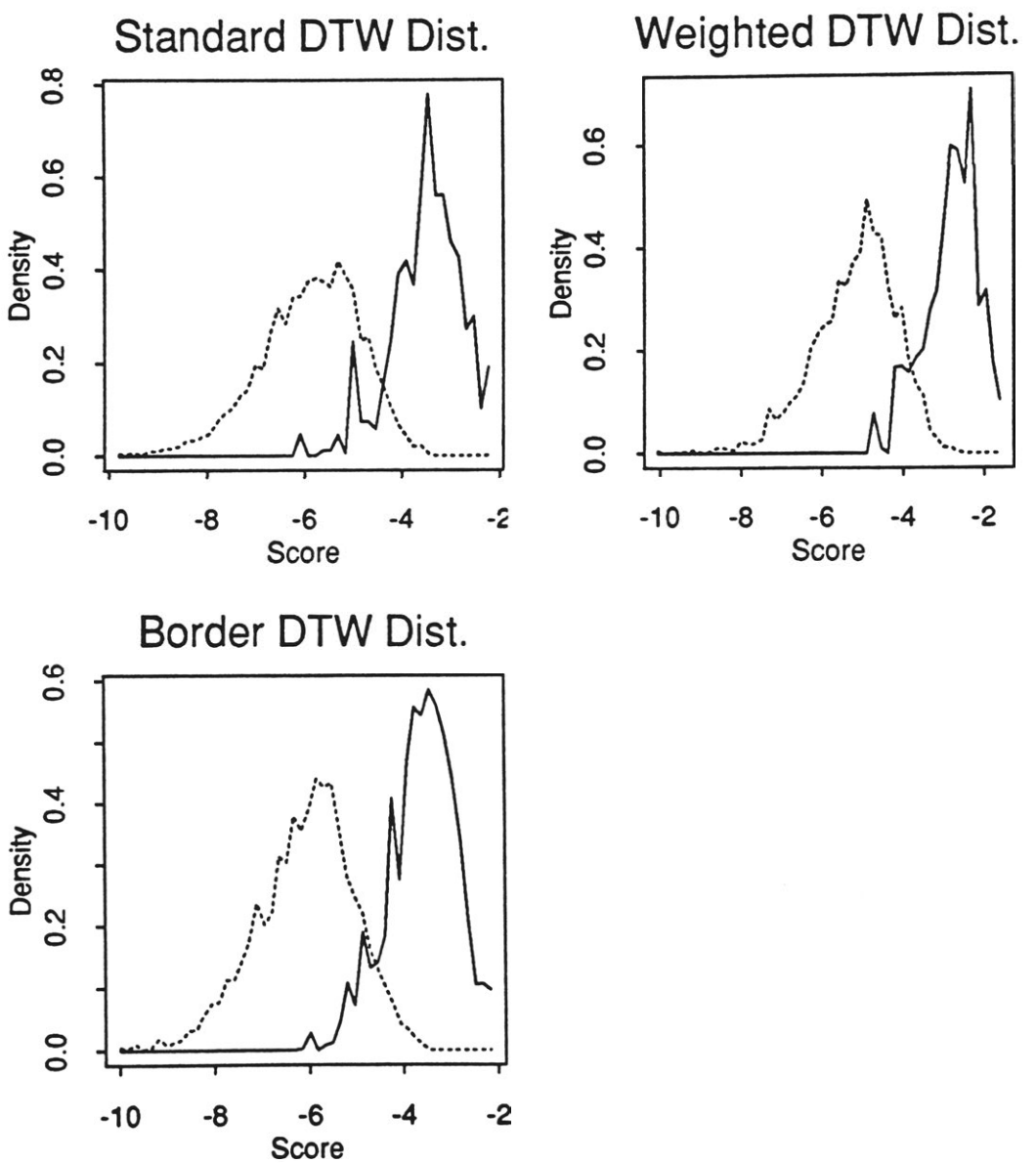


Figure 6.46: Speaker Identity Discriminate Plot - Comparison of the 3 forms of the DTW Distance measure; utilising all 4 sentences. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant measures for the four speech parameters—E, F_0, Vuv, and Zc, across all 4 sentences.

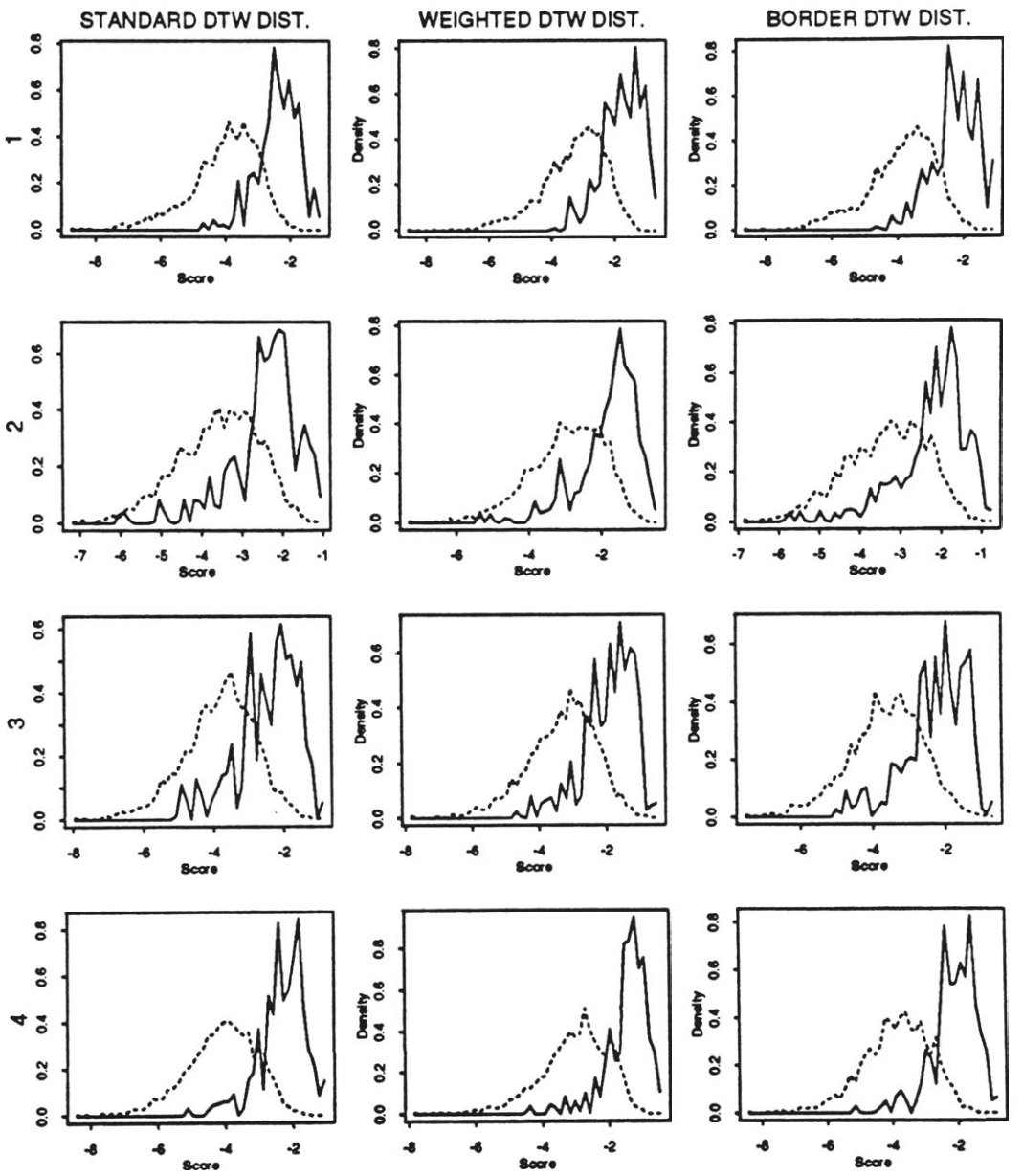


Figure 6.47: Speaker Identity Discriminate Plot - Comparison of the 3 forms of the DTW Distance measure; for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution. Figures are derived by combining the relevant measure of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

The differences between the different measures, either for a single sentence or for the four-combined-sentence case are not significant at the 5% level. However, under all circumstances the weighted DTW distance measure has an inferior discrimination ability than either of the other two DTW distances. Thus while warp path measures alone are good discriminators of speaker identity, when incorporated into the distance metric they lead to a reduction in discriminant ability.

6.7.2 Speaker Sex

Figures 6.48 and 6.49 combined with Table 6.23 show the results of the analysis of the three variants of the DTW distance measure with regard to their speaker sex discrimination ability.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Simple Distance	91.0	86.3	85.7	89.6	88.2	92.8
Weighted Distance	85.0	81.6	84.2	82.1	83.2	91.7
Border Distance	90.7	85.4	85.6	88.1	87.4	92.9

Table 6.23: Speaker Sex Discrimination Rates comparing the three quantifications of the DTW distance that were examined. The four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate were used across all 4 sentences.

On a single sentence basis it may be seen that the weighted distance measure is significantly (*1% level*) inferior to either of the other two. When the four sentences are combined the weighted distance still is inferior to the other two measures though the difference is not significant at the 5% level. There is no significant difference between the default DTW distance and the border distance.

6.7.3 Speaker Dialect

Figures 6.50, and 6.51 together with Table 6.24 show the results of the analysis of the three variants of the DTW Distance measure with respect to their correlation to the dialect-difference-score.

Measure Type	*Sentence*					
	1	2	3	4	Mean	Combined
Simple Distance	.184	.223	.151	.096	.164	.301
Weighted Distance	.177	.215	.146	.110	.162	.301
Border Distance	.179	.220	.152	.106	.164	.309

Table 6.24: Speaker Dialect Correlation Scores comparing the three quantifications of the DTW distance that were examined. The four speech parameters Energy, F_0, Voicing, and Zero Crossing Rate were used across all 4 sentences.

Comparing the correlation values for each of the three distance measures, either on a single sentence, or when all four sentences are combined, there appears no significant difference between them, though the border distance is marginally superior for the four-combined-sentences case.

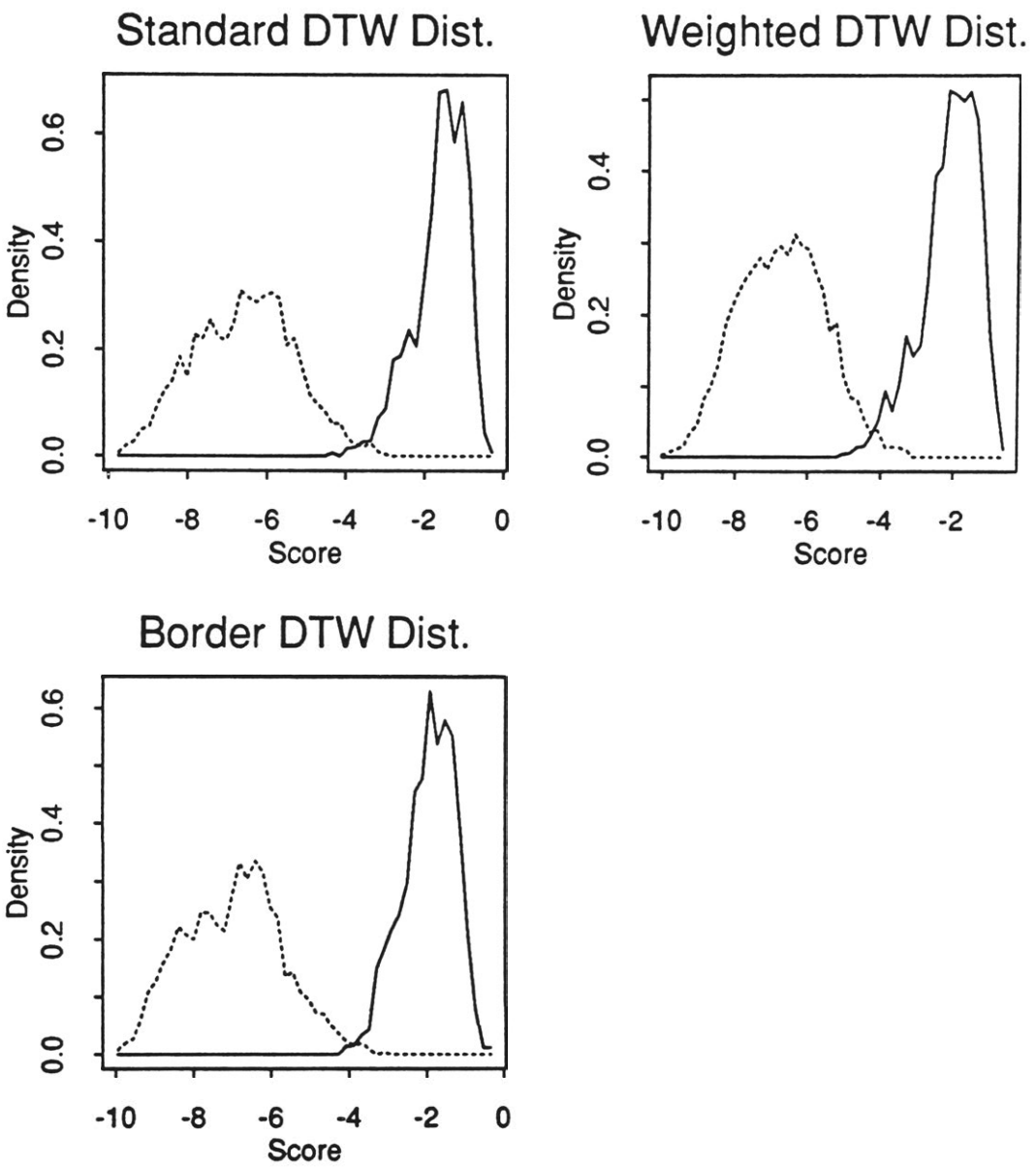


Figure 6.48: Speaker Sex Discriminate Plot - Comparison of the 3 forms of the DTW Distance measure; utilising all 4 sentences. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant measure for the four speech parameters—E, F_0, Vuv, and Zc, across all 4 sentences.

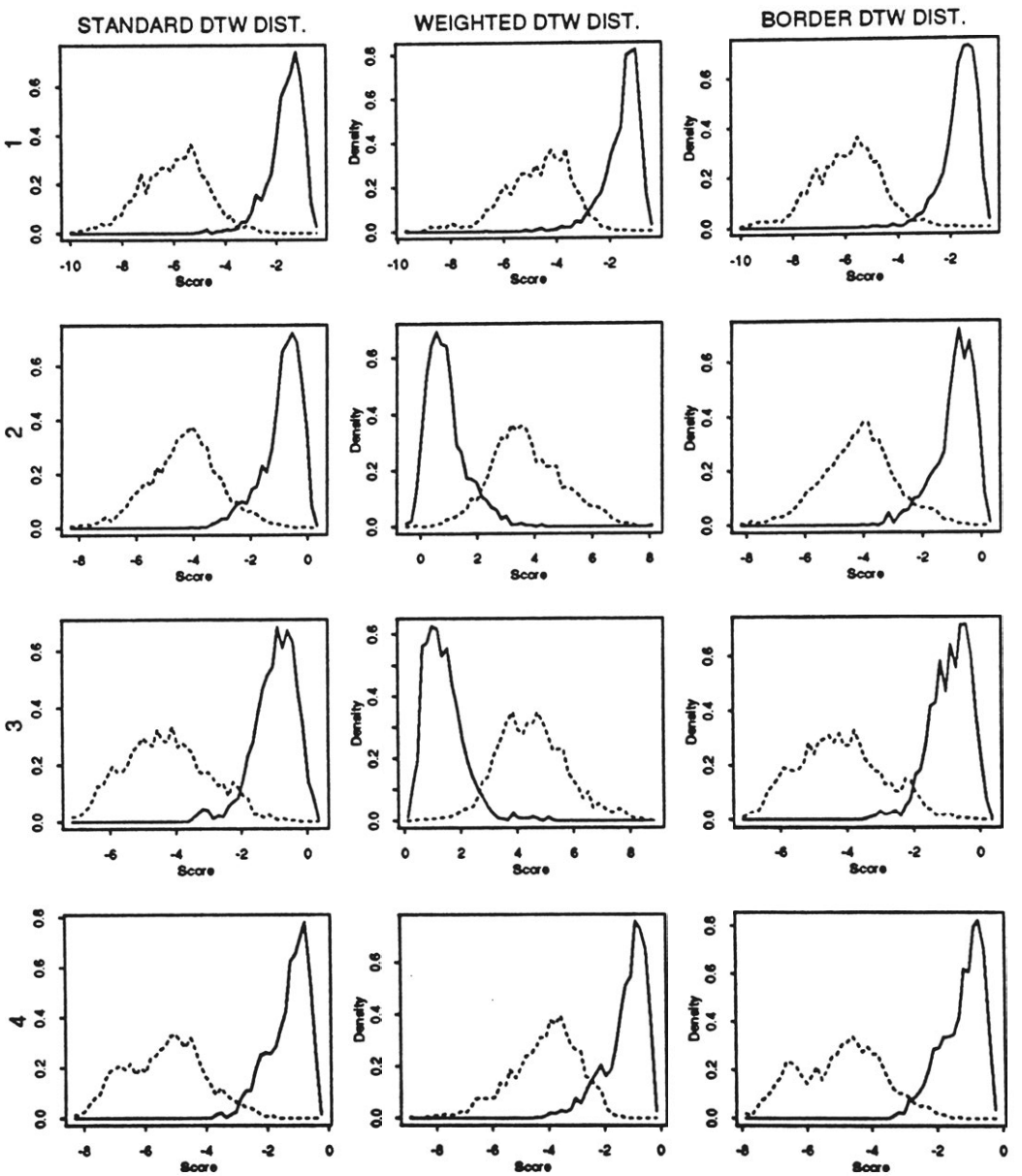


Figure 6.49: Speaker Sex Discriminate Plot - Comparison of the 3 forms of the DTW Distance measure; for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution. Figures are derived by combining the relevant measure of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

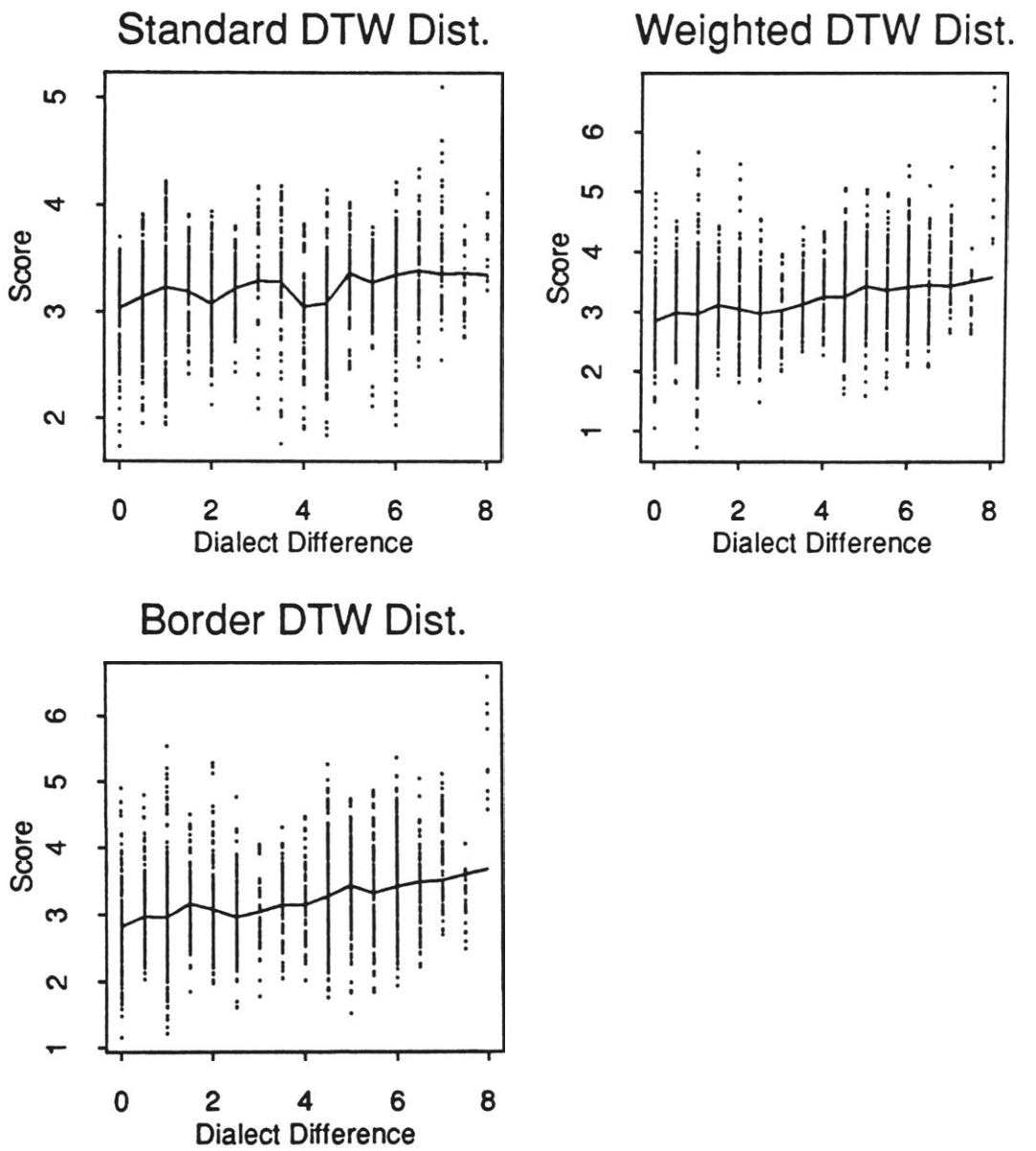


Figure 6.50: Speaker Dialect Least-Squares-Fit Scatter Plot - Comparison of the 3 forms of the DTW Distance measure, utilising all 4 sentences. Figures are derived by combining the relevant measure for the four speech parameters—E, F_0, Vuv, and Zc, across all 4 sentences.

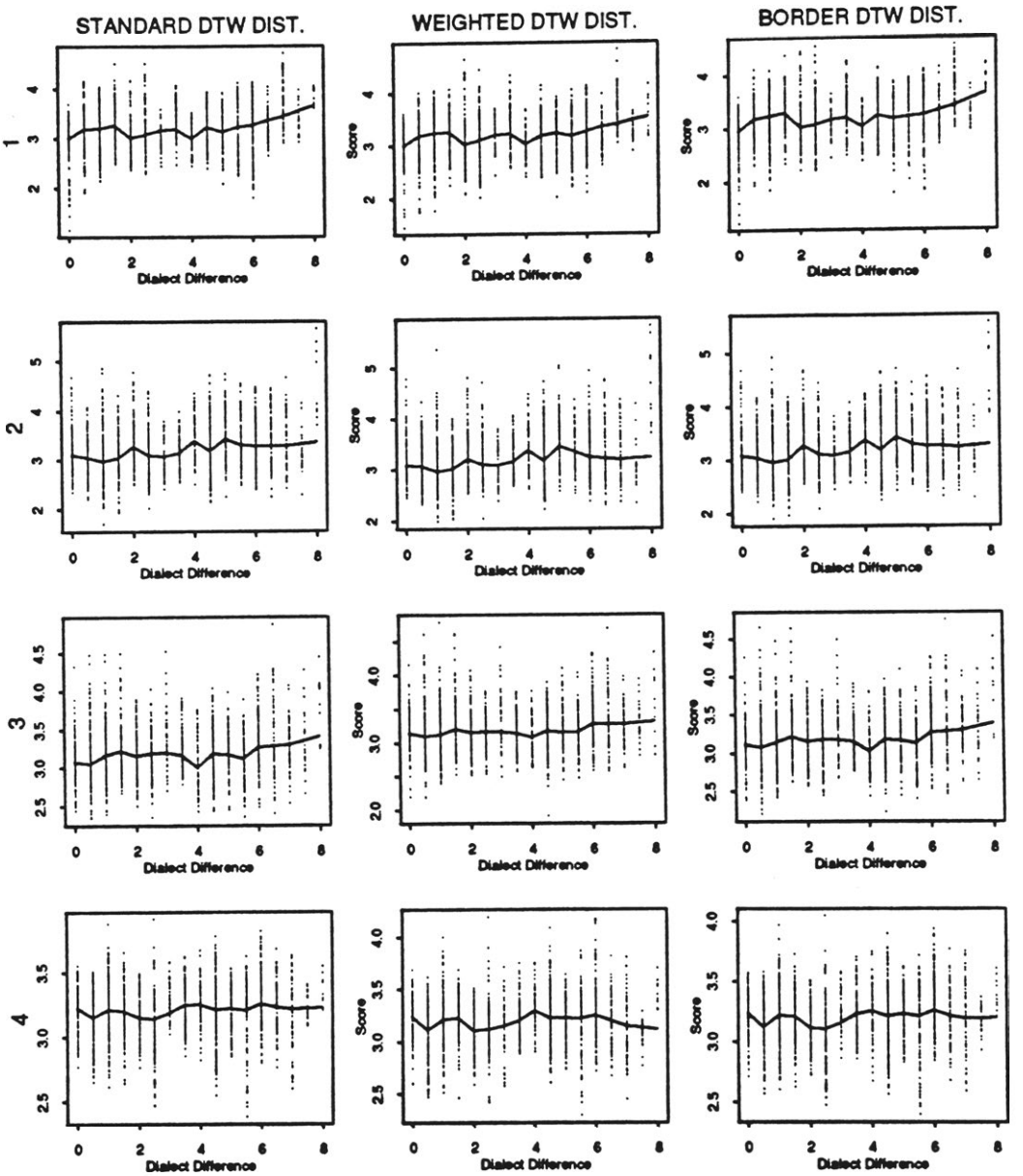


Figure 6.51: Speaker Dialect Least-Squares-Fit Scatter Plot - Comparison of the 3 forms of the DTW Distance measure, for each of the 4 sentences in turn. Figures are derived by combining the relevant measure of the four speech parameters—E, F_0, Vuv, and Zc, for each of the 4 sentences.

It appears worthy of note that the correlation scores for sentence four are significantly (*1% level*), by up to a factor of 2, worse than any of the other three sentences. There appears no ready explanation for this lower correlation value; sentence 4 is both the longest (most speech material) and all-voiced (importance of F_0). It appears that there are other constraints dictating the degree of dialect encoding within a sentence.

6.8 Evaluation of Individual Measures

To this stage little attention has been paid to the twenty one measures individually, rather they have been considered as logical groupings or a homogeneous whole, in the process of examining the speaker characteristics. For practical applications such as recognition systems a lack of focus upon the individual measures is quite acceptable as it is the overall system performance that is of interest, not any one individual component. However, attention to the individual measures may yield information pertaining to the specific forms of encoding that the speaker characteristic takes, yielding data of general theoretical importance and possibly illuminating areas where further speaker characteristic features may be extracted.

There are several different ways to view the wealth of available data on the correlation between measures and speaker characteristics. Appendix C contains a complete breakdown of the computed correlation or ω^2 values for all experiments. That is for the three speaker characteristics; the four sentences; the four speech parameters, including the four representations of F_0; both normalised and non-normalised parameters; and both dynamic and static measures. However, in order to summarise this information for the present the distribution of these correlation or ω^2 values will be examined for each measure across the four basic speech parameters and for the four sentences combined.

Due to the large number of measures examined, it was necessary to label the measures on the plots numerically rather than descriptively. The static measures were labelled as such and numbered 1 to 7, while the dynamic were labelled dynamic and numbered 1 to 14. The correspondence between numeric and descriptive labels for the static measures is:

1. Mean
2. Standard Deviation
3. Minimum
4. Maximum
5. Range
6. Mean absolute Rate of Change
7. Length or Duration

Similarly for the dynamic measures:

1. DTW Distance

2. Weighted DTW Distance
3. 'Border' DTW Distance
4. Ratio of Warp Path Length to Maximum Contour Length
5. Fraction of Vertical Transitions on Warp Path
6. Fraction of Horizontal Transitions on Warp Path
7. Fraction of Diagonal Transitions on Warp Path
8. Number of Vertical Excursions/Warp Path Length
9. Number of Horizontal Excursions/Warp Path Length
10. Number of Diagonal Excursions/Warp Path Length
11. Maximum Length Vertical Excursions/Warp Path Length
12. Maximum Length Horizontal Excursions/Warp Path Length
13. Maximum Length Diagonal Excursions/Warp Path Length
14. Off Diagonal Warp Path Distance

6.8.1 Speaker Identity

Figure 6.52 presents various statistics upon the 'performance' of of the various measures as to their speaker discriminant potential.

From the figure it is clear that for all but one measure there is an extremely wide distribution of ω^2 values for each individual measure. Such a result indicates that no single measure is highly correlated with speaker identity in all circumstances, parameters and sentences, and tends to vindicate the approach of using a melange of measures. In terms of 'consistent' performance the DTW distance measures and the *Off Diagonal Warp Path Distance* appear to be the best regular indicators of speaker identity. However, if extreme correlation values are considered then the two measure: *Number of Vertical Excursions*, and *Number of Diagonal Excursions* have individual correlations within their distributions which exceed 0.1 in value, and hence, under those particular circumstances, appear to be strong measures of speaker identity.

Many researchers [Dod71b, NND89] have shown the significance of the DTW distance for speaker recognition and it needs no elaboration here. The *Off Diagonal Warp Path Distance* is an overall measure of how close the calculated warp stayed to a hypothetically optimal path of a straight diagonal linking the start and ends of both contours (linear time warping). We can reasonably expect that intra-speaker comparisons would deviate far less from this course than would inter-speaker comparisons [SF78]. The *Number of Vertical Excursions* and *Number of Diagonal Excursions* both measure the amount of micro timing adjustment and changes being done, implying that in general intra-speaker comparisons require far less adjustment than do inter-speaker comparisons.

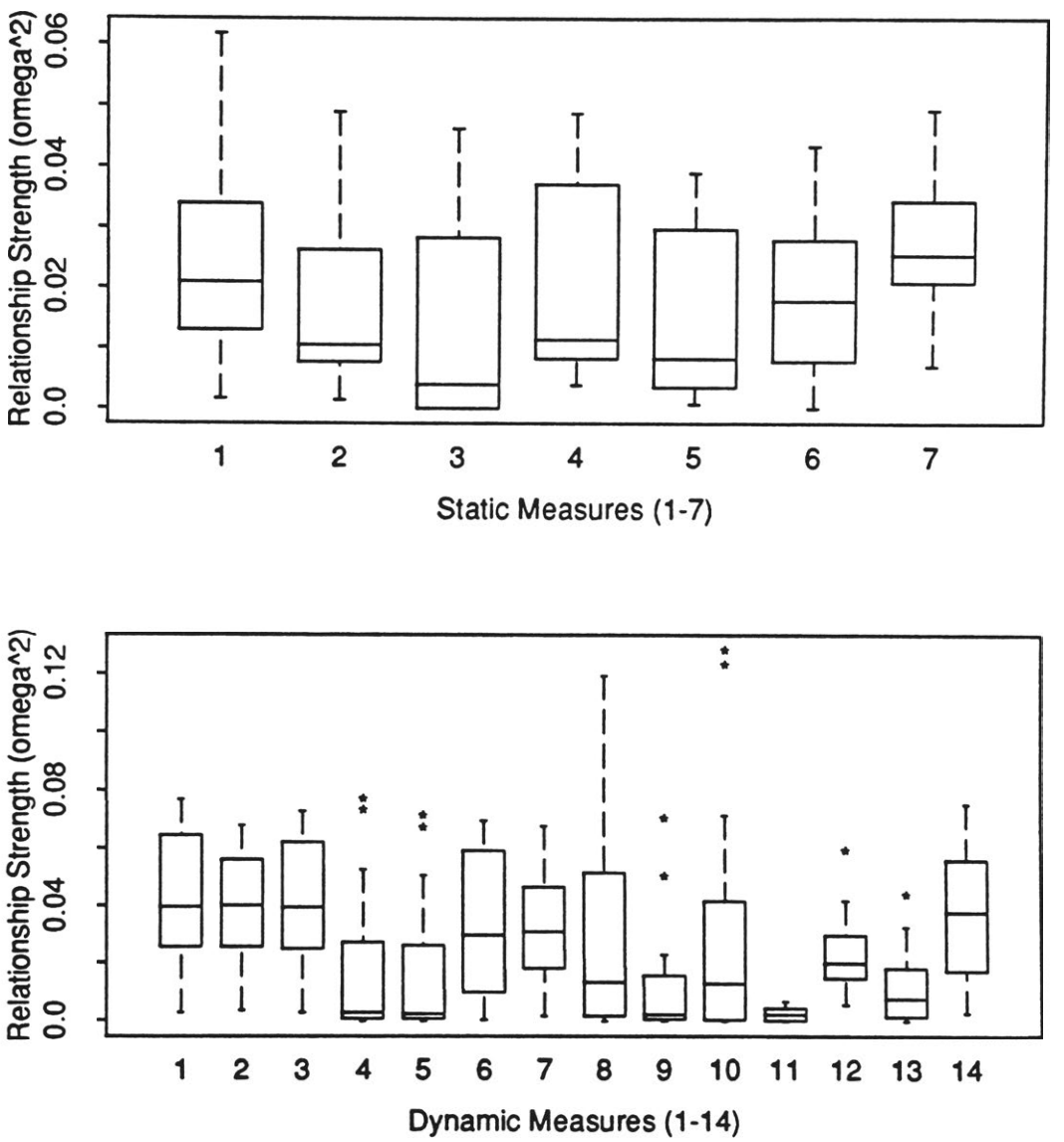


Figure 6.52: Speaker Identity: Boxplot of distributions of the 21 measure correlation scores for all 4 sentences and the 4 speech parameters E, F_0, Vuv, and Zc. Individual boxes show the distribution for each of the measures, indicating median value and quartile distribution.

6.8.2 Speaker Sex

Figures 6.53, and 6.54 present statistics regarding the individual measures and their ability to discriminate sex. Due to the difference in speaker sex discrimination level between F_0 and the other three parameters the measure distributions have been split into two—one for the F_0 parameter and the second for the three 'weaker' speech parameters.

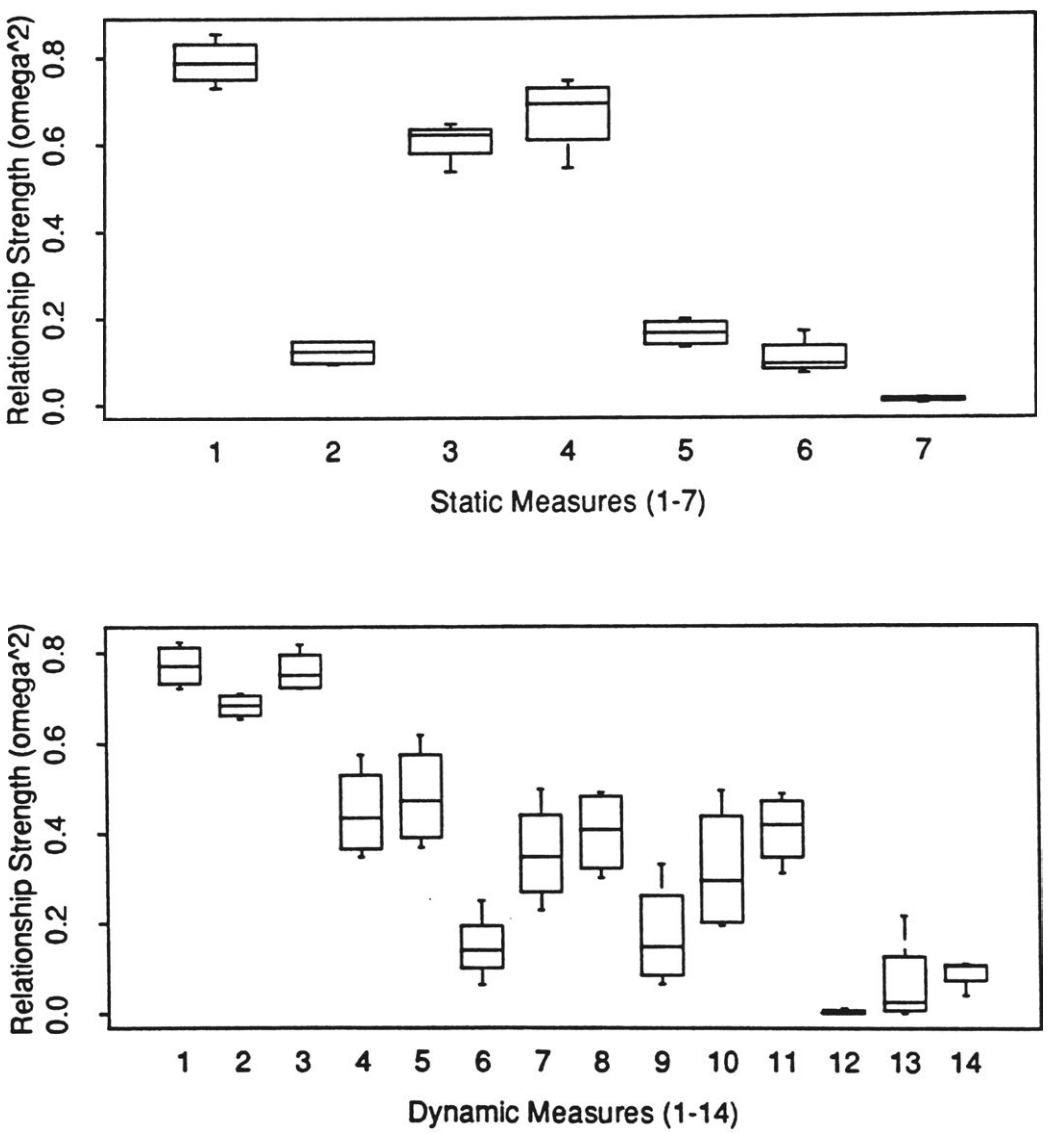


Figure 6.53: Speaker Sex: Boxplot of distributions of the 21 measure correlation scores for all 4 sentences and the speech parameter F_0 (by far the highest correlated of the 4 basic parameters examined). Individual boxes show the distribution for each of the measures, indicating median value and quartile distribution.

Analysis of Figure 6.53, showing the distribution of sex correlation values for measures of the F_0 parameter show that the three static measures *mean*, *max*, and *min*, together with the three DTW distance measures are the most strongly correlated, with average scores well in

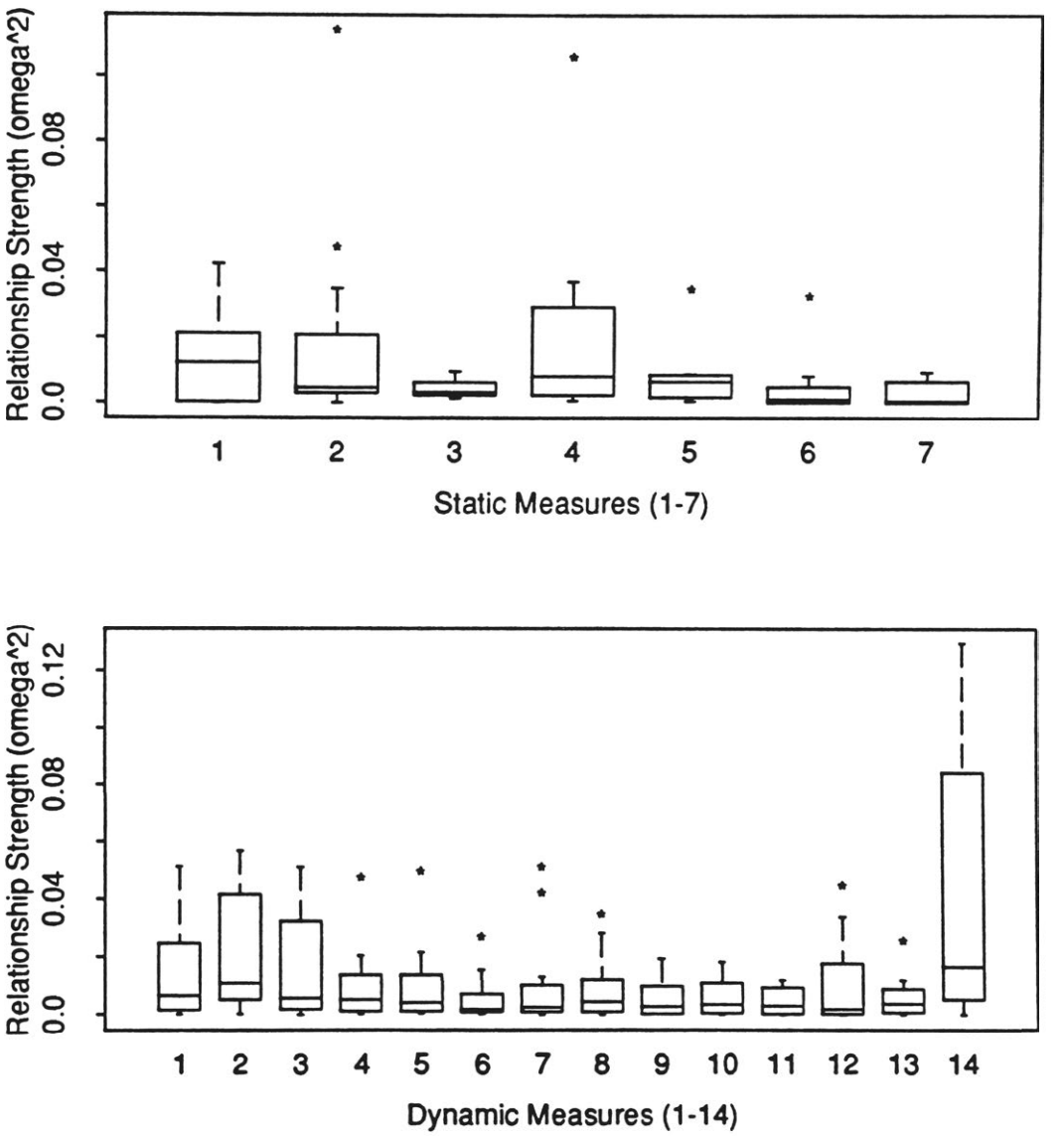


Figure 6.54: Speaker Sex: Boxplot of distributions of the 21 measure scores for all 4 sentences and the 3 speech parameters E, Vuv, and Zc. Individual boxes show the distribution for each of the measures, indicating median value and quartile distribution.

excess of 0.6, once again showing the importance of static F_0 for sex discrimination [Tit89]. However the larger portion of the DTW *transition* and *excursion* measures are also correlated with speaker sex to a non-trivial extent (medians ranging from 0.15 up to 0.45), showing the influence of these differences in mean F_0 upon the calculated warp path.

Comparing Figure 6.54 with the previous Figure 6.53 it is clear that measures of the non F_0 parameters are vastly inferior to those of F_0 for discriminating speaker sex. Nevertheless, all three of the measures *standard deviation, maximum*, and *Off diagonal warp path distance* contain individual correlation values that exceed 0.1, indicating that non-trivial indications of speaker sex are obtainable from individual measures of the three parameters energy, voicing, and zero crossing rate.

6.8.3 Speaker Dialect

Figure 6.55 shows the distribution of the correlations of the individual measures to the dialect-difference-score. The absolute correlation, sign less, is being used to simplify the presentation and in order that correlations of opposite sign derived from different speech parameters do not 'mask' each other.

Examination of the figure reveals that without fault all measures are correlated to lesser and greater extents to speaker dialect. The greater portion of measures have a median correlation value of 0.04 or more and the measures *duration* and *Fraction of horizontal transitions* have 'consistently' strong correlations, markedly in excess of those for the other measures. Further, the greater portion of measures, over 60%, have individual correlation scores in excess of 0.1 indicating far stronger correlation to dialect under particular constraints of sentence and speech parameter. Hence timing across the entirety of an utterance—duration and *Fraction of horizontal transitions*: a sort of 'catch-up' measure of shorter contour being stretched to match longer—appears to be the strongest indication of speaker dialect though all measures give some indication.

6.9 Examination of Sentences

In many of the previous sections the results of the analysis have been subdivided on the basis of the four sentences. Such an approach can show differences between the four sentences as to the varying degrees that speaker characteristics may be encoded within an utterance of the given sentence. Information such as this may prove useful in applications where it is necessary to choose appropriate utterances in order to maximise speaker characteristic discrimination, for example a speaker recognition system. Table 6.25 lists properties of each of the four sentences that may influence the discrimination or correlation rate for that sentence. Shown for each sentence are the mean duration, number of phonemes composing the utterance, mean voicing level (fraction of utterance that is voiced), and mean voiced duration.

There appears to be little previous research on the selection of suitable utterances for speaker recognition/verification systems. More generally, the lack of work to define suitable utterances is true for all speaker characteristics. It is beyond the scope of this work to derive such selection

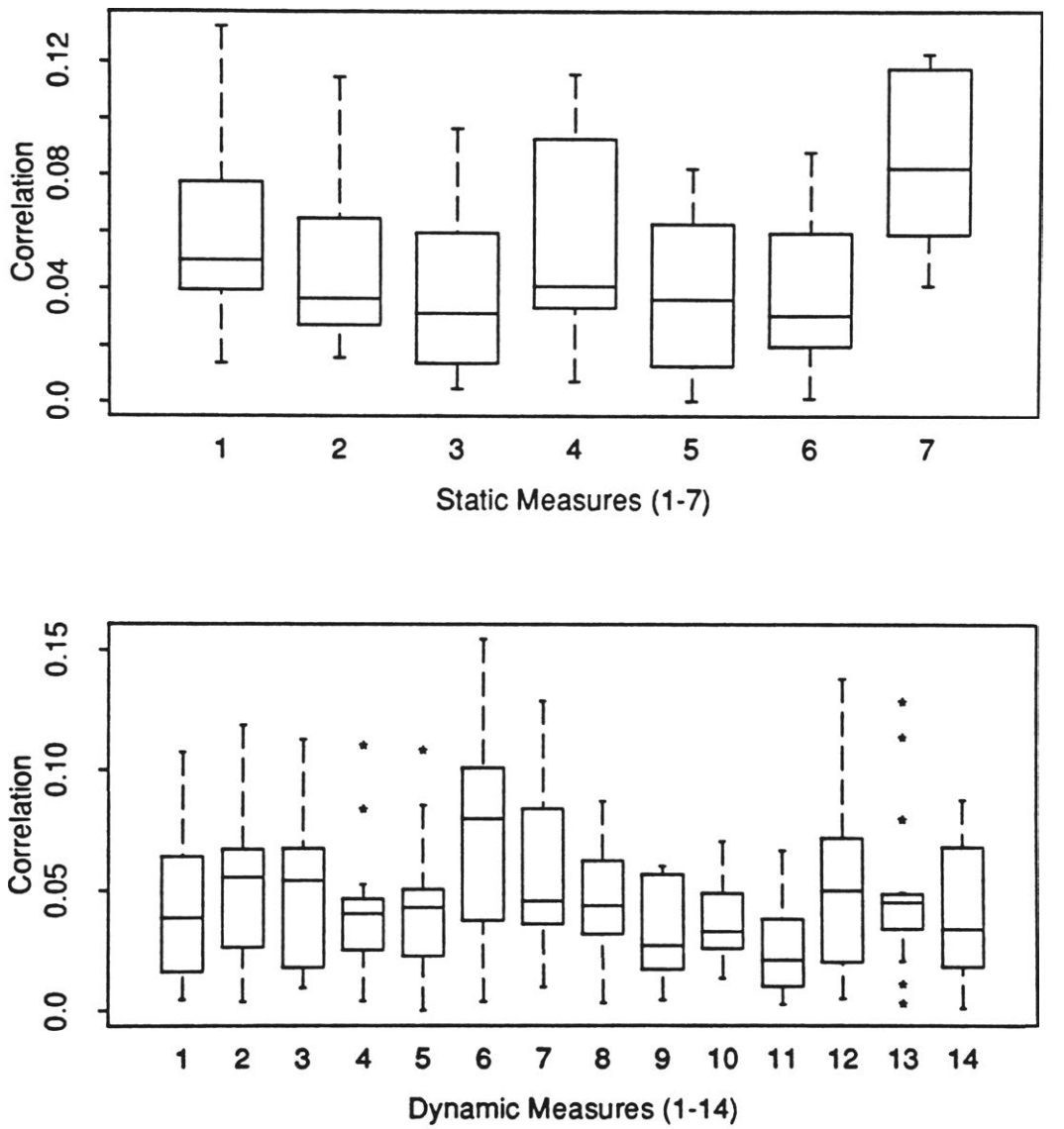


Figure 6.55: Speaker Dialect: Boxplot of distributions of the 21 measure scores for all 4 sentences and the 4 speech parameters E, F_0, Vuv, and Zc. Individual boxes show the distribution for each of the measures, indicating median value and quartile distribution.

Number	Text	Mean Duration [s]	App. Number of Phonemes	Mean Voicing Level	Mean Duration Voiced [s]
1	I cannot remember it.	1.16	15	0.61	0.70
2	How do you know?	0.81	8	0.82	0.67
3	We are firm.	0.89	6	0.71	0.63
4	We were away a year ago.	1.58	13	0.82	1.30

Table 6.25: Properties of the four analysis sentences. Listed for each of the four sentences are its number, the text of the sentence, its mean duration across all utterances, number of phonemes composing the utterance, the mean voicing level across all utterances, and the mean voicing duration across all utterances.

criteria. However, though four sentences are too few to adequately allow analysis of the various properties of sentences some parallels between the results obtained and properties of the four sentences will be drawn in this section. The following chapter, Chapter 7, will briefly outline a possible experiment for the establishment of such citeria.

An additional means of analysing sentence differences is to examine the distribution of the correlation scores for the individual measures. It might reasonably be expected that a sentence illustrating a distribution of higher measure correlation scores would make a better choice for overall discrimination performance than one displaying a lower distribution. Hence, for each of the three speaker characteristics, the distribution of the measure correlation scores, regardless of the particular measure itself or the speech parameter it was derived from, will be examined on a sentence basis. Distributions of correlation scores, and the discrimination or correlation scores for each sentence may then be related to properties of the sentences, Table 6.25, and some tentative parallels drawn.

6.9.1 Speaker Identity

Figure 6.56 shows the results of the analysis of the measure correlations, ω^2, values on the basis of sentence.

Based on the results it appears that sentences 1 and 4 yield markedly higher measure correlation scores in terms of general distribution, median and quartiles, and upper extremes, than do the other two sentences: 2 and 3. Such data correlates well with the identity discrimination experiments for each sentence (Section 6.1), where it was found that sentences 1 and 4 had significantly higher discrimination rates than either of sentences 2 or 3.

Both sentences 1 and 4 are greater in mean duration, number of phonemes, and have a higher mean voicing duration than either of 2 or 3. Given the demonstrated significance of F_0, and increased discrimination performance with increased material, the reason for the better performance of sentences 1 and 4 may be attributed to any, and likely all, of these differences.

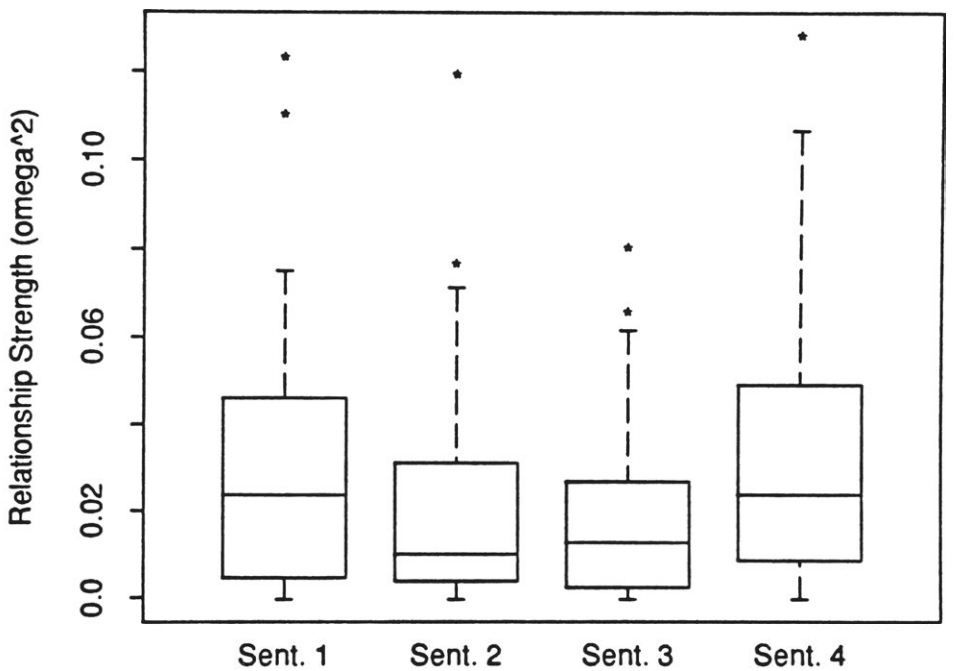


Figure 6.56: Speaker Identity: Boxplot of measure correlation Distributions for the 4 sentences examined. For all 21 measures of the 4 basic speech parameters E, F_0, Vuv, and Zc the correlation values are grouped according to the 4 sentences.

6.9.2 Speaker Sex

Figure 6.57 shows the results of the analysis of the measure correlation values on the basis of sentence for the speaker characteristic sex. As for the previous analysis of measures and their sex discrimination levels the plot has been split in two, the first plot represents the F_0 parameter, while the second plot is for the three other parameters combined.

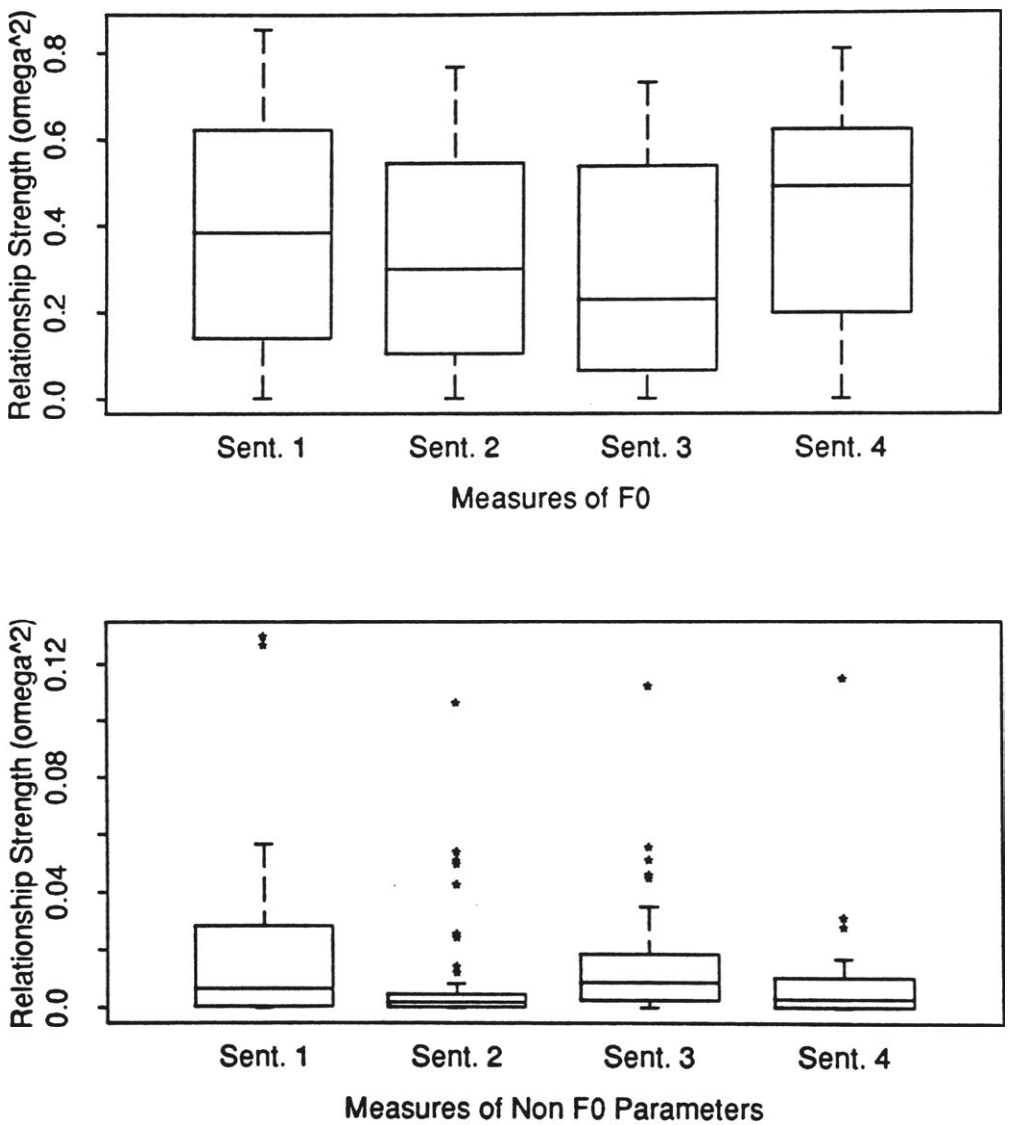


Figure 6.57: Speaker Sex: Boxplot of measure correlation Distributions for the 4 sentences examined. Two plots are shown, one for the speech parameter F_0, the other for the three speech parameters E, Vuv, and ZC. For each plot all 21 measure correlation values of the appropriate speech parameter(s) are grouped according to the 4 sentences.

From the plot for the F_0 parameter it is clear that both sentences 1 and 4 yield markedly higher measure correlation values, both in terms of general distribution, mean, and upper extremes, than either of sentences 2 or 3. Again, this better distribution of measure correlation

values for sentences 1 and 4 corresponds to the higher sex discrimination rates achieved for those same sentences in Section 6.1.

Interestingly, the same two sentences found to be most effective for speaker sex discrimination, namely 1 and 4, were also found to be the most effective sentences for speaker identity discrimination.

For the three other speech parameters—energy, voicing, and zero crossing rate, sentence 1 appears the best selection in terms of distribution of measure correlation values.

As mentioned in the previous discussion of the identity results sentences 1 and 4 have both a greater mean duration, number of phonemes, and a higher mean voicing duration than either of sentences 2 or 3. Any, and all of these reasons may be the cause for the better sex discrimination performance.

6.9.3 Speaker Dialect

Figure 6.58 shows the results of the analysis of the measure correlation values on the basis of sentence for the dialect-difference-score.

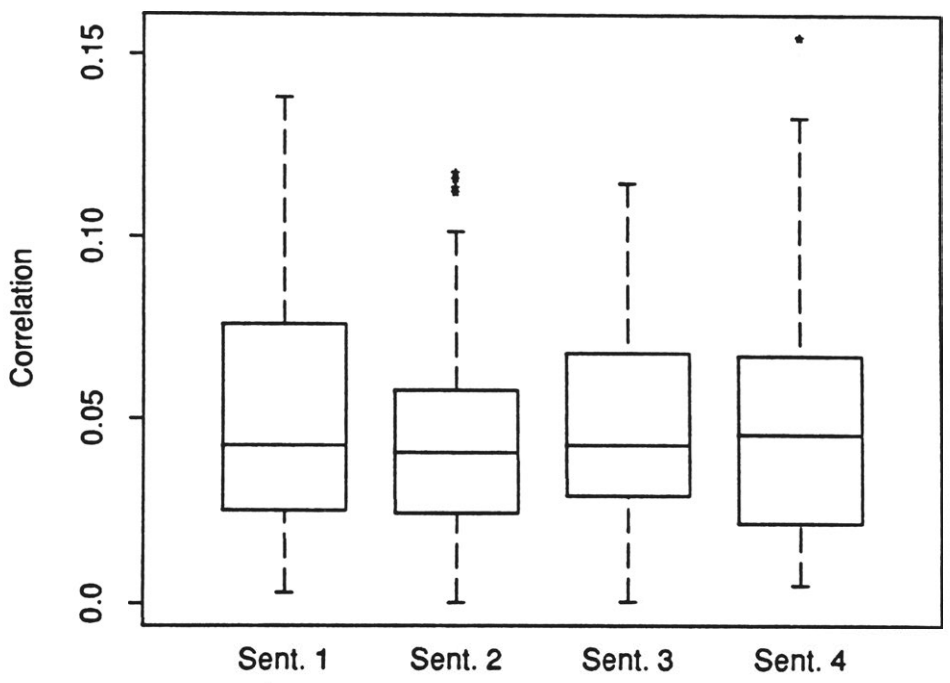


Figure 6.58: Speaker Dialect: Boxplot of measure correlation Distributions for the 4 sentences examined. For all 21 measures of the 4 basic speech parameters E, F_0, Vuv, and Zc the correlation values are grouped according to the 4 sentences.

Based on the figures and tables it appears difficult to reliably select any one sentence as potentially better for speaker dialect analysis. It is known from the results of Section 6.1 that sentences 1 and 2 have the highest dialect correlation scores, significantly better than either

sentence 3 or 4, but it is hard to find a parallel to such a result in Figure 6.58.

Further, it is hard to base these differences on properties of the sentences. For both speaker identity and sex results it was found that sentences 1 and 4 had better performance and this was assigned to the greater average duration of the sentences, higher number of phonemes, or the greater voiced duration of the two sentences. However the better dialect correlation performance of sentences 1 and 2, and poorer performance of 4, does not appear to be attributed to these reasons. In fact no grouping of the properties presenting in Table 6.25 appears to explain the better performance of sentences 1 and 2 as compared to sentences 3 and 4. Therefore there are other 'parameters' governing the degree of encoding of speaker dialect within a particular sentence.

6.10 Individual Speaker Effect

Up until this point the results have been analysed as to the various factors of the experimental design, such as static versus dynamic measures, or comparing speech parameters or other factors. In all such analysis the contribution and effect of any *single* speaker has been ignored and subsumed in the whole, the results for all speakers being grouped together and, in the cases of speaker sex and identity, divided into meta-classes such as inter-sex or intra-speaker.

In general the division of results on a single speaker basis is a time consuming process. However such a division of the results may show several interesting items of data hidden in the larger scale results. Often in recognition systems it is a small number of speakers that are responsible for the major portion of errors. Detecting such speakers may allow the system to be 'tuned' to account for these special speakers. Secondly, by grouping all speakers together for identity and sex discrimination experiments, there is an underlying assumption that there is a single common threshold for all speakers at which a 'cut-off' point between intra-class and inter-class distributions may be imposed. However, it may be that such a 'cut-off' point is speaker-dependent. Finally, grouping all speakers together shows overall trends and results for the entire speaker population, yet disguises the individual variance from this 'mean' corresponding to each speaker.

Many researchers [LKE85, LKW85, SMB81, Nod89, CF89, Dom90] have shown that speaker identity encoding in parameters is speaker dependent. Few if any have sought to examine this speaker specific manifestation for the other speaker characteristics. The breakdown of results on an individual speaker basis seeks to explore this phenomenon.

Hence, for each of the three speaker characteristics examined the results of the 'discrimination' experiment, when all four sentences with all measures of the four basic parameters are utilised, are split on the basis of speaker identity. For the speaker sex and identity experiments a boxplot is used to show the distribution of individual speaker 'scores' (the weighted sum of the individual measures). For each of the speakers two boxes were used to show distribution. The first 'narrow' box represents intra-class, sex or identity, comparisons. The second 'wider' box represents the inter-class comparisons for that speaker. For example, in the speaker identity plot, Figure 6.59, Speaker 0's distribution of 'scores' may be seen by examining the first two

boxes of the plot. The first 'narrow' box shows the distribution of all intra-speaker comparisons; i.e., the comparisons between all of Speaker 0's utterances. The second, 'wider' box shows the distribution of all inter-speaker comparisons for Speaker 0; i.e., comparisons between one of Speaker 0's utterances and one of those of the other speakers. For speaker dialect, the scatter plot, as previously encountered, was used to display the results for each individual speaker.

6.10.1 Speaker Identity

Figure 6.59 is a boxplot showing the intra- and inter-speaker comparisons for each of the eighteen speakers in the experiment.

Several items of information may be garnered from the plot. Firstly, the appears no single uniform distribution for intra- or inter-speaker values, and hence no single decision threshold may be selected for all speakers that will yield optimal discrimination performance. Rather, the selection of such a threshold is speaker dependent, i.e., individual for each speaker, and may only be made after examining a distribution such as the one shown. based on the figure it would appear that if such a policy were implemented then the discrimination rate would improve markedly.

Secondly, there appears wide variance between individual speakers as to the separation of intra-speaker and inter-speaker distributions, in effect the discrimination rate. While speakers such as 1, 3, 10 and 11 appear, with the selection of the appropriate individual thresholds, to have 100% discrimination rates, others, such as speakers 6, 12, and 15 show considerable overlap between intra-speaker and inter-speaker distributions, such that no matter the selection of threshold, 100% discrimination is unobtainable. It is to these later 'trouble' speakers that further attention, in any speaker recognition system, must be paid, either by further processing, or speaker training etc. It is worth noting that in those cases of overlap it is due to a wider intra-speaker distribution, in effect a lack of consistency in repetitions of the utterance by the speakers in question.

Hence, as other researchers as Shridar et. al. [SMB81] or Noda [Nod89] have shown, encoding of speaker identity within speech parameters is speaker dependent.

6.10.2 Speaker Sex

Figure 6.60 is a boxplot showing the distribution of inter-sex and intra-sex comparisons for each of the eighteen speakers used in the experiments.

It may be seen that there is a degree of variance between the individual speaker distributions, implying that the imposition of a single 'catch-all' threshold to differentiate inter- and intra-sex comparisons will lead to sub-optimal results. In fact close observation of the data reveals that if individual thresholds for each speaker are selected then a sex identification rate of 100% is obtainable for all speakers. Hence, manifestation or encoding of speaker sex within the parameters energy, F_0, voicing and zero crossing rate is speaker dependent to a minor extent.

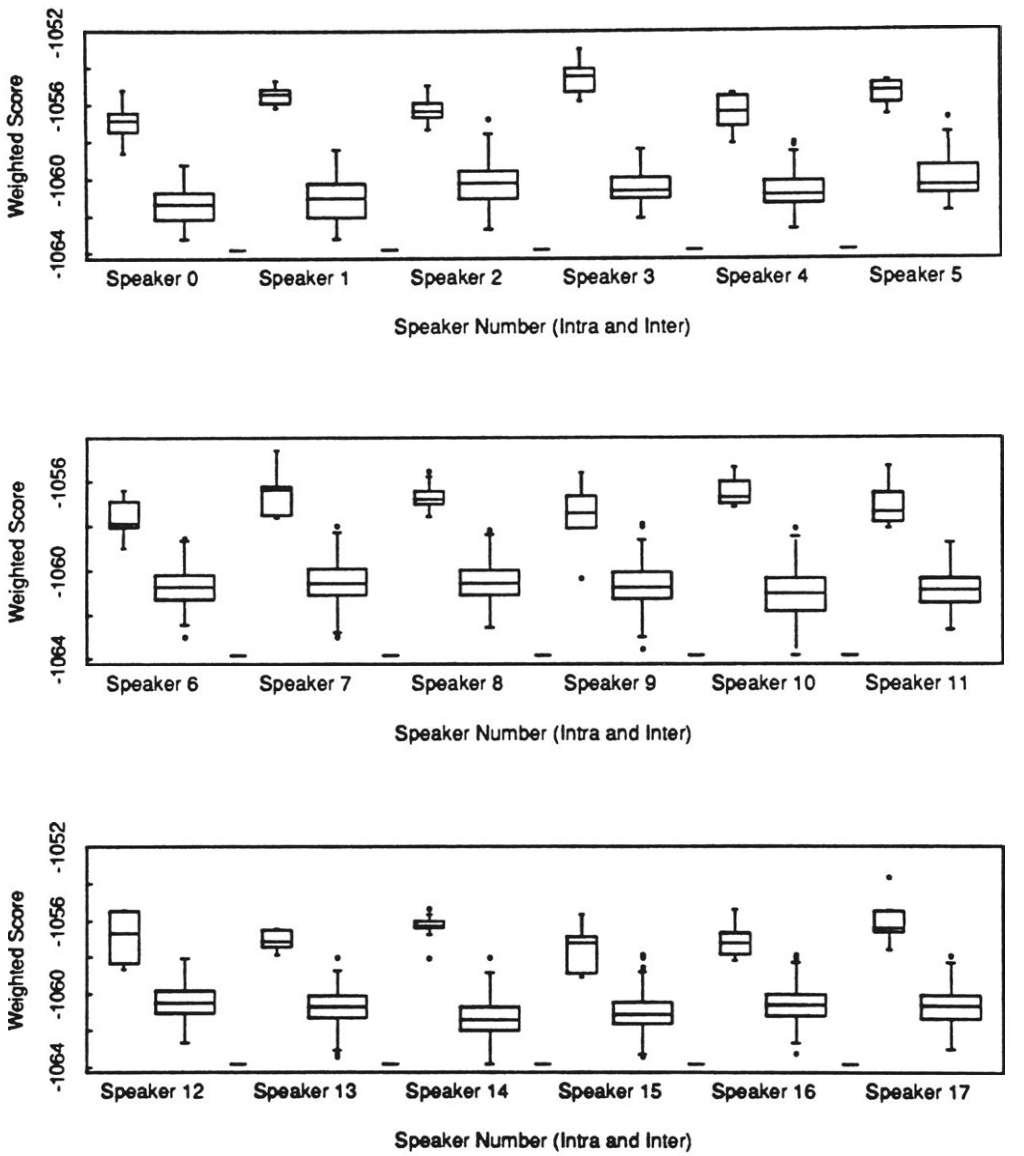


Figure 6.59: Speaker Identity: Boxplot showing the distribution of comparison 'scores' for each individual speaker. For each speaker two boxes are plotted. The first narrow box shows intra-speaker comparisons, while the second wider box shows the distribution of inter-speaker comparisons. The results are from the speaker discrimination run using all 4 sentences and all measures of the 4 basic speech parameters.

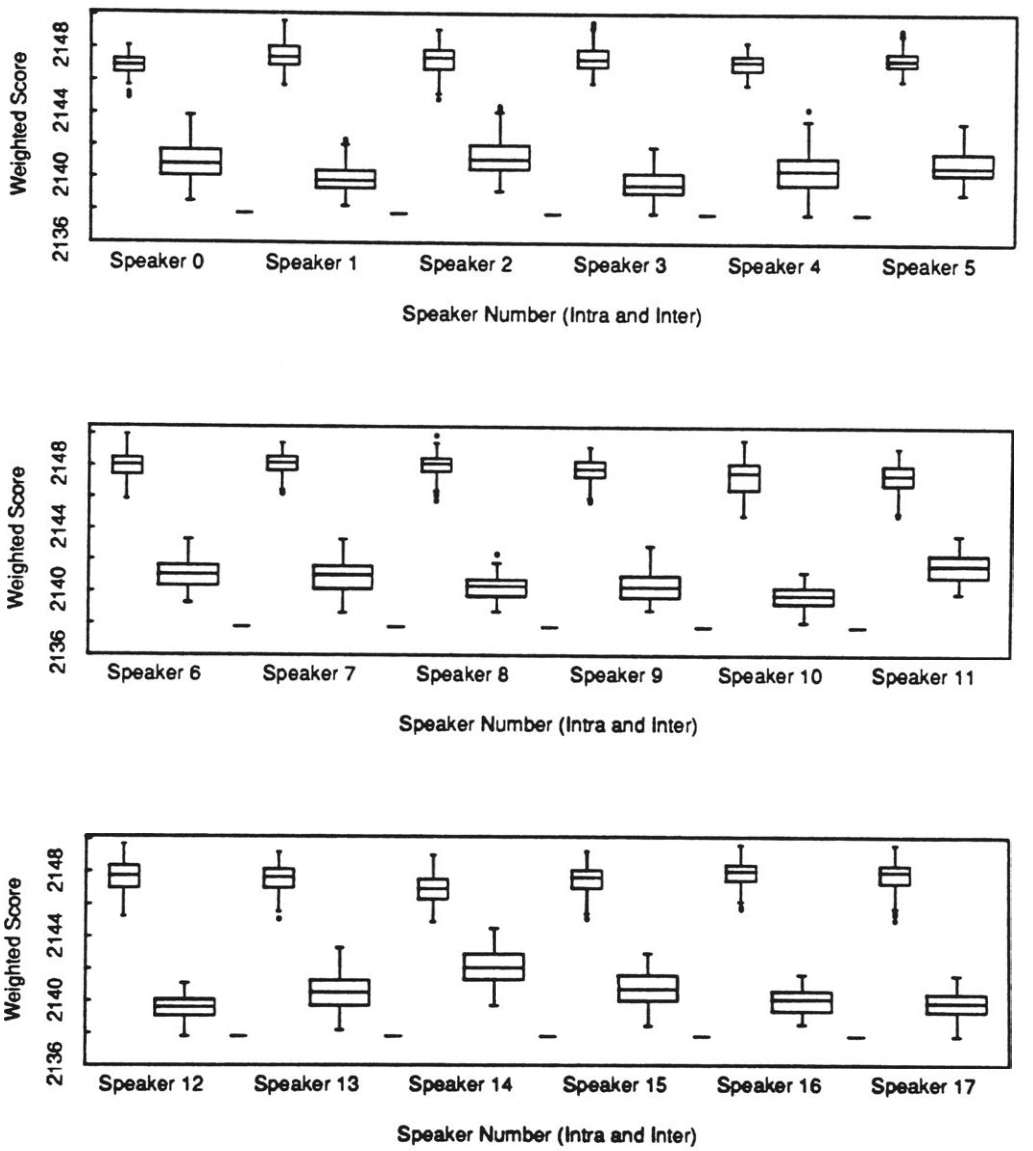


Figure 6.60: Speaker Sex: Boxplot showing the distribution of comparison 'scores' for each individual speaker. For each speaker two boxes are plotted. The first narrow box shows intra-sex comparisons, while the second wider box shows the distribution of inter-sex comparisons. The results are from the sex discrimination run using all 4 sentences and all measures of the 4 basic speech parameters.

6.10.3 Speaker Dialect

Figures 6.61 and 6.62 show the scatter plot of the dialect-difference-score against the calculated least-squares-fit, for each of the individual speakers. Its is immediately apparent that while for all speakers taken together there appears an overall trend or relationship between the measured parameters and the dialect-difference-score, for individual speakers there are wide deviations from this pattern.

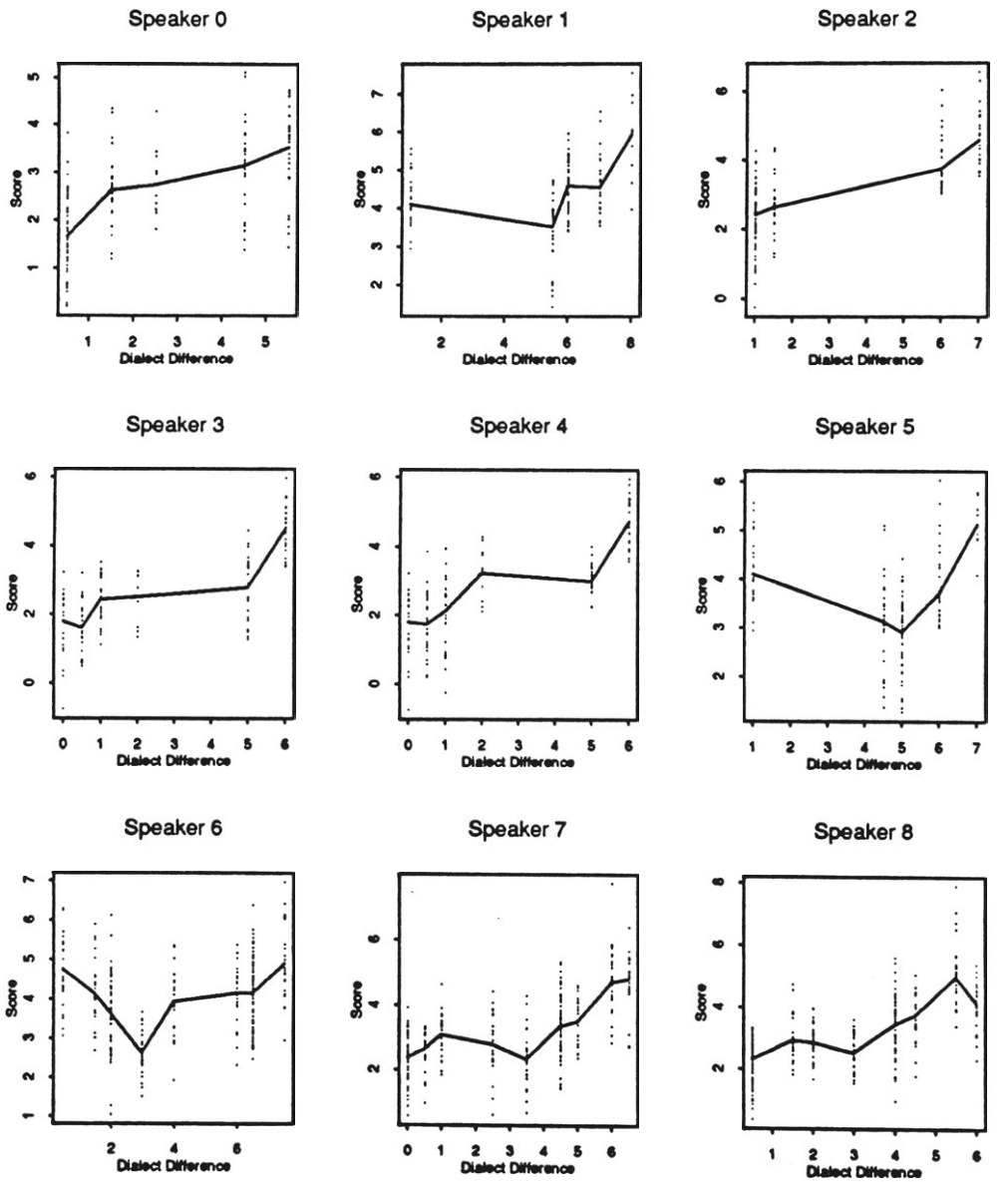


Figure 6.61: Speaker Dialect Least-Squares-Fit Scatter Plot: Individual scatter plots for the first 9 speakers. The results are from the experimental run using all 4 sentences and all measures of the 4 basic speech parameters.

Such a result seems to imply that while there are general dialect imposed trends in prosodic parameters for a large population of speakers, for individual speakers there may be large individual variances from this mean pattern. In effect, individual speaker effect may greatly

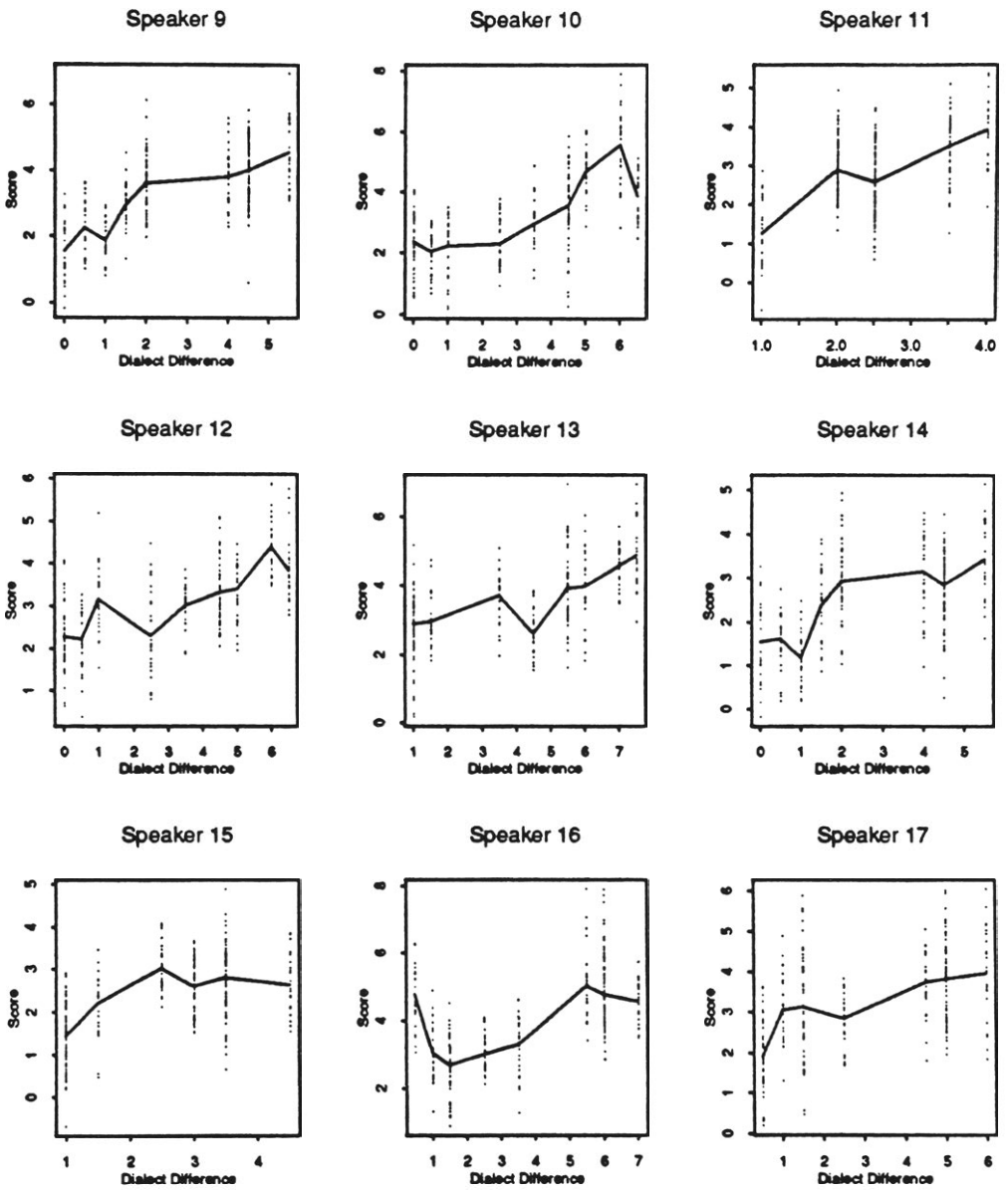


Figure 6.62: Speaker Dialect Least-Squares-Fit Scatter Plot: Individual scatter plots for the second 9 speakers. The results are from the experimental run using all 4 sentences and all measures of the 4 basic speech parameters.

influence such dialect imposed trends and make a general population model of prosody and dialect less than applicable to individual members of the population; i.e., encoding of speaker dialect in prosodic parameters is speaker dependent and highly variable.

Chapter 7

Discussion – Analysis Experiments

The previous two chapters have described and presented the results of an investigation of the three speaker characteristics identity, sex, and dialect by a direct analysis scheme. This chapter will discuss the results from several perspectives. Firstly the results for each of the three characteristics in turn will be discussed followed by a discussion of the technique and broad implications of the results regardless of specific characteristic.

7.1 Speaker Identity

Speaker discrimination experiments were carried out using four different sentences, and the four prosodic parameters F_0, voicing, energy, and zero crossing rate.

For a single sentence, with a speaker population of nineteen, a mean speaker discrimination rate of approximately 60% was achieved, rising, apparently 'logarithmically', to over 75% when all four sentences were combined. The discrimination rates may appear low at first glance but the nature of discrimination analysis - simple two class discrimination, where a single overlap in intra-speaker and inter-speaker distributions results in less than 100% discrimination - is more restrictive and 'tighter' than the standard decision algorithms [Dod85] used in speaker recognition. A single sentence appears insufficient to capture all possible speaker specific information for the four parameters and the addition of further sentences increases discrimination rate. Further, discrimination rate varies markedly between the individual sentences and appears adequately modelled, as a growth function, by both the duration and extent of voicing of the sentence. However a growth function on the basis of speech material alone was found to be inadequate to accurately estimate optimal discrimination performance for an unlimited amount of speech material.

In contrasting dynamic (time varying) measures of the four parameters with static (time invariant) measures, as to their discriminant ability, it was found that dynamic measures were significantly better. For the four sentences combined, dynamic measure discrimination rate was

74.6%; only 0.6% less than the combined discrimination rate of 75.2%; as opposed to the static measure discrimination rate of 54.8%. A difference between static and dynamic measures of just under 20%. Clearly static measures have little to add to dynamic measures in regards speaker discrimination, and, taken alone, dynamic measures are far superior to static measures for speaker discrimination. This result shows that dynamic measures alone are adequate to encapsulate the speaker specific information in the four parameters examined.

Normalisation experiments were carried out in order to further differentiate static properties of a contour from its dynamic properties. Discrimination rates dropped by an order of 10% for a single sentence but by only 5% for the four sentences combined, as compared to that for non-normalised parameters. On a further breakdown of results it was shown that the discrimination rate based on static measures was more adversely affected than that for dynamic measures. These impressive results for normalisation, 70.5% discrimination of speaker for normalised parameters show the importance of the dynamic properties of the contours examined over their static properties. Contrasting discrimination rates for static measures of non-normalised parameters with dynamic measures of normalised parameters it was found that the rates for dynamic measures of normalised parameters were significantly higher. On this basis it appears that speaker identity is more strongly encoded in the time varying properties, of the parameters examined, than in the time invariant properties.

Examining each of the four parameters separately it was found that each carried speaker specific information though to greater and lesser extents. For the four sentences combined F_0 proved the 'superior' parameter, with a discrimination rate of 64.6%, followed by energy at 58.2%, zero crossing, and down to voicing at 47.6%. Breaking the results down, for each parameter, on the basis of static versus dynamic measures it was found that dynamic based discrimination rates were an order of 10% to 20% higher than the corresponding static measures. It is worth noting that the F_0 discrimination rate is still over 10% below the discrimination rate when all parameters are combined, showing, that while there is an overlap in speaker specific information between the parameters, encoding of speaker identity is distributed over the four parameters such that no single parameter is sufficient to encapsulate all encoded information.

While F_0 had hitherto been represented in the linear-concatenated form, three other variants, being linear-interpolated, log-concatenated and log-interpolated were also examined as to their speaker discriminant properties. It was found that of the four possibilities the log-interpolated version of F_0 yielded the highest discriminant score of 66.9%; 2.3% higher than the control linear-concatenated F_0. This result has bearing for earlier and subsequent results as in all cases the linear-concatenated version of F_0 is used. Therefore it could reasonably expected that if the log-interpolated F_0 was substituted these discrimination results would be a minimum, with the actual result being up to 2.3% higher than previously stated. Further it was found that the hierarchy of results could be explained by the properties of the four alternate representations. Results for log representations were higher than the equivalent linear version. Results for interpolated representations were higher than the equivalent concatenated version. Given these results it would appear that a log representation of F_0, where values have been interpolated for unvoiced frames, is the best representation of F_0 to obtain optimal speaker

discrimination. Interestingly when results were examined on a static and dynamic basis it was found that log interpolated gave the highest static discrimination rate while linear interpolation gave the highest dynamic. Possibly there is scope to include both representations of F_0 in a speaker recognition scheme.

As defined, dynamic measures are logically divided into the DTW distance measure, a measure of the difference between two contours once their relative dynamics have been normalised or equated as thoroughly as possible, and warp path measures, measures of the relative dynamics of the two compared contours. These logical grouping may be contrasted as to their efficacy to discriminate speakers. For a single sentence the DTW distance discriminates speakers with a mean of 37.2% as opposed to a rate of 51.6% for warp path measures. For all four sentences combined DTW distance has a discrimination rate of 54.2% as opposed to 72.0% for warp path measures. Clearly, the warp path measures contain more speaker specific information than the DTW distance. Obviously the performance of many DTW based speaker recognition systems could be significantly improved by using the calculated warp path!

A simple Euclidean fixed end-point DTW distance was used for all dynamic measures. However two alternative distances, one weighted based on the warp path and the other allowing a degree of freedom in endpoint checking were also tried. It was found that the weighted DTW distance was inferior to both of the other two. Thus, while there is speaker specific information in the warp path this should be used as an additional vector for a recognition system, rather than attempting to incorporate it as weighting in the DTW distance. There appeared no significant difference between the other two measures.

A total of twenty one measures, seven static and fourteen dynamic, were used as part of the discrimination experiments. These measures were individually contrasted as to their speaker discrimination for the four prosodic parameters used. All bar a single measure showed a wide distribution of ω^2 values (ANOVA derived estimates of correlation) implying that no single measure performed consistently well for all parameters and sentences but conversely under the right circumstances each was significant in contributing to the overall discrimination rate. Such a result shows that no single or small set of these measures is sufficient to achieve optimal speaker discrimination and a large suite or number of measures is necessary. However, of the set, the DTW distance measures and the off-diagonal-warp-path-distance were the strongest consistent measures; while number of vertical excursions and number of diagonal excursions yielded the best individual extreme (high) ω^2 values.

Finally, the speaker discrimination results were analysed on the basis of each individual speaker. It was found that there was a wide individual variance between individual speakers' intra-speaker and inter-speaker distributions such that no common threshold could be assigned to divide intra and inter-speaker distributions that was applicable to all speakers. Further, it was found that speakers could be split into two categories; those who had clear divisions between intra and inter-speaker distributions and hence are easy to discriminate; and those with overlapping intra and inter-speaker distributions, and hence are more difficult to discriminate. It was found that those speakers with overlapping distributions characteristically had wider, or more diffuse, intra-speaker distributions, implying a greater degree of variability in utterance

productions. As could be expected it is this small group of 'trouble' speakers who account for a great proportion of discrimination errors and to whom most attention needs to be paid to improve overall system performance. Noda [Nod89] has examined this phenomenon by investigating the uniqueness of speaker utterances in the N-dimensional parameter space. However it appears to be very much a 'chicken-and-egg' problem because uniqueness is defined by the choice of parameters, hence it appears to be difficult to detect trouble speakers ahead of time.

7.2 Speaker Sex

Speaker sex discrimination experiments were carried out using four different sentences and the four prosodic parameters F_0, voicing, energy, and zero crossing rate.

The mean rate for sex identification using a single sentence was found to be 93.1%, rising to 96.2% for all four sentences. Discrimination rate was shown to rise as the number of sentences used and hence duration, increased. However the correlation between discrimination and duration does not appear to be simple as, for example, a discrimination rate of 95.2% was obtained for a single sentence; only 1% less than that for four sentences combined. Based on the results for the F_0 parameter it appears more likely that the extent of voicing in the sentence has a strong effect.

When the measures are split into dynamic and static categories it was found that there was little to differentiate the two sets. Both discriminated sex at a mean level of 91% for a single sentence and for four sentences combined, dynamic measures discriminated at a rate of 95.8% as opposed to 94.7% for static measures. The four-sentence dynamic-measure discrimination rate is only 0.4% less than that for both measure categories combined. Therefore it appears that dynamic measures alone are sufficient to capture the sex specific information in the four parameters examined.

Parameter contours were normalised into the range 0 to 1 in order to examine whether speaker sex could still be discriminated with any degree of accuracy. Surprisingly, while discrimination levels dropped markedly it was still significantly above chance, and a rate of 77.7% was achieved when all four sentences were combined using normalised parameters. Therefore speaker sex is encoded in more than the mean values of these parameters; there is a significant amount of information in the dynamics of the contours. However, contrasting discrimination rates for static measures of the non-normalised parameters, with dynamic measures of the normalised parameters, it was found that static measures of non-normalised parameters yielded a significantly higher discrimination rate. On this basis it appears that speaker sex is more strongly encoded in the time invariant properties of the parameters examined, rather than time varying properties.

Examining the four parameters individually it was found that, as expected, F_0 was the superior parameter with a discrimination rate of 95.5% for the four sentences combined. However, all four parameters appear to carry some degree of sex specific information; with voicing at a surprising 71.8% for the four sentences. The F_0 alone discrimination rate of 95.5% is only 0.7%

less than for all four parameters combined; showing the other three parameters have little contribution to sex specific information that F_0 does not already encapsulate. For all parameters bar zero crossing rate dynamic measures were stronger discriminants than static measures.

As well as the base linear-concatenated representation of F_0, three other representations were examined as to their speaker sex discrimination. These were linear-interpolated, log-concatenated, and log-interpolated. It was found that there was little to differentiate the two linear representations. Both log versions of F_0 proved to be approximately 30% worse at discriminating sex, at a mean of 66% than the linear version. This trend is expected as the formula (Equation 5.5, page 59) for log F_0 involves the subtraction of a sex specific constant. Therefore it is surprising that log F_0 is able to discriminate sex at all, and this may be a reflection upon the choice of F_{0min} in the log formula.

Earlier it was shown that the dynamic measure set is superior to the static measure set for sex discrimination. The dynamic measures may be logically divided into the DTW distance measure and measures of the warp path. Examining this subdivision it was found that for a single sentence DTW distance discriminated sex with a mean of 88.2% as opposed to the warp measure rate of 80.3%. For all four sentences combined DTW distance discrimination rate was 92.8% was opposed to 91.7%. Thus, DTW distance is superior to the warp measures though the warp measures are surprisingly significant, and the DTW distance rate is 3% less than the combined dynamic rate; indicating that warp path measures do make a contribution not encapsulated by the DTW distance.

Evaluating the three variants for the DTW distance it was found that the weighted distance was markedly inferior to the other two. Incorporation of warp path information in the DTW distance appears to adversely affect sex discrimination rates. There appears little difference between the other two measures.

Comparisons of each of the twenty one measures were made as to their utility for speaker discrimination. Given the marked superiority of F_0, measures of F_0 were analysed separately to that of the other three parameters. Superior F_0 measures were found to be the mean, minimum, maximum and the DTW distances; and 60% of the warp path measures showed significantly large, 0.15 to 0.6, ω^2 values. Measures of the other three parameters were found to be markedly inferior to those of F_0, though based upon extreme values the off-diagonal-warp-distance, standard deviation and maximum appeared the best.

Finally, sex discrimination results were analysed on the basis of the individual speakers making up the test set. It was found that there was a high degree of individual variance between speakers as to the distribution of intra-sex scores. This result shows the influence of the individual speaker upon the results, indicating that individual thresholds will yield the best sex discrimination results, rather than a single threshold for the entire speaker population.

7.3 Speaker Dialect

Speaker dialect experiments were carried out using least-squares-fit analysis to correlate difference in dialect scores with measures of the four parameters F_0, voicing, energy, and zero

crossing rate; for four sentences.

Utilising all four parameters, a single sentence mean correlation value of 0.38 was achieved, rising to 0.58 for four sentences. Clearly and significantly speaker dialect is encoded in the four prosodic parameters. Correlation rate was plotted versus number of sentences used and modelled by a growth function of duration or number of sentences. Sentence choice and number of sentences, rather than duration, appear to be the most significant factors in increased correlation performance and it appears that major increases in correlation are possible with the addition of yet more sentences. Therefore, with sufficient data it would seem that Australian speaker dialect may be determined with a significant degree of accuracy based solely on prosodics.

Dividing measures on the logical basis of static versus dynamic it was found that dynamic measures were significantly more highly correlated with dialect than static measures. For the four-combined-sentence case static correlation value was 0.45 as opposed to the dynamic value of 0.56. Further this dynamic alone value of 0.56 is barely 0.02 less than the correlation value for dynamic and static sets combined showing that dynamic measures are virtually a superset of static with regard to dialect encoded information.

In an attempt to further explore the importance of the dynamics of the four parameters normalisation was applied prior to least-squares-fit analysis. For a single sentence the mean correlation rate dropped from 0.38 to 0.33; and for the four sentences combined from 0.58 to 0.54. The small drop in correlation score from un-normalised to normalised parameters clearly shows the importance of the dynamics of a contour in determination of speaker dialect. Contrasting the correlation value for static measures of non-normalised parameters, with that of dynamic measures of normalised parameters, it was found that dynamic measures of normalised parameters yielded significantly higher values. On this basis it appears that speaker dialect is more strongly encoded in the time varying properties of the parameters examined, rather than in the time invariant properties.

Analysing each of the four parameters individually it was found that all carried dialect specific information, the two parameters energy and F_0 being the most significant, each with a four combined sentence correlation value of 0.36, zero crossing rate of 0.31 and voicing of 0.26. Clearly there is a hierarchy of parameters in terms of utility for dialect determination; but even the best two parameters are markedly less (0.2) correlated with dialect, when taken alone, than when all four are combined. Thus it appears that no single parameter suffices to capture all the dialect information encoded in each of the four prosodics, but that multiple parameters lead to enhanced performance.

Together with the base linear-concatenated representation of F_0 three other representations were tested. These were linear-interpolated, log-concatenated, and log-interpolated. It was found that of the four, measures of the log-interpolated were the most correlated, a marked 0.032 higher than the default linear-concatenated version. Based on this result we could reasonably expect that if the log-interpolated version of F_0 were substituted for the default linear-concatenated, used in all multi parameter experiments, that quoted correlation rates would be even higher. Further a relational hierarchy was found amongst the four versions of F_0. Log

versions had higher correlation values than the corresponding linear version, and interpolated versions were more highly correlated than the equivalent concatenated version.

A further breakdown of dynamic measures is that between the DTW distance and the measures of the warp path. When these were examined separately the warp path measures were found to be significantly more correlated than the DTW distance alone: for the four-combined-sentence case 0.450 versus 0.301. This result shows the importance of the relative dynamics of the contours for the encoding of dialect information. However the warp-path measures' correlation value of 0.45 is markedly less than the dynamic measure correlation value of 0.56 showing that the DTW distance and warp path measures combined carry more dialect information than either alone.

Twenty one different measures of the four parameters were examined as to their correlation to the dialect difference. It appears that every single measure elicits a substantial portion of dialect information. In fact over 60% of measures had individual correlations exceeding 0.1 in value. Based on this spread of correlations across many measures the strategy for optimal dialect correlation must be to use all of them as no single one or small subset suffices to capture all the information. Significantly and strongly correlated measures include duration, mean, maximum, fraction-of-horizontal-transitions, and maximum-length-vertical-excursion.

Finally, the four sentence-four-parameter results were broken down on an individual speaker basis. It is immediately apparent that individual speaker correlations diverge markedly from the total population model of the relationship of prosodics and dialect. Hence, while there is a total population model of prosodics and dialect, individual speaker effect greatly influences this model; that is: there appear to be *general* population trends correlating dialect and prosodics but individual speakers may vary greatly from this model. As a corollary of this result it is of utmost importance to gather a large speaker set in order to model the relationship between prosody and dialect.

Throughout the investigation a direct linear relationship in dialect difference was implicitly assumed. Thus the difference between a general and cultivated speaker is assumed to be similar to that between a broad and general speaker. This may however be an over-simplification. In order to explore this area using the dialect-difference paradigm it would be necessary to attempt various forms of curvilinear least-squares analysis [HW71] in order to determine if these yield higher correlations. Much work may yet be done in this area.

7.4 General Issues

The analysis experiments introduced the notion of a number of measures of the properties of prosodic contours. In particular, as an expansion to Saito and Furui's [SF78] work, a number of measures of properties of the dynamic time warp warp-path were designed and extracted. Taken as a set these measures were found to be more strongly correlated with dialect, better discriminators of identity, and marginally worse discriminators of sex, than the DTW distance. This result has important implications for recognition systems based upon time alignment, where generally the warp path or its equivalent is discarded after calculation and not utilised.

By the incorporation of properties of the warp path in the vector for recognition, significant, and possibly major, improvements in recognition levels might be expected. Further, while the technique was only applied to prosodic parameters there appears no reason why it should not also be applicable to spectral parameters. In fact, taking the progression one step further, it appears that the technique could be taken beyond the analysis of speaker characteristics and applied to that of speech recognition where improvements in recognition rate might also be obtained due to temporal differences between the targeted speech 'elements'.

A recurring theme of the analysis has been the comparison of the encoding of speaker characteristics in dynamic or static properties of the parameters involved. Parameter normalisation, and splitting the measures into dynamic and static sets were both used to examine the encoding of speaker characteristics. Speaker identity and dialect were both found to be more strongly encoded in the dynamics of the parameters, whereas speaker sex was most strongly encoded in the static properties of F_0, though still to a lesser extent in the dynamics of the parameters. Therefore the importance of the dynamic properties of parameters should not be overlooked. Additionally, these invariant properties of dynamic parameters may prove 'useful' to recognition systems where external agencies may alter the static value of extracted parameters, for example recognition over noisy phone lines, or processing the voices of deep sea divers.

Examining the four parameters F_0, voicing, energy and zero crossing rate separately it was found that for all three speaker characteristics each parameter carried relevant information. That is, that for all possible parameter-speaker characteristic pairings there was information relevant to the speaker characteristic encoded in the parameter. This result further strengthens the conclusion made in the literature review of the many-to-many relationship between speech parameters and speaker characteristics. Of the four parameters, F_0 was found to contain the most encoded speaker-related (as opposed to message-related) information, regardless of the particular speaker characteristic; showing, as many other investigators have [Dod71b, Wol72, Wag78, Tit89] the significance of F_0 in speaker characteristic analysis. Additionally, for both speaker identity and dialect, a log representation of F_0 was found to give significantly better results than a linear scale F_0, implying the utility of a log version of F_0 over linear for any recognition system seeking to employing F_0 as an input parameter.

As previously stated the 2 class discriminant analysis applied for both speaker identity and speaker sex is 'stricter' or yields lower 'scores' than most recognition system designs and classifiers [Ata76, O'S86]. In fact it is expected that, for the current speaker population and parameters, 100% identification rates would be obtained for many trials if even a simple k-nearest-neighbour decision algorithm was used. However, this very strictness of the discriminant algorithm is beneficial. Firstly it sets a lower bound on performance that could be achieved for a recognition system using similar input parameters. Secondly, many experimental systems with small speaker populations obtain very high recognition rates. Under such circumstances it is difficult to determine the impact of alterations in the experimental format, such as parameter set or measures, as all recognition rates are high. By setting stricter experimental criteria the impact of the changes upon a system under 'real world' conditions may possibly be more clearly seen.

A suite of twenty one measures of properties of the four prosodics, was trialed as to its utility in discriminating the three speaker characteristics. No single measure of the set was found to be the best or strongest under all conditions of parameter and speaker characteristic. In fact all measures appeared relevant to one or more speaker characteristics.

Four different sentences were used throughout the experimentation. Definite and significant differences were found between the sentences as to the inherent coding of speaker characteristics. In particular sentence number 1: "I cannot remember it." was found to be the 'best' for all three speaker characteristics. This result shows that the choice of spoken material used in experiments has a strong influence upon the results. Correlation and discrimination rates for the individual sentences were examined against properties of the fours sentences such as duration, or voicing level. No fixed relationship applicable to all results (speaker characteristic/sentence combinations) was found though duration (whether measured in seconds or number of phonemes) and voicing duration were in general good indicators.

Ideally, a comprehensive and well designed study is required to abstract a set of guidelines or rules for choosing utterances for a speaker characteristic recognition system; e.g. speaker verification. A large corpus of different utterances from a number of speakers must be collected, and a number of properties of the different utterances defined and measured. Recognition experiments may then be run for the different utterances and the results correlated with the measured properties of the utterances. Based on these correlations, rules guiding the selection of utterances may be derived. Many factors must be considered in the design of such an experiment—speaker characteristic(s) examined, parameters extracted, utterance properties defined, utterance set, speaker set, type of recognition trial, etc.

Increasing discrimination and correlation performance, as a function of increasing speech material, was modelled by growth curves. For all three speaker characteristics, and amount of speech material represented by mean duration in seconds or number of sentences, the model was found to satisfactorily represent the relationship. However the models could not satisfactorily represent all variance in obtained discrimination and correlation values, and gave unreliable estimates of upper limits upon performance. Clearly, as mentioned above, other factors than simply the amount of speech material affect performance and would need to be incorporated into any model if accurate judgements of performance for different experimental parameters (e.g., unlimited speech material) are required.

Results for each of the three speaker characteristics were also examined on the basis of the individual speakers that comprised the speaker set. It was found that there was considerable variance in discriminant and correlation levels between particular speakers. For example there were speakers whose every utterance could clearly be recognised as theirs while for others it was not always possible (some inter-speaker measures being smaller than intra-speaker). Clearly then the choice of speaker set has major ramifications for the results obtained. A set of 'good' speakers will yield high correlation and recognition rates while a number of 'bad' speakers may greatly reduce such rates. Therefore there is a need to evaluate speaker sets to determine their uniqueness [Nod89] to better quantify the test-bed of speech data, and in order that these 'trouble' speakers may be detected and suitable strategies designed to improve recognition for

these speakers, as it would appear that this is where considerable improvements in *total* system performance could be achieved.

Chapter 8

Method – Perceptual Experiments

In the previous three chapters the method and results of an analytic investigation of the three speaker characteristics:- identity, sex, and dialect; have been presented and discussed. A complementary approach, that was alluded to in Chapter 3 is that of using human perception to examine the acoustic correlates of speaker characteristics.

Fundamental to the concept of such a scheme is a process of presenting speech material to a set of human listeners who provide judgements of the material. Said judgements are then compared and correlated with known qualities and parameters of the presented speech.

Such a scheme is subjective by its very nature—listener judgements— and hence serves as a complement to the purely analytical approach. Moreover such an approach has direct application in areas such as speech synthesis and speech coding, whereas the analysis scheme is more applicable to speech or speaker recognition systems.

This and the following two chapters shall present and discuss the method and results of a series of perceptually based experiments. The basic mechanism utilised in the examination of all three speaker characteristics was to compute a *'composite'* utterance comprising two or four individual speakers. Individual parameters of the composite utterance were then systematically altered and correlated to listener judgements of the presented utterances.

Following sections of this chapter will detail the experimental method—including the speech material, analysis-resynthesis system, listening experiments, and parameter alterations.

8.1 Analysis-Resynthesis Scheme

In order to compute a composite utterance from a number of speakers and allow the systematics alteration of prosodic speech parameters a means of analysing and resynthesising utterances was required.

To this end the linear predictive [MAHG76] source-filter model of speech production was adopted. The model is well known and leads to a strong separation between the parameters of

speech investigated—F_0, voicing, and energy— and the segmental or spectral parameters.

The autocorrelation linear predictive algorithm [MAHG76] was applied to derive a 20-th order linear prediction spectrum, a voicing, F_0, and energy term, for each frame of 25ms, with a 12.5ms overlap. The voicing, F_0 and energy terms extracted in the analysis experiments were substituted for those derived by the linear prediction; both due to the greater accuracy of the parameter values and in order that analysis and perception experiments parallel each other as much as possible.

8.2 Speech Material

A single sentence was selected to be used for all perception experiments. Based on the results of Chapter 6 the sentence: "I cannot remember it." - number 1, was selected because of its consistently high discrimination and correlation scores. A single utterance, that of each speaker's second recording session, was employed.

All utterances used in experiments were hand segmented into the phonetic sequence:-

[aɪ <sil> k æ nɒ t rə' m ɛ m bə ɪt]

each segment being transformed to an integer frame number corresponding to frames of the linear predictive analysis.

8.3 Speaker Characteristics

As for the previous analysis experiments the three speaker characteristics, identity, sex, and dialect, were selected for investigation. In order to eliminate or minimise the contribution of the other two characteristics when a single characteristic was examined, speaker groups were selected for each characteristic such that unexamined speaker characteristic variance within the set was minimised.

Hence, for the speaker identity trial two male speakers, numbers 7 and 12, were selected; both speakers having the same dialect score. For the speaker sex trials four speakers were selected, 2 female:- 3 and 4; and 2 male:- 9 and 14. Again, all four speakers had the same dialect score. Finally, for the dialect experiment, four male speakers:- 6, 10, 13 and 16 were selected, two from either end of the dialect scale.

Besides the above criteria of minimising the influence of non-examined characteristics for each experiment other conditions were utilised in the selection of speakers. Firstly, no speaker used in the investigation of one characteristic was to be used in the examination of the other two characteristics, on the basis that association by the listeners could influence their responses. Secondly, based on the results of Section 6.10, speakers were selected, where possible, who were highly discriminable or correlated with the model for that particular characteristic. For example, speakers 3 and 4, used in the sex perception experiments, were 100% discernible as females in the sex analysis experiments. Based on these selection criteria the previously mentioned ten speakers were selected.

8.4 Composite Model

Two significant problems need to be addressed when considering a perceptual investigation of the correlates of speaker characteristics to prosodic parameters. Firstly a method is required to neutralise or normalise the contribution of segmental or spectral[1] information in an utterance so that listeners' utilisation of prosodic parameters alone may be addressed. Secondly, for speaker characteristics other than identity, there is the problem of creating a single 'archetype' of that characteristic free of the influences of a single speaker. For example, all of speakers 0 through 5 are female, yet they are each female with their own individuality and not *solely* female.

A means of addressing these two problems is to generate a composite utterance, consisting of a number of component utterances from different speakers. Such a method 'averages' the spectral properties of the utterances while prosodic parameters may still be varied to that of one or the other of the component utterances. Further, combining a number of speakers, all with a single fixed speaker characteristic, leads to an 'archetype' utterance for that characteristic free of the influences of any single speaker. Alternatively, combining utterances from speakers of opposed or balanced characteristics, such as male and female speakers, may lead to a neuter utterance for that characteristic. Adjustments to parameters of the neuter utterance may then be correlated with listener alterations in perception. Figure 8.1 is a visual presentation of the scheme.

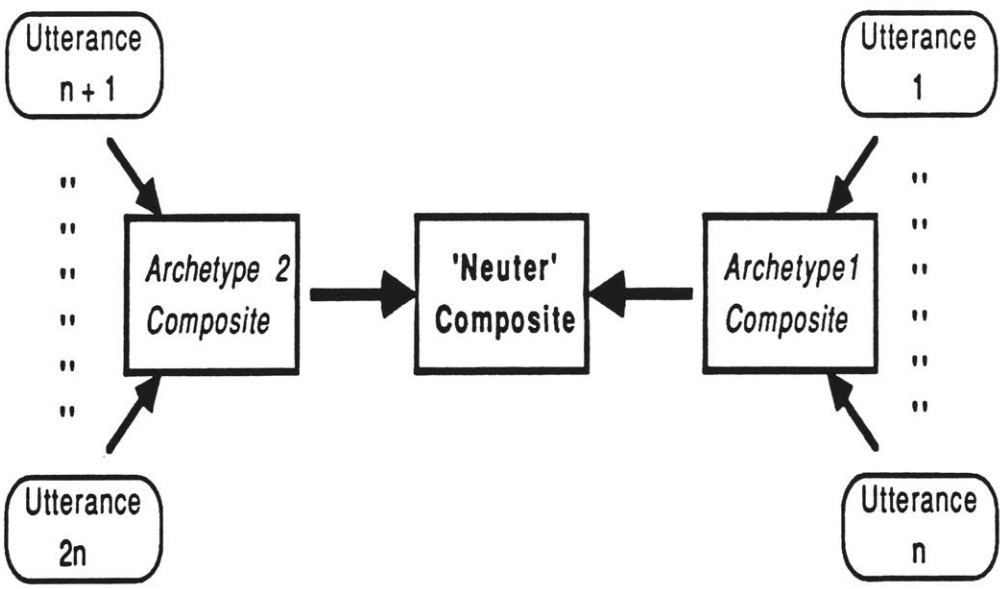


Figure 8.1: Composite Model of Resynthesis scheme. Archetype utterances are composed via the incorporation of a number of utterances; each representing the characteristic to be modelled (e.g., utterances from a number of female speakers to create an archetype female utterance). A 'neuter' utterance is then composed by combining two archetype utterances.

[1] Using the source-filter model of speech production associated with linear prediction the spectral information may be regarded as the co-efficients of the vocal filter.

8.5 Listener Experiments

Three separate listening experiments were conducted in sequence, using the same set of listeners, one experiment for each of the three characteristics investigated.

A total of sixteen listeners, eight male and eight female, ranging in age from 20 to 38 year took part in the experiments. No listeners reported any hearing defects and all bar two listeners were native speakers of Australian English. The two exceptions were speakers of British English and their judgements of speaker dialect were not incorporated into the results.

Listeners were verbally presented with a series of instructions, reproduced in Appendix F, and invited to question any instruction that was unclear.

As stated, three separate listener experiments were conducted, each consisting of between 7 and 15 minutes of listening time. Between any two experiments a rest period of approximately 5 minutes was imposed during which time listeners were encouraged to relax.

All experiments were forced-decision, double-blind trials presented in a random order; each sample being presented twice during the listening experiment. A period of 2.5 seconds was inserted between the end of one listening unit and the start of the next in which time listeners were forced to make a decision; either by marking an appropriate box as for the identity and sex trials, or marking a point on a line for the dialect trials.

Listeners were presented utterances in a micro-computer laboratory over stereo headphones (Dick Smith C-4101), a software script driving the presentation of the digitised utterances.

8.6 Speech Alteration

A number of alterations were made to the linear predicted speech parameters, both in order to generate composite utterances and to judge the effect of systematic parameter alteration upon listeners' perception of speaker characteristic.

8.6.1 Piecewise Segmental Interpolation

In order to compose a composite utterance it is necessary to time align utterances and parameters of different durations. Such an alignment may be achieved by performing a piecewise linear alignment for each of the segmental durations as derived from hand segmentation of the utterances. The basic algorithm, excluding special cases, for such a process is:-

Given:

x[1...N], the original contour

S, the number of segments

oldsize[i], the original number of values in the i'th segment

newsize[i], the target number of values for the i'th segment

Calculate:

y[1...M], the adjusted contour

oldindex ← 1

```
newindex ← 0
for i ← 1 to S do
        interpolation_factor ← oldsize[i]/newsize[i]
        weighting ← −½ interpolation_factor
        for j ← 1 to newsize[i] do
                weighting ← weighting + interpolation_factor
                while weighting > 1 do
                        oldindex ← oldindex + 1
                        weighting ← weighting - 1
                newindex ← newindex + 1
                y[newindex] ← weighting * x[oldindex+1]
                        + (1-weighting) * x[oldindex]
```

In order to maintain filter stability the 20 linear prediction co-efficients representing each frame were transformed to reflection co-efficients (RC) using the STEPUP/STEPDN algorithm [MAHG76] for the interpolation process.

8.6.2 Direct Parameter Substitution

Fundamental to the listener experiments is the concept of directly substituting one or more prosodic parameter contours onto an extant utterance. Such a scheme is achieved by aligning the contour with the target utterance via the above interpolation method, followed by replacing the appropriate field in the linear prediction file. Such a process is applicable for the F_0, voicing, and energy parameters.

8.6.3 Parameter Warping

It is clear from the results of the analysis section that the warp path parameters, and hence relative dynamics of parameter contours carry significant speaker characteristic information. One means of investigating the perceptual significance of such information is to use the DTW mechanism to warp a contour to conform as closely as possible to another. Listener responses to the warped contour utterance may then be elicited and contrasted to responses for the original and target contour utterances.

Figure 8.2 shows the results of warping the F_0 contours of two speakers towards each other. Speaker 7's F_0 is the solid line; while speaker 12's is the broken. The first plot shows both speakers' F_0 unaligned but transformed to the composite utterance duration via segmental interpolation. The second plot shows speaker 12's F_0 warped to conform to speaker 7's; while the third plot shows speaker 7's warped to conform to speaker 12's. At no point does the warping process introduce new values, or discard old values; only compressing, by averaging multiple values together; and stretching, by repeating a single value a number of times.

A simplified algorithm, showing the transformation to the 'horizontal' contour, of a warp matrix, alone is shown below. All notation used is that consistent with the explanation of DTW

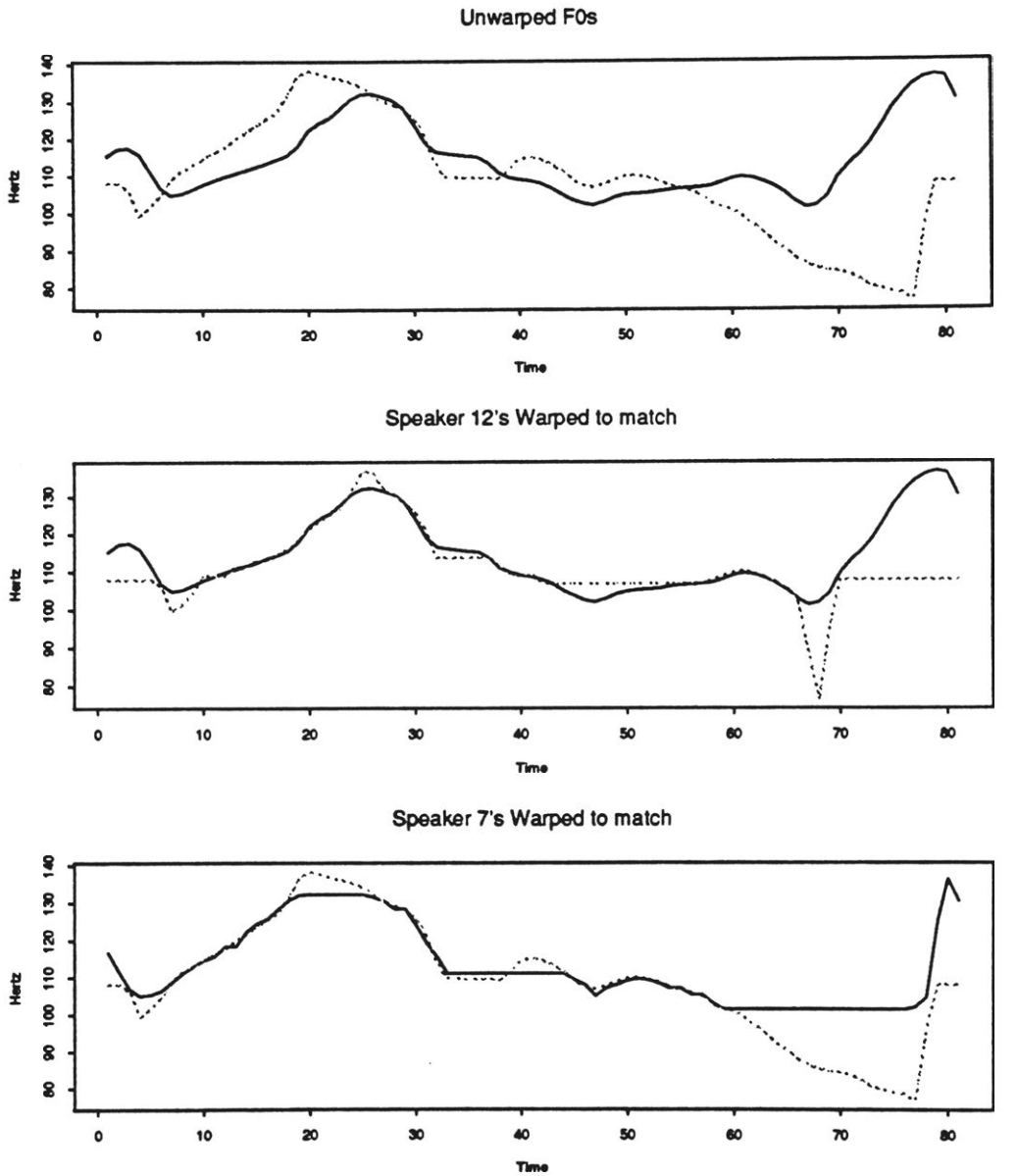


Figure 8.2: Original and warped F_0 contours from speakers 7, solid line, and 12, broken line, for the sentence: "I cannot remember it.". First plot shows both F_0s unwarped. Second is speaker 12's warped to match 7's, and third is 7's warped to match 12's.

in Chapter 5. It should be noted that the transformation algorithm for the vertical contour is isomorphic with the following algorithm.

Given:

w[k], the warp path transitions
K, the number of warp path transitions
a[i], the horizontal contour of length N

Calculate:

anew[j], the warped contour of a, of length M

```
k ← 1
i ← 1
j ← 1
while k ≤ K do
    sum ← 0
    count ← 0
    while w[k] = HORIZONTAL do
        i ← i + 1
        sum ← sum + a[i]
        count ← count + 1
        k ← k + 1
    if count ≠ 0 then
        anew[j] ← sum/count
        j ← j + 1
    while w[k] = DIAGONAL do
        i ← i + 1
        anew[j] ← a[i]
        j ← j + 1
        k ← k + 1
    while w[k] = VERTICAL do
        anew[j] ← a[i]
        j ← j + 1
        k ← k + 1
```

Such a transformation may be applied to the parameters F_0, voicing and energy.

8.6.4 Linear Parameter Alteration

Another possible adjustment to a prosodic contour is to add or multiply each value on the contour by a fixed amount. Such an approach will be used in the speaker sex investigations where F_0 contours will be linearly shifted to a mean of 165Hz. Such a scheme was accomplished by finding the mean of the original contour and comparing it with the new desired mean. Additive shift, which maintains absolute range and contour shape, was achieved by adding

the difference between the two means to each new contour value. Multiplicative shift, which compresses or expands contour shape and range, was performed by multiplying individual contour values by the ratio between the two means. Figure 8.3 shows speaker 3's original F_0 contour for the sentence: "I cannot remember it.", and as shifted to a mean of 165Hz via the two methods.

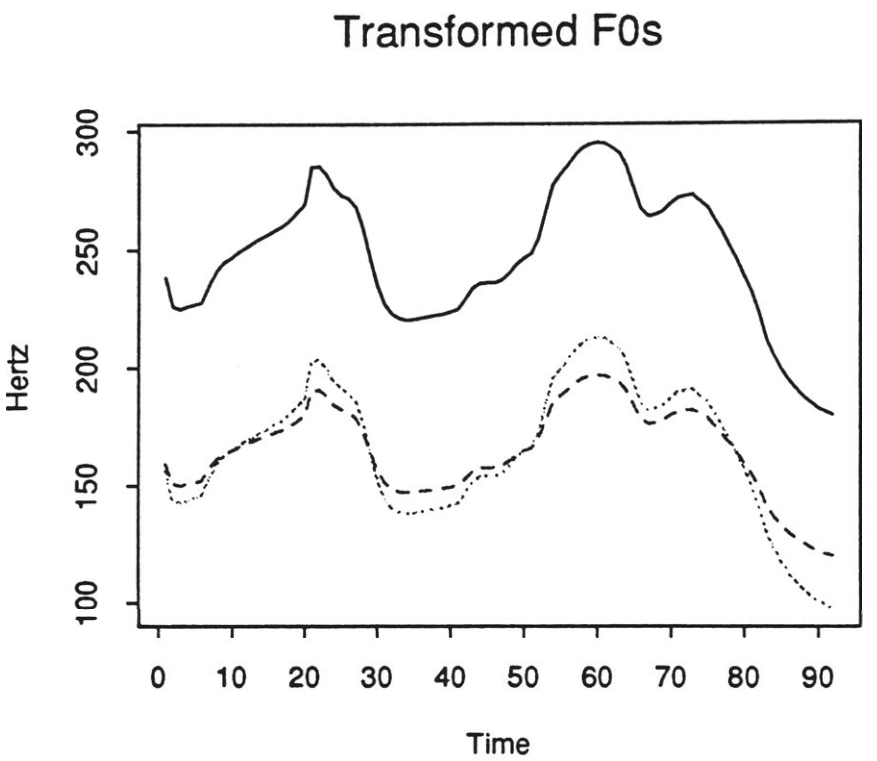


Figure 8.3: Linearly transformed F_0 contours for the sentence: "I cannot remember it." as spoken by speaker 3. The unbroken line is the original contour, the 'dotted' line is the contour shifted to a mean of 165Hz by addition; while the 'dashed' line is the contour shifted to a mean of 165Hz by multiplication.

Chapter 9

Results – Perceptual Experiments

The previous chapter outlined the method used in the conduct of the experiments to investigate the relationship between listener perception of speaker characteristics and prosodic acoustic parameters. To reiterate, composite utterances will be generated and played to listeners. Parameters of the utterances—timing, F_0, voicing, and energy, will be systematically altered and the new utterance played to listeners to ascertain the importance of that parameter for listener perception of the characteristic examined.

This chapter shall present the results of those experiments in three sections. Each section will correspond to a meta-class of experiments, one for each of the three speaker characteristics examined, namely:- speaker identity, speaker sex, and speaker dialect.

9.1 Speaker Identity

For the speaker identity experiments a single pair of speakers was selected in order that all parameter combinations could be thoroughly checked. The speakers selected, 7, known to the listeners as Alan and hereafter referred to as A, and 12, known to the listeners as Peter and hereafter referred to as B, were both male and had dialect scores of 2.5 (tending towards cultivated). For the utterance of "I cannot remember it.", used in the experiments, speaker A's utterance had a duration of 0.93 seconds and a mean F_0 of 114Hz; while speaker B's utterance was 1.20 seconds in length and had a mean F_0 of 108Hz.

From the utterances of A and B a single composite utterance was created. One concern with using such a model is the relative contribution of segmental or spectrum information versus prosodic as to listener's judgement of identity. Therefore an initial investigation was conducted where prosodics were held fixed at the mean of the two speakers, while spectral information, the linear prediction co-efficients, was varied from 25% to 75% on a linear scale between the two speakers. For a small group of listeners, none of whom took part in the main experiments, it was found that an equal weighting, 50%, of A and B's LPC to an equal distribution in listener

responses to the question of who the speaker was.

For the experiments four prosodic parameters were examined. These were F_0, voicing, energy, and segmental timing. Direct substitution of these parameters, both individually and in combination, was performed upon the composite utterance. Further, warping of the three parameters F_0, energy, and voicing (see Section 8.6.3) was performed and these alone, and in combination with each other and the segmental timing, were translated onto the composite utterance. Figures 9.1, 9.2 and 9.3 show the original and warped F_0, energy and voicing contours for the two speakers.

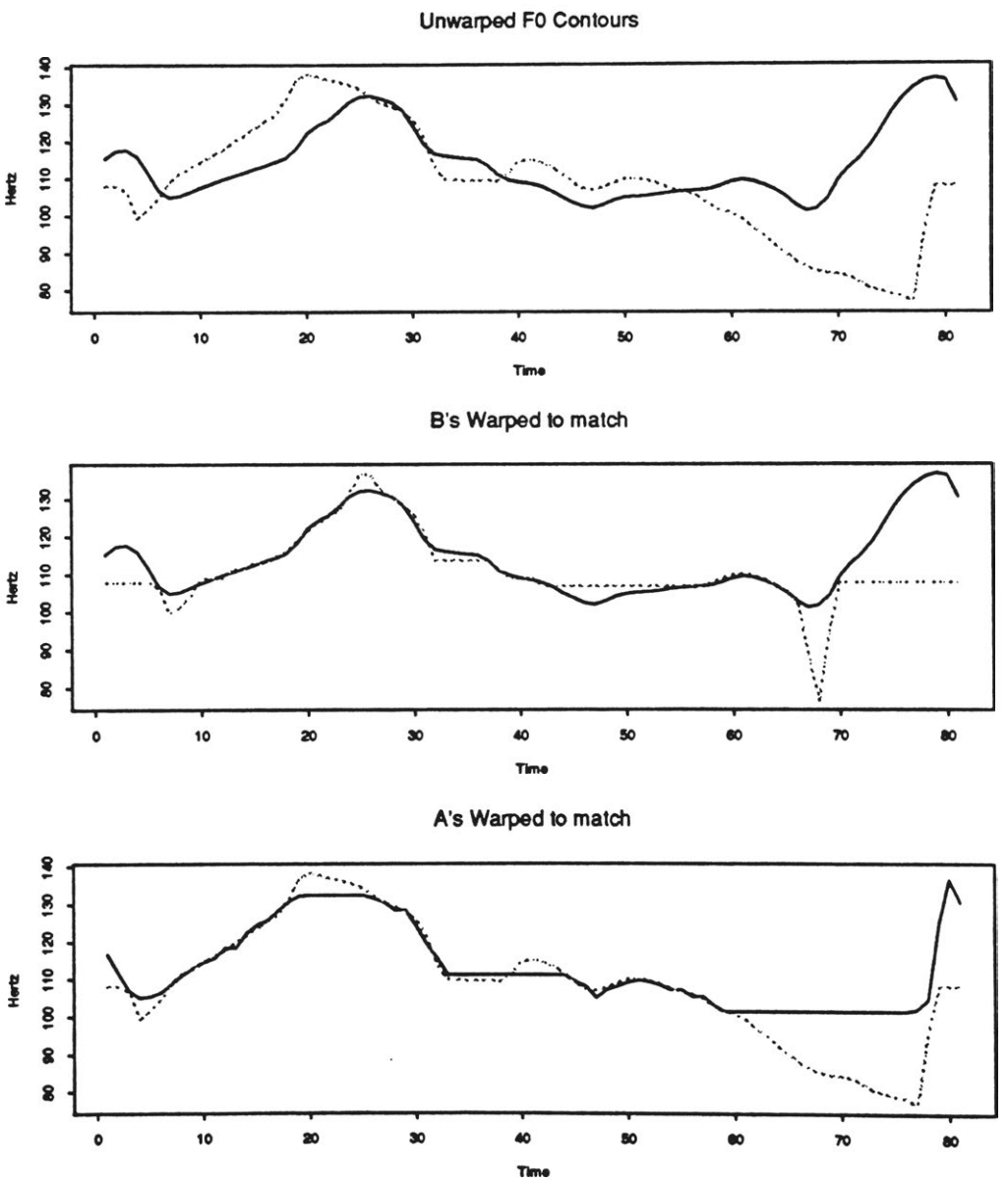


Figure 9.1: Original and warped F_0 contours for speaker A, solid line, and B, broken line; uttering the sentence: "I cannot remember it.". First plot shows both F_0s unwarped. Second is speaker B's warped to matched A's; while the third is A's warped to match B's.

For all result figures a barplot is used to show listener response distribution between speakers

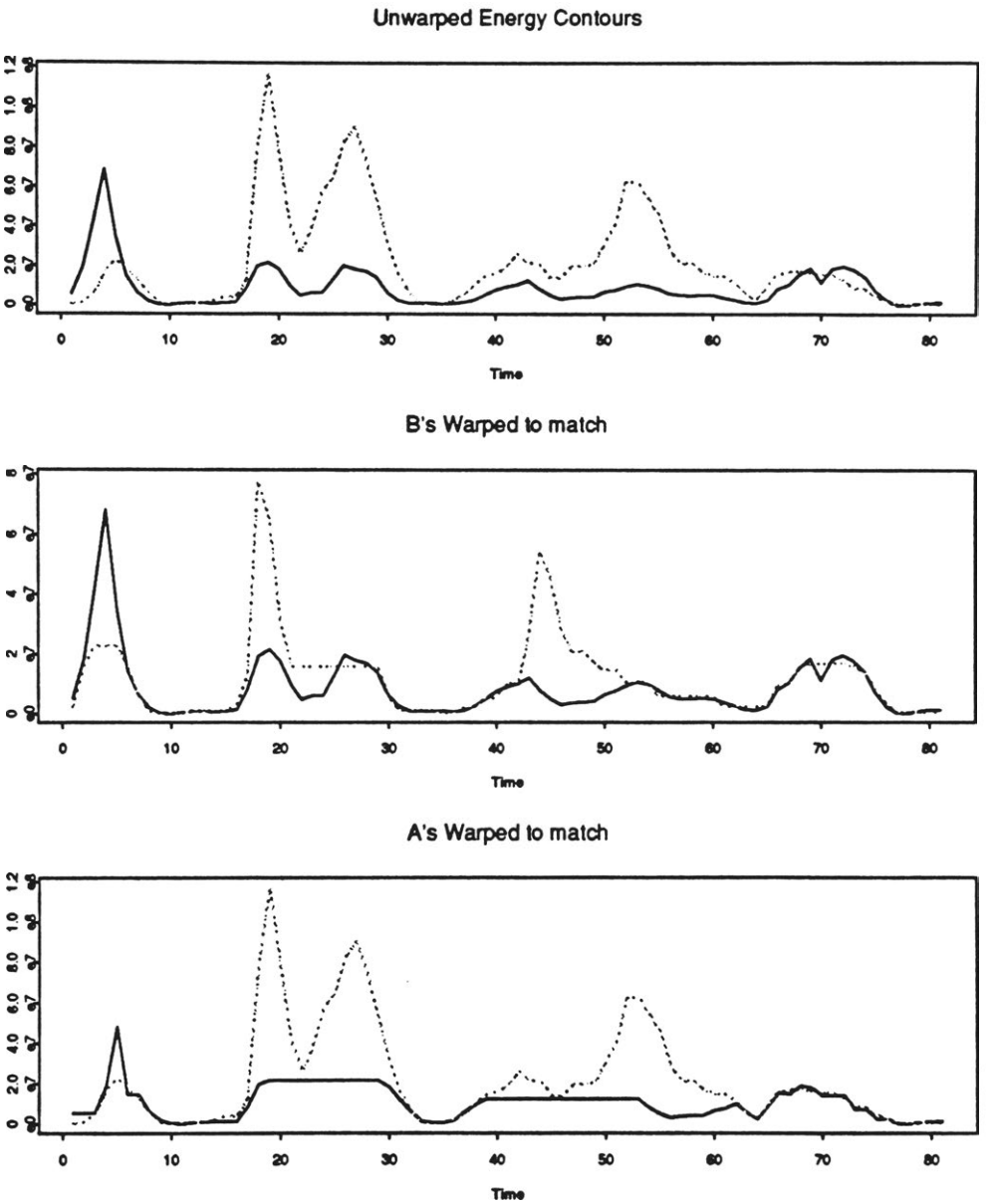


Figure 9.2: Original and warped Energy contours for Speaker A, solid line, and B, broken line; uttering the sentence: "I cannot remember it.". First plot shows both energies unwarped. Second is speaker B's warped to matched A's; while the third is A's warped to match B's.

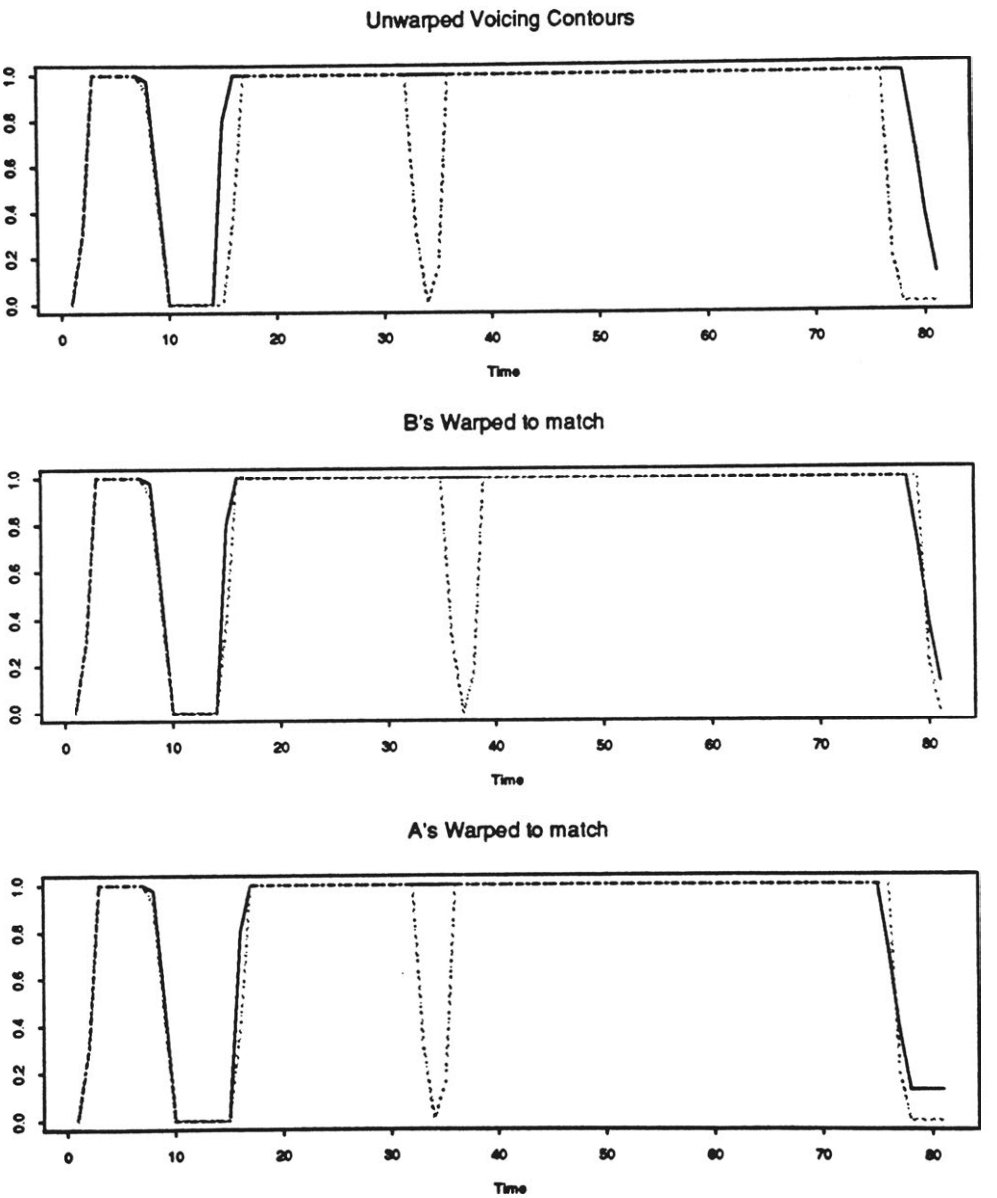


Figure 9.3: Original and warped Voicing contours for speaker A, solid line, and B, broken line; uttering the sentence: "I cannot remember it.". First plot shows both voicings unwarped. Second is speaker B's warped to matched A's; while third is A's warped to match B's.

A and B. Central to each plot is the grouped listener response to the 'totally averaged' composite utterance; that is an utterance composed of 50% LP co-efficients and 50% prosodics from both speakers. This is the control against which listener responses to the alterations may be measured. For the result tables, the shift in response from that of the control utterance is shown, rather than absolute listener response. A chi-squared test of proportion [FW80] is used to determine the significance of the results.

9.1.1 Parameter Substitution

This section will present the results of the direct substitution of the non-warped parameters onto the composite utterance.

Figure 9.4 and Table 9.1 show the results of listener responses to the substitution of a single prosodic parameter, from either speaker, onto the composite utterance.

"Encoded" Type		*Perception*	
		Shift to A	Shift to B
A	F_0	5	
	Energy	33	
	Voicing	8	
	Seg. Timing	23	
B	F_0		14
	Energy	5	
	Voicing	10	
	Seg. Timing	0	0

Table 9.1: Listener perception of identity based on a single encoded parameter. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance. The left-most column indicates the speaker from whom the encoded parameter was derived.

Two important facts may be noted from the figure and table. Firstly, the four prosodic parameters are not equivalent in their effect of listener perception of identity. While parameters such as energy or segmental timing appear to significantly (*5% level*) alter listener perceptions, others such as voicing appear to have no significant effect upon identity perception. Secondly, the significance of a parameter does not appear to be necessarily bi-directional. That is that while encoding A's segmental timing upon the composite utterance significantly alters listener perception 'towards' A; encoding B's timing does not significantly alter the results. Thus, as others such as Van Lancker et. al. [LKE85, LKW85], and Takagi and Kuwabara [TK86] have shown, the significance of parameters in listener perception of identity is speaker dependent.

While the alteration of a single prosodic parameter alone may give some indications of the cues utilised by listeners there are many other 'competing' cues which the listener may use. Consideration of two parameters both encoded to match a particular speaker's may yield further data. Figure 9.5 and Table 9.2 represent the results for the two parameter combinations examined.

Contrasting the results with those for a single parameter it maybe seen that all paired combinations show as high as, and in all cases bar 1 higher, shifts in perception towards the

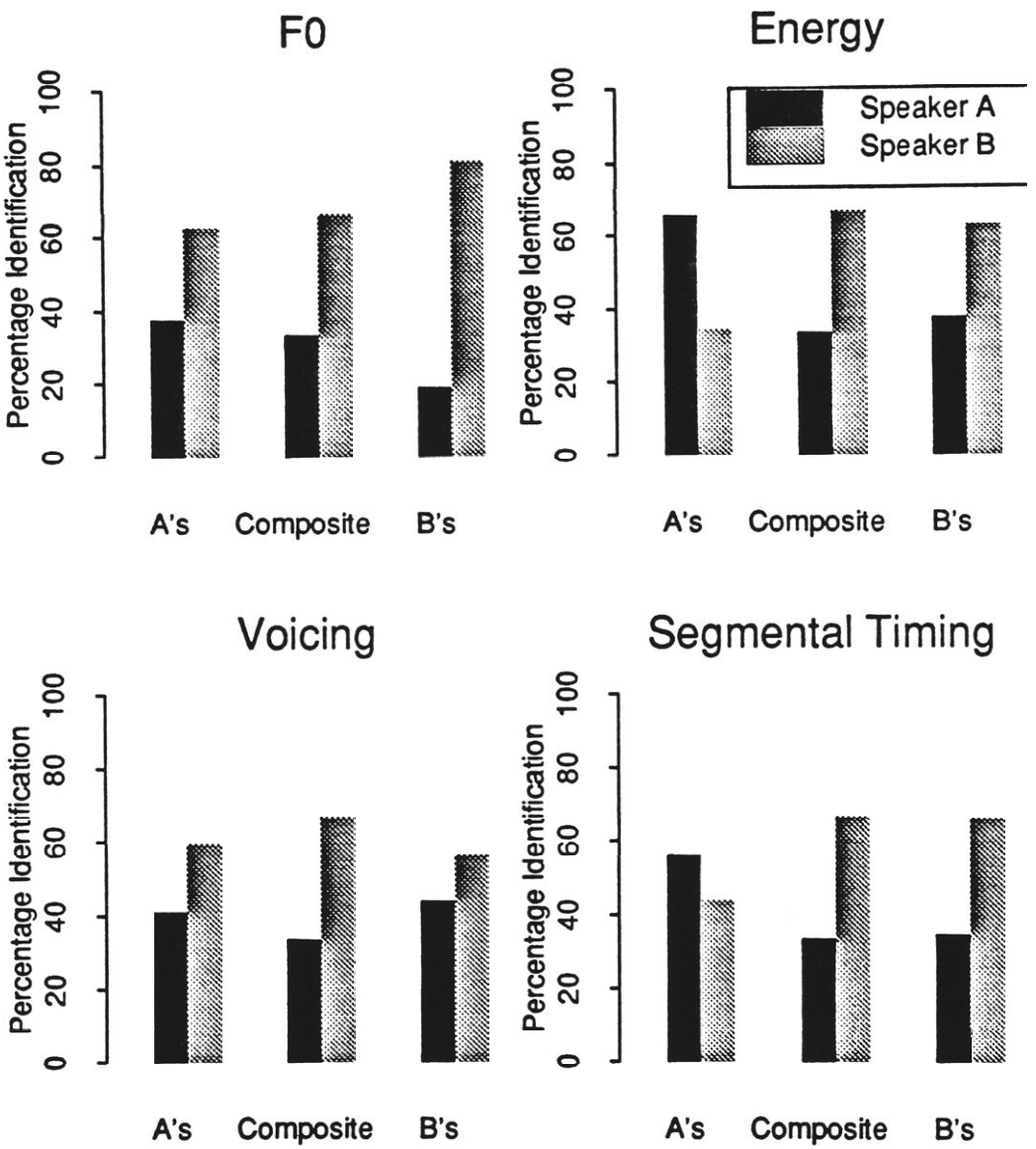


Figure 9.4: Listener perceptions of identity based on a single encoded parameter. Each plot present the results for a different encoded parameter. The left-most pair of bars on a plot show listener responses to an utterance containing speaker A's encoded parameter. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when speaker B is the originator of the encoded parameter.

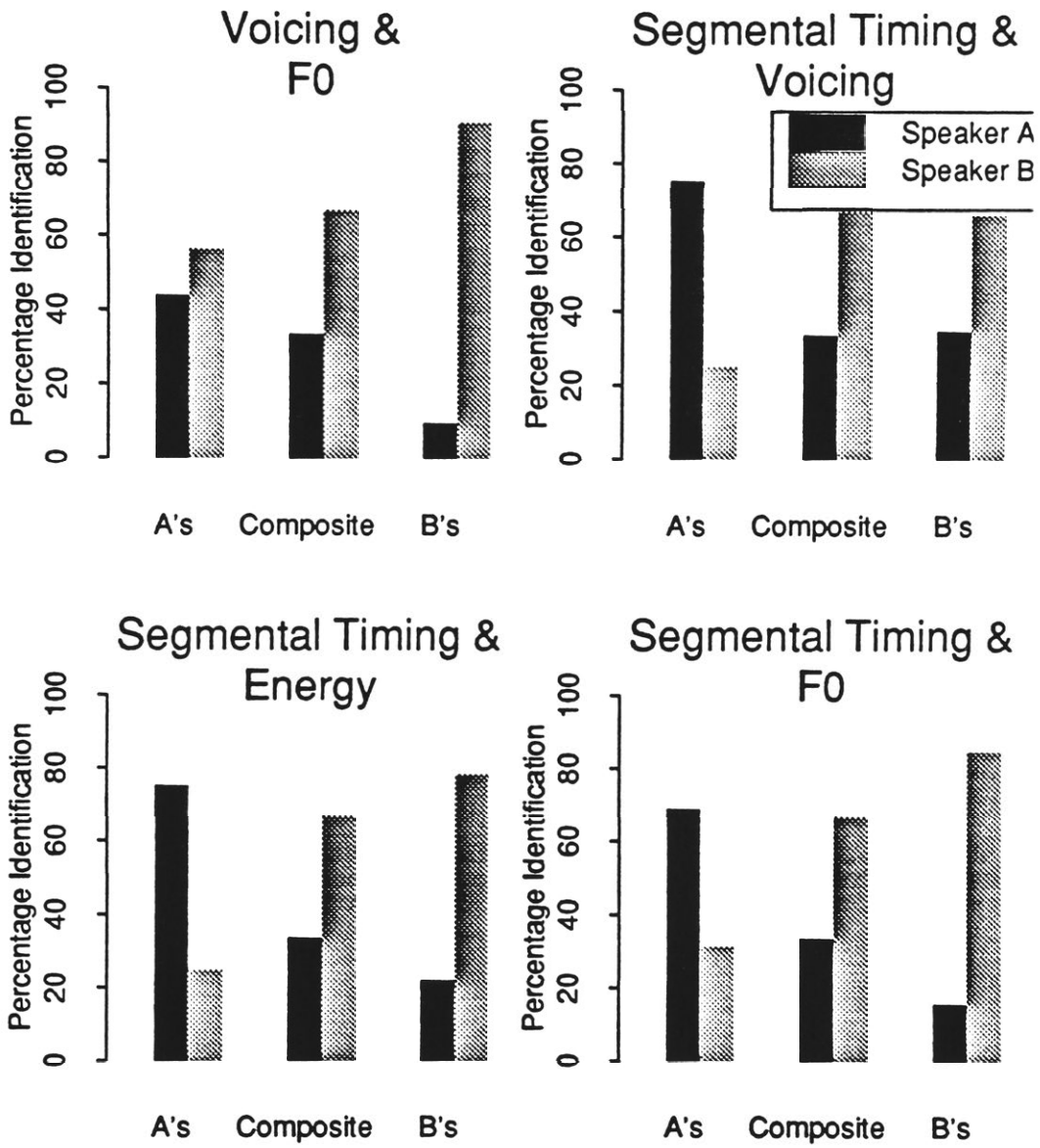


Figure 9.5: Listener perceptions of identity based on two encoded parameters. Each plot present the results for a different encoded pair of parameters. The left-most pair of bars on a plot show listener responses to an utterance containing speaker A's encoded parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when B is the originator of the encoded parameters.

"Encoded" Type		*Perception*	
		Shift to A	Shift to B
A	Voicing & F_0	11	
	Voicing & Timing	42	
	Energy & Timing	42	
	F_0 & Timing	36	
B	Voicing & F_0		24
	Voicing & Timing	1	
	Energy & Timing		11
	F_0 & Timing		17

Table 9.2: Listener perception of identity based on a two encoded parameters. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance. The left-most column indicates the speaker from whom the encoded parameters were derived.

originator of the parameter than either single parameter alone. All parameter combinations show a significance at the 5% level, though again; such a shift in perception is not bi-directionally significant.

Finally, Figure 9.6 and Table 9.3 show the shifts in listener perception when all four prosodic parameters of one or the other speaker is encoded onto the composite utterance.

"Encoded" Type	*Perception*	
	Shift to A	Shift to B
All speaker A	51	
All speaker B		21

Table 9.3: Listener perception of identity based on all four simultaneously encoded parameters. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance.

Clearly, when all four prosodics are combined the greater weight, over 80%, of responses as to speaker identity favour the originator of the prosodics; indicating their significance in listener perception of identity.

Based on these results it appears that all parameters examined make contributions to listener perception of identity. All parameters, however, are not equal in perceptual weighting, and their significance appears to be speaker dependent, voicing appearing to be the one consistently weak parameter.

9.1.2 Warped Parameter Substitution

This section presents the results of listener perceptions of substitution of warped parameters onto the composite utterance.

Figure 9.7 and Table 9.4 present the results when a single parameter; F_0, voicing or energy; from a speaker is warped to match the corresponding contour from the other speaker and encoded in the composite utterance.

While only marginal change is evident for F_0 and voicing, substitution of energy yields

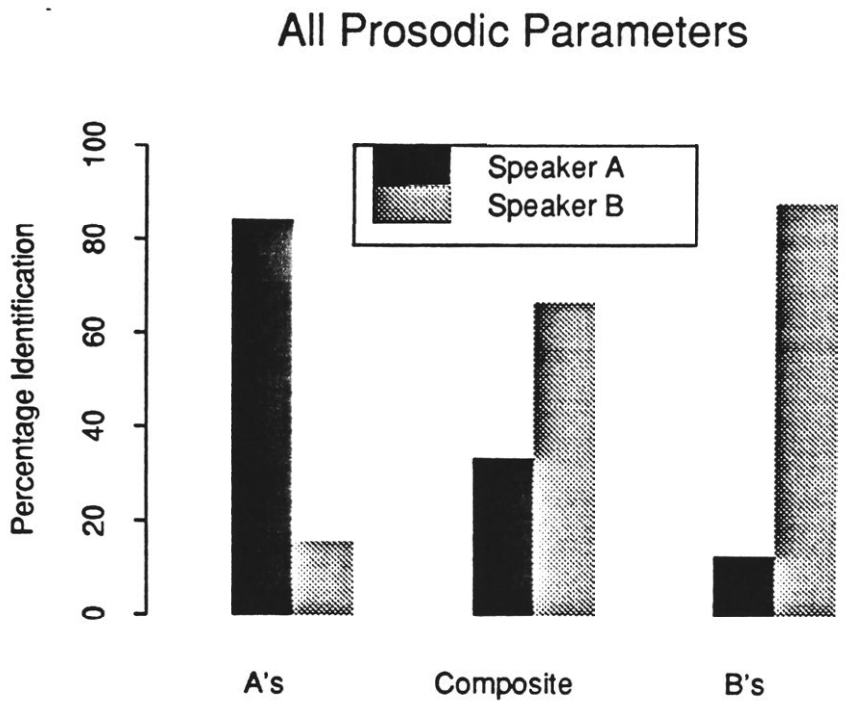


Figure 9.6: Listener perceptions of identity based on all four simultaneously encoded parameters. The left-most pair of bars on a plot show listener responses to an utterance containing all speaker A's encoded parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when B is the originator of the encoded parameters.

"Encoded" Type	*Perception*	
	Shift to A	Shift to B
A's F_0	12	
A's Energy	33	
A's Voicing	4	
B's F_0	4	
B's Energy	17	
B's Voicing	1	

Table 9.4: Listener perception of identity based on a single encoded *warped* parameter. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance.

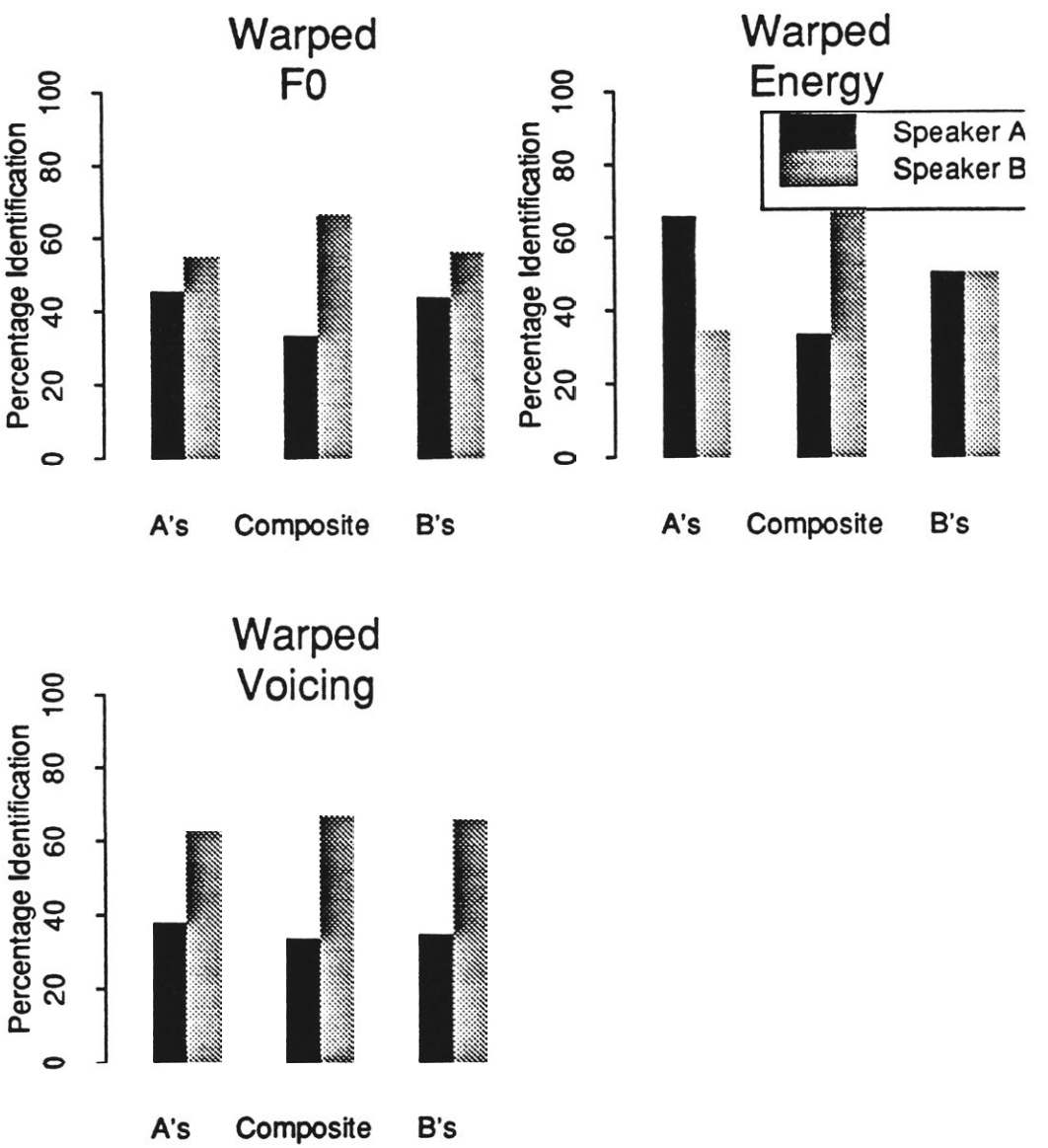


Figure 9.7: Listener perceptions of identity based on a single encoded *warped* parameter. Each plot present the results for a different encoded warped parameter. The left-most pair of bars on a plot show listener responses to an utterance containing speaker A's encoded parameter. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when B is the originator of the encoded parameter.

significant (*5% level*), and interesting, results. Whether speaker A or B's energy is the original, substituting either warped contour leads to an increase in identifications of A as the speaker. Such a result may indicate a uniqueness in A's energy contour that warping will not 'disguise', though B's may be warped to 'sound like' A. In general all listener perceptions shifted to some degree towards speaker A.

Figure 9.8 and Table 9.5 show the results when two warped parameters or one speaker's warped parameter and the other's segmental timing are combined together. Significant results, at the 5% level are obtained for both timing and energy, and timing and F_0 combinations. It appears that while parameters of speaker B's speech may be warped to such an extent that many listener responses indicate A as the speaker the reverse is not true of parameters of A's original utterance. The reason for this unidirectional shift in perception is unclear.

"Encoded" Type		*Perception*	
		Shift to A	Shift to B
A	Voicing & F_0	7	
	Voicing + P. Timing		5
	Energy + P. Timing	18	
	F_0 + P. Timing		11
B	Voicing & F_0		2
	Voicing + A. Timing	17	
	Energy + A. Timing	48	
	F_0 + A. Timing	22	

Table 9.5: Listener perception of identity based on two warped parameters or a warped parameter and the other speaker's segmental timing. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance. The left-most column indicates the speaker from whom the encoded warped parameter was derived.

Figure 9.9 and Table 9.6 show the result of listener responses to composite utterances with the three warped parameters of a speaker encoded along with the other speaker's timing.

"Encoded" Type	*Perception*	
	Shift to A	Shift to B
All A + B Timing	14	
All B + A Timing	48	

Table 9.6: Listener perception of identity based on the encoding of three warped parameters from one speaker and the other speaker's timing. Rates shown are positive shift in percent, towards one speaker or the other, from the response to the control-composite utterance.

Clearly it may be seen that using A's segmental timing and the warped F_0, energy and voicing contours of speaker B over 80% of responses indicate a listener perception of the speaker as A. Further, the shift to A is 25% higher than using A's timing alone, a result which shows the significance (*5% level*) of warping in altering listener perception of identity. However the 'transformation' does not apply in the other direction. Taking speaker B's timing and A's three warped parameters, listener response actually shifts further towards A, rather than towards B.

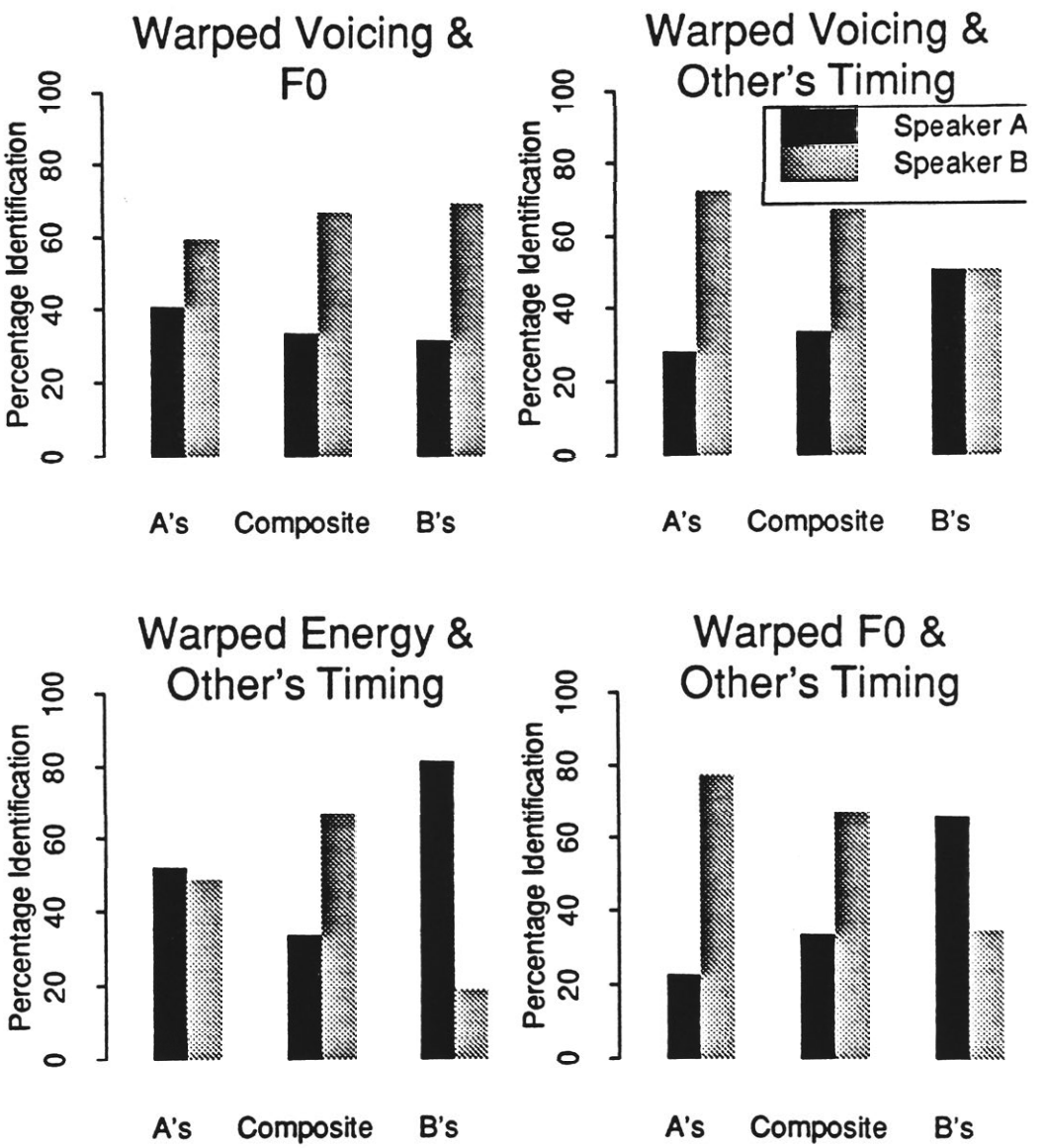


Figure 9.8: Listener perceptions of identity based on two warped parameters or a warped parameter and the other speaker's segmental timing. Each plot present the results for a different pair of parameters. The left-most pair of bars on a plot show listener responses to an utterance containing speaker A's encoded warped parameter. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when B is the originator of the encoded warped parameter.

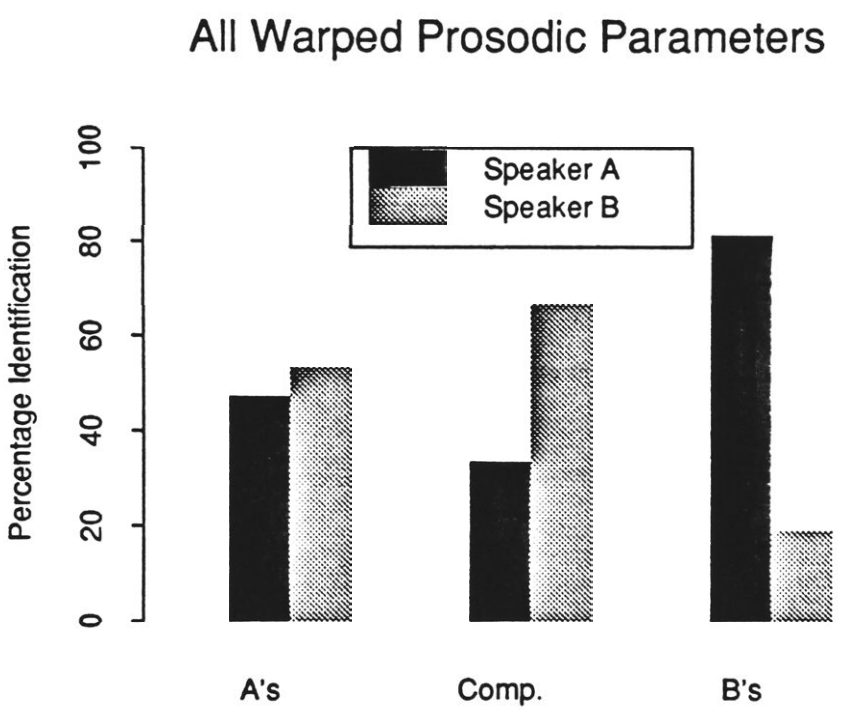


Figure 9.9: Listener perceptions of identity based on the encoding of three warped parameters from one speaker and the other speaker's timing. The left-most pair of bars on a plot show listener responses to an utterance containing all speaker A's encoded warped parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show listener response when B is the originator of the encoded warped parameters.

9.2 Speaker Sex

For the speaker sex perception experiments four speakers, two male and two female, were selected to comprise the speech material. The sentence: "I cannot remember it." was again selected; and the four speakers; 3, 4, 9, and 14; all had dialect scores of 7 (this condition of having equal dialect scores limited the number of speakers that could be used to compose the composite utterance). The mean F_0 of the two male speakers, designated male 1 and male 2, were 116Hz and 133Hz respectively; while that of the two female speakers, designated female 1 and 2, were 209Hz and 247Hz.

A single composite or "androgynous" utterance was composed of equal weightings, 25%, of the utterances of the four speakers; which were segmentally aligned via the process of Section 8.6.3. Due to the well known significance of F_0 in the perception of speaker sex [Col71, LHB+76] an initial experiment was carried out with a small number of listeners who did not participate in the primary experiment. Using the androgynous utterance a flat F_0 varying from 130 to 200Hz was synthesised. It was found that the point of equal distribution, between male and female, of listener responses, was at 165Hz. Based on such a result each speaker's F_0 was also shifted to a mean of 165Hz via the method of Section 8.6.4.

For each of the four speakers the three prosodic parameters F_0, energy and voicing were encoded alone and in combination upon the composite utterance and listener responses were gauged. Further the 'shifted' F_0 parameters were encoded alone and in combination with the other two parameters from the same speaker. The results were then grouped on the basis of the sex of the speaker.

Presentation of results follows the format of those for the speaker identity experiments, and a chi-squared based test of proportion differences [FW80] is used to determine the significance of the results.

9.2.1 Original Parameters

This section will present the results for the original three parameters as taken from each of the four speakers. Figure 9.10 and Table 9.7 show the results for the encoding of a single parameter upon the composite utterance.

"Encoded"	*Perception*	
	Shift to Male	Shift to Female
Male F_0	24	
Male Energy		3
Male Voicing		7
Female F_0		25
Female Energy	4	
Female Voicing	7	

Table 9.7: Listener perception of sex based on a single encoded parameter. Rates shown are positive shift in percent, towards one sex or the other, from the response to the control-composite utterance. The left-most column indicates the parameter and its source.

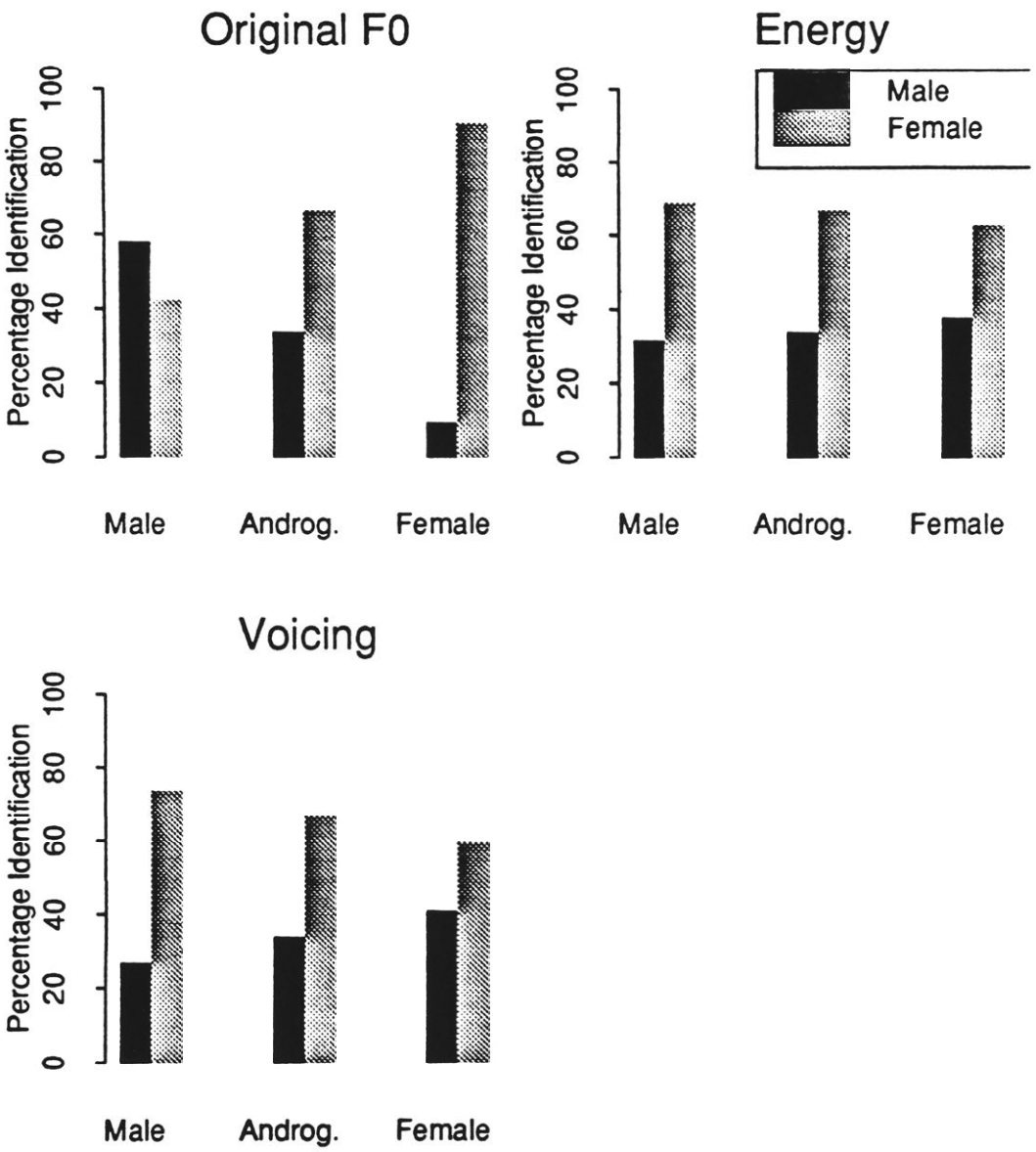


Figure 9.10: Listener perceptions of sex based on a single encoded parameter. Each plot present the results for a different encoded parameter. The left-most pair of bars on a plot show mean listener response to utterances containing male derived parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show mean listener response to female derived parameters.

Clearly, the addition of F_0 to the composite significantly (*5% level*) influences listener perception such that in all cases the majority of listener decisions favour the sex of the originating F_0. The other two parameters, energy and voicing, appear to have little impact upon listener perception of sex. For these parameters the listener response varies little from that of the purely 'average' or androgynous utterance in which case listener response was of the ratio of 2 to 1 in favour of a female speaker.

Figure 9.11 and Table 9.8 indicate listener responses when all three parameters; F_0, voicing, and energy; from a particular speaker are encoded upon the composite utterance.

"Encoded"	*Perception*	
	Shift to Male	Shift to Female
All Male	10	
All Female		22

Table 9.8: Listener perception of sex based on the encoding all all three parameters from each of the four speakers utilised in the experiment and grouped on the basis of sex. Rates shown are positive shift in percent, towards one sex or the other, from the response to the control-composite utterance.

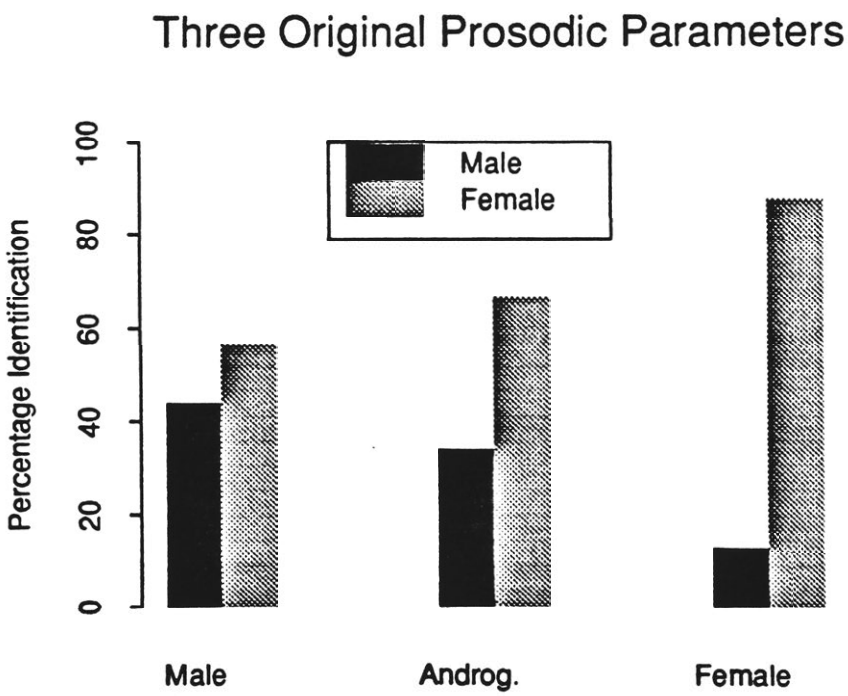


Figure 9.11: Listener perceptions of sex based on the encoding of all three parameters from each of the four speakers utilised in the experiment and grouped on the basis of sex. The left-most pair of bars on a plot show mean listener response to utterances containing male derived parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show mean listener response to female derived parameters.

It may be seen that listener responses vary from that of the androgynous utterance in relation

to the sex of the originator of the encoded contours. Utterances based on female prosodics appear clearly discernible as females to the listeners. Male derived utterances, however, are still perceived as being more female than male; even though there has been a minor increase in male responses, from the control androgynous utterance. The result appears further perplexing when contrasted with that of F_0 alone encoded, Figure 9.10, where clearly that is sufficient to alter the majority of sex perceptions to the sex of the originator.

9.2.2 Linear Shifted F_0

This section will present the results of the listener sex perceptions for the 'shifted' F_0 contours alone and in combination with the other two parameters from each speaker. Three methods were used to shift the mean F_0 of all speakers to 165Hz. Additive - adding a constant ; multiplicative - multiplying by a constant; and combined additive and multiplicative - additive. to $\frac{mean+165}{2}$ then multiplicative.

Figure 9.12 and Table 9.9 show the listener perception responses for the various transformation of F_0.

"Encoded"	*Perception*	
	Shift to Male	Shift to Female
Original Male F_0	24	
Additive Male F_0		3
Mult. Male F_0		6
Add.+Mult. Male F_0		14
Original Female F_0		25
Additive Female F_0		14
Mult. Female F_0		11
Add.+Mult. Female F_0		19

Table 9.9: Listener perception of sex based on the encoding of F_0, and shifted (to a mean of 165Hz) F_0, from two male and two female speakers. Rates shown are positive shift in percent, towards one sex or the other, from the response to the control-composite utterance. The left-most column indicates the parameter and its source.

It is readily apparent that irrespective of the transformation or sex of the originator of the contour that listener response varies little from the 70% to 80% weighting towards female. Such a result indicates the significance of *mean* F_0 in speaker sex perception. Further, as Spencer [Spe88] showed there appears a strong discontinuity in listener perception of sex based on mean F_0, such that utterances below the 'transition' were perceived as male, and those above it female. Therefore, it would appear that in this case for at least 70% of the listener group their 'transition' point for sex perception based on mean F_0 was at a value below 165Hz.

Figure 9.13 and Table 9.10 are presentations of the listener perceptions when the transformed F_0 are combined with the energy and voicing of their original speaker.

Contrasting the results with those of the F_0 alone there appears an extremely high correlation. Regardless of transformation or sex of originator; listener response varies from 60% to 80% towards a female speaker. Therefore the addition of energy and voicing contours has altered

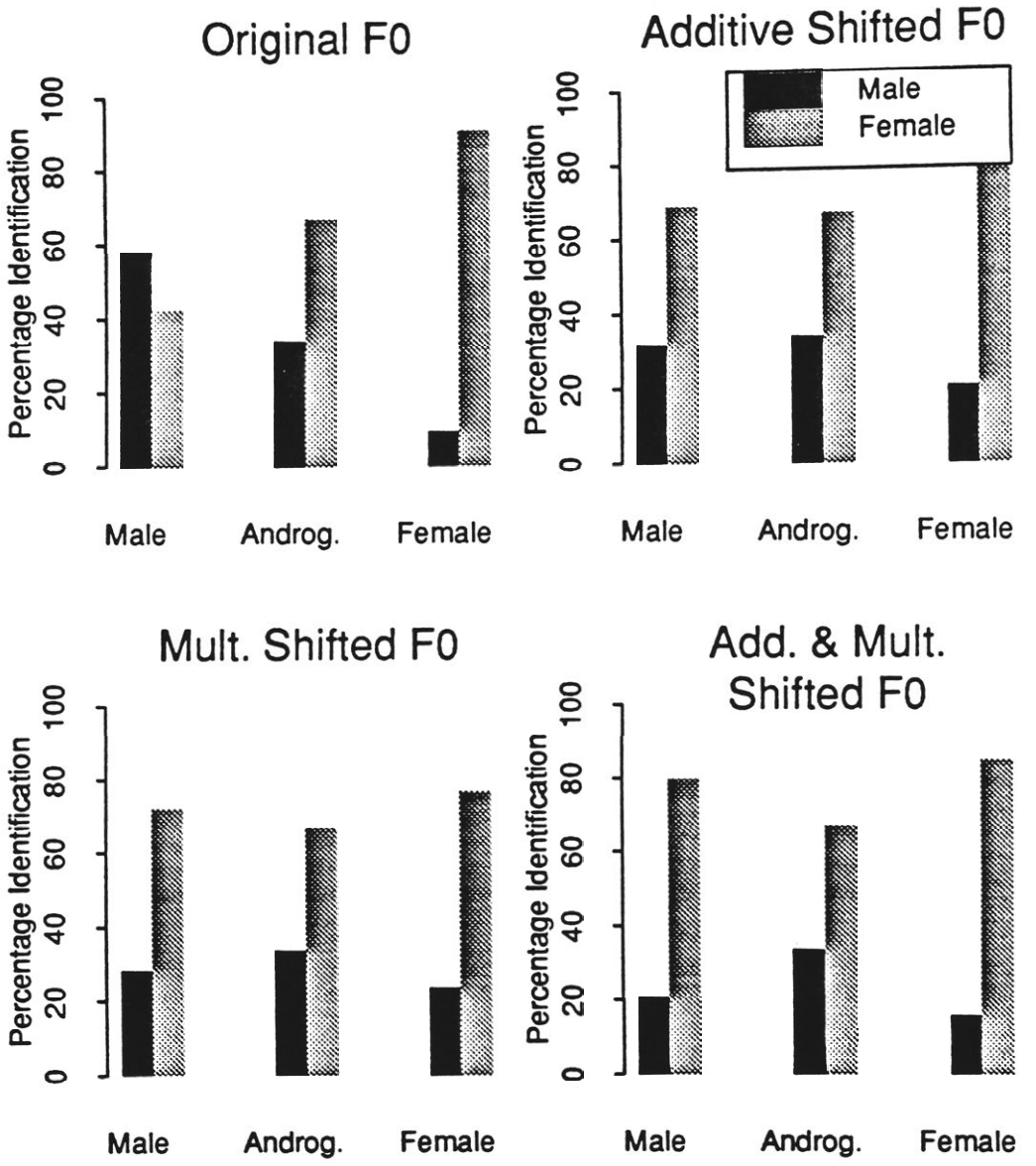


Figure 9.12: Listener perceptions of sex based on the encoding of F_0, and shifted (to a mean of 165Hz) F_0, from two male and two female speakers. Each plot present the results for a different encoded representation of F_0. The left-most pair of bars on a plot show mean listener response to utterances containing male derived parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show mean listener response to female derived parameters.

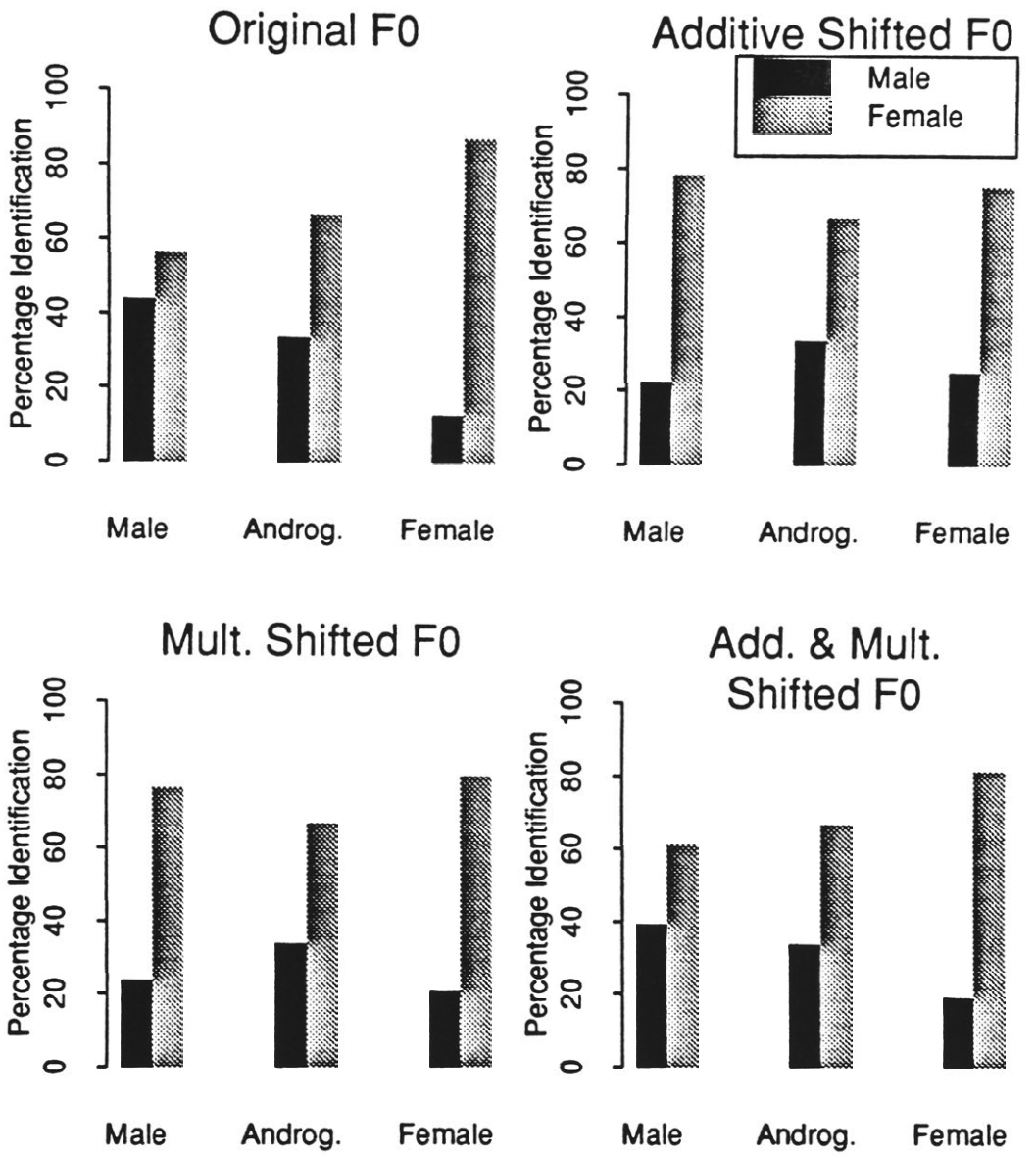


Figure 9.13: Listener perceptions of sex based on the encoding of F_0, or shifted (to a mean of 165Hz) F_0, with energy and voicing, from two male and two female speakers. Each plot present the results for a different encoded representation of F_0. The left-most pair of bars on a plot show mean listener response to utterances containing male derived parameters. The central pair of bars show listener response to the control composite utterance; while the right-most pair show mean listener response to female derived parameters.

"Encoded"	*Perception*	
	Shift to Male	Shift to Female
Original Male F_0	10	
Additive Male F_0		12
Mult. Male F_0		11
Add.+Mult. Male F_0	5	
Original Female F_0		22
Additive Female F_0		9
Mult. Female F_0		14
Add.+Mult. Female F_0		16

Table 9.10: Listener perception of sex based on the encoding of F_0, and shifted (to a mean of 165Hz) F_0, with energy and voicing; from two male and two female speakers. Rates shown are positive shift in percent, towards one sex or the other, from the response to the control-composite utterance. The left-most column indicates the F_0 representation and its source.

listener perception of sex very little, indicating the subordinate role of energy and voicing to F_0 in listener judgements of speaker sex.

9.3 Speaker Dialect

For the speaker dialect experiments utterances of four speakers, all male, were selected for the speech material. Two speakers, 10 and 13 with dialect scores of 1.5 and 2.5, were selected from the cultivated end of the dialect spectrum; while two others, 6 and 16 with dialect scores of 9 and 8.5, were selected from the broad end of the dialect spectrum. As for previous perception experiments the sentence: "I cannot remember it." was selected as the base utterance and a composite or "general" utterance was composed of equal elements of a single utterance from each of the four speakers. The four component utterances were segmentally aligned prior to the composition of the general utterance.

For each of the four speakers the three prosodic parameters F_0, energy, and voicing were encoded alone and in combination upon the composite utterance and listener responses were gauged. Further, the general utterance was time altered, duration increased and decreased, by values of 25% and 50% of its original value and played to listeners.

Listener response to the verbal stimulus was to mark a point on a line indicating position of speaker/utterance on the dialect spectrum. The marked point was then converted into a distance from the base-point for a cultivated utterance and divided by the length of the original line to yield a value between 0 and 1. Results in the following sections are shown by plotting the distribution of listener response values for various logical groupings of encodings upon the general utterance; and tables showing the mean and standard deviation of the distributions. The statistical significance of the results are evaluated by performing an F-test [HW71] to compare variances, and a t-test [HW71] to compare means. In all comparisons the was found to be no significant, at the 5% level, difference in the variance of the distributions under comparison.

9.3.1 Listener Response Consistency

This section explores two questions. Firstly, are untrained listeners capable of judging dialect with any degree of accuracy or consistency. Secondly, did the listeners adopt an absolute scale for dialect or one relative to the presented material.

Brennan et. al. [BRD75] showed that naive listeners appeared consistent and accurate in the judgement of the accentedness of Spanish-English bilingual speakers. To test the consistency of the current listener set the distribution of listener responses to the general utterance was plotted.

Further, to test the question of 'relative' or 'absolute' responses, the responses of each listener were linearly transformed, or normalised, such that their 'most cultivated' response was scored as 0; while the 'broadest' response was scored as 1; i.e. ensuring the entire range of values is used for each listener. The distributions of normalised and non-normalised listener responses to the general utterance could then be compared and contrasted. Figure 9.14 and Table 9.11 present these results.

Response Type	Mean	Standard Deviation
Un-normalised	.50	.19
Normalised	.52	.23

Table 9.11: Mean and standard deviation of listener dialect responses to the general utterance when responses are taken 'as is', or when normalised for each listener. Low response scores represent the perception of cultivated dialect by the listeners.

Examining the figure it may be seen that the greater portion of listener responses to the general utterance are clustered about the centre of the response score spectrum; indicating a degree of consistency and accuracy between listeners. Therefore, though the utterance was synthetic in nature and the hence the results must be regarding with some caution, it appears that within the context of this experiment naive Australian born listeners are consistent and reasonably accurate judges of Australian dialect. A far more thorough test with natural utterances is required, however, to ascertain the general validity of this claim.

Comparison of the two distributions show that there is no statistical significance at the 5% level between the variances of the two distributions, given the current number of responses. However the un-normalised responses do have a marked lower variance, and hence appear a preferable means of representing listener responses. Therefore, for all subsequent results, the non-normalised listener response will be used.

9.3.2 Parameter Encoding

This section presents the results of listener responses to the encoding of the three parameters F_0, voicing and energy, alone and in combination; from each of the four speakers. Results are grouped on the basis of the dialect of the originator of the encoded parameter(s). Figure 9.15 and Table 9.12 present analysis of the listener responses to the encoding of a single parameter.

Statistical analysis of the results represented by the table and figure show that encoding of energy is the only parameter of the three that shows a significant difference (*5% level*)

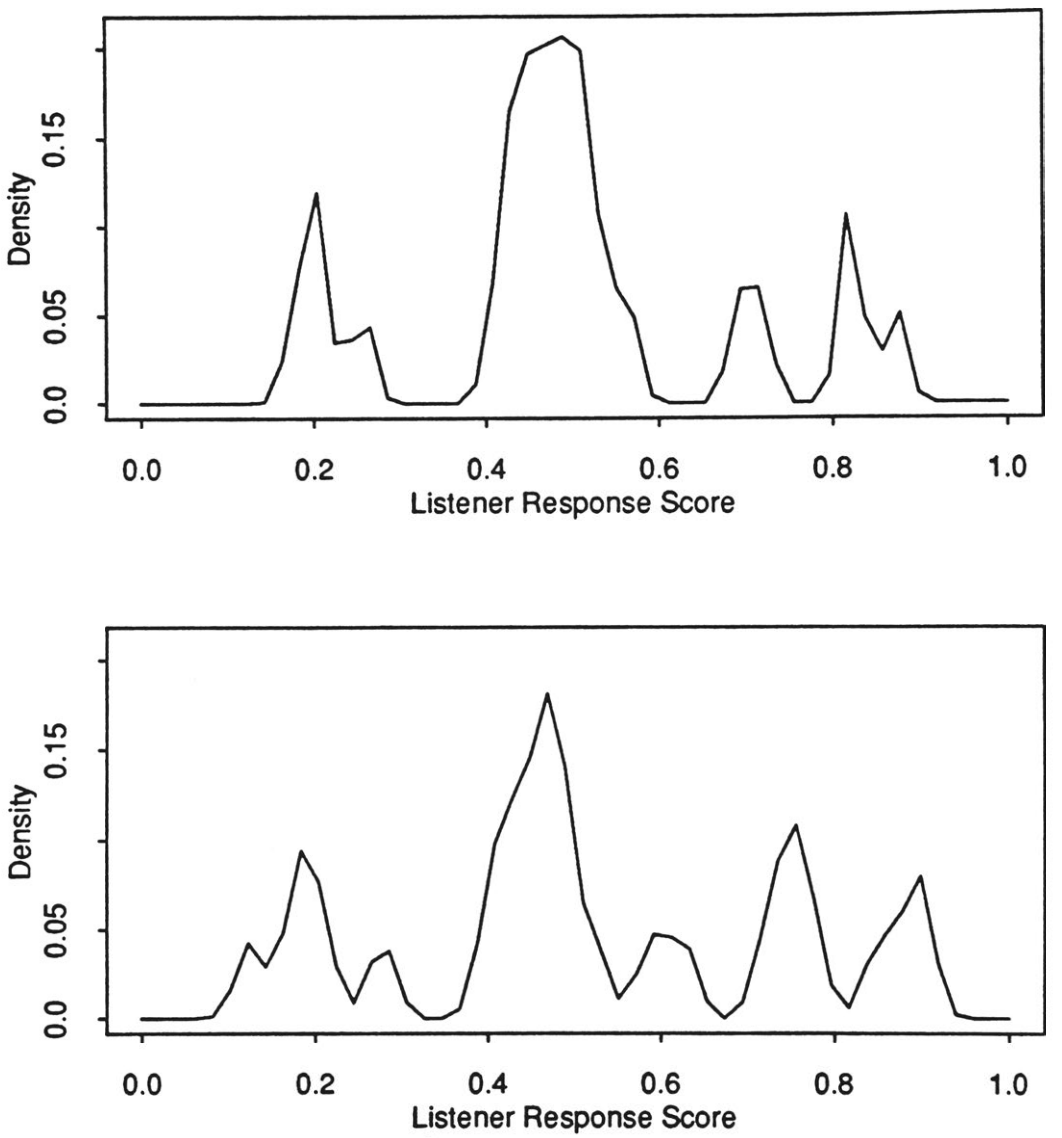


Figure 9.14: Distribution of listener dialect responses to the general utterance. Top plot is listener responses taken 'as is'; while lower plot is responses normalised for each listener. Low response scores represent the perception of cultivated dialect by the listeners.

Parameter	Cultivated Source		Broad Source	
	Mean	Std. Dev.	Mean	Std. Dev.
Energy	.50	.19	.41	.17
Voicing	.47	.17	.49	.19
F_0	.42	.22	.41	.18

Table 9.12: Mean and standard deviation of listener dialect responses to the encoding of a single parameter, from the two cultivated and two broad dialect speakers, upon the composite utterance. Low response scores represent the perception of cultivated dialect by listeners.

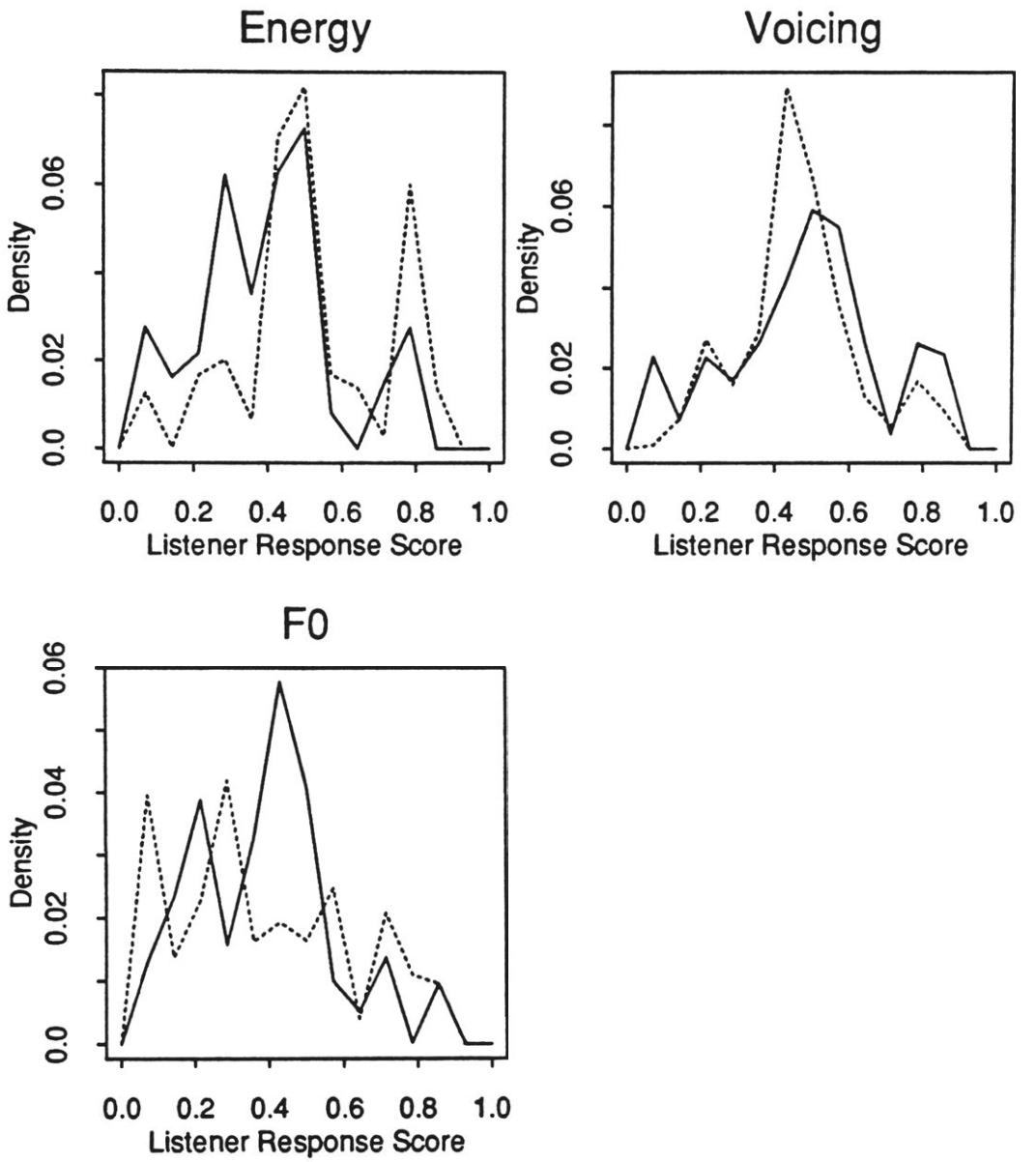


Figure 9.15: Distribution of listener dialect responses to encodings of a single parameter upon the general utterance. Each plot represents a different parameter. The solid line represents utterances from broad dialect speakers, the broken line from cultivated dialect. Low response scores represent the perception of cultivated dialect by listeners.

between broad and cultivated sources. However the shift in dialect perception is opposite to that which might reasonably be expected. Cultivated source derived utterances have a mean around the centre of the dialect spectrum while broad source derived utterances have a mean shifted significantly towards the cultivated end of the spectrum. Examination of the four speaker's energy contours showed that listener perception appeared correlated with maximum intensity—high maximum energy perceived as broader, while lower maximum energy perceived as more cultivated. However other dynamic properties of the contours may be attributable as the cause of these results.

Clearly no single parameter, encoded upon the general utterance, gives clear and unambiguous perceptual indications of the dialect of the originator.

Figure 9.16 and Table 9.13 show listener perception of dialect when all three prosodics from each speaker in turn are encoded.

Source	Mean	Standard Deviation
Cultivated	.36	.18
Broad	.43	.20

Table 9.13: Mean and standard deviation of listener dialect responses to composite utterances with energy, voicing, and F_0 encoded from a single speaker, and grouped on the basis of the dialect of the speaker. Low response scores represent the perception of cultivated dialect by the listeners.

While the table shows a minor difference in means there is no significant difference, at the 5% level, between the two distributions. Further listener judgements of the two utterances are required to thoroughly examine this difference. Hence it may be concluded that under the conditions of the current experiment the prosodic parameters were insufficient cues for the listeners to discern the dialect of the originating speaker.

9.3.3 Time Alteration

This section will present an analysis of the listener perceptions of the general utterance when it is piecewise linearly (segment by segment) increased and decreased in duration by 25% and 50% of its original value. Figure 9.17 and Table 9.14 present the results.

Percentage Alteration	Duration Decreased		Duration Increased	
	Mean	Std. Dev.	Mean	Std. Dev.
25%	.40	.21	.61	.25
50%	.42	.23	.67	.28

Table 9.14: Mean and standard deviation of listener dialect responses to time alterations to the general utterance. The general utterance was duration increased and decreased by values of 25% and 50%. Low response scores represent the perception of cultivated dialect by listeners.

Examination of the results shows that there is a discernible and significant (*5% level*) difference in the distributions for duration increased against duration decreased utterances, whether by 25% or 50%. It may be seen that in general listeners perceive a shorter utterance as originating from a more cultivated dialect speaker; while a longer utterance is perceived as

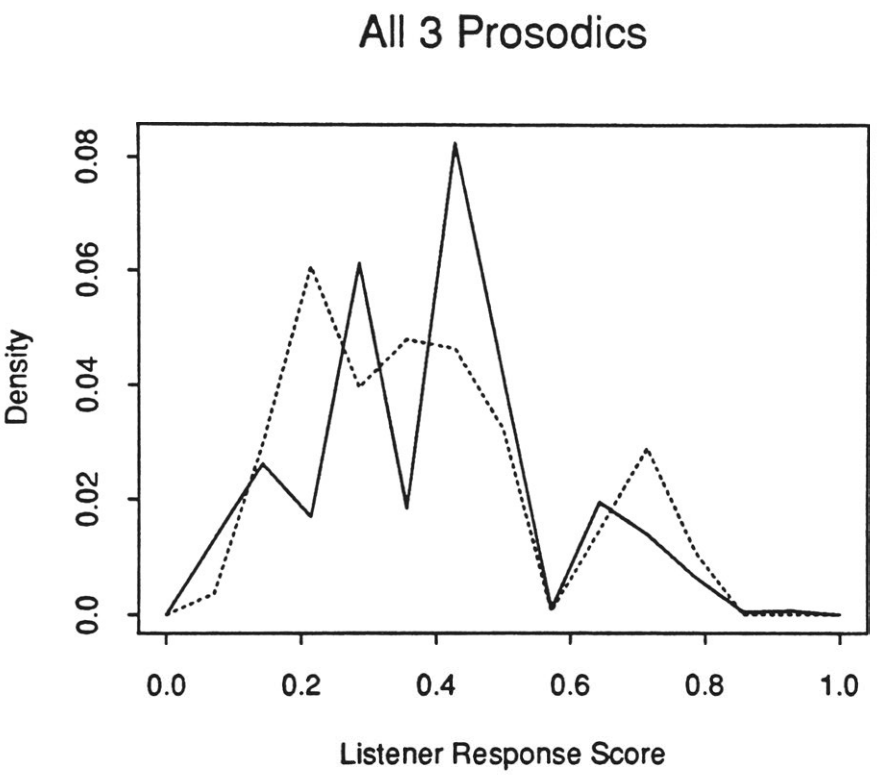


Figure 9.16: Distribution of listener dialect responses to composite utterances with energy, voicing, and F_0 encoded from a single speaker, and grouped on the basis of the dialect of the speaker. Low response scores represent the perception of cultivated dialect by the listeners.

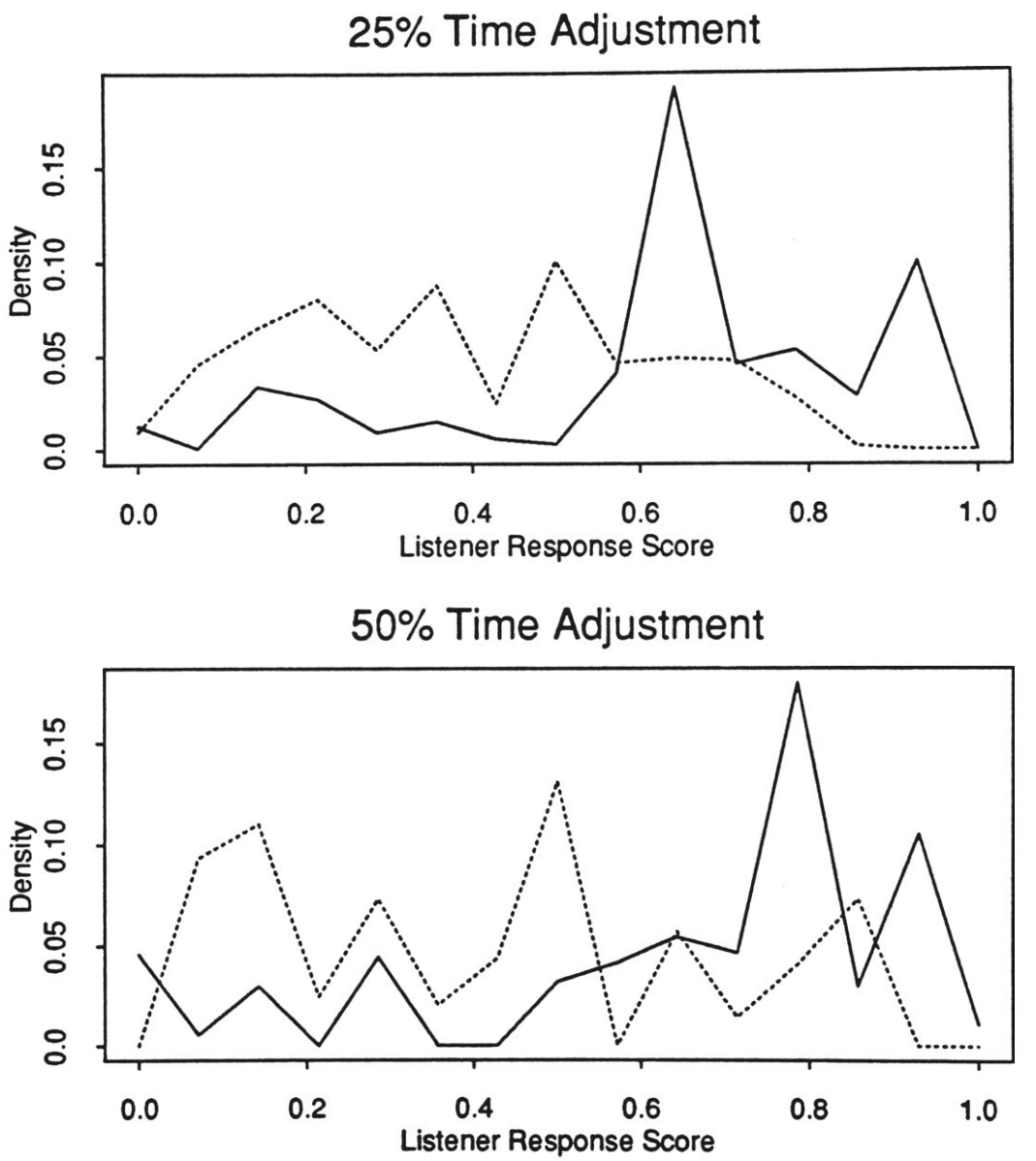


Figure 9.17: Distribution of listener dialect responses to linear adjustments, of 25% and 50%, to the duration of the general utterance. Responses to lengthened utterances are represented by the solid line; while shortened utterances are represented by the broken line. Low response scores represent the perception of cultivated dialect by listeners.

originating from a speaker of broader dialect. Shortening the utterance from 75% to 50% of its original length did not appreciably alter listeners' perception of dialect, whereas increasing duration from 125% to 150% of original did make the utterance 'broader' to listeners, though not significant at the 5% level.

Clearly, duration is a factor in naive listener perception of dialect. This result is further supported by Lass et. al. [LMK78], who showed that rate alteration adversely affected correct listener identification of speaker 'race', indicating listener utilisation of cues based on speaking rate.

Chapter 10

Discussion – Perception Experiments

The previous two chapters introduced and applied the novel 'composite model' approach of analysis-synthesis to the investigation of the perception of speaker characteristics. In particular three speaker characteristics: identity sex and dialect; and the prosodic parameters: F_0, voicing, energy, and segmental timing were examined.

The speaker identity experiments were conducted using two speakers to build a composite utterance upon which each speaker's prosodics were encoded. Results for the encoding of a single parameter showed that all parameters are not equivalent in their degree of discrimination between the two speakers, at least from the listener perspective. Further the importance of a parameter varied according to the speaker. That is that while energy and segmental timing were significant in shifting perception towards speaker A when encoded from speaker A; the opposite was not true for the same parameters from speaker B; for whom F_0 was the most significant. Such a result indicates that cue utilisation in the perception of identity by listeners is speaker dependent:- the significance of an acoustic parameter in determining a speaker's identity is a function of the speaker.

When two parameters from a single speaker were both encoded upon the composite utterance the shifts in listener perception became greater in magnitude than for a single parameter alone; and as previously of varying degrees dependent upon the parameter pair and the speaker they were encoded from. When all four prosodics from either speaker were encoded upon the composite utterance there was strong agreement, $> 80\%$, across all listeners that the speaker was the originator of the four prosodics.

Such a result indicates that the four parameters encoded upon a base composite utterance are sufficient to allow listeners to discriminate between two speakers with a high degree of accuracy. The three parameters: segmental timing, energy, and F_0 all appeared to carry significant speaker specific information; while voicing appeared to make no significant contribution.

In addition to the substitution of unaltered parameters, parameters time warped so as to

closely match that of the other speaker, were encoded upon the composite utterance. This approach intuitively sought to measure whether it was the 'shape' of a contour that listeners used; and was motivated by the results of Chapter 5 where it was found there was significant speaker specific information in the relative dynamics of a contour. As for the unaltered parameters; results for the warped parameters were not symmetrical between speakers. For a single warped parameter, no matter the parameter nor speaker from whom it was derived listener perception shifted towards speaker A. In particular energy showed the strongest shift. Examination of both speakers warped and unwarped energy contours show that speaker A's contour is particularly flat and monotone as compared to speaker B's.

Results are altered somewhat when two parameters are encoded. Generally the shift in perception is stronger, though now not all shifts are towards speaker A. When all warped parameters from speaker B are encoded with A's timing there is a high degree of agreement amongst the listeners, $> 80\%$, that the speaker is A. However using A's parameters and B's timing listener perception shifts towards A. These results indicate that warping may alter listener perception of speaker identity and hence the significance of the contour dynamics or 'shape' as opposed to its mean. However, warping is not a guarantee of switching in listener perception; in this case the uniqueness of speaker A's energy contour and its difference from B's meant that it could not be warped to 'sound like' speaker B.

One fact worthy of note is the listener perception of the 'mean' or composite utterance. If individual speaker's acoustic parameters used to comprise the composite were equally 'strong' or 'unique' it could be expected that listener perception of identity would be divided approximately equally between A and B. However this is not the case with listener perception favouring speaker B at a 2:1 ratio. It may be that an 'extreme' or 'unique' parameter when averaged with a less 'extreme' parameter is still sufficiently strong to alter listener perception. More work is required to examine this phenomenon.

For the speaker sex experiments utterances from four speakers; two female, two male; were combined to form a composite 'neuter' utterance. As for the previous identity experiment listener response to the composite utterance strongly favoured one source over the other; in this case female by 2:1.

Encoding of a single parameter upon the composite showed, as expected, that F_0 strongly influenced listener perception, towards the sex of the originator. The other two parameter; energy and voicing had no significant effect upon listener perception. When all three parameters were encoded together it was again found that listener perception shifted towards the sex of the originator. Interestingly, for male speakers the shift was significantly less than for that achieved by F_0 alone; and in fact absolute perception of sex favours female by approximately 10%. It is unclear as to why this 'masking' of the effects of F_0 should occur when energy and voicing are added.

When F_0 contours were linearly altered, by multiplication and/or addition to a mean of 165Hz and encoded alone or in combination with energy and voicing it was found that listener response remained fundamentally invariant, regardless of the source of the parameters, at 70% to 80% in favour of female. This result combined with the above results for the unaltered F_0

show the importance of mean F_0, as opposed to its dynamics, in listener perception of sex.

Speaker dialect perception experiments were carried out by composing a composite or 'general' utterance from the utterances of four male speakers, two from either end of the dialect scale. Listener response to this general utterance showed a general Gaussian distribution, with 'sidelobes', about the centre of the spectrum, indicating a degree of consistency and agreement between the listeners. When individual listener responses were normalised to cover the entire spectrum it was found that response to the general utterance was more diffuse, and hence less consistent between speakers.

Encoding of a single parameter from each of the four speakers showed inconsistent and unreliable results. Little difference between broad and cultivated speaker distributions exists for the parameters F_0 and voicing; while for energy the shift in perception was opposite to that of the originator. Examination of the energy contours shows that listener perception appears correlated to maximum intensity: higher intensity being perceived as broader, lower intensity as more cultivated. However for both energy and F_0 listener responses were widely distributed across the dialect spectrum implying a lack of common agreement amongst listeners and thus casting doubt upon the veracity of the results.

When all three parameters from each speaker are encoded together there appears a minor difference in the distribution of listener responses: cultivated source composed utterances being perceived as more cultivated than broad source. However the difference is minimal and both distributions are shifted to the cultivated from that of the general utterances. This 'undecided' result is in no way surprising for, as was shown in the analysis investigation of dialect, while there appear general population trends in dialect as related to prosodics wide variations from this 'standard' may exist on an individual speaker basis. Clearly more data, both in terms of speaker and listener numbers are needed to thoroughly explore this topic.

The general utterance was also linearly (segment by segment) increased and decreased in duration by 25% and 50% of its original. Listener response was clearly correlated with duration such that shorter utterances were perceived as cultivated and longer as broader. A decrease in duration to 50% of original was not perceived significantly differently to the 75% duration utterance while the 150% duration utterance was significantly broader than the 125% duration utterance. The question of whether the perceived broadness of dialect for longer utterances versus cultivation for shorter, is a fact garnered from experience by listeners, or simply a stereotype of the slow drawling country speaker versus the faster city resident is unclear. However it may be seen that this is a *perceived* characteristic of dialect for the general listener.

Unfortunately due to limits of time and speaker data it was impossible to obtain sufficient data to construct 'true' characteristic archetypes, nor to investigate multiple speaker pairings for speaker identity. However the above results appear to show the viability of the 'composite model' analysis-resynthesis scheme for the examination of speaker characteristics. Clearly other speaker characteristics; such as emotions, age or health; and other acoustic parameters including spectral characteristics could be investigated via this method. There is obviously a need for more speech data in order to utilise this scheme thoroughly.

In informal discussion following experimentation many listeners made similar comments

which may shed some further light upon the results. Typical paraphrased comments include: "... part way through I found I changed what I was listening for..." [to determine the characteristic], and, "It was really hard to decide...". Apparently, a number of listeners altered the cues they utilised to determine identity, sex, or dialect, somewhere during the conduct of the experiment. This phenomenon of cue swapping requires further investigation. Secondly many decisions were not simple and this is likely to be attributable to the conflicting acoustic cues giving different messages as to speaker characteristic.

Finally, no analysis of results on the basis of characteristics of the listener was performed, due both to time constraints and the limited size of the listener set. However an informal examination of listener responses showed that there were marked differences between listeners as to their perception of the three speaker characteristics. Therefore the effect of listener upon results should not be discounted and it would be desirable with a larger listener set to examine results on characteristics of the listener set such as age and sex.

The previous results barely 'scratch the surface' of the application of the 'composite model' method and leave as many new unanswered questions as those to which answers were sought initially. Obviously much work still needs to be done to explore the potential and limitations of this general resynthesis approach to the perception of speaker characteristics.

Chapter 11

Conclusion

A database of sentence-long utterances from nineteen adult speakers of Australian English was collected. Four prosodic parameters— energy, fundamental frequency, voicing, and zero crossing rate were extracted and analytical and perceptually based investigations of the parameters' correlations to the speaker characteristics identity, sex, and dialect were carried out. A number of previously known and new results regarding the relationship of the parameters to the speaker characteristics were discovered.

Analytical experiments were conducted using four of the fifteen different sentences recorded by speakers. Discriminant analysis was applied to examinations of the characteristics identity and sex, and least-squares-fit analysis for speaker dialect. Twenty-one measures of the properties of each of the parameters were examined, the measures being logically split into two groups:- dynamic—measures of the time varying properties of the parameter contours, and static— measures of the time invariant properties of the parameters.

It was found that identity, sex, and dialect could be detected to significant degrees based on the parameters and sentences used:- identity and sex discrimination at 75% and 96% respectively, and dialect correlated at 0.58. It is clear therefore that identity, sex, and dialect information is encoded in the prosodics examined.

Comparison of dynamic measure based, and static measure based discrimination and correlation rates showed that for all three speaker characteristics the dynamic measure set performance was equal to or superior to that of the static, though combined performance exceeded that of either alone. Clearly the dynamic measures extract more speaker-related information than the static measures, though dynamic measures do not encapsulate all of the information extracted by the static measures.

Normalisation—linear shifting of parameter contours into the range 0–1 was used in order to 'distill' the dynamic properties. Discrimination rates for identity, and correlation for dialect, dropped little following normalisation, while sex discrimination rates dropped sharply. Discrimination or correlation rates for each characteristic were contrasted between static measures of the non-normalised parameters and dynamic measures of the normalised parameters. Significant differences were found in all cases such that speaker identity and dialect were found to be more strongly encoded in the time varying properties of the contours, while speaker sex was

more strongly encoded in the time invariant properties.

A novel extension of the dynamic time warp algorithm was employed, where measures of the calculated warp path between two contours were examined for speaker characteristic encoded information. Most DTW based schemes implicitly calculate the warp path and discard it after deriving the DTW distance. It was found that both speaker identity and dialect were strongly encoded in the warp path measures, to such an extent that they were significantly 'better' than the DTW distance. For speaker sex the warp path measures were marginally inferior to the DTW distance. Clearly the DTW warp path—a measure of the relative dynamics of two contours, encodes temporal related speaker characteristic information to a high degree, which may be used to discriminate the speaker characteristics.

The four basic parameters— energy, fundamental frequency, zero crossing rate, and voicing were individually investigated for encoded speaker characteristic information. *All* four parameters were found to have encoded information relating to *each* of the three characteristics. Fundamental frequency was the 'best' (highest degree of encoding) parameter for all three speaker characteristics, though degree of encoding, and in general the relative 'worth' of parameters was speaker characteristic dependent.

Four variant representations of fundamental frequency, based on log versus linear scale and interpolation across unvoiced or concatenated voiced only, were examined. For both speaker identity and dialect the log representation of F_0 was found to be markedly superior to the linear scale. In all cases interpolation was equal or superior to concatenation.

Discriminant or correlation performance for the three speaker characteristics was adequately modelled as a growth function of the amount of speech material used. However other factors than amount of speech material appear to influence discriminant/correlation levels so that accurate estimates of optimal performance were not possible, nor general guidelines regarding choice of utterance for text-dependent recognition systems.

The twenty-one measures were individually compared and contrasted. It was found that all measures extracted some encoded speaker-related information and that no single measure stood as consistently strong for all combinations of characteristic-parameter. Clearly, the form or nature of encoding of speaker characteristics is parameter and speaker characteristic dependent.

Discriminant and correlation results were analysed on the basis of the individual speakers that comprised the speaker population. For all three speaker characteristics results were found to be variable between individual speakers, and in particular highly variable for speaker dialect (showing prosodic correlates of dialect are general population 'trends', not firm constraints). That is, that for all three speaker characteristics their encoding within the parameters was speaker dependent.

Perceptual experiments were conducted using a single sentence and a different subset of the speaker population for each of the three speaker characteristics. A novel method of analysis-resynthesis using linear prediction to construct a composite utterance from a number of utterances, and allowing the individual manipulation and alteration of energy, fundamental frequency, voicing, and timing was devised and used to evaluate listener utilisation of acoustic cues to speaker characteristics.

Identity perception experiments revealed that listeners used prosodic parameters to identify speaker and were capable of identifying speaker with a high degree of accuracy based on a combination of the four examined parameters. Weighting of parameters as perceptual cues to identity was found to vary between the parameters and be dependent upon speaker. That is, that listener cue utilisation was speaker dependent. In a parallel of the analytical examination of the warp path (dynamic differences) parameters from speakers were warped to match that of the other speaker's. In some parameter-speaker combinations listener perception was altered significantly based on the warping— showing listener utilisation of the dynamics of a contour for identity perception; while in others little or no alteration occurred. Clearly in some cases at least listeners use the dynamic (time varying) properties of a parameter more than its static (time invariant) properties to form judgements of speaker identity.

Sex perception experiments showed the significance of mean fundamental frequency in listener perception of sex. Different dynamics of F_0 and the parameters energy, and voicing were found to have no significant influence upon listener perception of sex.

Dialect perception experiments were conducted using utterances of speakers from either end of the dialect spectrum. Naive listener response was generally found to be consistent, though under certain conditions of parameter alteration it became highly variable. Encoding of the parameters energy, fundamental frequency, and voicing showed no significant shift in listener perception consistent with the dialect of the originator of the parameter. Alterations in duration of utterance were found to significantly influence results such that shorter utterances were perceived as more cultivated while lengthened utterances were perceived as broader. Whether this listener perception is an externally imposed stereotype (e.g., media influence) and not a true representation of prosodic correlates of dialect for the Australian population, or a model drawn from listener experience is unclear, though results of the analysis section did show the significance of duration for speaker dialect.

The results highlight several areas in which further work may be carried out. Both analytical and perceptual techniques may be applied to other speaker characteristics, such as emotion, and to other speech parameters, such as spectral parameters. The perceptual experiments were limited in scope, the addition of more speakers and listeners would greatly strengthen and add to the results already achieved. Further examination of the phenomenon of listener perception of warped parameters is required. The assumption of the dialect difference as a linear scale may be an over-simplification and non-linear transformations may yield better results. Further investigation is required in order to determine what constitutes a good utterance for speaker recognition systems, and build accurate mathematical models of recognition performance based on parameters of the experiment— utterance, acoustic parameters, number of speakers etc. Following on from this, individual speaker variance in encoding of speaker characteristics requires further examination both so that databases may be accurately quantified and compared, and so that existing recognition systems may be 'fine tuned' by targeting 'trouble' speakers. A means of quantifying the dynamics of a contour (DTW warp path) that did not require the comparison of two contours would be advantageous. A less restricted data set, where utterances were sampled under normal conversational conditions, and hence are more variable and

dynamic appears desirable. Finally, implementing measures of the warp path in an existing recognition system would allow their evaluation under practical, application conditions.

Appendix A

Sentence Set

1. "Cool shirts please me."
2. "Pay the man first please."
3. "I cannot remember it."
4. "Papa needs two singers."
5. "A few boys bought them."
6. "Cash this bond please."
7. "How do you know?"
8. "A boy played a tune."
9. "June danced hard."
10. "The bear chews his paw."
11. "Today I auctioned beer."
12. "There she sits."
13. "We are firm."
14. "Are you poor?"
15. "We were away a year ago."

Appendix B

Speaker Information

All speakers recorded for experimentation were asked to provide personal details of relevance to their language development. Due to different forms of communication:- electronic mail for some participants and physical mail for the others, no single format for a form was used. However the following information was provided by all talkers:

Age As at November 1988.

Gender

Education Number of years formal education, and 'highest' educational qualification.

Occupation

Parental Information Occupation, country of birth, 1'st Language/Dialect, and age on coming to Australia (where applicable) of both parents.

Places of Residence For each town or city in which the speaker had lived for six months or more the following information: Age during residence, town/city name, state, and country.

Other Details Items as judged by speakers as having been of relevance to their spoken language development. Examples of singing and elocution lessons were given.

Appendix C

Correlation Tables

This appendix presents ω^2 values—ANOVA derived estimates of correlation between measure values and categorical speaker characteristics [HW71], for speaker identity and sex, and correlation values for speaker dialect. These results show the relationship between all twenty one measures of all examined parameters—four parameters, with three alternate variants on F_0 (7 total), for both un-normalised and normalised forms of the parameters, and each of the four sentences ($\approx$ 3 characteristics $\times$ 21 measures $\times$ [6 parameters $\times$ 2 treatments + 1 parameter $\times$ 1 treatment[1]] $\times$ 4 sentences = 3276 separate results). The absolute magnitude of individual measures—for speaker identity and sex 0–1, and for dialect -1–1 —show the strength of relationship, and hence importance, of that particular parameter-measure-treatment-sentence combination for differentiating the speaker characteristic.

The results are broken down into a hierarchical structure:- the three characteristics representing the major division with a subdivision into dynamic measures and static measures. Under these categories a number of tables, one for each parameter-treatment combination present the results as to individual sentence values. Rows of tables correspond to individual measures, with columns corresponding to sentences.

C.1 Speaker Identity

The Tables C.1–C.26 present the ω^2 values for all measures of all treatments of all parameters for each of the four test sentences. In the individual tables all ω^2 values greater than or equal to 0.05 are emphasised in an attempt to highlight stronger, or 'better' measures.

C.1.1 Dynamic Measure Correlation Tables

Tables C.1–C.13 show the individual ω^2 values for dynamic measures of all parameters from each sentence.

[1] Voicing is not normalised as it already lies within the range 0–1.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.07*	0.03	0.04	*0.07*	*0.05*
Weighted DTW Dist.	*0.07*	0.03	0.04	*0.06*	*0.05*
Border DTW Dist.	*0.07*	0.03	0.04	*0.07*	*0.05*
Warp Len./Contour Len.	0.01	0	0.01	0	0.01
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	*0.07*	0.02	0.04	*0.07*	*0.05*
Diag. Trans.	*0.06*	0.01	*0.05*	0.04	0.04
Num. Vert. Excn	0	0	0	0.01	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0.02	0.01
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.03	0.01	0.02	0.03	0.02
Max. Len. Diag. Excn	0	0	0.01	0	0
Off Diag. Wpath Dist.	*0.06*	0.03	0.04	*0.07*	*0.05*

Table C.1: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.07*	0.03	0.04	*0.06*	*0.05*
Weighted DTW Dist.	*0.06*	0.03	*0.05*	*0.05*	*0.05*
Border DTW Dist.	*0.07*	0.03	0.04	*0.05*	*0.05*
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	*0.06*	0.03	0.04	*0.06*	*0.05*
Diag. Trans.	0.04	0.02	0.04	0.03	0.03
Num. Vert. Excn	0.01	0	0	0.01	0.01
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0.01	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.02	0.02	0.02	0.02	0.02
Max. Len. Diag. Excn	0	0	0	0	0
Off Diag. Wpath Dist.	*0.05*	0.04	0.03	*0.05*	0.04

Table C.2: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.06*	0.03	*0.06*	*0.08*	*0.06*
Weighted DTW Dist.	*0.05*	0.03	*0.05*	*0.06*	*0.05*
Border DTW Dist.	*0.05*	0.03	*0.06*	*0.07*	*0.05*
Warp Len./Contour Len.	*0.05*	*0.08*	*0.05*	*0.07*	*0.06*
Vert. Trans.	*0.05*	*0.07*	*0.05*	*0.07*	*0.06*
Horiz. Trans.	0.01	0	0	0.01	0.01
Diag. Trans.	0.03	0.02	0.02	0.04	0.03
Num. Vert. Excn	*0.11*	*0.12*	*0.08*	*0.11*	*0.11*
Num. Horiz. Excn	*0.05*	0.02	0.02	*0.07*	0.04
Num. Diag. Excn	*0.12*	*0.07*	*0.07*	*0.13*	*0.1*
Max. Len. Vert. Excn	0	0.01	0	0	0
Max. Len. Horiz. Excn	0.01	0.02	0.01	0.02	0.02
Max. Len. Diag. Excn	0.04	0.02	0.03	0.03	0.03
Off Diag. Wpath Dist.	0.02	0.01	0	*0.05*	0.02

Table C.3: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.03	0.02	0.01	0.02	0.02
Weighted DTW Dist.	0.03	0.02	0.01	0.02	0.02
Border DTW Dist.	0.03	0.02	0.01	0.01	0.02
Warp Len./Contour Len.	0.01	0.01	0.02	0.01	0.01
Vert. Trans.	0.01	0.01	0.02	0.01	0.01
Horiz. Trans.	0	0.01	0	0.01	0.01
Diag. Trans.	0	0	0	0	0
Num. Vert. Excn	0.02	0.02	0.01	0.02	0.02
Num. Horiz. Excn	0.01	0	0	0	0
Num. Diag. Excn	0.02	0.01	0.01	0.01	0.01
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0	0.01	0	0.01	0.01
Max. Len. Diag. Excn	0.01	0	0	0	0
Off Diag. Wpath Dist.	0	0.01	0	0.02	0.01

Table C.4: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.05*	0.03	*0.06*	*0.08*	*0.06*
Weighted DTW Dist.	0.04	0.03	*0.05*	*0.07*	*0.05*
Border DTW Dist.	*0.05*	0.03	*0.06*	*0.08*	*0.06*
Warp Len./Contour Len.	*0.05*	*0.08*	*0.05*	*0.08*	*0.07*
Vert. Trans.	*0.05*	*0.07*	*0.05*	*0.07*	*0.06*
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0.02	0.02	0.01	0.03	0.02
Num. Vert. Excn	*0.13*	*0.13*	*0.07*	*0.11*	*0.11*
Num. Horiz. Excn	0.03	0.01	0.02	0.04	0.03
Num. Diag. Excn	*0.1*	*0.07*	*0.06*	*0.12*	*0.09*
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.01	0.01	0.03	0.02
Max. Len. Diag. Excn	0.04	0.03	0.03	0.03	0.03
Off Diag. Wpath Dist.	0.02	0.01	0	*0.07*	0.03

Table C.5: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.05*	*0.05*	0.03	*0.09*	*0.06*
Weighted DTW Dist.	*0.05*	*0.05*	0.02	*0.07*	*0.05*
Border DTW Dist.	*0.05*	*0.05*	0.02	*0.07*	*0.05*
Warp Len./Contour Len.	0.03	0.04	0.02	0.03	0.03
Vert. Trans.	0.03	0.04	0.02	0.03	0.03
Horiz. Trans.	0	0	0	0.01	0
Diag. Trans.	0.01	0.01	0	0	0.01
Num. Vert. Excn	*0.06*	*0.08*	0.03	*0.05*	*0.06*
Num. Horiz. Excn	0.01	0	0	0.01	0.01
Num. Diag. Excn	0.04	0.04	0.02	0.04	0.04
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.01	0.01	0.02	0.01
Max. Len. Diag. Excn	0.03	0.01	0	0.01	0.01
Off Diag. Wpath Dist.	0.02	0.02	0.01	*0.06*	0.03

Table C.6: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.1*	*0.06*	*0.06*	*0.06*	*0.07*
Weighted DTW Dist.	*0.08*	*0.07*	0.04	0.04	*0.06*
Border DTW Dist.	*0.09*	*0.06*	*0.05*	*0.05*	*0.06*
Warp Len./Contour Len.	*0.05*	*0.07*	*0.05*	*0.06*	*0.06*
Vert. Trans.	*0.05*	*0.06*	*0.05*	*0.06*	*0.06*
Horiz. Trans.	0.01	0	0	0	0
Diag. Trans.	0.03	0.01	0.02	0.03	0.02
Num. Vert. Excn	*0.11*	*0.12*	*0.09*	*0.11*	*0.11*
Num. Horiz. Excn	0.04	0.02	0.03	*0.07*	0.04
Num. Diag. Excn	*0.11*	*0.08*	*0.08*	*0.13*	*0.1*
Max. Len. Vert. Excn	0	0.01	0	0	0
Max. Len. Horiz. Excn	0	0.02	0.02	0.02	0.02
Max. Len. Diag. Excn	0.04	0.03	0.03	0.04	0.04
Off Diag. Wpath Dist.	0.02	0.01	0	*0.06*	0.02

Table C.7: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.06*	0.04	0.03	*0.07*	*0.05*
Weighted DTW Dist.	*0.05*	0.04	0.03	*0.05*	0.04
Border DTW Dist.	*0.05*	0.04	0.03	*0.06*	*0.05*
Warp Len./Contour Len.	0.02	0.03	0.01	0.01	0.02
Vert. Trans.	0.02	0.03	0.01	0.01	0.02
Horiz. Trans.	0	0.01	0	0.01	0.01
Diag. Trans.	0.01	0	0	0	0
Num. Vert. Excn	*0.07*	*0.07*	0.03	*0.05*	*0.06*
Num. Horiz. Excn	0.02	0.01	0.01	0.02	0.02
Num. Diag. Excn	*0.06*	*0.05*	0.02	*0.06*	*0.05*
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.02	0.01	0.01	0.01
Max. Len. Diag. Excn	0.02	0.01	0.01	0.01	0.01
Off Diag. Wpath Dist.	0.02	0.02	0.01	*0.05*	0.03

Table C.8: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.08*	*0.06*	*0.05*	*0.06*	*0.06*
Weighted DTW Dist.	*0.07*	*0.05*	0.03	0.04	*0.05*
Border DTW Dist.	*0.08*	*0.06*	0.04	*0.05*	*0.06*
Warp Len./Contour Len.	*0.05*	*0.08*	*0.05*	*0.08*	*0.07*
Vert. Trans.	*0.05*	*0.07*	*0.05*	*0.07*	*0.06*
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0.02	0.02	0.01	0.03	0.02
Num. Vert. Excn	*0.18*	*0.18*	*0.07*	*0.11*	*0.11*
Num. Horiz. Excn	0.03	0.01	0.02	0.04	0.03
Num. Diag. Excn	*0.1*	*0.07*	*0.06*	*0.11*	*0.09*
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.01	0.01	0.03	0.02
Max. Len. Diag. Excn	0.04	0.03	0.03	0.03	0.03
Off Diag. Wpath Dist.	0.02	0.01	0	*0.07*	0.03

Table C.9: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.05*	0.04	0.02	*0.07*	*0.05*
Weighted DTW Dist.	*0.05*	*0.05*	0.02	*0.05*	0.04
Border DTW Dist.	*0.05*	0.04	0.02	*0.05*	0.04
Warp Len./Contour Len.	0.02	*0.05*	0.01	0.03	0.03
Vert. Trans.	0.02	0.04	0.01	0.02	0.02
Horiz. Trans.	0	0	0.01	0.01	0.01
Diag. Trans.	0.01	0.01	0	0	0.01
Num. Vert. Excn	*0.06*	*0.07*	0.02	*0.05*	*0.05*
Num. Horiz. Excn	0.01	0.01	0	0.01	0.01
Num. Diag. Excn	0.04	0.04	0.01	0.04	0.03
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.01	0.01	0.02	0.01
Max. Len. Diag. Excn	0.03	0.01	0	0.01	0.01
Off Diag. Wpath Dist.	0.02	0.02	0	*0.05*	0.02

Table C.10: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.02	0	0.01	0.02	0.01
Weighted DTW Dist.	0.02	0	0.01	0.02	0.01
Border DTW Dist.	0.02	0	0.01	0.02	0.01
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	*0.06*	0.03	0.03	*0.07*	*0.05*
Diag. Trans.	*0.06*	0.03	0.03	*0.07*	*0.05*
Num. Vert. Excn	0	0	0	0.01	0
Num. Horiz. Excn	0	0.01	0	0.01	0.01
Num. Diag. Excn	0	0.01	0	0.01	0.01
Max. Len. Vert. Excn	0	0	0	0.01	0
Max. Len. Horiz. Excn	0.04	0.03	0.02	*0.06*	0.04
Max. Len. Diag. Excn	0.01	0.01	0.01	0.01	0.01
Off Diag. Wpath Dist.	0.04	0.02	0.01	*0.05*	0.03

Table C.11: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.07*	*0.05*	0.03	0.04	*0.05*
Weighted DTW Dist.	*0.07*	0.04	0.03	0.04	*0.05*
Border DTW Dist.	*0.06*	*0.05*	0.03	0.04	*0.05*
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	*0.05*	0.01	0.02	*0.06*	0.04
Diag. Trans.	0.04	0	0.01	0.03	0.02
Num. Vert. Excn	0.02	0.01	0.02	0.02	0.02
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0.02	0.01	0.02	0.01	0.02
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.03	0.01	0.02	0.02	0.02
Max. Len. Diag. Excn	0	0.01	0	0	0
Off Diag. Wpath Dist.	*0.08*	0.04	*0.05*	*0.06*	*0.06*

Table C.12: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.03	0.01	0.01	0.01	0.02
Weighted DTW Dist.	0.03	0.01	0.01	0.01	0.02
Border DTW Dist.	0.03	0.01	0.01	0.01	0.02
Warp Len./Contour Len.	0	0	0	0.01	0
Vert. Trans.	0	0	0	0.01	0
Horiz. Trans.	0.04	0.02	0.02	0.03	0.03
Diag. Trans.	0.02	0	0.01	0	0.01
Num. Vert. Excn	0.01	0.01	0.02	0.01	0.01
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0.01	0	0.01	0	0.01
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.02	0.01	0.01	0.01	0.01
Max. Len. Diag. Excn	0	0	0	0	0
Off Diag. Wpath Dist.	*0.05*	0.01	0.02	0.03	0.03

Table C.13: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Zero Crossings.

C.1.2 Static Measure Correlation Tables

Tables C.14–C.26 show the individual ω^2 values for static measures of all parameters from all sentences.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.02	0.01	0.01	0.02	0.02
Std. Dev.	0.01	0	0.01	0.02	0.01
Min.	0.01	0	0	0	0
Max.	0	0	0.01	0.01	0.01
Range	0	0	0	0	0
Mean Rate Change	0.02	0.03	0.03	0.02	0.03
Length	0.03	0.02	0.02	*0.05*	0.03

Table C.14: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0	0	0	0
Std. Dev.	0.01	0.01	0.01	0.02	0.01
Mean Rate Change	0	0.02	0.01	0.01	0.01
Length	0.03	0.02	0.02	*0.05*	0.03

Table C.15: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.04	0.03	*0.05*	*0.06*	*0.05*
Std. Dev.	0.03	0.03	0.02	*0.05*	0.03
Min.	*0.05*	0.03	0.03	0.03	0.04
Max.	0.02	*0.05*	0.04	0.04	0.04
Range	0.02	0.04	0.01	0.04	0.03
Mean Rate Change	0.02	0.04	0.01	0.04	0.03
Length	0.01	0.03	0.01	0.02	0.02

Table C.16: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0.01	0.01	0.01	0.01
Std. Dev.	0	0.01	0	0.01	0.01
Mean Rate Change	0.01	0.01	0	0	0.01
Length	0.01	0.03	0.01	0.02	0.02

Table C.17: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.04	0.03	*0.06*	*0.06*	*0.05*
Std. Dev.	0.03	0.03	0.03	*0.05*	0.04
Min.	*0.05*	0.03	0.03	0.03	0.04
Max.	0.02	*0.05*	0.04	0.04	0.04
Range	0.02	0.04	0.01	0.04	0.03
Mean Rate Change	0.03	0.04	0.02	0.04	0.03
Length	0.01	0.03	0.02	0.04	0.03

Table C.18: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0.01	0.01	0.01	0.01
Std. Dev.	0	0	0	0.01	0
Mean Rate Change	0.01	0.01	0	0	0.01
Length	0.01	0.03	0.02	0.04	0.03

Table C.19: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.04	0.04	0.03	*0.05*	0.04
Std. Dev.	0.04	0.03	0.02	0.03	0.03
Min.	0.01	0.01	0	0.01	0.01
Max.	0.02	*0.05*	0.03	0.04	0.04
Range	0.02	0.02	0.01	0.02	0.02
Mean Rate Change	0.03	0.04	0.01	0.03	0.03
Length	0.01	0.04	0.01	0.02	0.02

Table C.20: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0.01	0.01	0.01	0.01
Std. Dev.	0	0.01	0	0.01	0.01
Mean Rate Change	0.01	0.02	0	0	0.01
Length	0.01	0.04	0.01	0.02	0.02

Table C.21: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.04	0.03	0.03	*0.05*	0.04
Std. Dev.	*0.05*	0.03	0.02	0.03	0.03
Min.	0.01	0	0	0.01	0.01
Max.	0.02	*0.05*	0.04	*0.05*	0.04
Range	0.03	0.02	0.01	0.03	0.02
Mean Rate Change	0.03	0.02	0.02	0.03	0.03
Length	0.01	0.03	0.02	0.04	0.03

Table C.22: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0.01	0	0.01	0.01
Std. Dev.	0	0	0	0	0
Mean Rate Change	0.02	0.01	0	0.01	0.01
Length	0.01	0.03	0.02	0.04	0.03

Table C.23: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0.01	0	0.01	0.01
Std. Dev.	0.01	0.01	0	0	0.01
Mean Rate Change	0.01	0.01	0	0	0.01
Length	0	0	0	0.01	0

Table C.24: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.03	0.04	0.02	0.03	0.03
Std. Dev.	0.04	0.01	0.01	0.01	0.02
Min.	0	0	0	0	0
Max.	0.04	0.01	0.01	0.01	0.02
Range	0.04	0.01	0.01	0	0.02
Mean Rate Change	0.03	0	0.01	0.01	0.01
Length	0.03	0.02	0.02	*0.05*	0.03

Table C.25: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0	0.01	0	0.01
Std. Dev.	0	0	0	0	0
Mean Rate Change	0	0	0.02	0	0.01
Length	0.03	0.02	0.02	*0.05*	0.03

Table C.26: Speaker Identity. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Speaker categories. A quantification of each individual measure's ability to predict identity. Parameter: Normalised Zero Crossings.

C.2 Speaker Sex

The Tables C.27–C.52 present the ω^2 values for all measures of all treatments of all parameters for each of the four test sentences. In the individual tables all ω^2 values greater than or equal to 0.05 are emphasised in an attempt to highlight stronger, or 'better' measures.

C.2.1 Dynamic Measure Correlation Tables

Tables C.27–C.39 show the individual ω^2 values for dynamic measures of all parameters from all sentences.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.02	*0.05*	0	0.01	0.02
Weighted DTW Dist.	*0.06*	*0.05*	0.01	0.02	0.03
Border DTW Dist.	0.03	*0.05*	0	0.01	0.02
Warp Len./Contour Len.	0.01	0.01	0.02	0.01	0.01
Vert. Trans.	0	0.01	0.02	0.01	0.01
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0.01	0.01	0	0	0.01
Num. Vert. Excn	0	0.01	0.01	0.01	0.01
Num. Horiz. Excn	0.01	0	0	0	0
Num. Diag. Excn	0.02	0.01	0	0.01	0.01
Max. Len. Vert. Excn	0.01	0	0.01	0.02	0.01
Max. Len. Horiz. Excn	0.03	0	0	0	0.01
Max. Len. Diag. Excn	0	0.03	0.01	0	0.01
Off Diag. Wpath Dist.	*0.13*	0.02	*0.06*	0.03	*0.06*

Table C.27: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0	0	0
Weighted DTW Dist.	0.01	0	0	0.01	0.01
Border DTW Dist.	0	0	0	0	0
Warp Len./Contour Len.	0	0.01	0.02	0.01	0.01
Vert. Trans.	0	0.01	0.02	0.01	0.01
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0	0.01	0	0	0
Num. Vert. Excn	0	0	0.01	0	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0	0
Max. Len. Vert. Excn	0	0	0.02	0.01	0.01
Max. Len. Horiz. Excn	0.01	0.01	0	0	0.01
Max. Len. Diag. Excn	0	0.01	0.01	0	0.01
Off Diag. Wpath Dist.	*0.09*	0.02	*0.06*	0.01	*0.05*

Table C.28: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.82*	*0.74*	*0.72*	*0.8*	*0.77*
Weighted DTW Dist.	*0.71*	*0.65*	*0.7*	*0.67*	*0.68*
Border DTW Dist.	*0.82*	*0.72*	*0.72*	*0.77*	*0.76*
Warp Len./Contour Len.	*0.49*	*0.35*	*0.38*	*0.57*	*0.45*
Vert. Trans.	*0.53*	*0.37*	*0.41*	*0.62*	*0.48*
Horiz. Trans.	*0.14*	*0.14*	*0.06*	*0.25*	*0.15*
Diag. Trans.	*0.38*	*0.31*	*0.23*	*0.5*	*0.36*
Num. Vert. Excn	*0.47*	*0.3*	*0.34*	*0.49*	*0.4*
Num. Horiz. Excn	*0.19*	*0.11*	*0.07*	*0.33*	*0.18*
Num. Diag. Excn	*0.38*	*0.21*	*0.2*	*0.49*	*0.32*
Max. Len. Vert. Excn	*0.45*	*0.31*	*0.38*	*0.49*	*0.41*
Max. Len. Horiz. Excn	0	0.01	0	0	0
Max. Len. Diag. Excn	0.04	0.01	0	*0.22*	*0.07*
Off Diag. Wpath Dist.	*0.11*	*0.11*	0.04	*0.11*	*0.09*

Table C.29: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0.01	0.03	0.01
Weighted DTW Dist.	0.01	0	0.01	0.04	0.02
Border DTW Dist.	0	0	0.01	0.02	0.01
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0.01	0	0	0.01	0.01
Horiz. Trans.	0.01	0	0	0.01	0.01
Diag. Trans.	0	0	0	0	0
Num. Vert. Excn	0	0	0.01	0.01	0.01
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0.01	0	0
Max. Len. Vert. Excn	0.01	0	0	0	0
Max. Len. Horiz. Excn	0.01	0.01	0	0	0.01
Max. Len. Diag. Excn	0	0	0	0	0
Off Diag. Wpath Dist.	0.02	0	0	0.03	0.01

Table C.30: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.83*	*0.75*	*0.73*	*0.81*	*0.78*
Weighted DTW Dist.	*0.72*	*0.65*	*0.68*	*0.61*	*0.67*
Border DTW Dist.	*0.82*	*0.73*	*0.73*	*0.78*	*0.76*
Warp Len./Contour Len.	*0.47*	*0.38*	*0.44*	*0.55*	*0.46*
Vert. Trans.	*0.51*	*0.4*	*0.47*	*0.6*	*0.5*
Horiz. Trans.	*0.26*	*0.12*	*0.1*	*0.33*	*0.2*
Diag. Trans.	*0.49*	*0.3*	*0.31*	*0.54*	*0.41*
Num. Vert. Excn	*0.42*	*0.32*	*0.34*	*0.48*	*0.39*
Num. Horiz. Excn	*0.23*	*0.08*	*0.11*	*0.33*	*0.19*
Num. Diag. Excn	*0.4*	*0.2*	*0.25*	*0.51*	*0.34*
Max. Len. Vert. Excn	*0.41*	*0.34*	*0.42*	*0.48*	*0.41*
Max. Len. Horiz. Excn	0.01	0.01	0	0	0.01
Max. Len. Diag. Excn	0.03	0.02	0	*0.22*	*0.07*
Off Diag. Wpath Dist.	*0.17*	*0.11*	*0.07*	*0.13*	*0.12*

Table C.31: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0	0	0
Weighted DTW Dist.	0	0	0.01	0	0
Border DTW Dist.	0	0	0	0	0
Warp Len./Contour Len.	0.01	0	0.01	0	0.01
Vert. Trans.	0.01	0	0.01	0	0.01
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0	0	0	0	0
Num. Vert. Excn	0	0	0.01	0	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0	0	0.01	0	0
Max. Len. Diag. Excn	0	0	0.01	0.01	0.01
Off Diag. Wpath Dist.	0	0.01	0.02	0	0.01

Table C.32: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.26*	*0.28*	*0.12*	*0.12*	*0.2*
Weighted DTW Dist.	*0.25*	*0.27*	*0.09*	*0.06*	*0.17*
Border DTW Dist.	*0.28*	*0.28*	*0.12*	*0.1*	*0.2*
Warp Len./Contour Len.	*0.08*	*0.08*	0.02	*0.08*	*0.07*
Vert. Trans.	*0.09*	*0.09*	0.02	*0.09*	*0.07*
Horiz. Trans.	0.01	*0.05*	0	0.01	0.02
Diag. Trans.	*0.06*	*0.1*	0.01	*0.05*	*0.06*
Num. Vert. Excn	*0.09*	*0.05*	0.02	*0.07*	*0.06*
Num. Horiz. Excn	0.03	0.02	0	0.02	0.02
Num. Diag. Excn	*0.08*	*0.05*	0.01	*0.05*	*0.05*
Max. Len. Vert. Excn	0.04	0.04	0.02	0.01	0.03
Max. Len. Horiz. Excn	0.03	0.01	0	0	0.01
Max. Len. Diag. Excn	0.02	0.02	0.01	0.01	0.02
Off Diag. Wpath Dist.	0	0	0	0.01	0

Table C.33: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0	0	0
Weighted DTW Dist.	0	0	0	0	0
Border DTW Dist.	0	0	0	0	0
Warp Len./Contour Len.	0	0	0	0.01	0
Vert. Trans.	0	0	0	0.01	0
Horiz. Trans.	0.02	0	0	0	0.01
Diag. Trans.	0.02	0	0	0	0.01
Num. Vert. Excn	0	0	0	0.01	0
Num. Horiz. Excn	0	0	0	0.01	0
Num. Diag. Excn	0	0	0	0	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0.01	0	0	0	0
Max. Len. Diag. Excn	0.01	0	0	0	0
Off Diag. Wpath Dist.	0.01	0	0.01	0.03	0.01

Table C.34: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.28*	*0.23*	*0.13*	*0.12*	*0.19*
Weighted DTW Dist.	*0.25*	*0.2*	*0.1*	*0.05*	*0.15*
Border DTW Dist.	*0.3*	*0.23*	*0.12*	*0.1*	*0.19*
Warp Len./Contour Len.	*0.09*	*0.06*	*0.05*	*0.06*	*0.07*
Vert. Trans.	*0.09*	*0.06*	*0.05*	*0.06*	*0.07*
Horiz. Trans.	*0.07*	0.02	0	*0.05*	0.04
Diag. Trans.	*0.12*	*0.05*	0.02	*0.07*	*0.07*
Num. Vert. Excn	*0.07*	0.03	0.03	*0.05*	*0.05*
Num. Horiz. Excn	*0.06*	0	0.01	0.04	0.03
Num. Diag. Excn	*0.1*	0.01	0.02	*0.07*	*0.05*
Max. Len. Vert. Excn	0.03	0.03	0.04	0.01	0.03
Max. Len. Horiz. Excn	0	0	0	0	0
Max. Len. Diag. Excn	0.02	0	0	0.02	0.01
Off Diag. Wpath Dist.	0.01	0	0	0	0

Table C.35: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0	0	0
Weighted DTW Dist.	0	0	0.01	0	0
Border DTW Dist.	0	0	0	0	0
Warp Len./Contour Len.	0	0	0.01	0	0
Vert. Trans.	0	0	0.01	0	0
Horiz. Trans.	0	0	0	0.01	0
Diag. Trans.	0	0	0	0.01	0
Num. Vert. Excn	0	0	0.01	0	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0	0	0.01	0	0
Max. Len. Diag. Excn	0	0	0.01	0	0
Off Diag. Wpath Dist.	0	0.01	0.02	0	0.01

Table C.36: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0.03	0.01	0.01
Weighted DTW Dist.	0.01	0	*0.05*	0.01	0.02
Border DTW Dist.	0	0	0.03	0.01	0.01
Warp Len./Contour Len.	*0.05*	0	0.02	0.01	0.02
Vert. Trans.	*0.05*	0	0.02	0.01	0.02
Horiz. Trans.	0.01	0	0.03	0	0.01
Diag. Trans.	0.04	0	*0.05*	0	0.02
Num. Vert. Excn	0.03	0	0.01	0.03	0.02
Num. Horiz. Excn	0	0	0.02	0	0.01
Num. Diag. Excn	0	0	0.01	0	0
Max. Len. Vert. Excn	0.01	0	0.01	0	0.01
Max. Len. Horiz. Excn	0.02	0	0.04	0	0.02
Max. Len. Diag. Excn	0.01	0	0.01	0.01	0.01
Off Diag. Wpath Dist.	*0.13*	0	*0.11*	0.02	*0.07*

Table C.37: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.05*	0	0	0.01	0.02
Weighted DTW Dist.	0.04	0	0.01	0.01	0.02
Border DTW Dist.	0.04	0	0	0.01	0.01
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	0	0	0.02	0	0.01
Diag. Trans.	0	0	0.01	0	0
Num. Vert. Excn	0	0	0	0	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0.01	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0	0	0.01	0	0
Max. Len. Diag. Excn	0	0	0	0	0
Off Diag. Wpath Dist.	0.01	0	0.01	0.02	0.01

Table C.38: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0.02	0.02	0.01
Weighted DTW Dist.	0.01	0	0.02	0.03	0.02
Border DTW Dist.	0	0	0.02	0.02	0.01
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	0	0	0.02	0	0.01
Diag. Trans.	0	0	0.02	0	0.01
Num. Vert. Excn	0	0	0	0.01	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0.02	0.01
Max. Len. Vert. Excn	0	0	0.01	0	0
Max. Len. Horiz. Excn	0	0	0.01	0	0
Max. Len. Diag. Excn	0	0	0	0.01	0
Off Diag. Wpath Dist.	0.01	0	0.03	0.03	0.02

Table C.39: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of dynamic measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Zero Crossings.

C.2.2 Static Measure Correlation Tables

Tables C.40–C.52 show the individual ω^2 values for static measures of all parameters from all sentences.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.02	0.04	0	0.01	0.02
Std. Dev.	0	0	0	0	0
Min.	0	0.01	0	0	0
Max.	0.02	*0.11*	0	0	0.03
Range	0	0.01	0.01	0	0.01
Mean Rate Change	0.01	0	0	0	0
Length	0	0	0.01	0	0

Table C.40: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0	0.03	0.01
Std. Dev.	0.01	0	0	0	0
Mean Rate Change	*0.08*	0.01	0.01	0	0.03
Length	0	0	0.01	0	0

Table C.41: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.85*	*0.77*	*0.73*	*0.81*	*0.79*
Std. Dev.	*0.15*	*0.09*	*0.1*	*0.15*	*0.12*
Min.	*0.62*	*0.65*	*0.54*	*0.62*	*0.61*
Max.	*0.74*	*0.55*	*0.67*	*0.71*	*0.67*
Range	*0.18*	*0.13*	*0.15*	*0.2*	*0.16*
Mean Rate Change	*0.1*	*0.09*	*0.17*	*0.07*	*0.11*
Length	0.01	0	0.01	0.02	0.01

Table C.42: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0	0	0
Std. Dev.	0.01	0	0	0	0
Mean Rate Change	0	0	0	0.02	0.01
Length	0.01	0	0.01	0.02	0.01

Table C.43: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.86*	*0.77*	*0.74*	*0.82*	*0.8*
Std. Dev.	*0.15*	*0.11*	*0.09*	*0.14*	*0.12*
Min.	*0.62*	*0.65*	*0.54*	*0.62*	*0.61*
Max.	*0.74*	*0.55*	*0.67*	*0.71*	*0.67*
Range	*0.18*	*0.13*	*0.15*	*0.2*	*0.16*
Mean Rate Change	*0.15*	*0.08*	*0.19*	*0.12*	*0.14*
Length	0	0	0.01	0	0

Table C.44: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0	0	0
Std. Dev.	0.01	0	0	0	0
Mean Rate Change	0	0	0	0.01	0
Length	0	0	0.01	0	0

Table C.45: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.21*	*0.27*	*0.06*	*0.07*	*0.15*
Std. Dev.	0.01	0	0.01	0.01	0.01
Min.	0.03	0.01	0.03	0.03	0.03
Max.	*0.32*	*0.18*	*0.12*	*0.23*	*0.21*
Range	0.03	0.01	0.03	0.02	0.02
Mean Rate Change	0	0	0.03	0.01	0.01
Length	0.01	0	0.01	0.02	0.01

Table C.46: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0	0	0
Std. Dev.	0	0	0	0	0
Mean Rate Change	0	0.01	0	0	0
Length	0.01	0	0.01	0.02	0.01

Table C.47: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.23*	*0.19*	*0.05*	*0.07*	*0.14*
Std. Dev.	0.01	0	0.02	0.01	0.01
Min.	0.03	0	0.04	0.02	0.02
Max.	*0.33*	*0.18*	*0.14*	*0.25*	*0.23*
Range	0.02	0.01	0.03	0.02	0.02
Mean Rate Change	0.01	0	0.04	0.01	0.02
Length	0	0	0.01	0	0

Table C.48: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0	0	0
Std. Dev.	0	0	0	0	0
Mean Rate Change	0	0	0	0	0
Length	0	0	0.01	0	0

Table C.49: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0	0	0.02	0.03	0.01
Std. Dev.	0.01	0	0.03	*0.11*	0.04
Mean Rate Change	0.01	0	0.03	*0.11*	0.04
Length	0	0	0	0	0

Table C.50: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.02	0	0.01	0	0.01
Std. Dev.	*0.05*	0	0.01	0	0.02
Min.	0	0	0.01	0.01	0.01
Max.	0.04	0	0.01	0	0.01
Range	0.03	0	0.01	0	0.01
Mean Rate Change	0.01	0	0.03	0	0.01
Length	0	0	0.01	0	0

Table C.51: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.01	0	0.01	0	0.01
Std. Dev.	0	0	0	0.01	0
Mean Rate Change	0	0	0	0	0
Length	0	0	0.01	0	0

Table C.52: Speaker Sex. Estimated strength (ω^2) of relationship [correlation] of static measure values to intra/inter Sex categories. A quantification of each individual measure's ability to predict sex. Parameter: Normalised Zero Crossings.

C.3 Speaker Dialect

The Tables C.53–C.78 present the correlation values for all measures of all treatments of all parameters for each of the four test sentences. In the individual tables all correlation values greater than or equal to 0.05 in absolute magnitude are emphasised in an attempt to highlight stronger, or 'better' measures.

C.3.1 Dynamic Measure Correlation Tables

Tables C.53–C.65 show the individual correlation values for dynamic measures of all parameters from all sentences.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.12*	*0.2*	*0.12*	0.04	*0.12*
Weighted DTW Dist.	*0.1*	*0.2*	*0.11*	*0.05*	*0.12*
Border DTW Dist.	*0.12*	*0.21*	*0.12*	0.04	*0.12*
Warp Len./Contour Len.	*0.06*	-0.01	0	*0.05*	0.02
Vert. Trans.	*0.07*	-0.01	0	*0.06*	0.03
Horiz. Trans.	*-0.09*	*-0.08*	*-0.09*	*-0.11*	*-0.09*
Diag. Trans.	0.03	*0.06*	*0.08*	*0.05*	*0.05*
Num. Vert. Excn	*0.05*	0.03	-0.02	-0.02	0.01
Num. Horiz. Excn	*-0.08*	*-0.07*	-0.02	*-0.06*	*-0.06*
Num. Diag. Excn	*-0.06*	*-0.07*	-0.04	*-0.07*	*-0.06*
Max. Len. Vert. Excn	*0.07*	-0.03	0.02	*0.05*	0.03
Max. Len. Horiz. Excn	0.02	-0.01	*-0.09*	*-0.06*	-0.04
Max. Len. Diag. Excn	*0.07*	*0.1*	*0.08*	*0.07*	*0.08*
Off Diag. Wpath Dist.	0.01	*0.09*	0.03	*0.05*	*0.05*

Table C.53: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.1*	*0.24*	*0.09*	-0.02	*0.1*
Weighted DTW Dist.	*0.08*	*0.24*	*0.07*	-0.02	*0.09*
Border DTW Dist.	*0.1*	*0.24*	*0.08*	-0.02	*0.1*
Warp Len./Contour Len.	*0.07*	0	0.02	*0.05*	0.04
Vert. Trans.	*0.08*	0.01	0.02	*0.06*	0.04
Horiz. Trans.	*-0.09*	*-0.07*	*-0.08*	*-0.11*	*-0.09*
Diag. Trans.	0.02	*0.05*	*0.06*	*0.06*	*0.05*
Num. Vert. Excn	*0.07*	*0.06*	0	0.03	0.04
Num. Horiz. Excn	*-0.06*	*-0.1*	*-0.05*	0	*-0.05*
Num. Diag. Excn	-0.02	*-0.08*	*-0.05*	0.01	-0.03
Max. Len. Vert. Excn	0.03	-0.03	0	0.01	0
Max. Len. Horiz. Excn	0.01	0.02	*-0.05*	*-0.06*	-0.02
Max. Len. Diag. Excn	0.04	*0.05*	*0.05*	*0.05*	*0.05*
Off Diag. Wpath Dist.	0.04	*0.19*	0.03	0.01	*0.07*

Table C.54: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.02	0	0	*-0.08*	-0.01
Weighted DTW Dist.	-0.03	-0.01	-0.01	*-0.07*	-0.03
Border DTW Dist.	0.01	-0.01	-0.01	*-0.08*	-0.02
Warp Len./Contour Len.	0	*-0.07*	0	*0.08*	0
Vert. Trans.	0	*-0.06*	0	*0.08*	0
Horiz. Trans.	*-0.07*	*-0.08*	*-0.05*	0.02	-0.04
Diag. Trans.	*0.05*	*0.1*	0.04	*-0.05*	0.04
Num. Vert. Excn	*-0.05*	*-0.08*	-0.03	*0.07*	-0.02
Num. Horiz. Excn	*-0.07*	*-0.09*	0.01	0	-0.04
Num. Diag. Excn	*-0.07*	*-0.1*	-0.01	*0.05*	-0.03
Max. Len. Vert. Excn	0	-0.02	0.03	0.02	0.01
Max. Len. Horiz. Excn	0.01	*0.08*	*-0.05*	0.01	0.01
Max. Len. Diag. Excn	*0.07*	*0.05*	*0.07*	*-0.08*	0.03
Off Diag. Wpath Dist.	*-0.15*	0	0.03	-0.02	-0.04

Table C.55: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0	0	0
Weighted DTW Dist.	0	0	0	0	0
Border DTW Dist.	0	0	0	0	0
Warp Len./Contour Len.	0	0	0	0	0
Vert. Trans.	0	0	0	0	0
Horiz. Trans.	0	0	0	0	0
Diag. Trans.	0	0	0	0	0
Num. Vert. Excn	0	0	0	0	0
Num. Horiz. Excn	0	0	0	0	0
Num. Diag. Excn	0	0	0	0	0
Max. Len. Vert. Excn	0	0	0	0	0
Max. Len. Horiz. Excn	0	0	0	0	0
Max. Len. Diag. Excn	0	0	0	0	0
Off Diag. Wpath Dist.	0	0	0	0	0

Table C.56: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0	0	0.01	*-0.07*	-0.02
Weighted DTW Dist.	*-0.05*	-0.01	0.01	*-0.06*	-0.03
Border DTW Dist.	-0.01	-0.01	0.01	*-0.07*	-0.02
Warp Len./Contour Len.	-0.01	*-0.05*	0.01	*0.09*	0.01
Vert. Trans.	-0.01	*-0.05*	0.01	*0.09*	0.01
Horiz. Trans.	-0.03	*-0.08*	*-0.09*	-0.01	*-0.05*
Diag. Trans.	0.03	*0.09*	*0.06*	-0.04	0.04
Num. Vert. Excn	-0.04	*-0.07*	-0.04	*0.09*	-0.01
Num. Horiz. Excn	-0.04	*-0.12*	-0.04	0.04	-0.04
Num. Diag. Excn	*-0.05*	*-0.12*	*-0.06*	*0.09*	-0.03
Max. Len. Vert. Excn	-0.01	-0.02	0.03	0.03	0.01
Max. Len. Horiz. Excn	-0.01	*0.09*	-0.04	-0.03	0
Max. Len. Diag. Excn	0.04	*0.08*	*0.08*	*-0.08*	0.03
Off Diag. Wpath Dist.	*-0.16*	0.01	*0.05*	-0.04	-0.03

Table C.57: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.14*	*0.14*	0.01	0.03	*0.08*
Weighted DTW Dist.	*0.1*	*0.16*	0.02	0	*0.07*
Border DTW Dist.	*0.15*	*0.15*	0.02	0	*0.08*
Warp Len./Contour Len.	*-0.1*	*-0.09*	0.04	0.03	-0.03
Vert. Trans.	*-0.1*	*-0.09*	*0.05*	0.03	-0.03
Horiz. Trans.	*-0.1*	*-0.1*	*-0.07*	*-0.07*	*-0.09*
Diag. Trans.	*0.13*	*0.12*	0.03	0.03	*0.08*
Num. Vert. Excn	*-0.08*	*-0.09*	0.01	*0.07*	-0.02
Num. Horiz. Excn	*-0.09*	*-0.11*	-0.03	-0.01	*-0.06*
Num. Diag. Excn	*-0.11*	*-0.13*	-0.02	0.04	*-0.05*
Max. Len. Vert. Excn	-0.04	-0.03	*0.08*	-0.02	0
Max. Len. Horiz. Excn	0.01	*0.08*	-0.01	0	0.02
Max. Len. Diag. Excn	*0.13*	*0.17*	0.04	0.03	*0.09*
Off Diag. Wpath Dist.	*-0.13*	0	-0.02	0.01	-0.03

Table C.58: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.08*	*0.07*	-0.01	*-0.09*	0.01
Weighted DTW Dist.	0.04	0.03	-0.02	*-0.09*	-0.01
Border DTW Dist.	*0.08*	0.04	-0.01	*-0.1*	0
Warp Len./Contour Len.	0.01	*-0.05*	0	*0.11*	0.02
Vert. Trans.	0.01	-0.04	0.01	*0.11*	0.02
Horiz. Trans.	*-0.06*	*-0.08*	*-0.05*	0.04	-0.04
Diag. Trans.	0.04	*0.08*	0.03	*-0.08*	0.02
Num. Vert. Excn	*-0.08*	*-0.05*	-0.03	*0.09*	-0.02
Num. Horiz. Excn	-0.04	*-0.09*	0.02	-0.01	-0.03
Num. Diag. Excn	*-0.07*	*-0.09*	0	*0.05*	-0.03
Max. Len. Vert. Excn	0.04	-0.02	0.01	0.04	0.02
Max. Len. Horiz. Excn	0.02	*0.08*	-0.04	0.01	0.02
Max. Len. Diag. Excn	*0.05*	*0.07*	*0.05*	*-0.07*	0.03
Off Diag. Wpath Dist.	*-0.1*	0.01	*0.05*	*-0.05*	-0.02

Table C.59: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.12*	*0.18*	0.04	0.02	*0.09*
Weighted DTW Dist.	*0.11*	*0.2*	*0.05*	0.02	*0.1*
Border DTW Dist.	*0.13*	*0.18*	*0.05*	0.01	*0.09*
Warp Len./Contour Len.	-0.04	*-0.1*	0.02	-0.02	-0.04
Vert. Trans.	-0.04	*-0.1*	0.02	-0.02	-0.04
Horiz. Trans.	*-0.1*	*-0.11*	-0.04	*-0.07*	*-0.08*
Diag. Trans.	*0.09*	*0.13*	0.02	*0.06*	*0.07*
Num. Vert. Excn	*-0.09*	*-0.11*	0.02	0.02	-0.04
Num. Horiz. Excn	*-0.08*	*-0.14*	0	-0.02	*-0.06*
Num. Diag. Excn	*-0.11*	*-0.16*	0	0	*-0.07*
Max. Len. Vert. Excn	*0.05*	-0.03	0.02	0	0.01
Max. Len. Horiz. Excn	0	*0.11*	0	-0.01	0.02
Max. Len. Diag. Excn	*0.07*	*0.14*	0.02	0.02	*0.06*
Off Diag. Wpath Dist.	-0.02	*0.05*	-0.01	0.01	0.01

Table C.60: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.06*	0.04	0	*-0.08*	0
Weighted DTW Dist.	-0.02	0	-0.02	*-0.07*	-0.03
Border DTW Dist.	0.04	0.01	-0.01	*-0.08*	-0.01
Warp Len./Contour Len.	-0.01	*-0.05*	0.01	*0.09*	0.01
Vert. Trans.	-0.01	*-0.05*	0.01	*0.09*	0.01
Horiz. Trans.	-0.03	*-0.08*	*-0.1*	-0.01	*-0.05*
Diag. Trans.	0.03	*0.09*	*0.06*	-0.04	0.04
Num. Vert. Excn	-0.04	*-0.06*	-0.04	*0.09*	-0.01
Num. Horiz. Excn	-0.04	*-0.12*	-0.04	0.04	-0.04
Num. Diag. Excn	*-0.05*	*-0.12*	*-0.06*	*0.09*	-0.03
Max. Len. Vert. Excn	0	-0.01	0.03	0.03	0.01
Max. Len. Horiz. Excn	0	*0.09*	-0.04	-0.03	0.01
Max. Len. Diag. Excn	0.04	*0.08*	*0.08*	*-0.08*	0.03
Off Diag. Wpath Dist.	*-0.16*	0.01	*0.05*	-0.04	-0.03

Table C.61: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*0.15*	*0.15*	0.02	0.03	*0.09*
Weighted DTW Dist.	*0.12*	*0.17*	0.03	0.01	*0.08*
Border DTW Dist.	*0.16*	*0.15*	0.03	0	*0.09*
Warp Len./Contour Len.	*-0.08*	*-0.08*	0.03	0.01	-0.03
Vert. Trans.	*-0.09*	*-0.08*	0.03	0.01	-0.03
Horiz. Trans.	*-0.09*	*-0.1*	*-0.08*	*-0.08*	*-0.09*
Diag. Trans.	*0.11*	*0.11*	0.04	*0.05*	*0.08*
Num. Vert. Excn	*-0.09*	*-0.09*	0.01	*0.06*	-0.03
Num. Horiz. Excn	*-0.08*	*-0.11*	-0.02	0.01	*-0.05*
Num. Diag. Excn	*-0.11*	*-0.13*	-0.01	*0.05*	*-0.05*
Max. Len. Vert. Excn	-0.02	-0.02	*0.06*	0	0
Max. Len. Horiz. Excn	0.01	*0.06*	-0.02	-0.02	0.01
Max. Len. Diag. Excn	*0.08*	*0.15*	0.04	0.04	*0.08*
Off Diag. Wpath Dist.	*-0.13*	0.01	-0.01	0.01	-0.03

Table C.62: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	0.04	*0.05*	*-0.06*	-0.01	0.01
Weighted DTW Dist.	-0.03	0.03	*-0.07*	-0.04	-0.03
Border DTW Dist.	0.02	0.03	*-0.06*	-0.03	-0.01
Warp Len./Contour Len.	0.02	-0.03	0.01	*0.1*	0.03
Vert. Trans.	0.02	-0.02	0.01	*0.11*	0.03
Horiz. Trans.	*-0.13*	*-0.07*	*-0.09*	*-0.09*	*-0.1*
Diag. Trans.	*0.11*	*0.07*	*0.09*	0.03	*0.08*
Num. Vert. Excn	0.03	0	0.02	*0.06*	0.03
Num. Horiz. Excn	*-0.1*	-0.04	-0.04	-0.04	*-0.05*
Num. Diag. Excn	*-0.07*	-0.04	-0.03	-0.03	-0.04
Max. Len. Vert. Excn	0.04	-0.01	0	*0.11*	0.03
Max. Len. Horiz. Excn	*-0.08*	*-0.07*	*-0.09*	*-0.07*	*-0.08*
Max. Len. Diag. Excn	*0.1*	*0.05*	*0.08*	*0.06*	*0.07*
Off Diag. Wpath Dist.	*-0.07*	-0.03	*-0.08*	-0.01	*-0.05*

Table C.63: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*-0.13*	*-0.09*	*-0.06*	-0.04	*-0.08*
Weighted DTW Dist.	*-0.13*	*-0.08*	-0.04	-0.03	*-0.07*
Border DTW Dist.	*-0.12*	*-0.08*	*-0.06*	-0.04	*-0.07*
Warp Len./Contour Len.	0.02	0.02	0.03	0.04	0.03
Vert. Trans.	0.03	0.03	0.03	*0.05*	0.03
Horiz. Trans.	*-0.12*	*-0.05*	*-0.08*	*-0.11*	*-0.09*
Diag. Trans.	*0.07*	0.02	*0.05*	*0.06*	*0.05*
Num. Vert. Excn	*0.09*	0.01	0.04	*0.06*	*0.05*
Num. Horiz. Excn	-0.01	0.03	0.03	-0.03	0
Num. Diag. Excn	*0.06*	0.02	*0.05*	0.03	0.04
Max. Len. Vert. Excn	-0.03	0.04	0.03	*0.06*	0.02
Max. Len. Horiz. Excn	*-0.08*	*-0.06*	*-0.07*	*-0.07*	*-0.07*
Max. Len. Diag. Excn	0.04	-0.01	-0.02	0.04	0.01
Off Diag. Wpath Dist.	*-0.11*	-0.04	-0.04	0	*-0.05*

Table C.64: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
DTW Dist	*-0.05*	-0.02	-0.03	0.04	-0.02
Weighted DTW Dist.	-0.04	-0.02	-0.03	0.04	-0.01
Border DTW Dist.	-0.04	-0.02	-0.03	0.04	-0.01
Warp Len./Contour Len.	0.01	0.02	*0.06*	0.03	0.03
Vert. Trans.	0.02	0.03	*0.06*	0.03	0.03
Horiz. Trans.	*-0.13*	*-0.05*	*-0.07*	*-0.12*	*-0.09*
Diag. Trans.	*0.09*	0.02	0.03	*0.08*	*0.06*
Num. Vert. Excn	*0.05*	0.01	0	0.01	0.02
Num. Horiz. Excn	-0.01	-0.02	0.03	*-0.08*	-0.02
Num. Diag. Excn	0.02	-0.01	0	*-0.05*	-0.01
Max. Len. Vert. Excn	0.02	0.03	*0.05*	*0.06*	0.04
Max. Len. Horiz. Excn	*-0.06*	-0.02	*-0.05*	-0.03	-0.04
Max. Len. Diag. Excn	*0.07*	0	0.02	*0.12*	*0.05*
Off Diag. Wpath Dist.	*-0.05*	0.04	0.01	0	0

Table C.65: Speaker Dialect difference scores correlated with dynamic measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Zero Crossings.

C.3.2 Static Measure Correlation Tables

Tables C.66–C.78 show the individual correlation values for static measures of all parameters from all sentences.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	-0.01	-0.04	-0.04	-0.03	-0.03
Std. Dev.	-0.02	-0.02	*-0.07*	*0.05*	-0.01
Min.	-0.03	-0.03	-0.03	*0.05*	-0.01
Max.	-0.04	*-0.12*	0.04	0.01	-0.03
Range	-0.04	*-0.07*	0	0.04	-0.02
Mean Rate Change	*-0.08*	*-0.07*	0.03	0.01	-0.03
Length	*0.12*	*0.06*	*0.08*	*0.12*	*0.09*

Table C.66: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.05*	-0.04	-0.02	0.04	0.01
Std. Dev.	*0.05*	*-0.05*	*-0.05*	*0.06*	0
Mean Rate Change	-0.04	*0.05*	0	*0.07*	0.02
Length	*0.12*	*0.06*	*0.08*	*0.12*	*0.09*

Table C.67: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Energy.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.06*	*0.05*	-0.04	*0.13*	*0.05*
Std. Dev.	-0.04	0.03	*-0.09*	0.03	-0.02
Min.	0.01	*-0.07*	0.03	0.02	0
Max.	0.03	0.04	0.02	*0.1*	*0.05*
Range	-0.02	0.01	*-0.05*	-0.01	-0.02
Mean Rate Change	*-0.09*	*0.06*	-0.03	-0.02	-0.02
Length	*0.06*	0.04	*0.05*	*0.05*	*0.05*

Table C.68: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.07*	-0.01	0.02	0.02	0.02
Std. Dev.	-0.01	-0.03	*-0.05*	0.02	-0.02
Mean Rate Change	0.01	*0.07*	0.02	*-0.05*	0.01
Length	*0.06*	0.04	*0.05*	*0.05*	*0.05*

Table C.69: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.09*	0.04	*-0.05*	*0.13*	*0.05*
Std. Dev.	*-0.06*	0.03	*-0.1*	*0.05*	-0.02
Min.	0.01	*-0.07*	0.03	0.02	0
Max.	0.03	0.04	0.02	*0.1*	*0.05*
Range	-0.02	0.01	*-0.05*	-0.01	-0.02
Mean Rate Change	*-0.08*	0.04	-0.03	-0.01	-0.02
Length	0.02	*0.05*	*0.1*	*0.08*	*0.06*

Table C.70: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.05*	-0.02	0.02	0.01	0.02
Std. Dev.	*-0.07*	-0.02	*-0.1*	0.04	-0.04
Mean Rate Change	-0.01	*0.07*	0.02	*-0.08*	0
Length	0.02	*0.05*	*0.1*	*0.08*	*0.06*

Table C.71: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.04	*0.08*	0	*0.14*	*0.06*
Std. Dev.	0	0.04	-0.02	-0.02	0
Min.	*0.07*	-0.02	*0.05*	0.03	0.03
Max.	0.03	0.04	0.01	*0.09*	0.04
Range	0	-0.02	-0.02	*-0.08*	-0.03
Mean Rate Change	0.04	*0.05*	0.04	*-0.05*	0.02
Length	*0.06*	*0.05*	*0.06*	*0.05*	*0.05*

Table C.72: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.06*	*-0.09*	-0.01	0.02	-0.01
Std. Dev.	-0.03	*-0.09*	-0.03	-0.02	-0.04
Mean Rate Change	0.02	*0.1*	0	*-0.05*	0.02
Length	*0.06*	*0.05*	*0.06*	*0.05*	*0.05*

Table C.73: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Log-Concatenated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.08*	*0.08*	-0.01	*0.12*	*0.07*
Std. Dev.	-0.02	0.02	-0.03	-0.01	-0.01
Min.	*0.06*	-0.04	*0.05*	0.03	0.02
Max.	0.02	0.04	0	*0.09*	0.04
Range	-0.03	*-0.05*	-0.01	*-0.11*	*-0.05*
Mean Rate Change	-0.01	0.01	0.01	*-0.06*	-0.01
Length	0.02	*0.05*	*0.1*	*0.08*	*0.06*

Table C.74: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	0.02	*-0.07*	0	0	-0.01
Std. Dev.	-0.02	*-0.05*	*-0.06*	0.01	-0.03
Mean Rate Change	0	*0.07*	-0.01	*-0.08*	0
Length	0.02	*0.05*	*0.1*	*0.08*	*0.06*

Table C.75: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Log-Interpolated F_0.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.05*	0.04	*-0.08*	0.04	0.01
Std. Dev.	0.04	0.03	*-0.11*	0.03	0
Mean Rate Change	0.04	0.03	*-0.11*	0.03	0
Length	-0.02	0.04	0	0.03	0.01

Table C.76: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Voicing.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.08*	*0.12*	*0.07*	*0.11*	*0.09*
Std. Dev.	*0.06*	*0.07*	0.03	0.02	0.04
Min.	*-0.1*	-0.01	*-0.07*	0	*-0.05*
Max.	*0.08*	*0.11*	0.04	*0.09*	*0.08*
Range	*0.08*	*0.08*	0.03	0.03	*0.05*
Mean Rate Change	0.03	*0.06*	0	*-0.06*	0.01
Length	*0.12*	*0.06*	*0.08*	*0.12*	*0.09*

Table C.77: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Zero Crossings.

Measures	*Sentences*				
	1	2	3	4	Mean
Mean	*0.05*	0.02	0.01	0.02	0.02
Std. Dev.	*0.08*	-0.01	0.01	-0.02	0.01
Mean Rate Change	*0.05*	0.01	*0.1*	-0.02	0.04
Length	*0.12*	*0.06*	*0.08*	*0.12*	*0.09*

Table C.78: Speaker Dialect difference scores correlated with static measures. A quantification of each individual measure's ability to predict dialect. Parameter: Normalised Zero Crossings.

Appendix D

Sentence-Parameter Pairing Results

This appendix presents the discriminant rates for speaker identity and sex experiments, and the correlation values for least-squares-fit analysis of speaker dialect, where individual sentence and parameter combinations are examined. The four basic parameters: energy, F_0 (all 4 versions), voicing, and zero crossing rate, both normalised and un-normalised, are examined on an individual sentence basis where static measures alone, dynamic measures alone, and combined static and dynamic measures are used in the discriminant or least-squares-fit analysis.

Results are subdivided on the basis of the three speaker characteristics identity, sex, and dialect. Within each section a figure and table present the results of the analysis of a single parameter (either normalised or un-normalised). The figure is composed of 12 individual plots: 4 sentences $\times$ 3 measure groupings. Similarly the table is 4 $\times$ 3 presenting the same data in a numeric format.

D.1 Speaker Identity

Figures D.1–D.13 and Tables D.1–D.13 present the results of the analysis of parameters on an individual sentence basis for speaker identity discrimination. All tabular values are percentages.

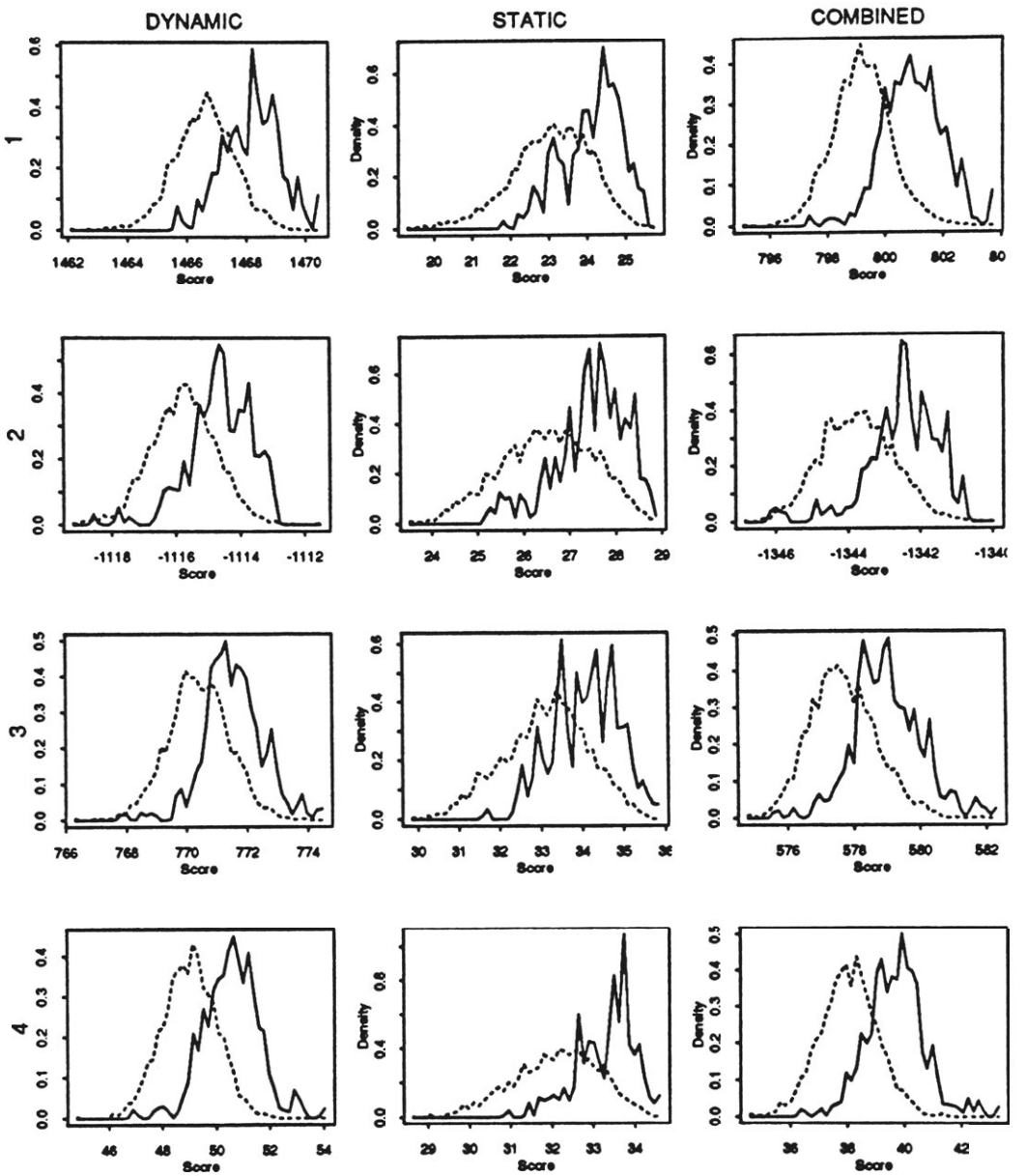


Figure D.1: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Energy* for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	39.3	27.4	43.3
2	29.5	24.8	32
3	30.2	25.2	34.2
4	39.1	30.5	40.7

Table D.1: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Energy* for each of the 4 sentences in turn. Values are in percent.

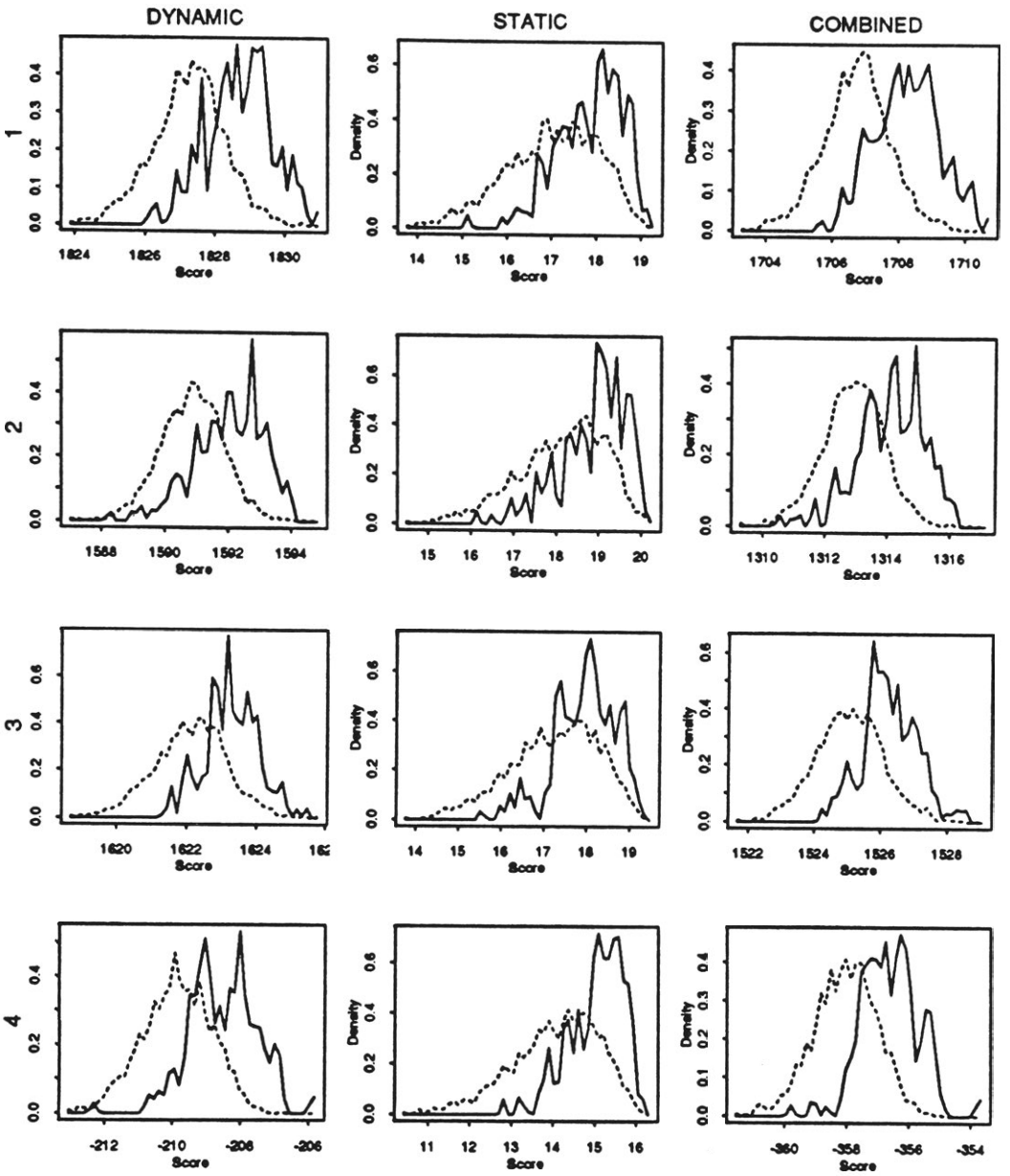


Figure D.2: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Energy* for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	36.3	21.9	37.8
2	29.6	19.4	30.6
3	29.5	18.4	30.7
4	34.5	25.3	36

Table D.2: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Energy* for each of the 4 sentences in turn. Values are in percent.

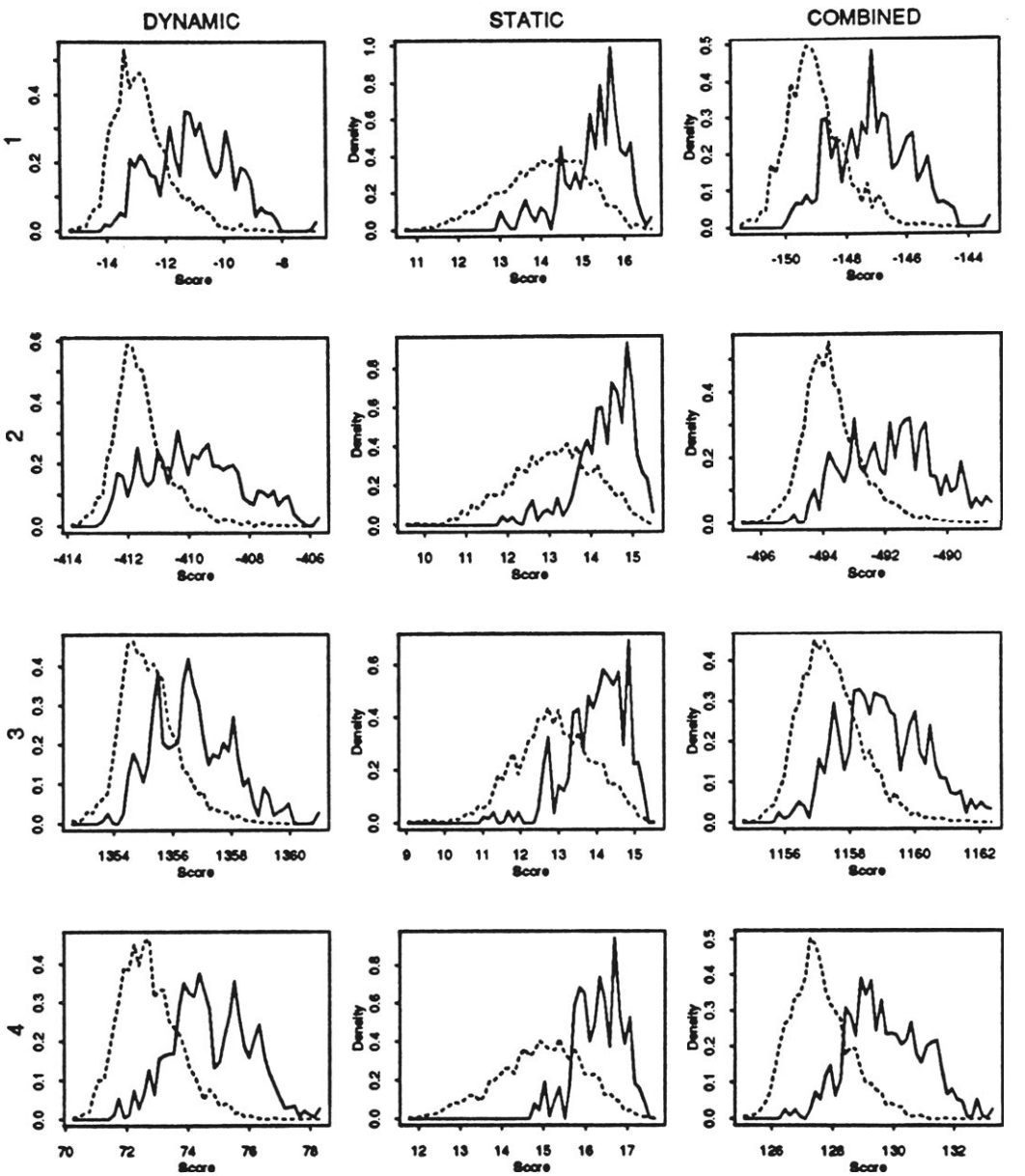


Figure D.3: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	42.3	30.8	44.8
2	42.3	32.5	47
3	36.8	29.1	39.2
4	47.8	35.5	50.1

Table D.3: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

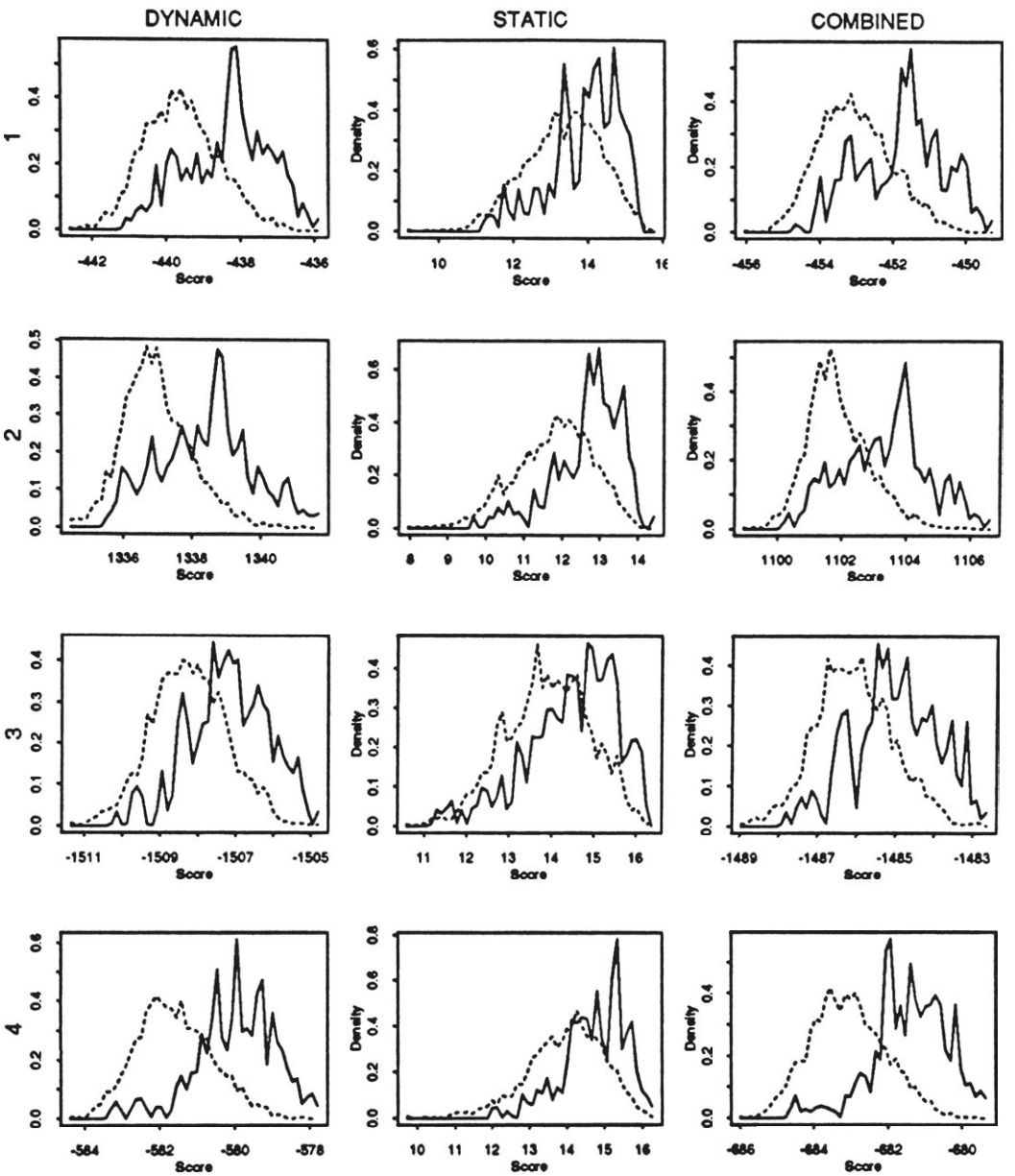


Figure D.4: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	31.8	16	32.4
2	36.1	23.9	37
3	27.1	15.3	27.5
4	40.6	20.3	41.1

Table D.4: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

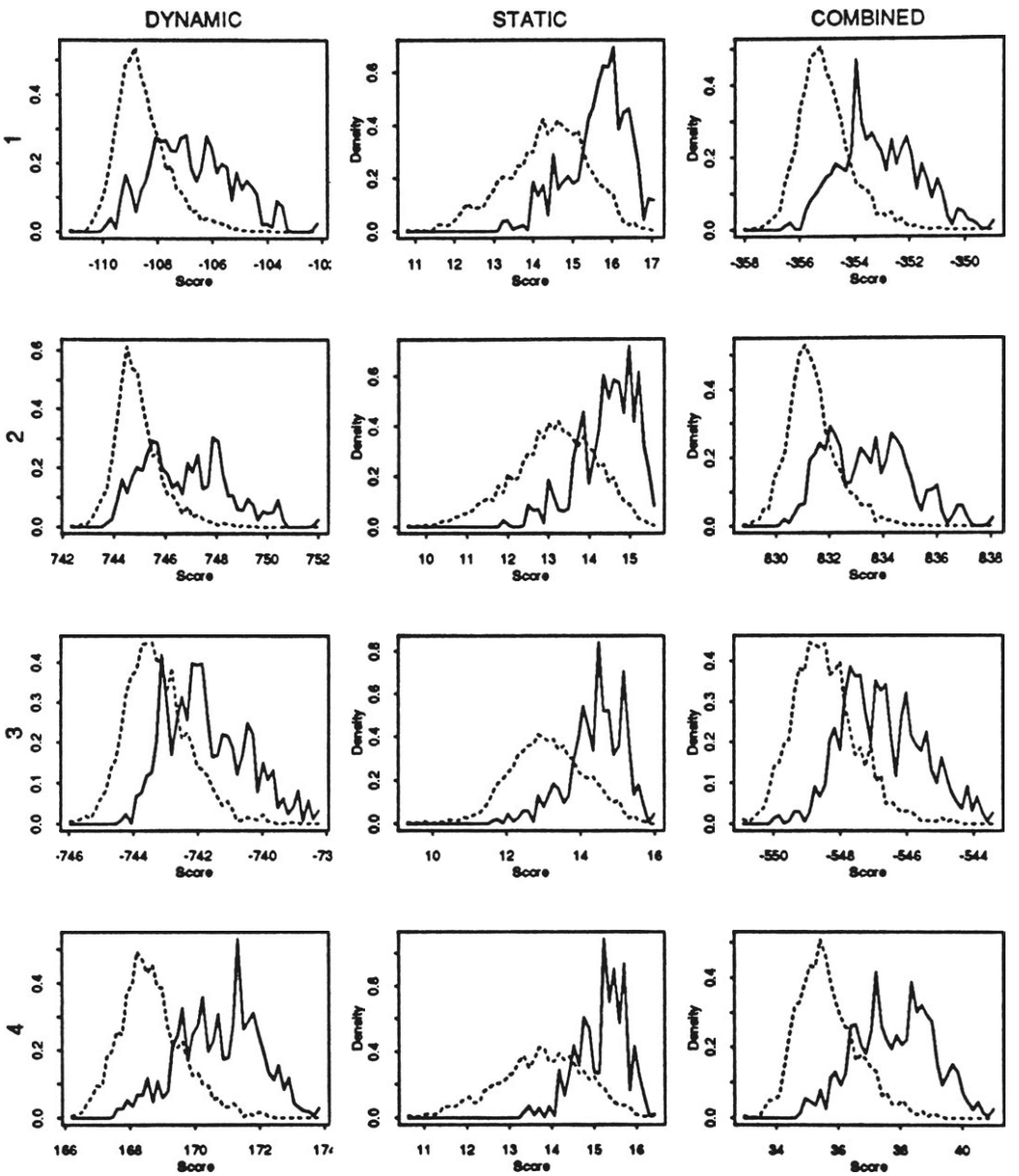


Figure D.5: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	43.1	32.7	46.7
2	44.5	33	48.8
3	38.4	33.2	42
4	49.1	36.1	50.8

Table D.5: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

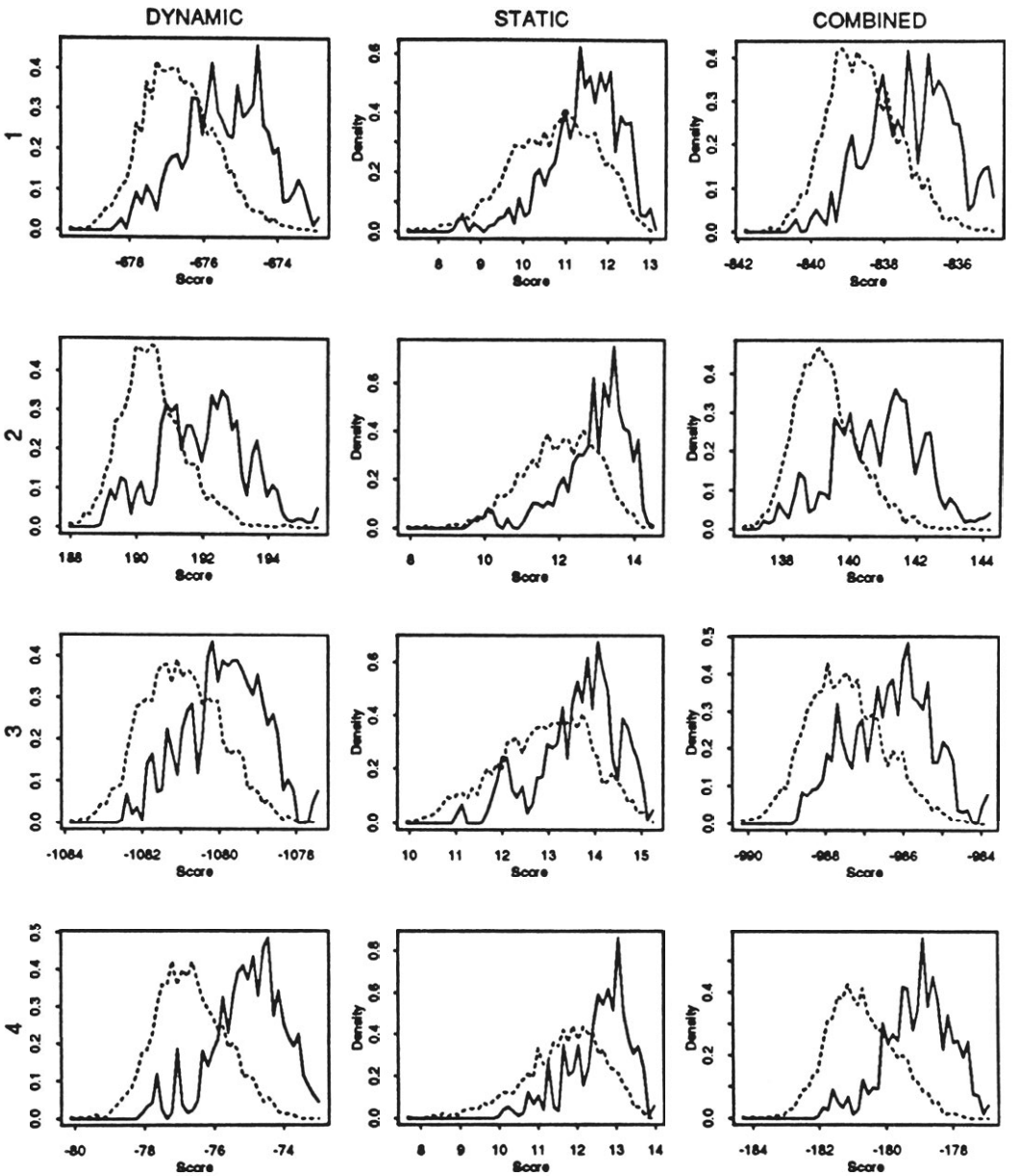


Figure D.6: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	31.6	18.6	32.8
2	37.8	23.6	38.5
3	28.1	19.3	28.7
4	41.3	23.4	42.2

Table D.6: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

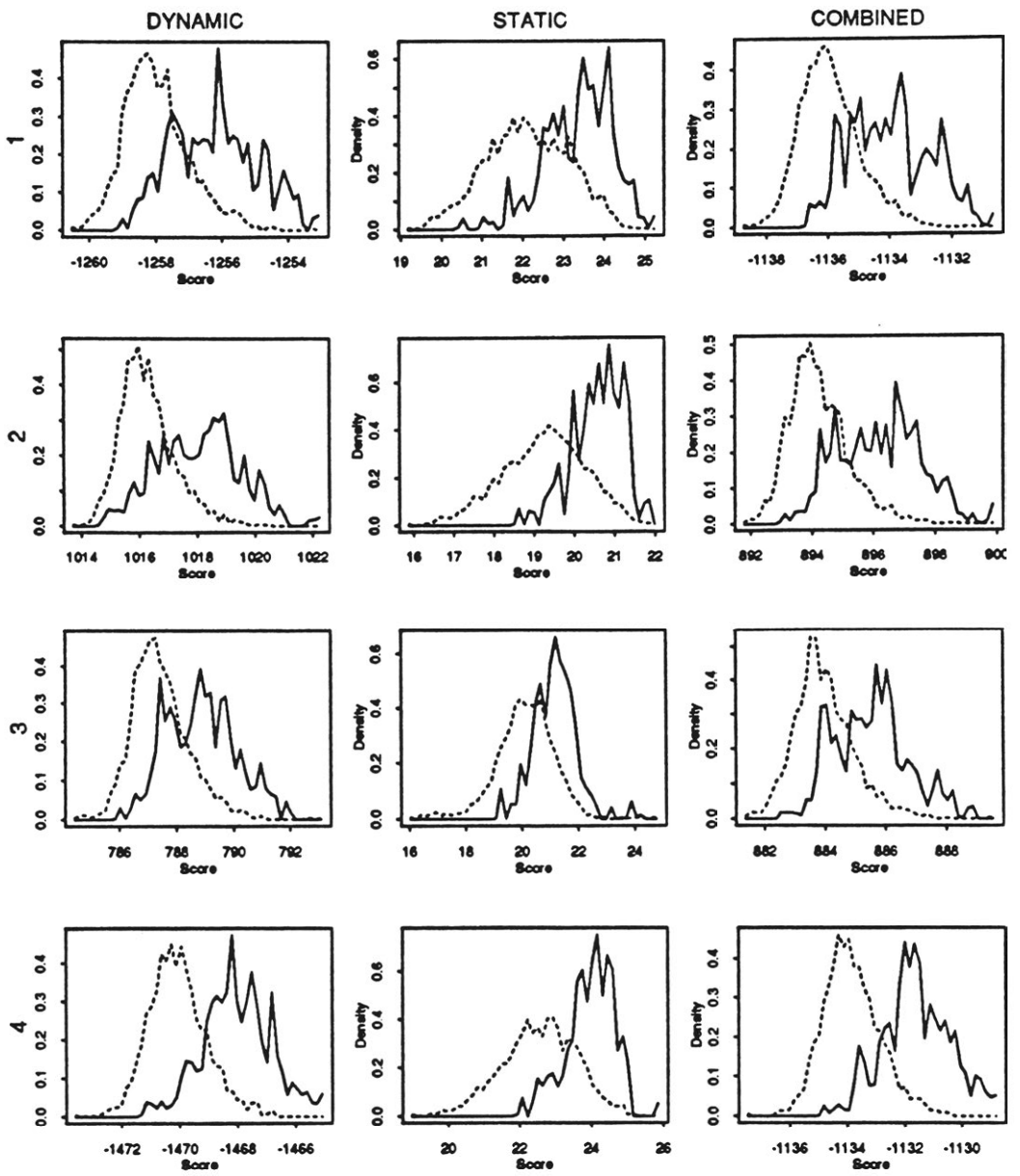


Figure D.7: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	42.5	32.3	46.7
2	42.9	34	47.4
3	36.4	27.4	38.6
4	48.8	35.8	52.1

Table D.7: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

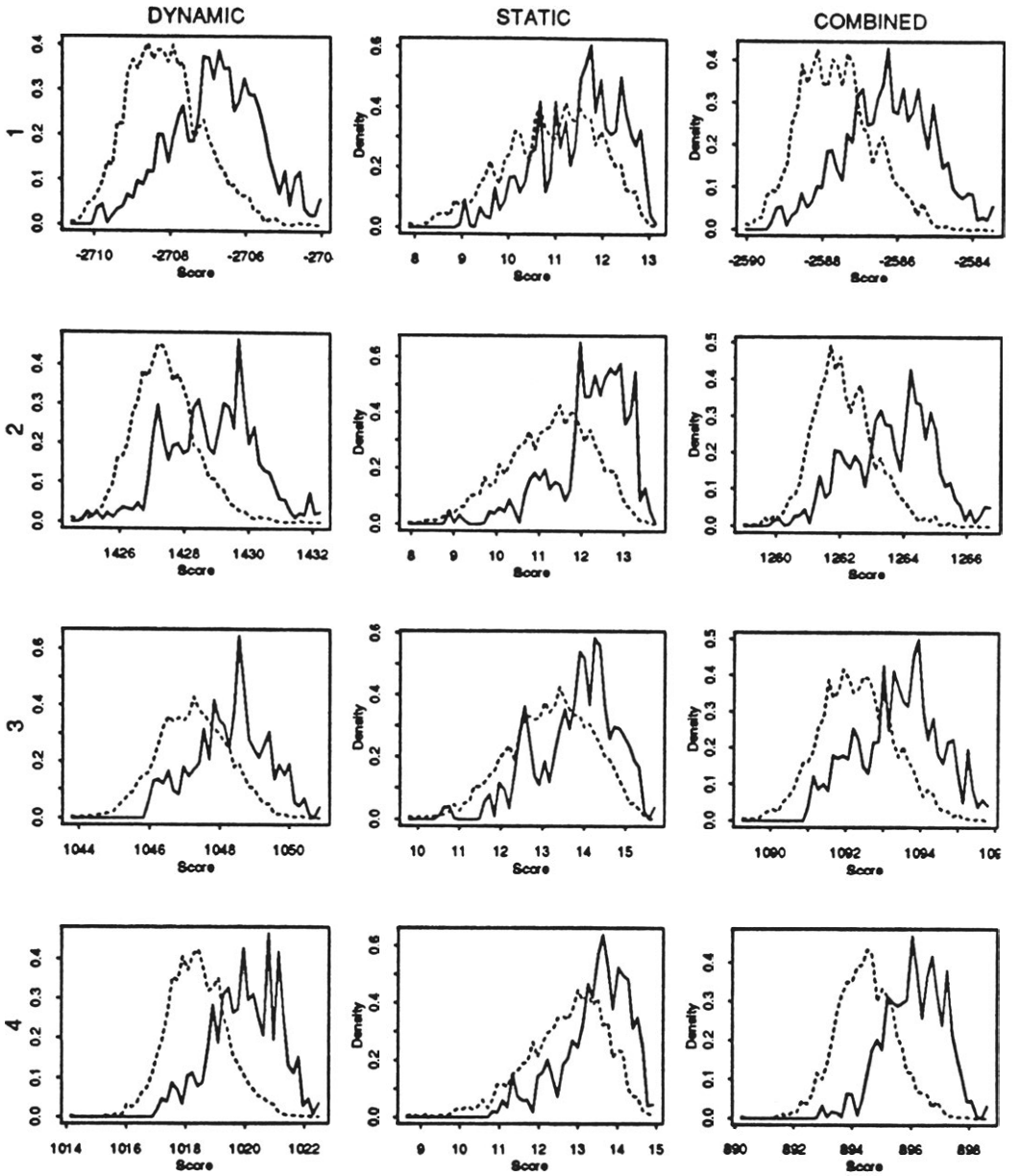


Figure D.8: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	34.2	15.9	34.8
2	36.4	25.2	38.1
3	27.7	16.1	28.4
4	39.2	19.2	40.5

Table D.8: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

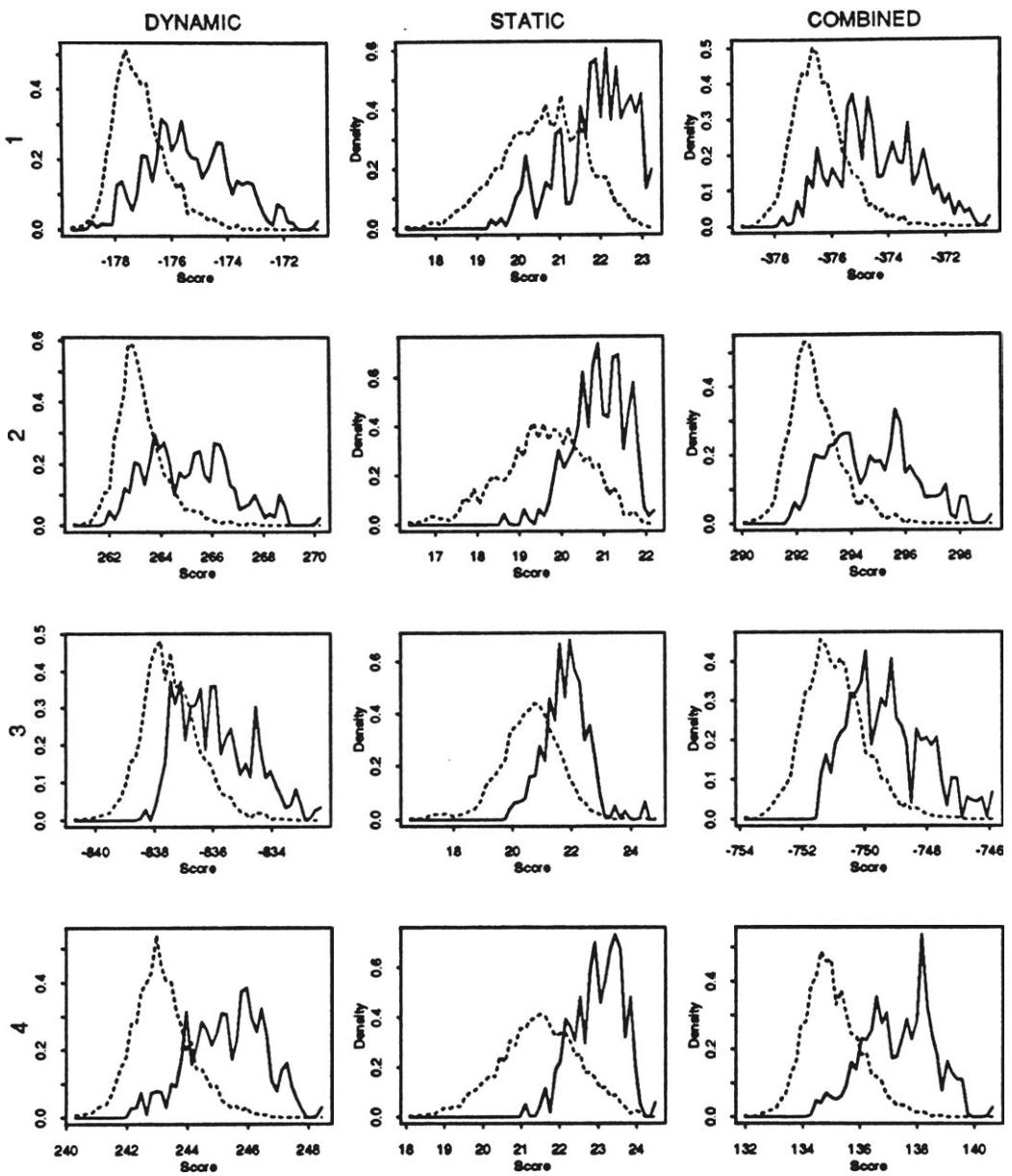


Figure D.9: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	43.1	32.5	46.6
2	45.2	32.9	48.9
3	38.1	30.8	41.2
4	49.9	38	53

Table D.9: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

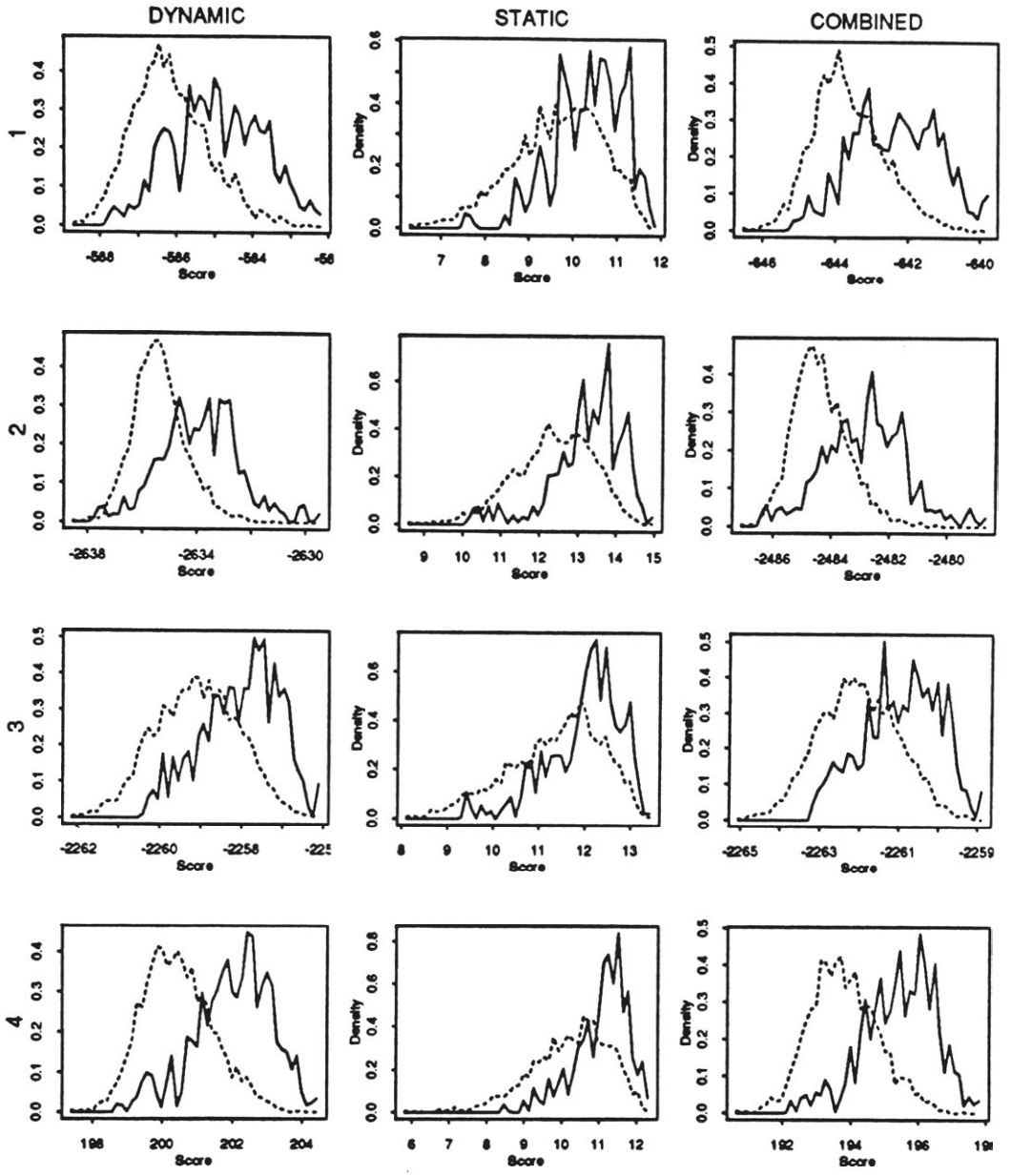


Figure D.10: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	31.9	18.4	33.3
2	37.4	23	38.3
3	27.5	16.7	27.8
4	39.1	21.5	40.1

Table D.10: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

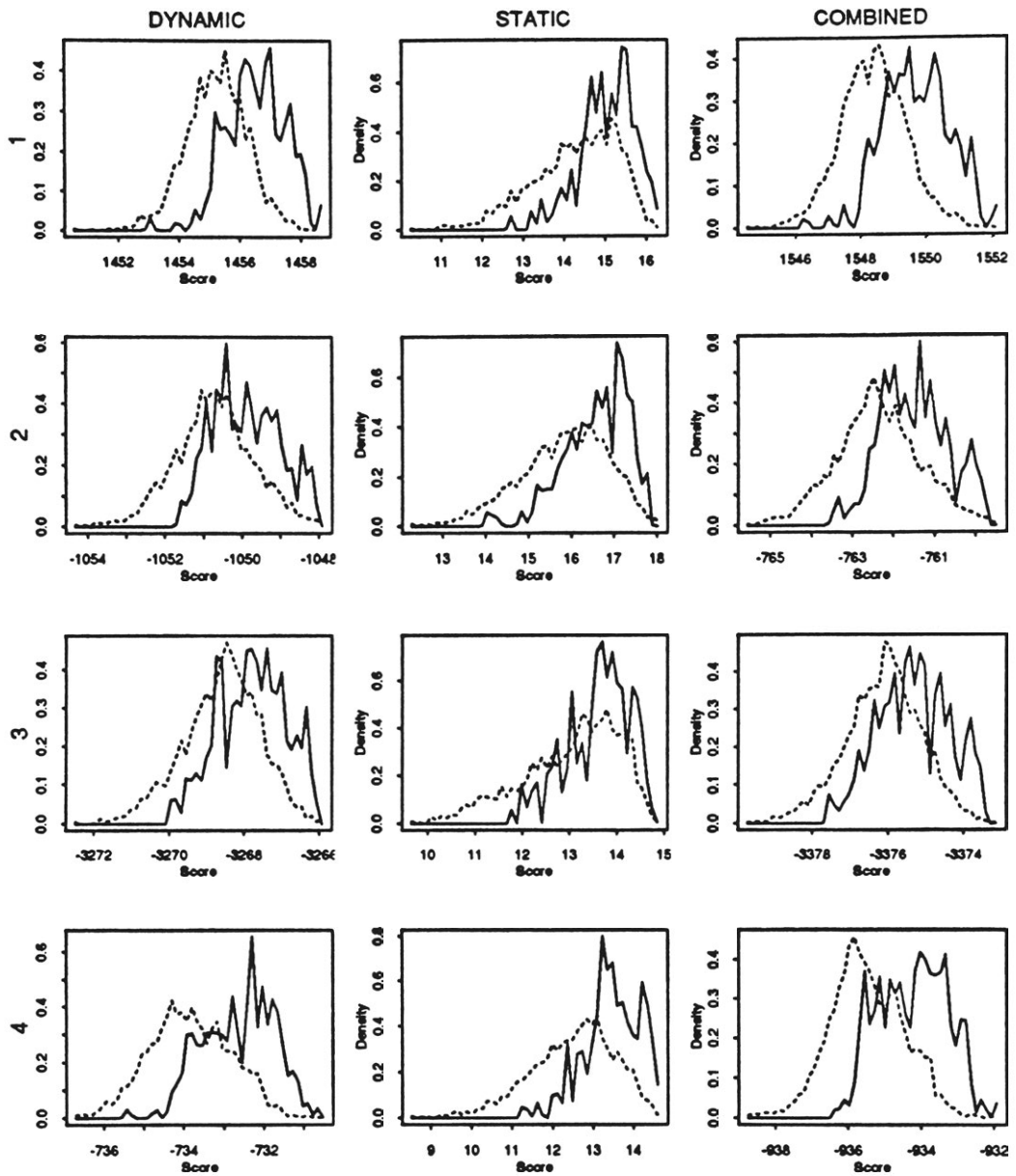


Figure D.11: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Voicing* for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	32.8	20.3	33.4
2	22.2	19.9	23.5
3	21	15.3	21.5
4	31	24.4	32

Table D.11: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Voicing* for each of the 4 sentences in turn. Values are in percent.

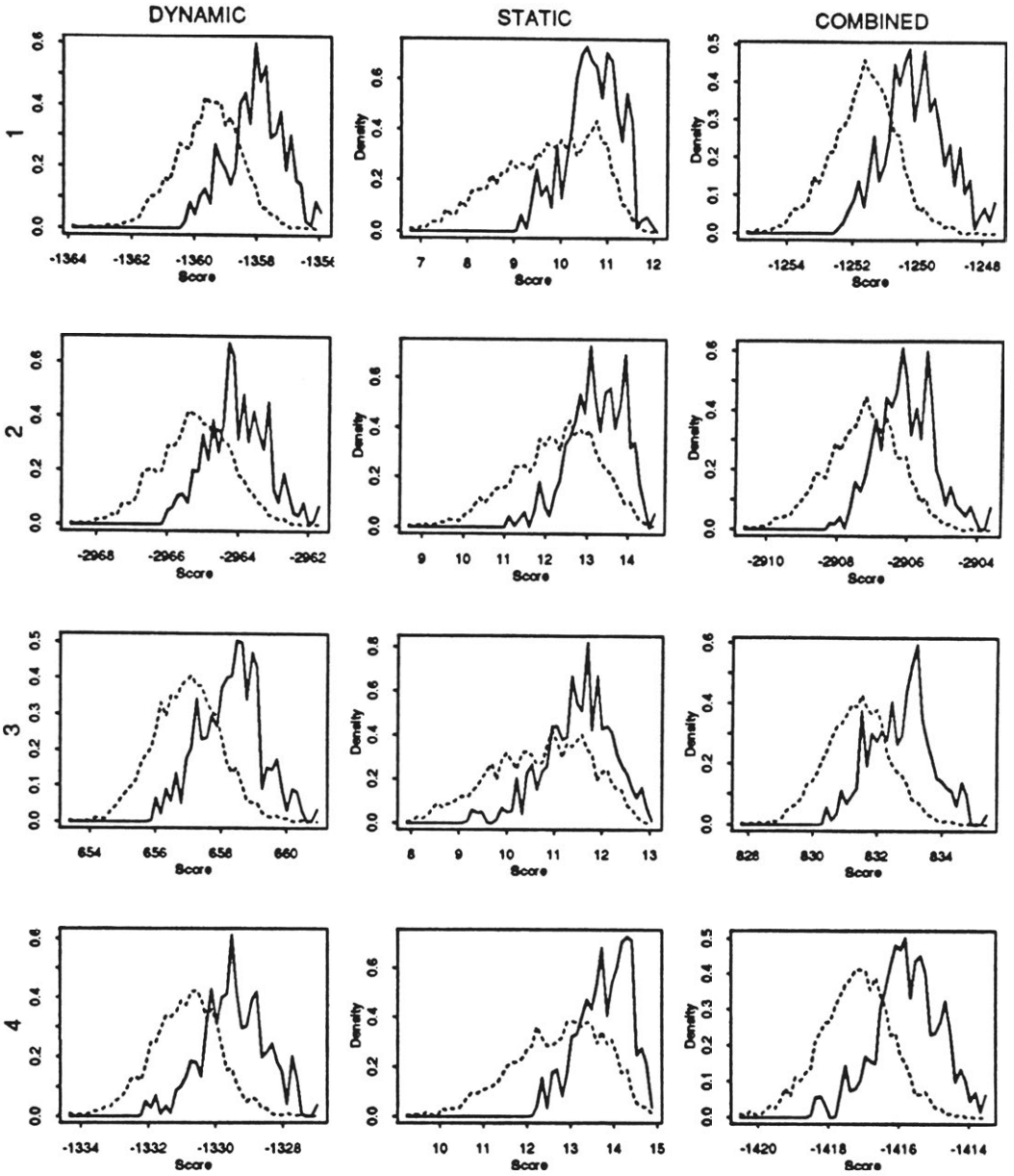


Figure D.12: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	38.1	25	39.4
2	30.6	25.6	31.1
3	34.3	20.6	34.6
4	36.4	27.4	36.9

Table D.12: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn. Values are in percent.

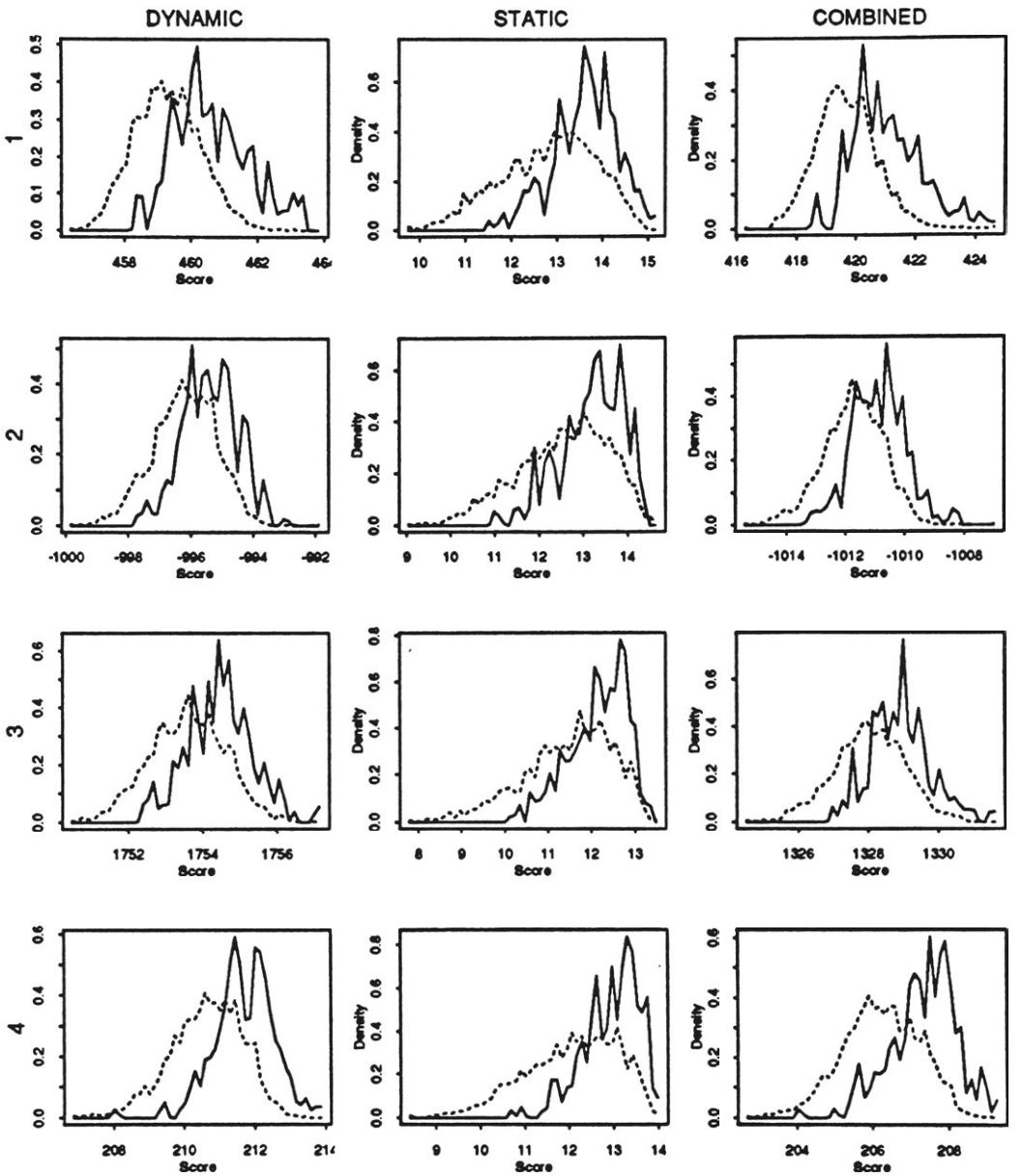


Figure D.13: Speaker Identity Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn. Intra-speaker (broken line) distribution is plotted against Inter-speaker (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	33.9	21.2	35.3
2	21.7	17.3	22.4
3	22.6	18.5	23.3
4	27.4	23	29.2

Table D.13: Speaker Identity Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn. Values are in percent.

D.2 Speaker Sex

Figures D.14–D.26 and Tables D.14–D.26 present the results of the analysis of parameters on an individual sentence basis for speaker sex discrimination. All tabular values are percentages.

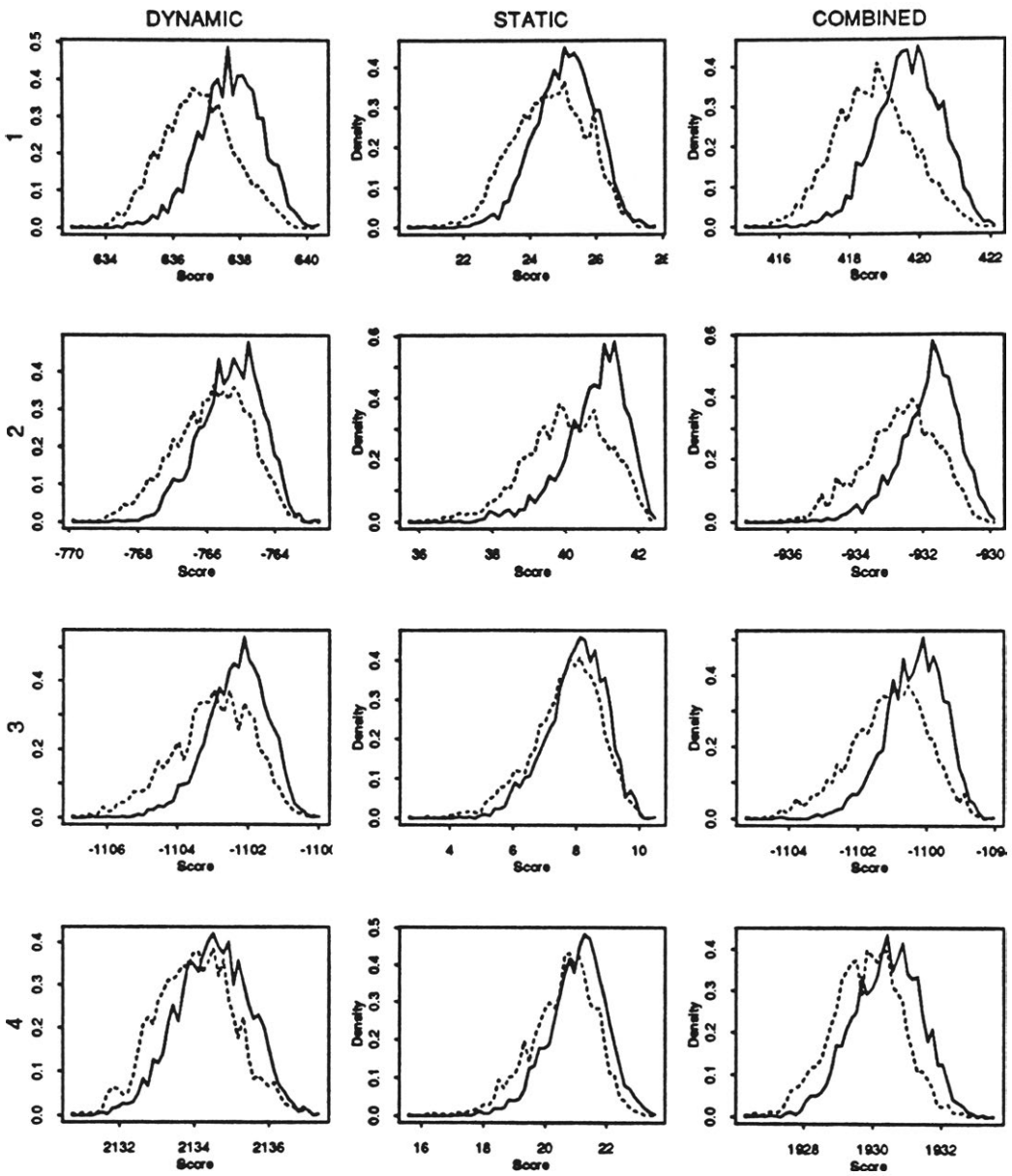


Figure D.14: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Energy* for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	42.8	23.1	44.1
2	27.3	35.6	39.7
3	32.7	15.6	35.5
4	23.1	23.1	28.2

Table D.14: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Energy* for each of the 4 sentences in turn. Values are in percent.

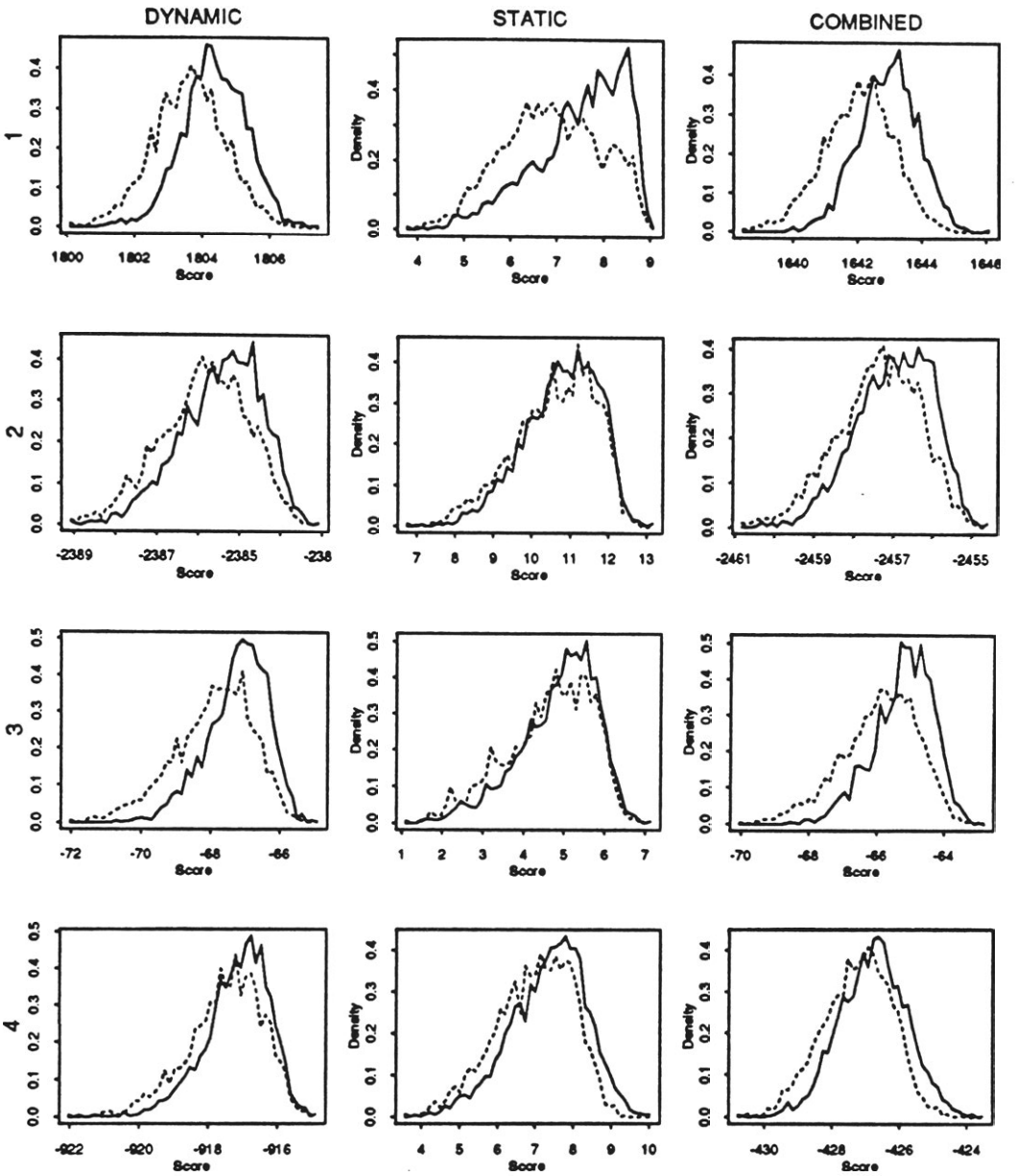


Figure D.15: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Energy* for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	35.8	28.8	41.7
2	19.8	8.6	20.4
3	32.5	11.4	32.7
4	16.6	18.6	22.6

Table D.15: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Energy* for each of the 4 sentences in turn. Values are in percent.

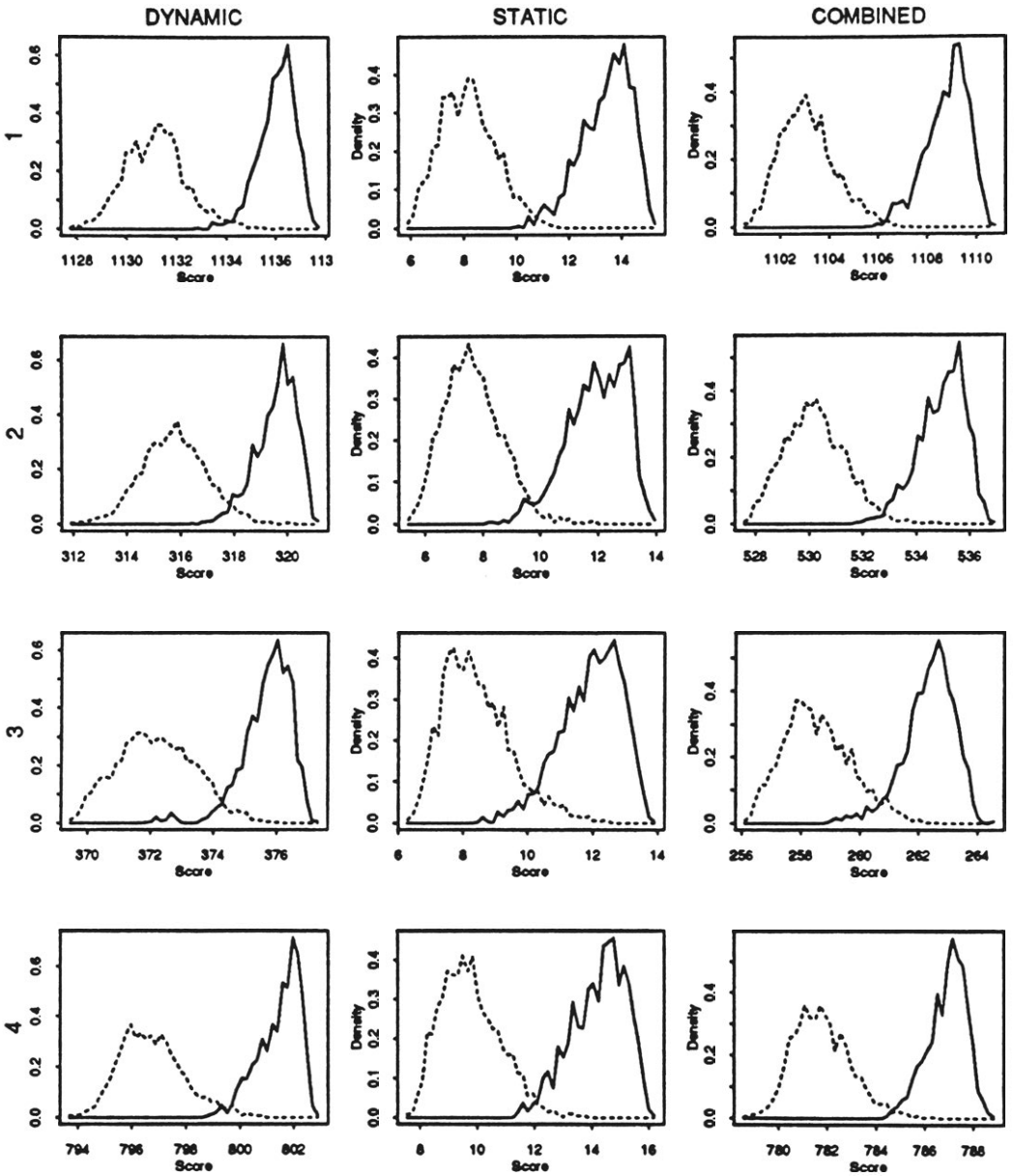


Figure D.16: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	92.5	93.3	94.2
2	88.6	90.6	92.1
3	87	86.9	88.3
4	92	91.3	93.1

Table D.16: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

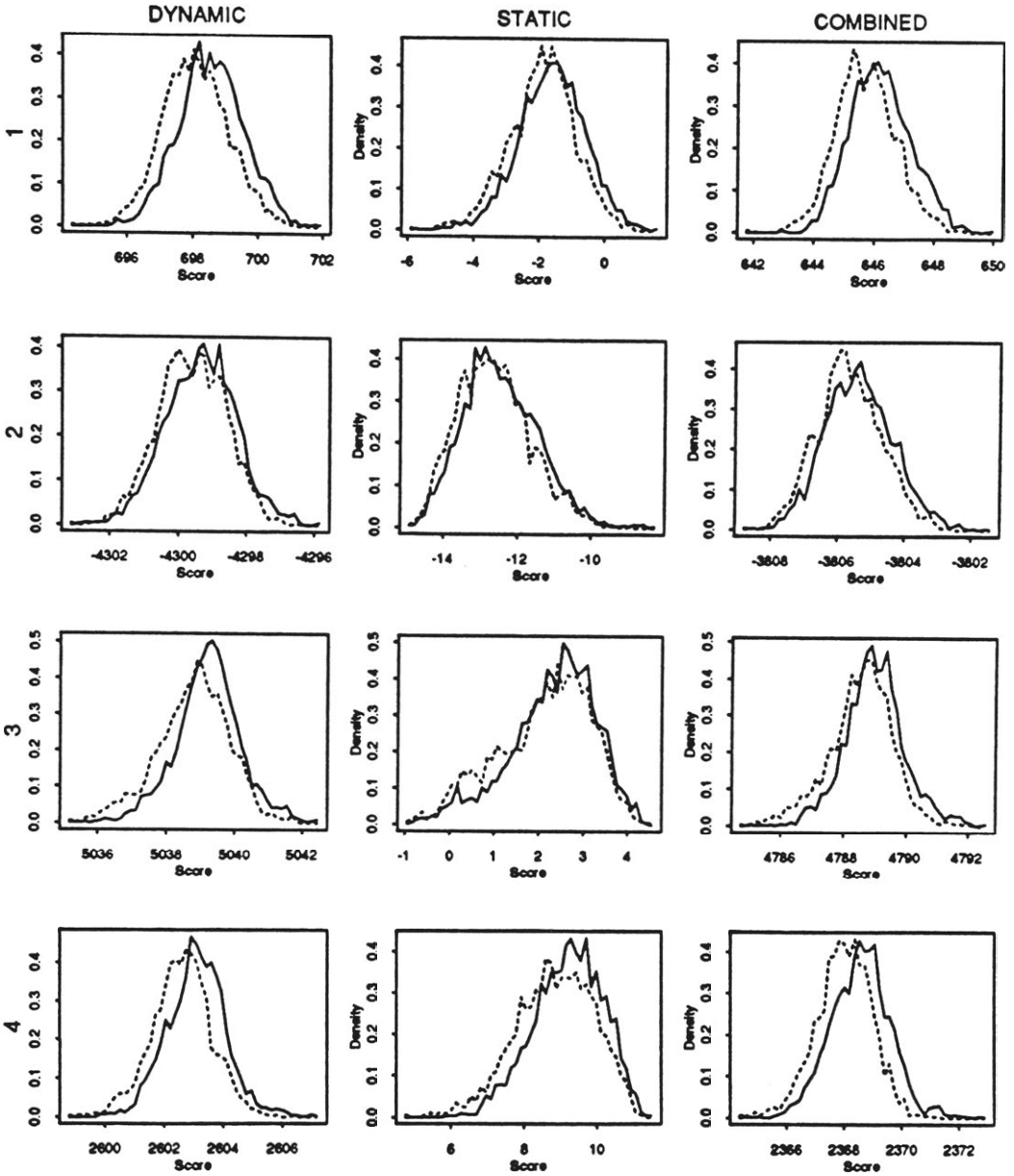


Figure D.17: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	23.3	15.1	24.3
2	12.8	9.2	14.3
3	20.8	10.3	22.1
4	24.3	18	28.1

Table D.17: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

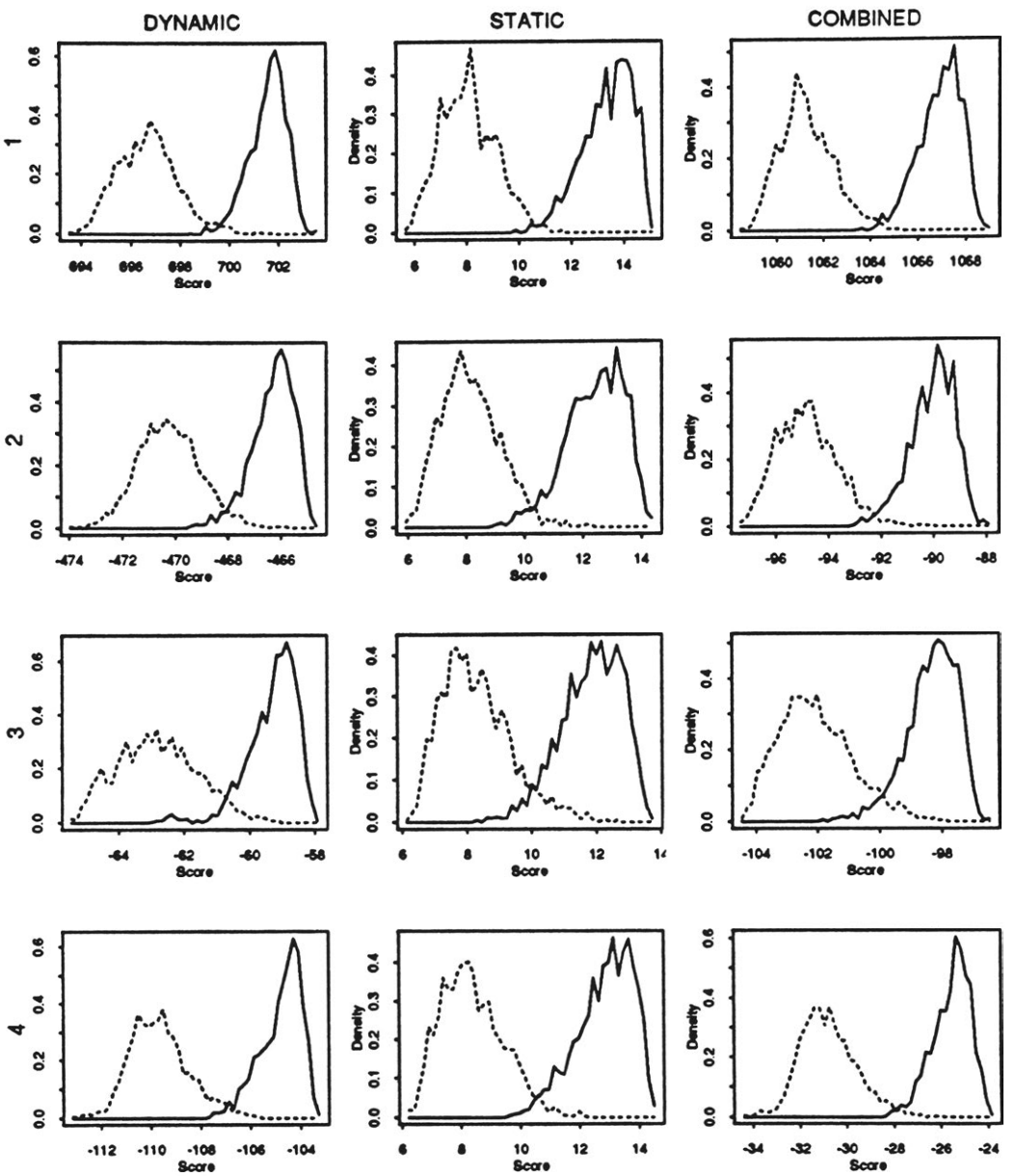


Figure D.18: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	92.5	93.6	94.2
2	89.1	90.5	92.2
3	86.9	87.2	88.3
4	92.6	91.2	93.5

Table D.18: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

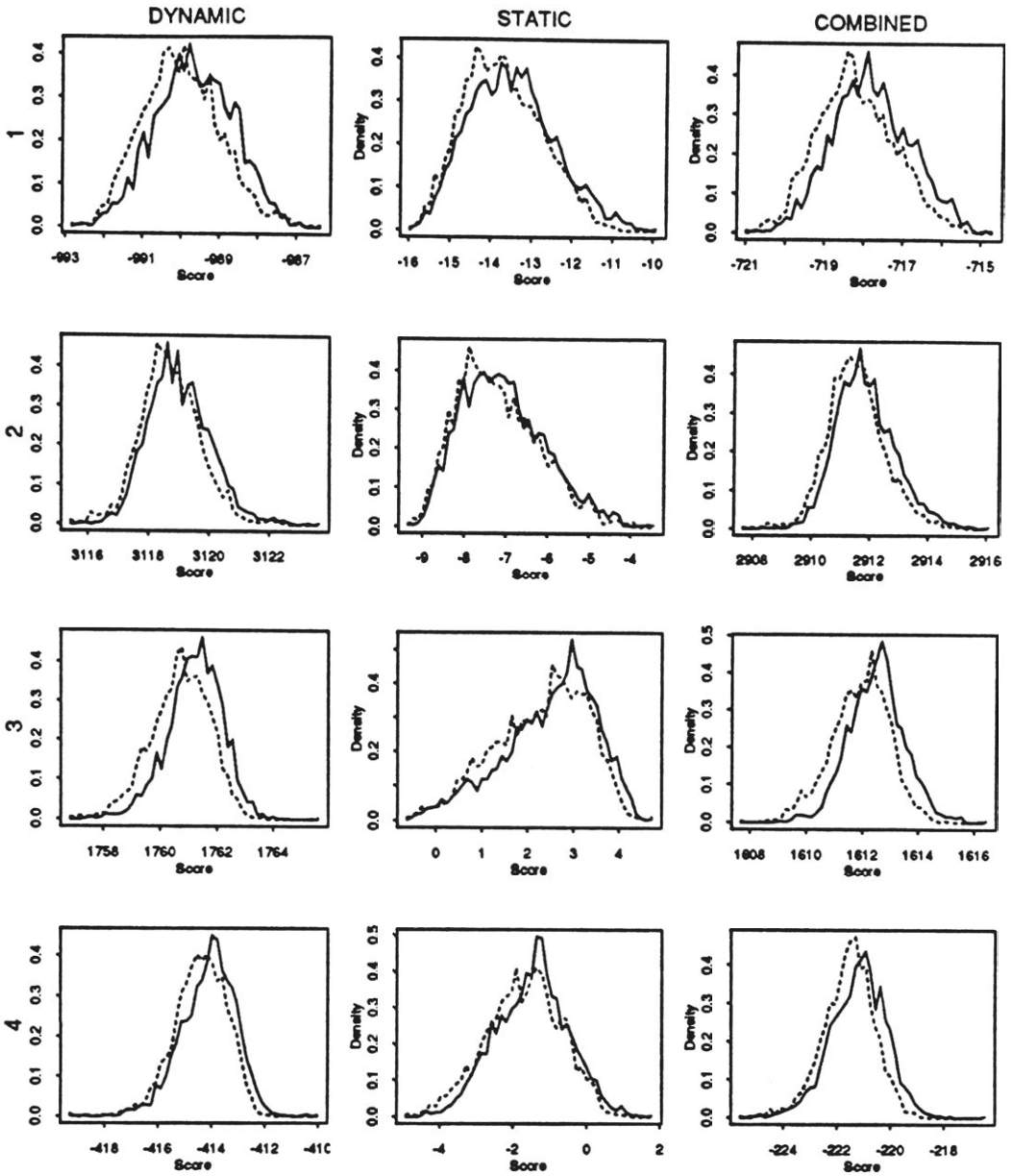


Figure D.19: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	16.7	11.1	18.9
2	15.5	6.7	16.6
3	24.6	10	27.5
4	14.8	11	18.1

Table D.19: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

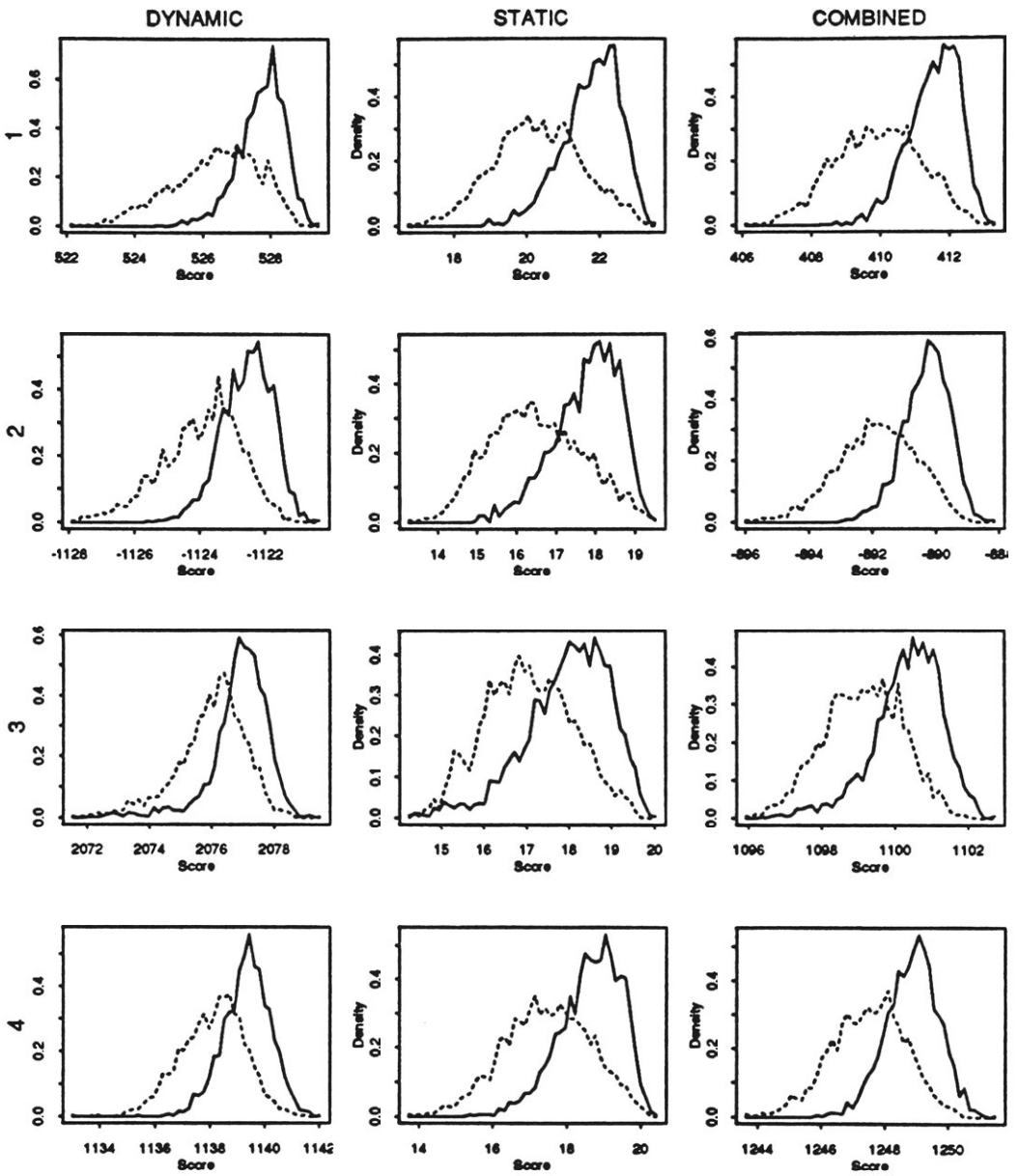


Figure D.20: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	55.8	59.1	63.1
2	57	53.9	61
3	41.5	41.7	51.8
4	52.6	50.7	56.9

Table D.20: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

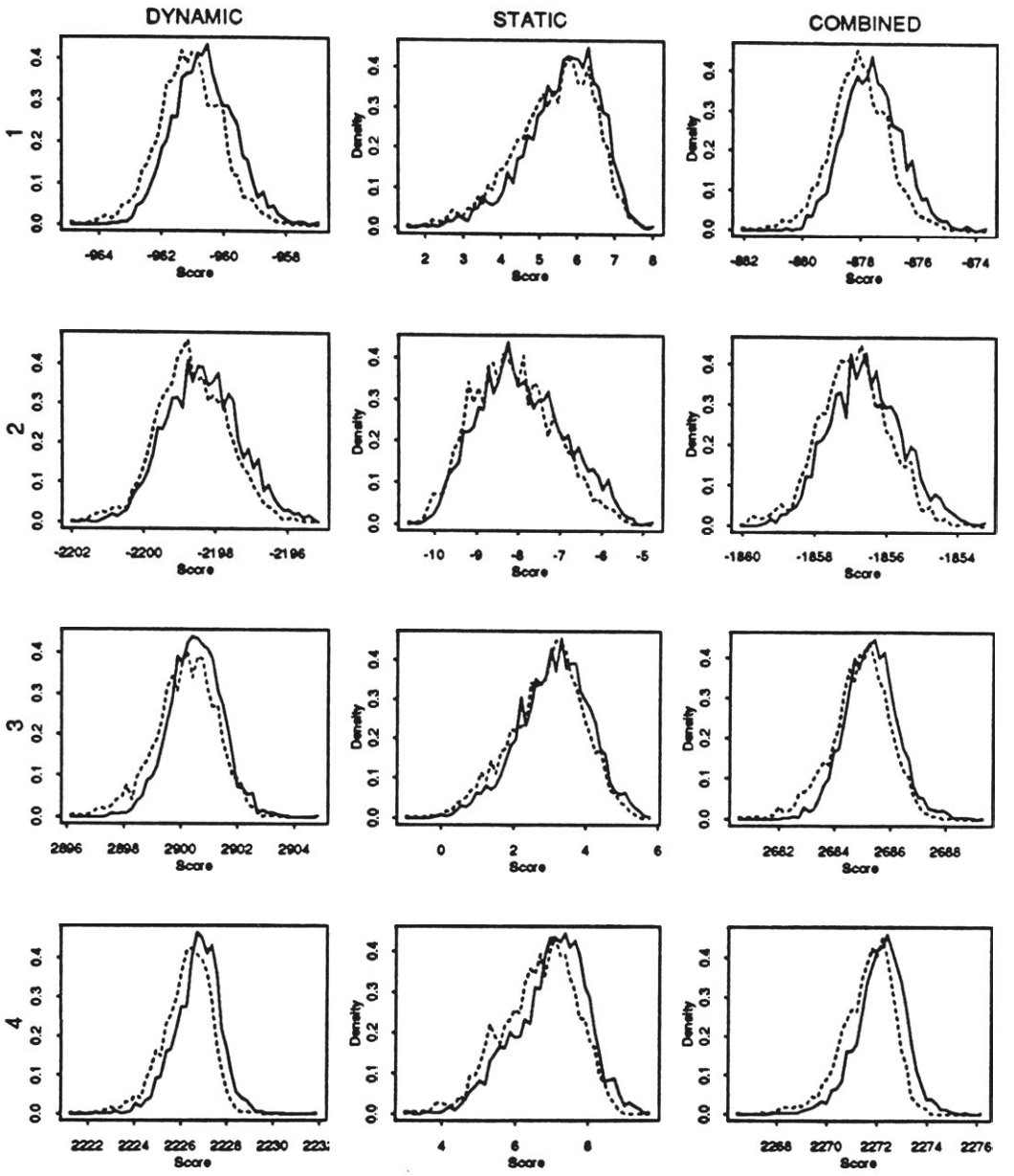


Figure D.21: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	22.3	10.8	22.5
2	14.9	10.4	17
3	17.4	9.6	19
4	23.8	16.6	25.3

Table D.21: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn. Values are in percent.

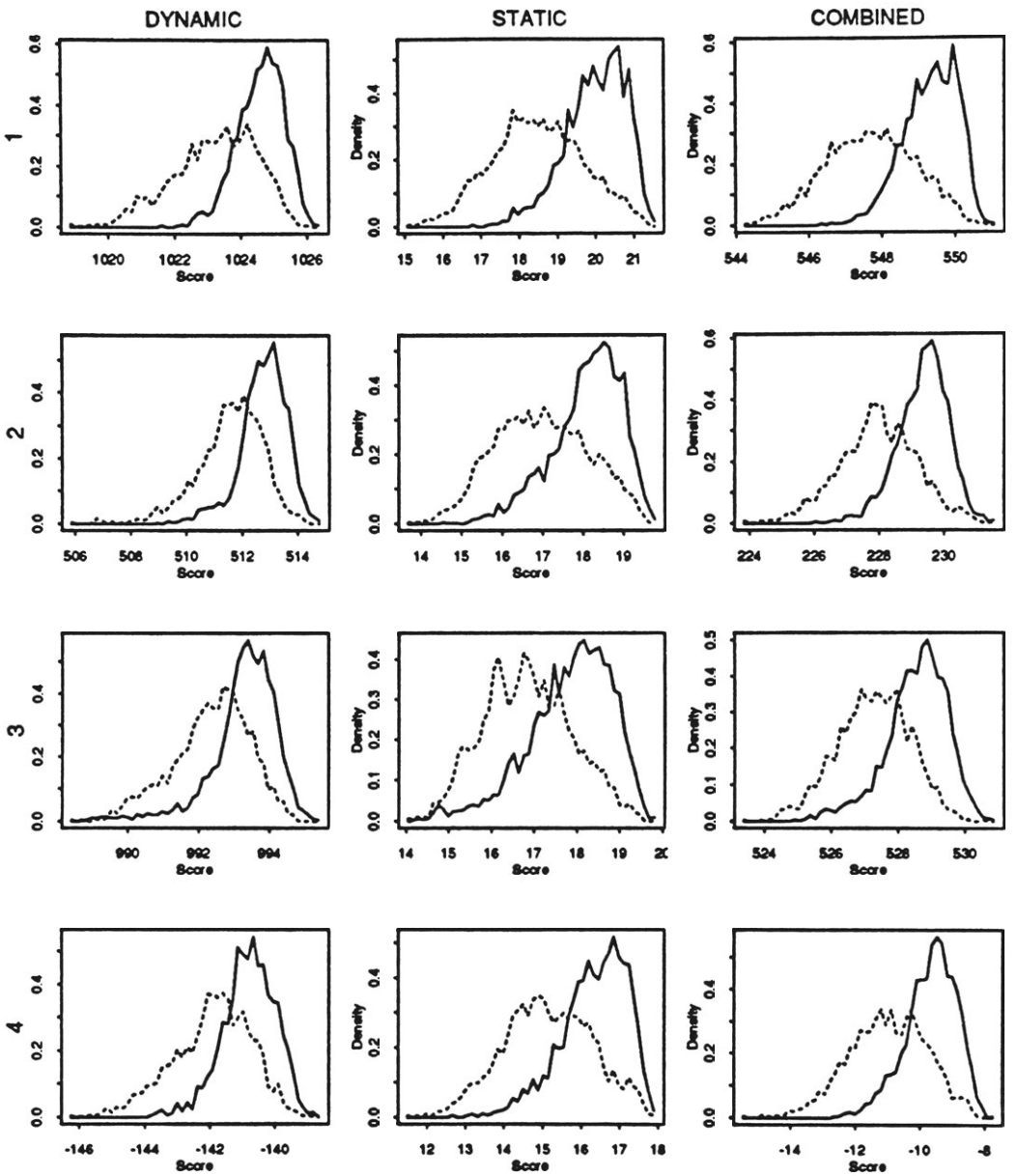


Figure D.22: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	56.4	60.1	63.9
2	51.4	48.5	57.1
3	41.7	44.8	54.3
4	49.1	51.4	54.8

Table D.22: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

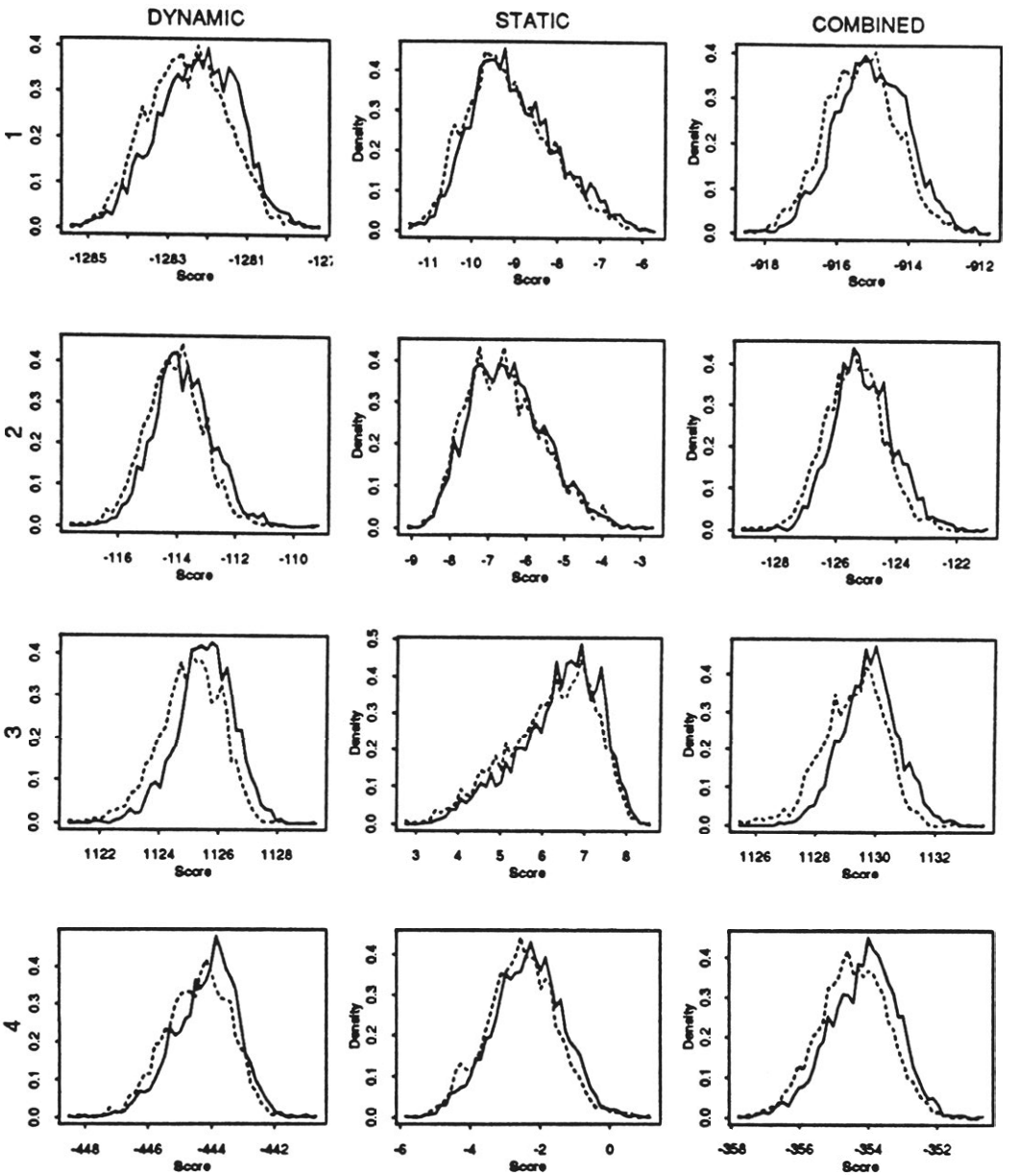


Figure D.23: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	13.5	6.9	15.2
2	15.9	4.1	16.2
3	24.2	9.9	26.9
4	13.4	9.3	15.3

Table D.23: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn. Values are in percent.

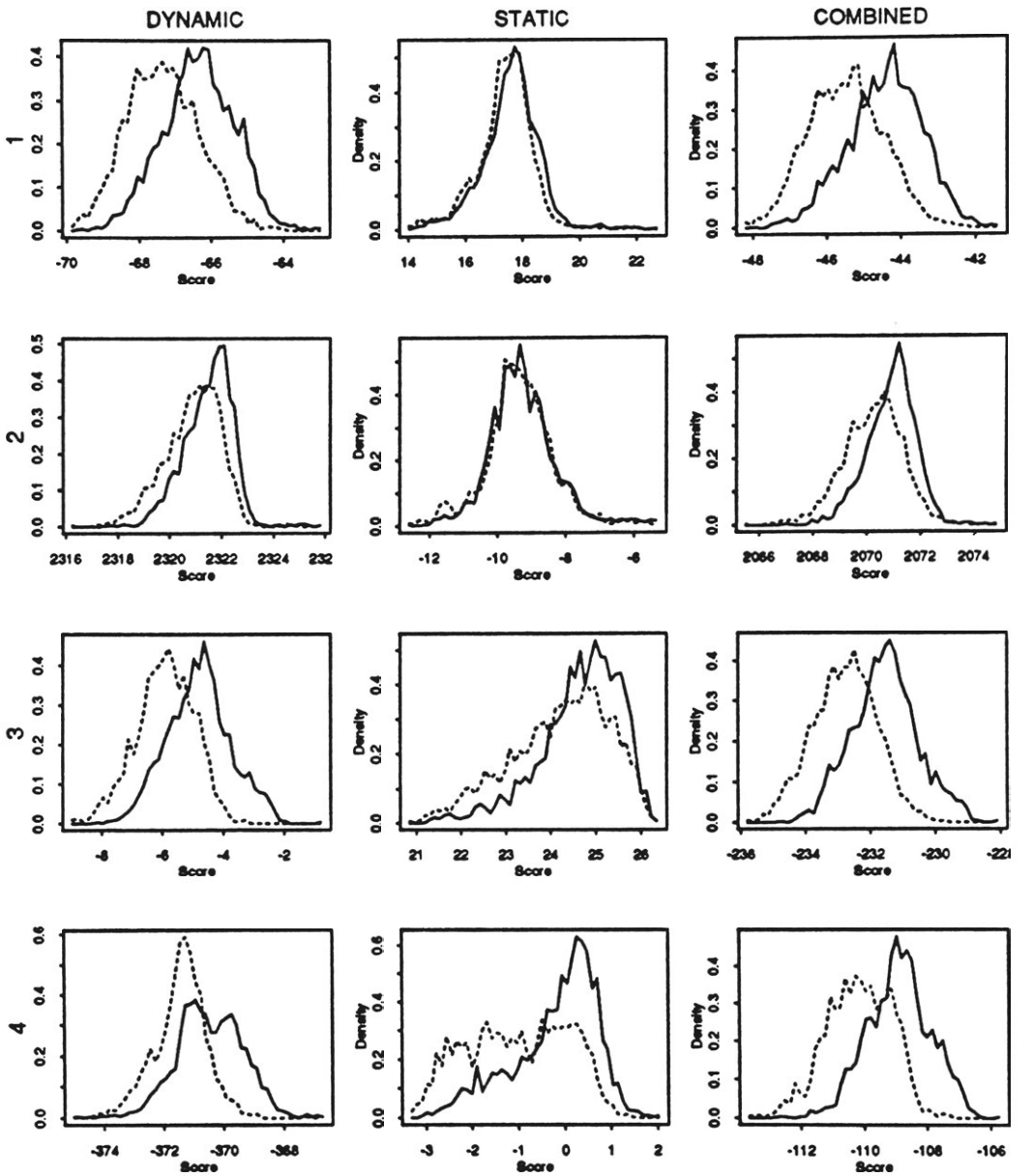


Figure D.24: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Voicing* for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	44.2	10.2	46.5
2	28.6	4.7	31.8
3	48.6	21.5	52.4
4	44.2	37.9	53.1

Table D.24: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Voicing* for each of the 4 sentences in turn. Values are in percent.

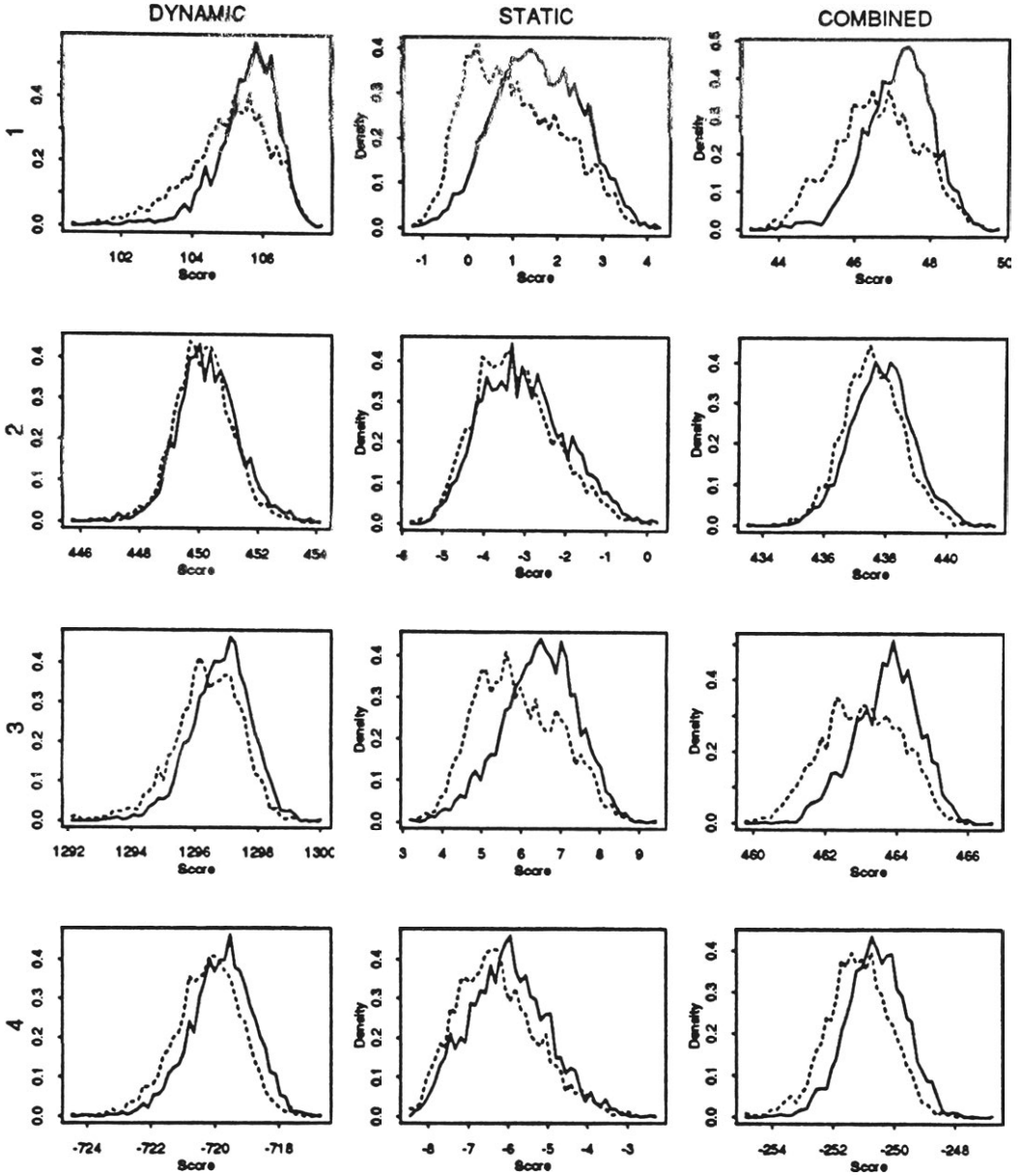


Figure D.25: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	24.3	24.4	27.8
2	9.2	10	14.8
3	21.3	26.1	32.3
4	20.7	13.5	28.1

Table D.25: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn. Values are in percent.

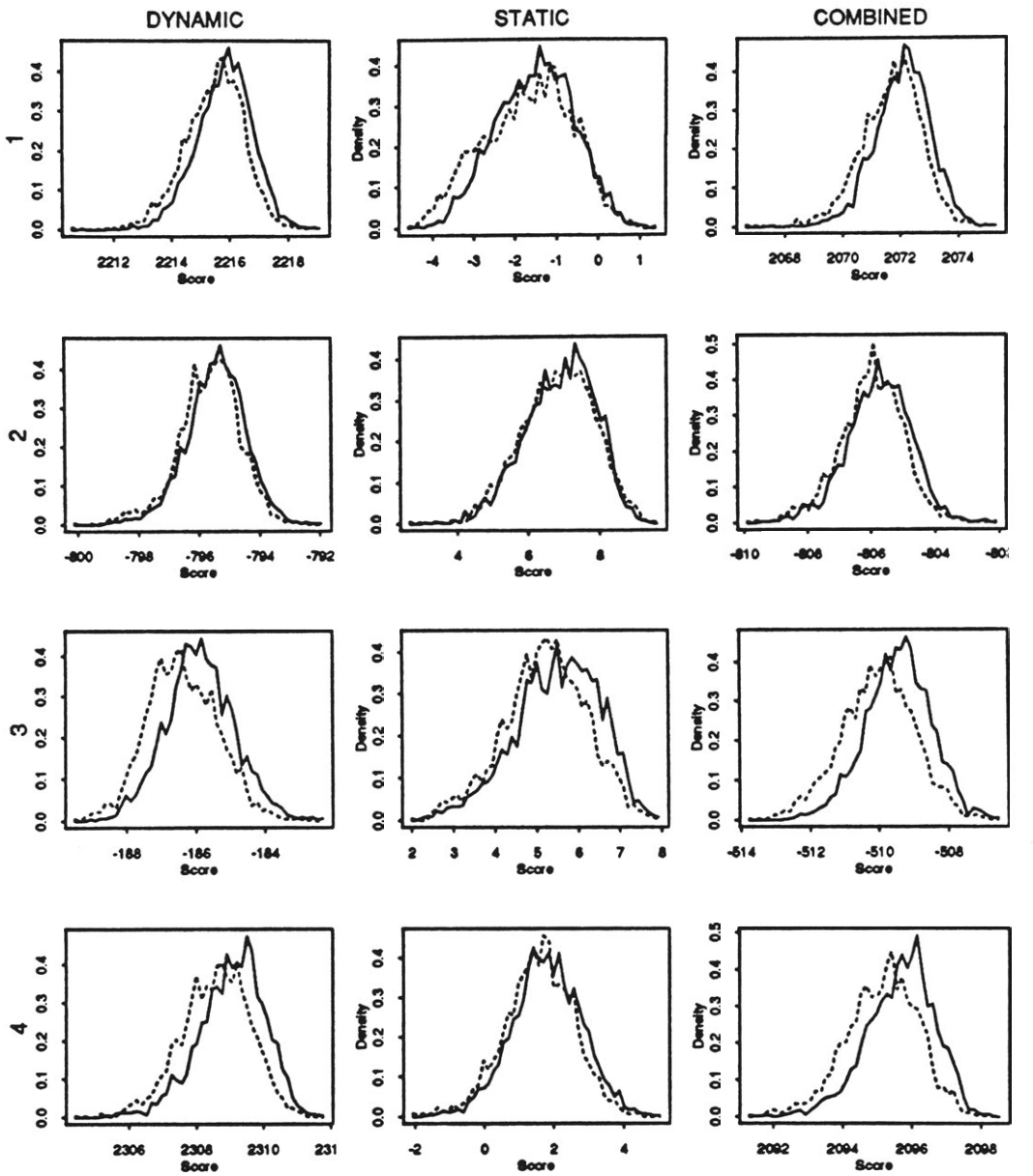


Figure D.26: Speaker Sex Discriminate Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn. Intra-sex (broken line) distribution is plotted against Inter-sex (unbroken line) distribution.

Sentence	*Measures*		
	Dynamic	Static	Both
1	17.2	8.7	18.8
2	10.8	2.6	11.8
3	24.4	14.4	28.2
4	23.1	11.5	25.7

Table D.26: Speaker Sex Discriminate Rates - Dynamic, Static, and Combined Dynamic-Static rates for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn. Values are in percent.

D.3 Speaker Dialect

Figures D.27–D.39 and Tables D.27–D.39 present the results of the analysis of parameters on an individual sentence basis for correlation to the dialect-difference score. All tabular values are correlations ranging from 0–1.

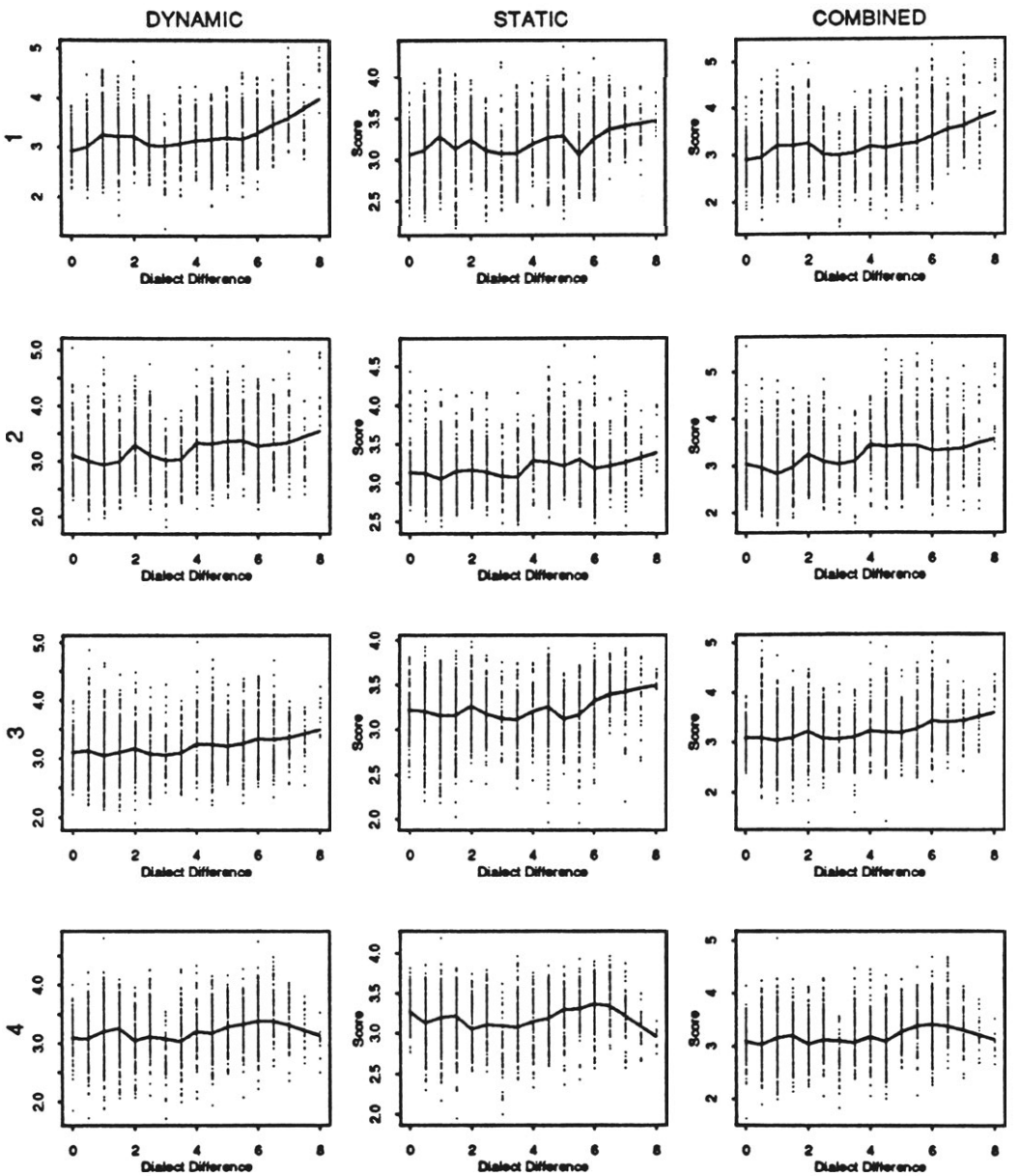


Figure D.27: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Energy* for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.222	0.161	0.258
2	0.237	0.156	0.288
3	0.2	0.148	0.231
4	0.19	0.141	0.202

Table D.27: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Energy* for each of the 4 sentences in turn.

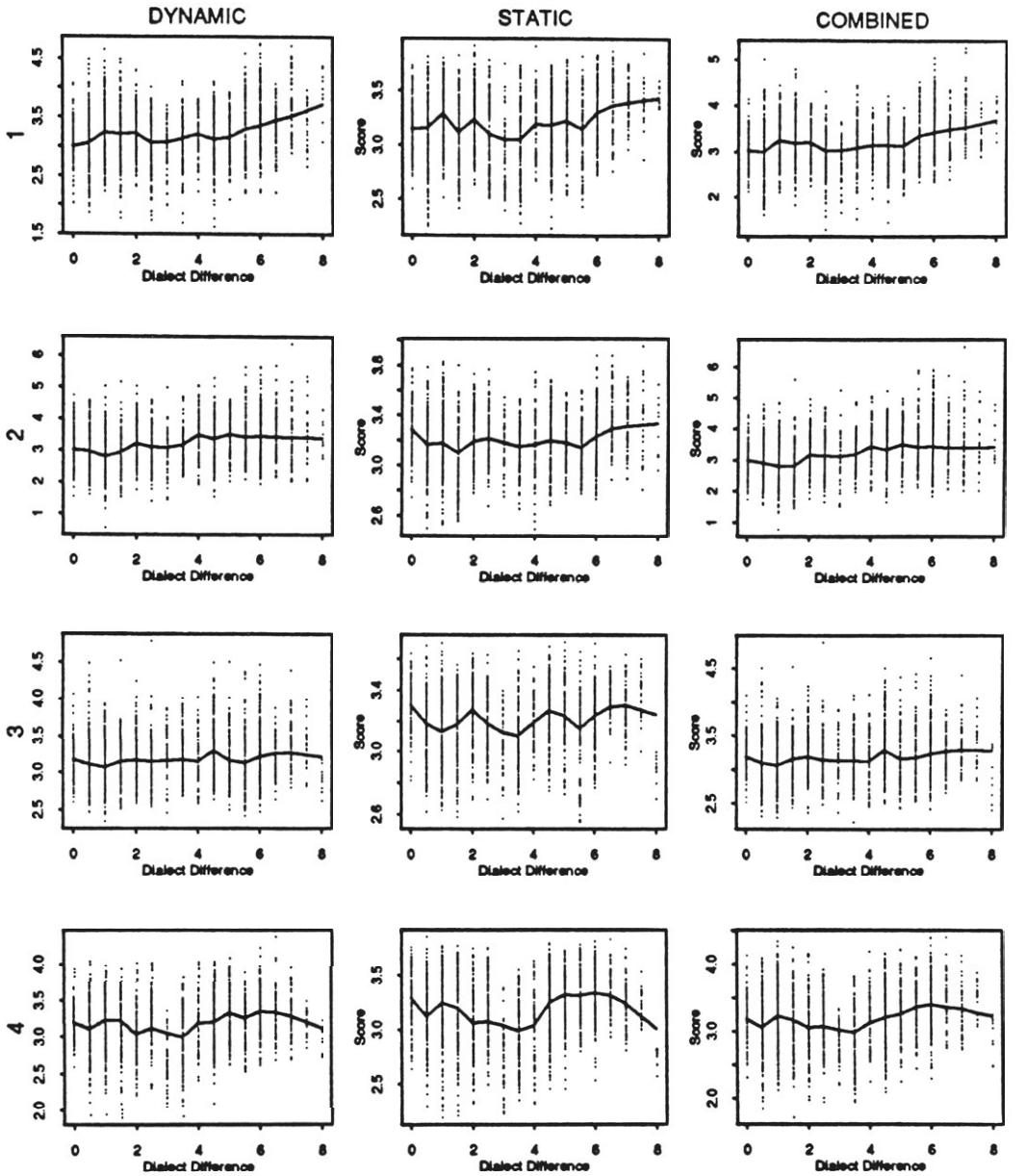


Figure D.28: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Energy* for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.211	0.139	0.234
2	0.318	0.101	0.331
3	0.154	0.096	0.162
4	0.163	0.138	0.193

Table D.28: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Energy* for each of the 4 sentences in turn.

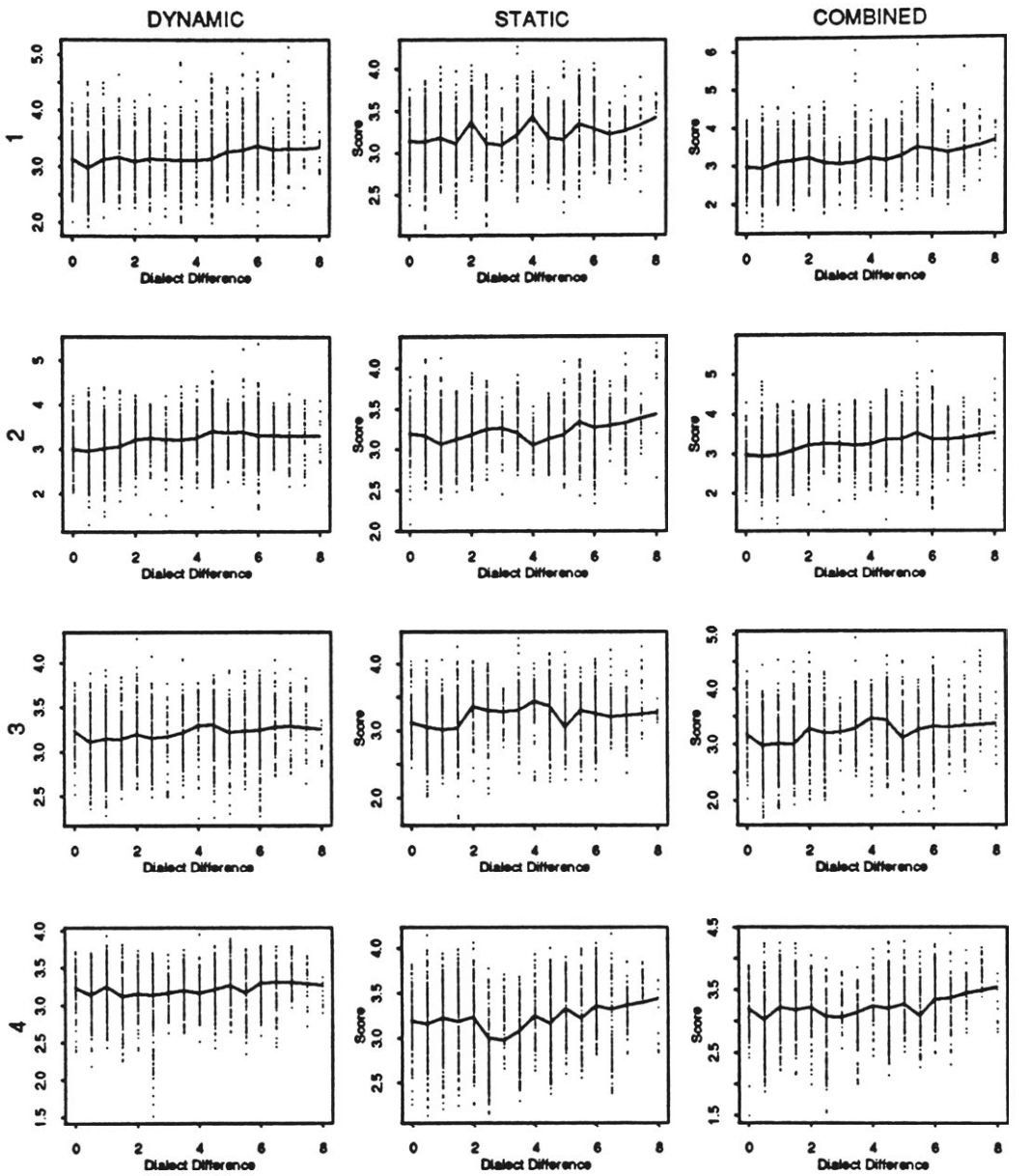


Figure D.29: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.214	0.148	0.263
2	0.24	0.141	0.264
3	0.138	0.171	0.227
4	0.138	0.17	0.197

Table D.29: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Linear Concatenated* F_0 for each of the 4 sentences in turn.

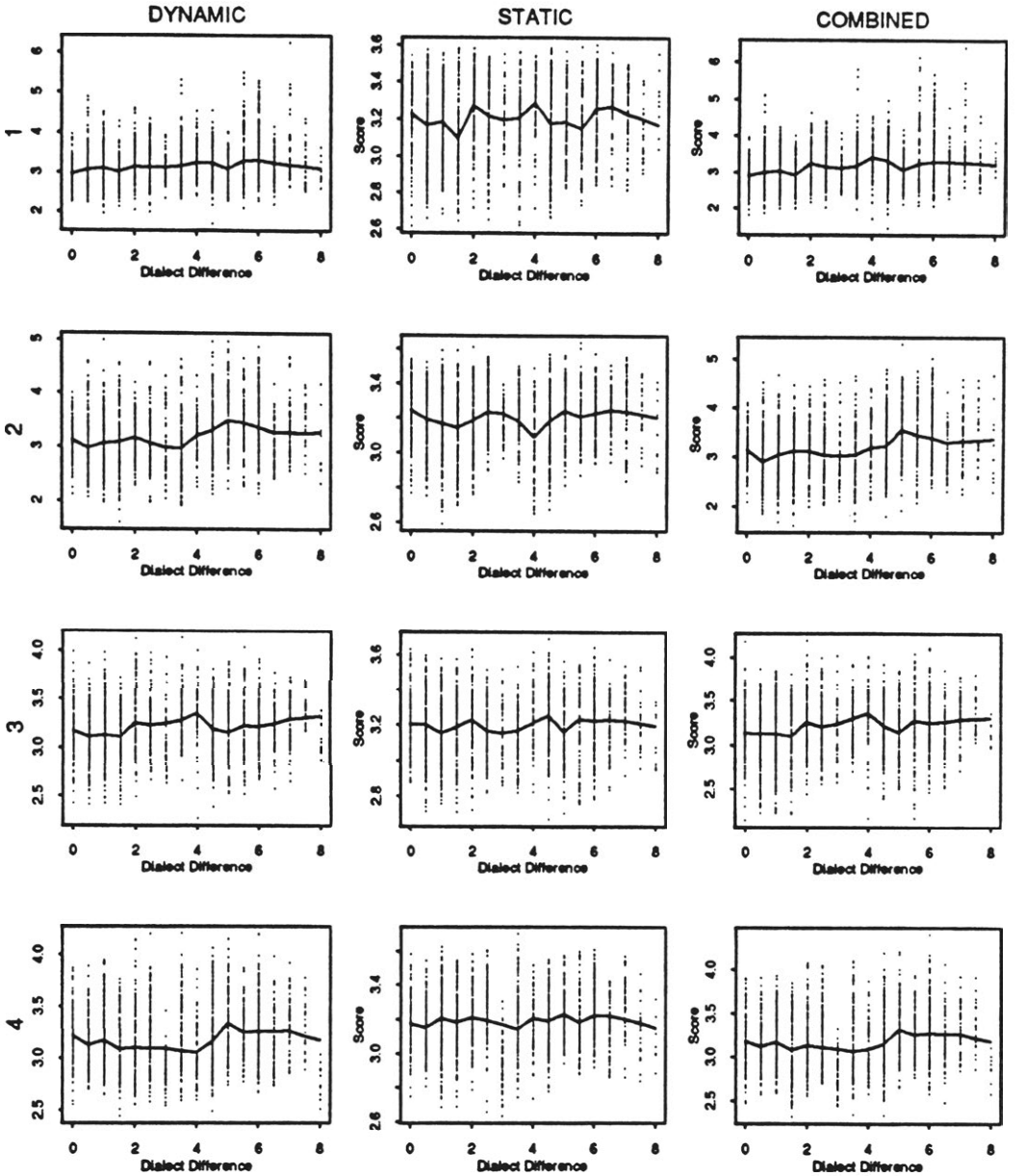


Figure D.30: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.227	0.091	0.262
2	0.239	0.079	0.26
3	0.126	0.077	0.144
4	0.134	0.077	0.145

Table D.30: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Linear Concatenated* F_0 for each of the 4 sentences in turn.

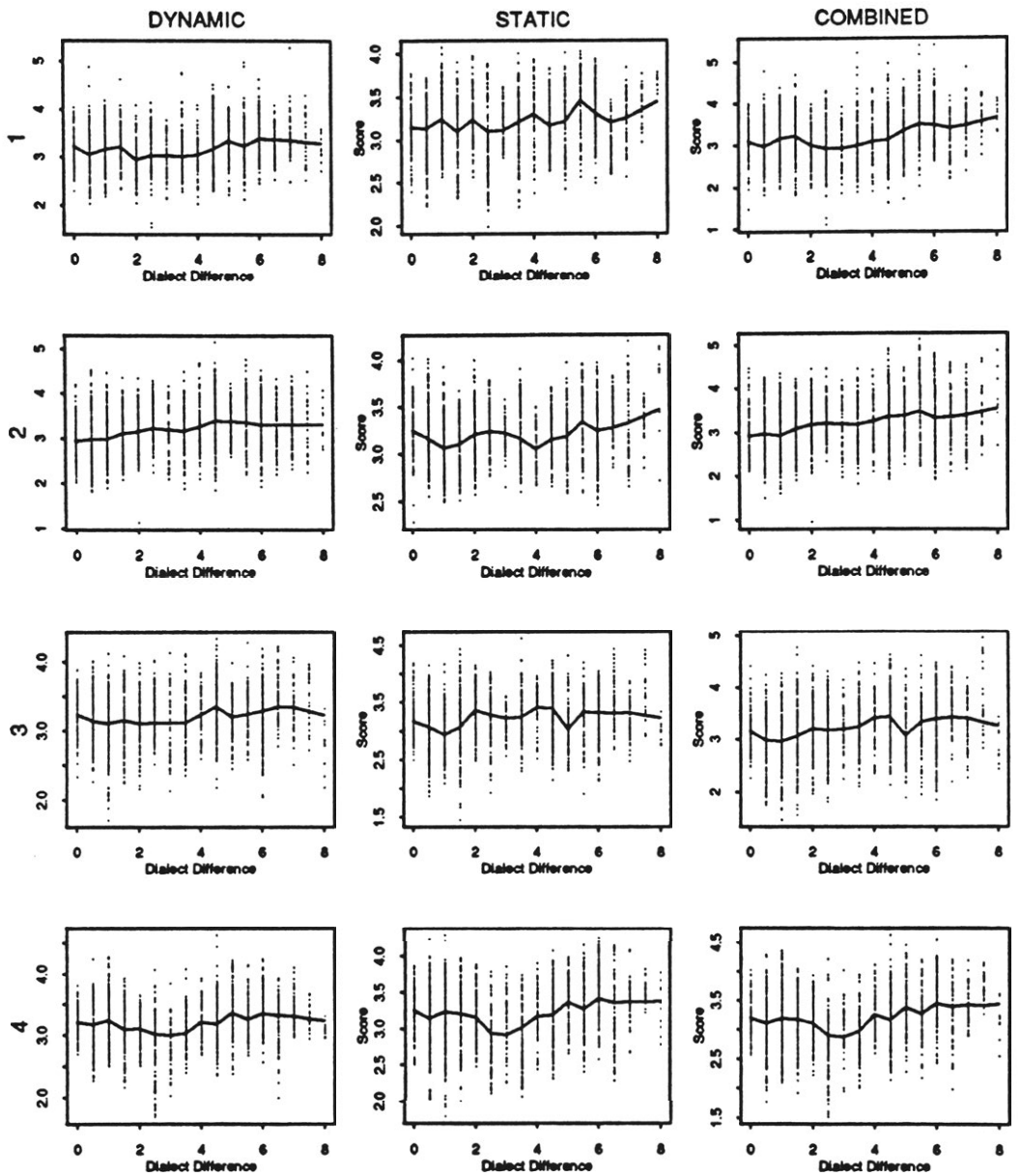


Figure D.31: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.199	0.153	0.256
2	0.242	0.133	0.266
3	0.168	0.2	0.246
4	0.164	0.191	0.216

Table D.31: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Linear Interpolated* F_0 for each of the 4 sentences in turn.

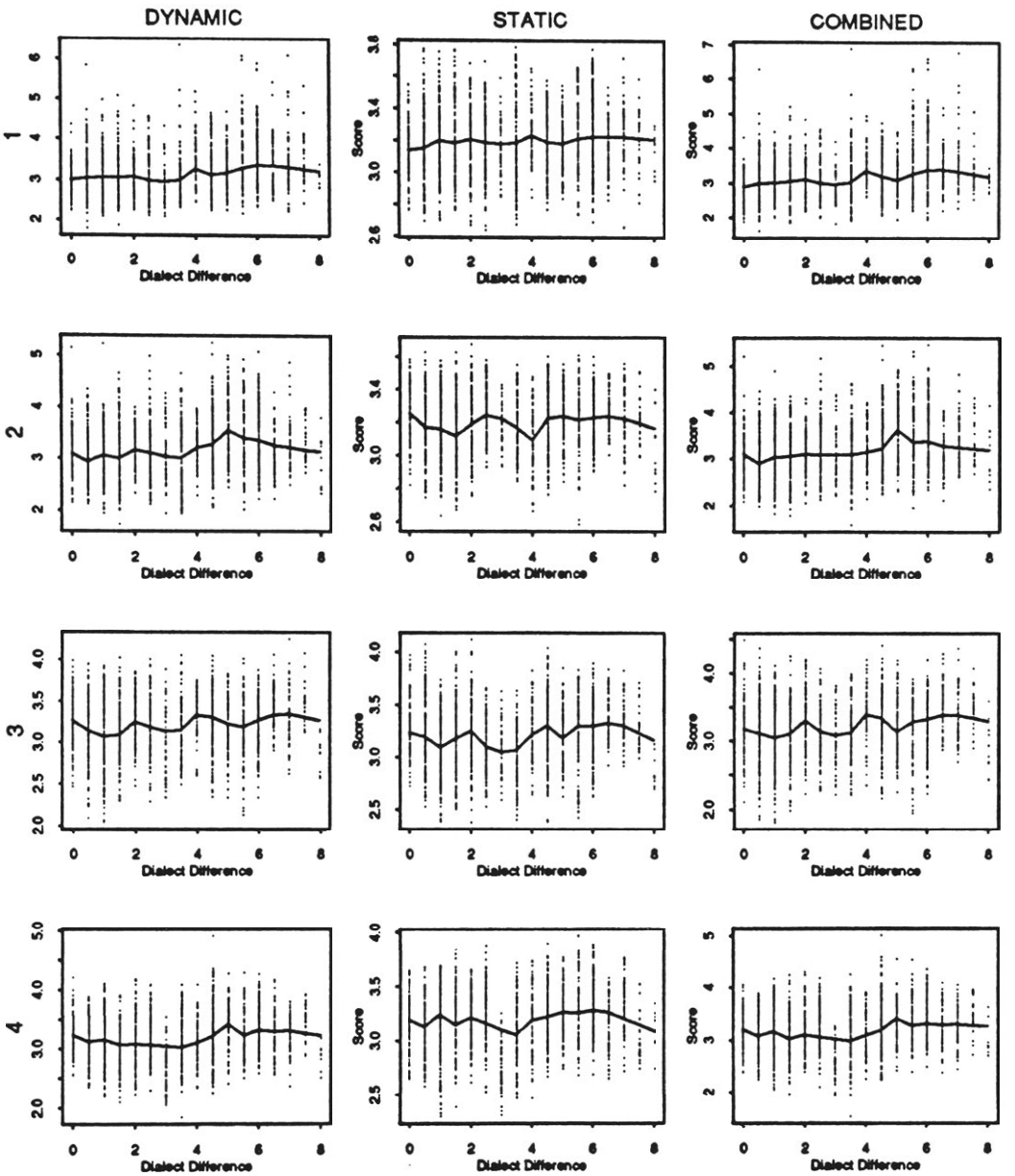


Figure D.32: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.25	0.092	0.285
2	0.241	0.082	0.26
3	0.155	0.134	0.19
4	0.176	0.122	0.192

Table D.32: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Linear Interpolated* F_0 for each of the 4 sentences in turn.

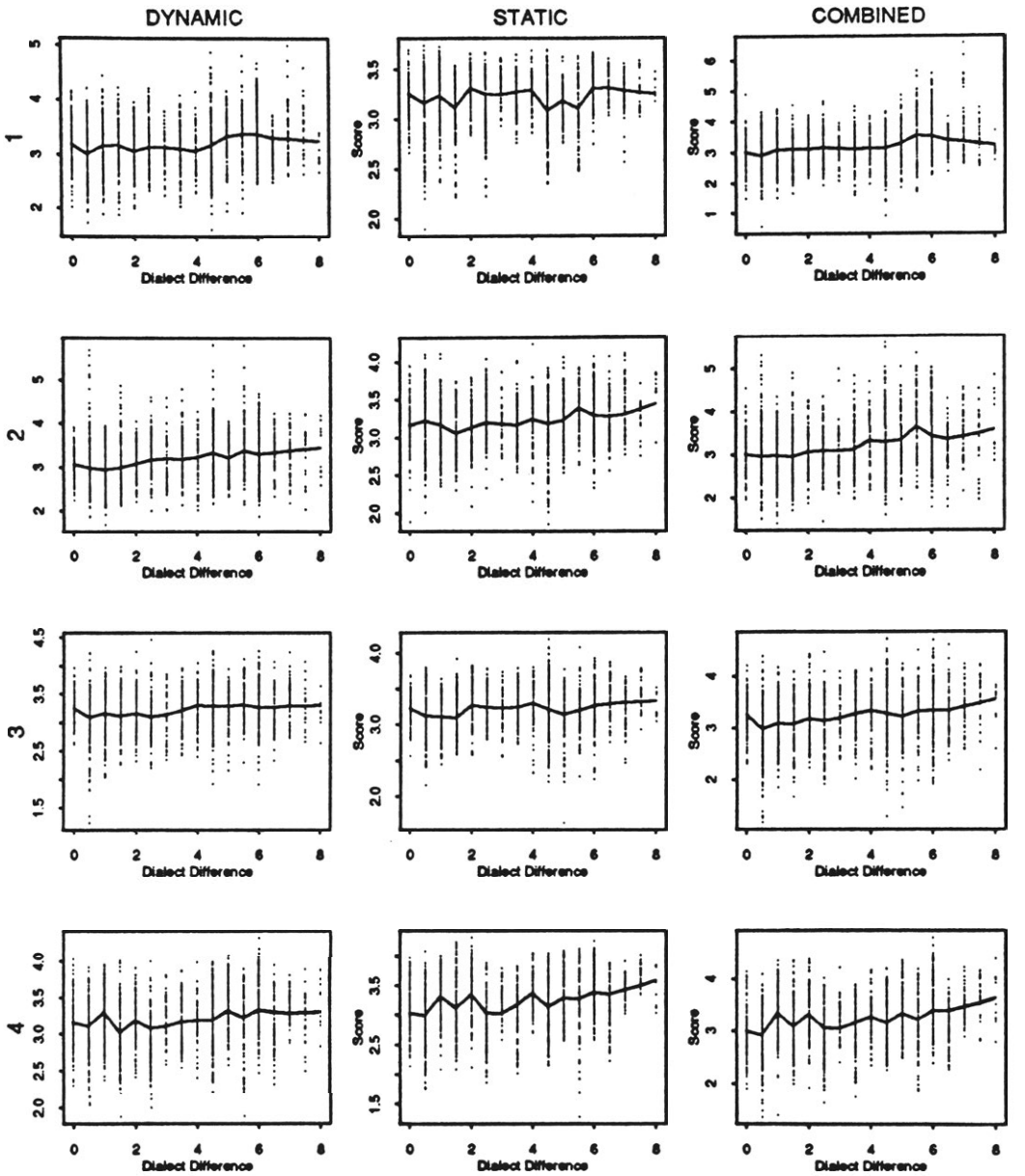


Figure D.33: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.209	0.117	0.285
2	0.24	0.15	0.285
3	0.168	0.13	0.221
4	0.162	0.202	0.226

Table D.33: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Log Concatenated* F_0 for each of the 4 sentences in turn.

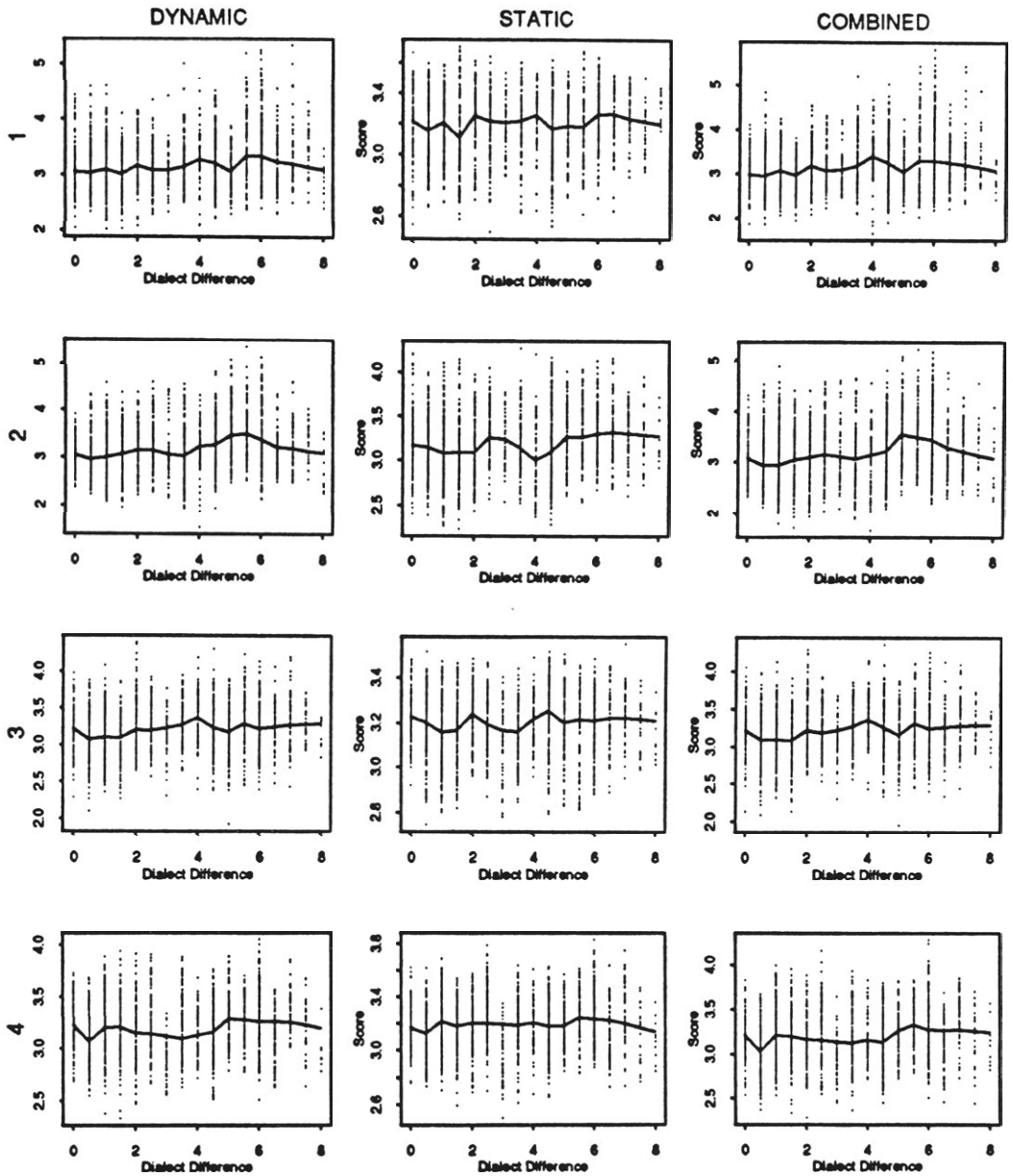


Figure D.34: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.21	0.093	0.255
2	0.246	0.158	0.268
3	0.149	0.064	0.154
4	0.121	0.086	0.136

Table D.34: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Log Concatenated* F_0 for each of the 4 sentences in turn.

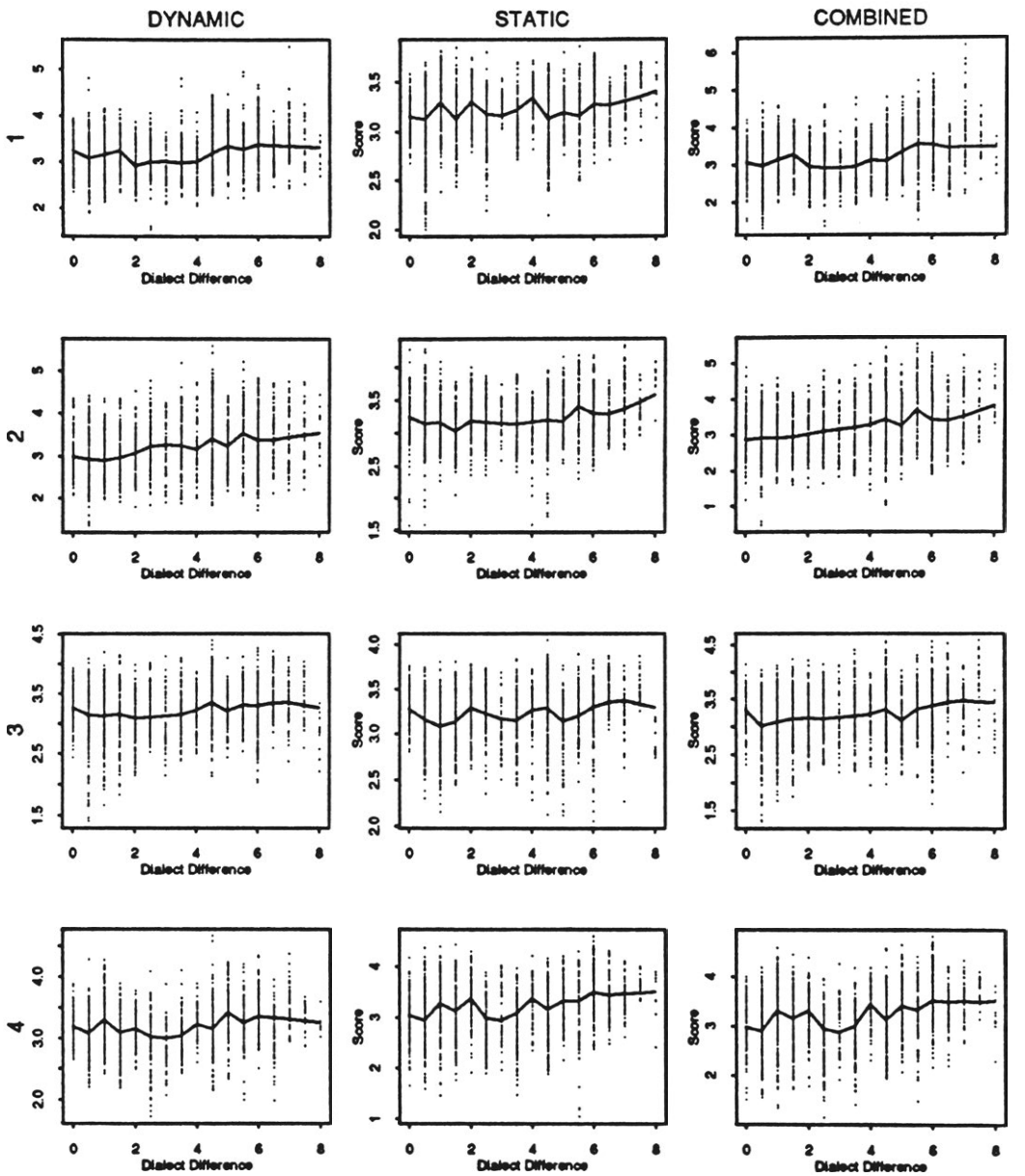


Figure D.35: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.201	0.127	0.278
2	0.287	0.162	0.324
3	0.185	0.138	0.217
4	0.172	0.243	0.266

Table D.35: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Log Interpolated* F_0 for each of the 4 sentences in turn.

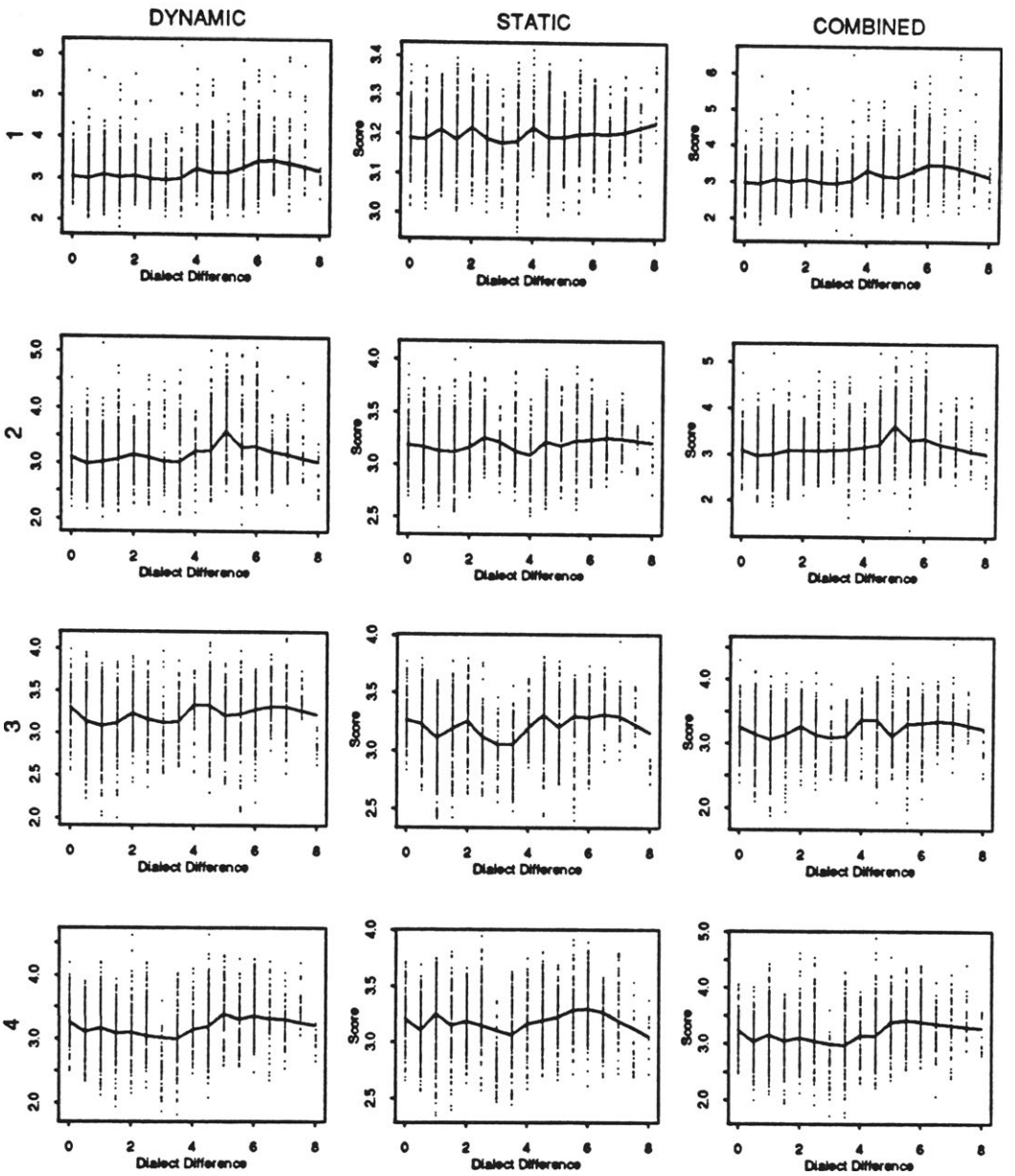


Figure D.36: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.259	0.034	0.283
2	0.228	0.114	0.248
3	0.151	0.117	0.168
4	0.181	0.12	0.205

Table D.36: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Log Interpolated* F_0 for each of the 4 sentences in turn.

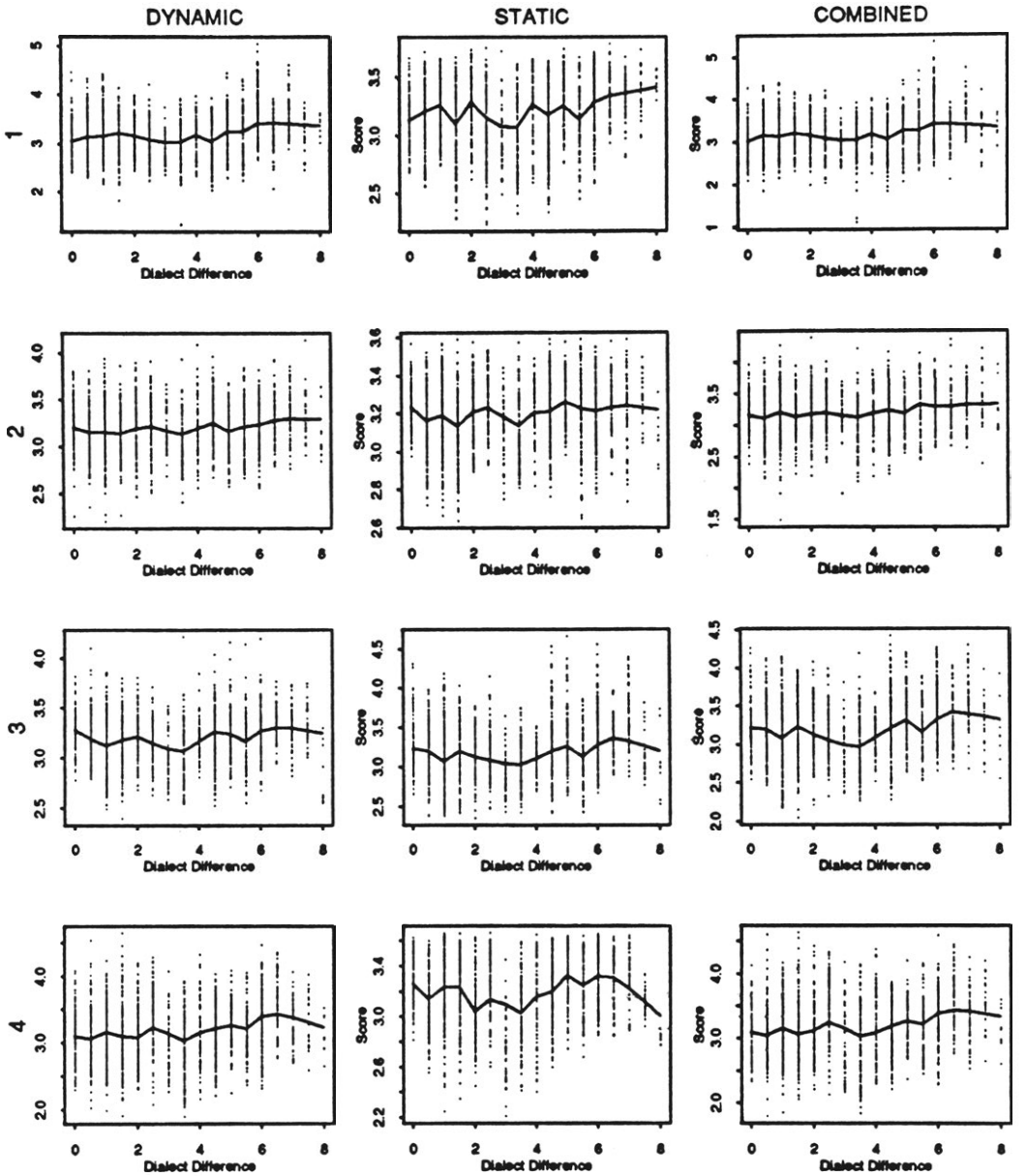


Figure D.37: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Voicing* for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.195	0.128	0.205
2	0.119	0.075	0.155
3	0.114	0.156	0.176
4	0.196	0.121	0.203

Table D.37: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Voicing* for each of the 4 sentences in turn.

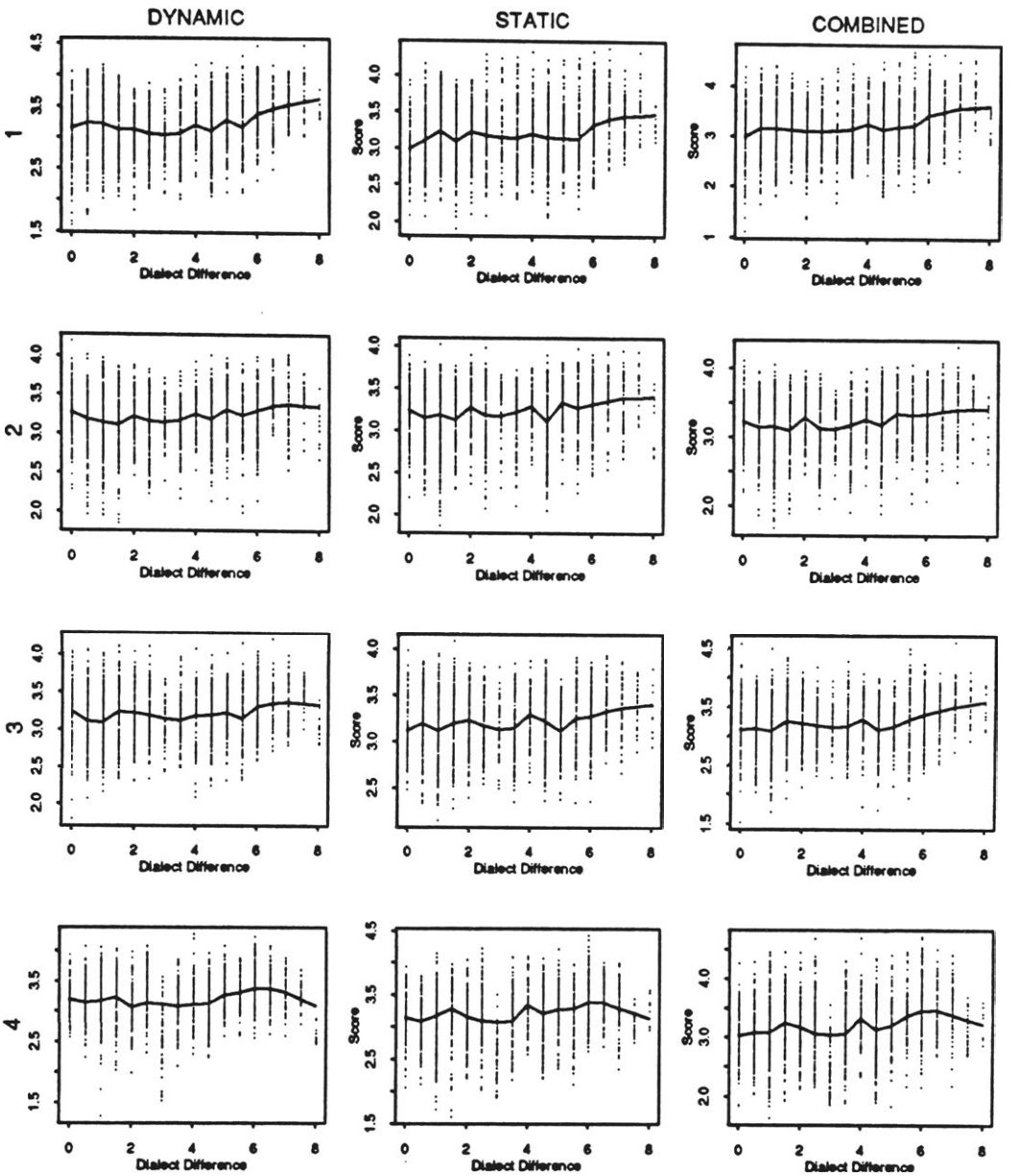


Figure D.38: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.201	0.185	0.24
2	0.159	0.157	0.181
3	0.159	0.147	0.198
4	0.165	0.189	0.228

Table D.38: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Zero Crossing Rate* for each of the 4 sentences in turn.

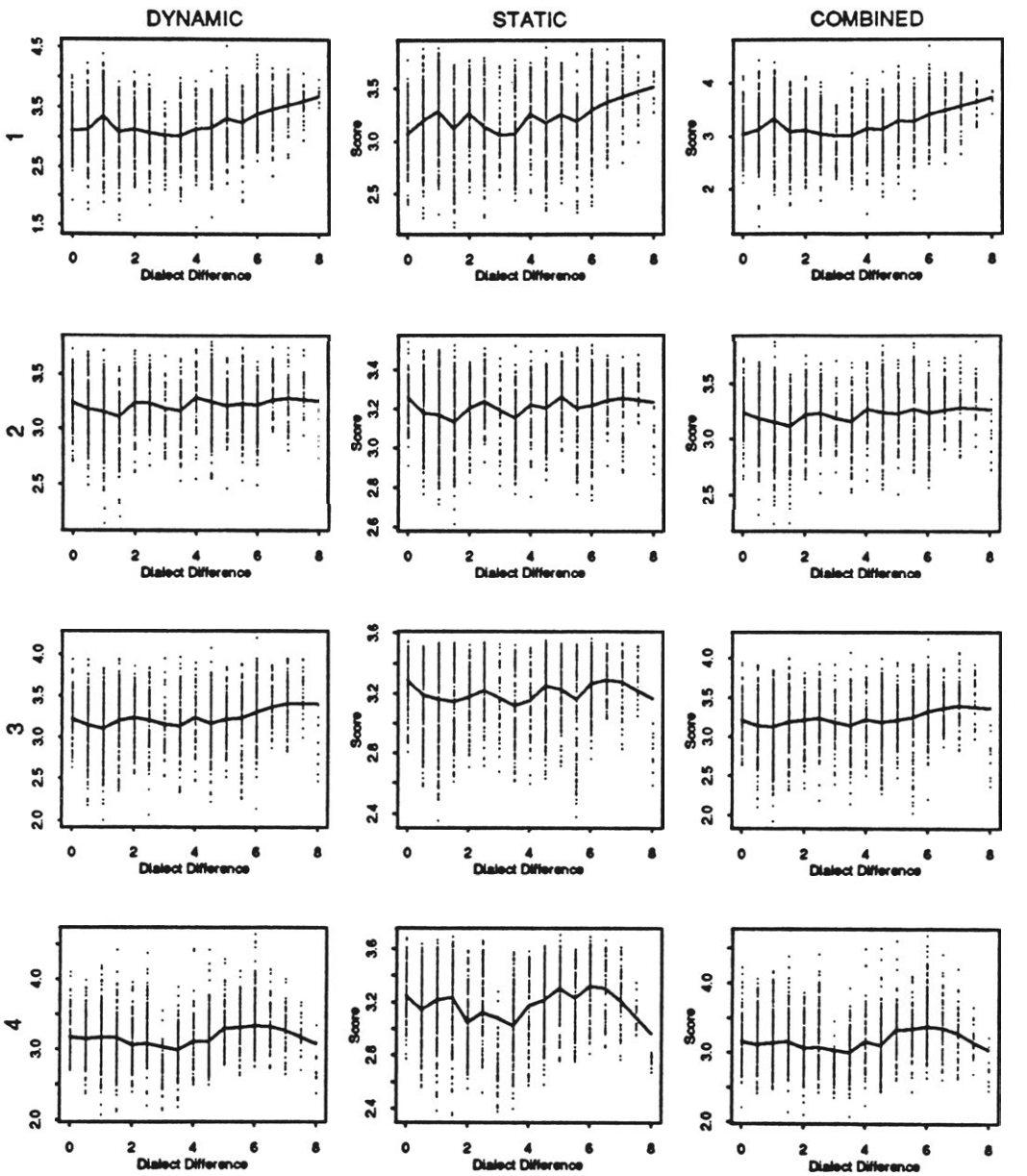


Figure D.39: Speaker Dialect Least-Square-Fit Scatter Plot - Dynamic, Static, and Combined Dynamic-Static plots for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn.

Sentence	*Measures*		
	Dynamic	Static	Both
1	0.198	0.147	0.212
2	0.104	0.072	0.111
3	0.15	0.104	0.158
4	0.158	0.119	0.17

Table D.39: Speaker Dialect Least-Square-Fit Correlation Values - Dynamic, Static, and Combined Dynamic-Static correlation values for the speech parameter *Normalised Zero Crossing Rate* for each of the 4 sentences in turn.

Appendix E

Principal Component Analysis of Measures

This appendix presents a principal component analysis [Dun84] of the twenty-one analysis measure set, for each of the four parameters— energy, F_0 (linear-concatenated), voicing, and zero crossing rate.

Principal component analysis may be viewed as a means of capturing the essential information in a large set of 'overlapping' variables in a new smaller set. These new variables—principal component variables are linear combinations of the original variables such that their variance is maximised and they are orthogonal to each other.

Hence, principal component analysis may reduce the dimensionality of inputs, to a recognition system, to a manageable size by deriving a smaller set of measures which capture the essence of the original data. Secondly, principal component analysis may, based on the calculated weights, show the underlying dimensionality and relationship between the original variables.

Each of the four basic parameters is analysed separately. The measure values for the four sentences were combined and each individual measure, m_{ij}, scaled as:

$$m'_{ij} = \frac{m_{ij} - \bar{m}_j}{\sqrt{(\sum_i m_{ij}^2)/(n-1)}} \tag{E.1}$$

Figure E.1 is a scree graph showing variance values of each of the principal components for the four individual parameters. For all four parameters a sharp decline in variance is evident after the first few principal components, showing that a smaller set of variables can encapsulate the greater variance of the original data.

The following four sections will present the correlations (loadings) of the *major* principal components for each of the four parameters. For the purposes of this work major is defined as having a variance greater than that of any of the original scaled measures—hence a threshold of 1. Thus 6 principal components for energy, 3 for F_0, 5 for voicing, and 6 for zero crossing rate, are shown. In all tables, correlation values greater than 0.3 in absolute terms are emphasised in order to help readability.

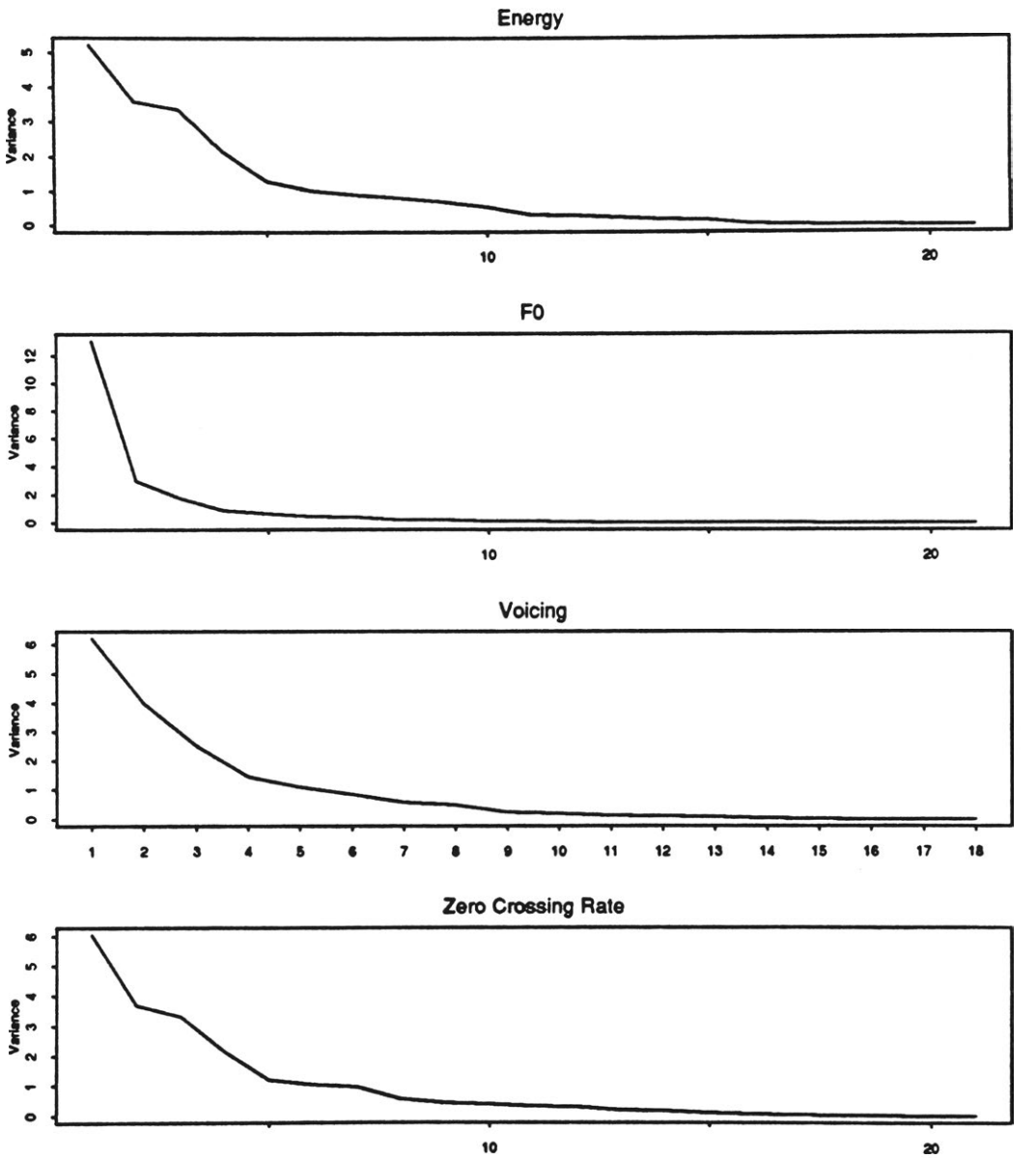


Figure E.1: Scree graph of principal component decomposition for the four parameters: energy, F_0, voicing, and zero crossing rate. The variance of the principal components for each parameter are plotted.

E.1 Energy

Table E.1 shows the correlations (loadings) of the 21 measures with the 6 principal components for energy. Combined the 6 principal components account for 79.7% of all variance in the original data.

Measures	*Principal Components*					
	1	2	3	4	5	6
DTW Dist	*0.41*	-0.09	0.02	-0.02	0.08	-0.16
Weighted DTW Dist.	*0.41*	-0.06	0.01	0.04	0.09	-0.17
Border DTW Dist.	*0.41*	-0.09	0.03	0.02	0.08	-0.18
Warp Len./Contour Len.	-0.03	-0.11	*0.52*	-0.01	0.04	0.03
Vert. Trans.	-0.04	-0.14	*0.52*	-0.02	0.05	0.02
Horiz. Trans.	0.09	*0.5*	0.08	0.01	-0.05	0
Diag. Trans.	-0.06	*-0.37*	*-0.36*	0	0.02	-0.01
Num. Vert. Excn	-0.19	-0.21	0.23	-0.17	*-0.32*	*-0.41*
Num. Horiz. Excn	-0.15	0.29	-0.1	-0.04	*0.57*	-0.12
Num. Diag. Excn	*-0.32*	0.1	0.07	-0.16	0.22	*-0.42*
Max. Len. Vert. Excn	0.06	-0.05	*0.35*	0.1	*0.39*	*0.43*
Max. Len. Horiz. Excn	0.17	0.27	0.15	0	*-0.52*	0.06
Max. Len. Diag. Excn	0.16	*-0.31*	-0.27	0.07	0.01	0.09
Off Diag. Wpath Dist.	0.28	0.13	0.04	0.07	0.06	-0.08
Mean	*-0.35*	0.04	-0.09	-0.02	-0.14	0.08
Std. Dev.	-0.07	0	0.02	*0.47*	0.05	-0.05
Min.	-0.1	0.02	0.03	*0.51*	-0.06	-0.24
Max.	-0.17	-0.02	-0.07	0.02	0.06	0.15
Range	-0.05	-0.01	0.03	*0.58*	0.02	-0.29
Mean Rate Change	-0.09	0	0	*0.32*	-0.17	*0.42*
Length	-0.1	*-0.48*	0.16	-0.01	0.06	0.01

Table E.1: Principal Component loadings of the 6 *major* principal components of Energy with the 21 original measures of Energy. Combined the 6 presented principal components account for 79.7% of all variance in the original measures of Energy.

Principal component (PC) one, which accounts for 24.8% of all variance, is most strongly correlated with the DTW distance measures and the mean. PC-2 accounts for 17.2% of variance and appears to be a measure of differences in total duration. PC-3 accounts for 16.1% of the variance and is highly correlated to measures differentiating original contour length from warped length. PC-4 accounts for 10.4% of variance and appears a measure of distribution (range). PC-5 accounts for 6.3% of the variance and appears a contrast of durations—horizontal excursions at a 'micro' and 'macro' level. Finally, PC-6 accounts for 5.0% of the total variance and appears a measure of the relative frame to frame variance of the energy contours under comparison.

E.2 F_0

Table E.2 shows the correlations (loadings) of the 21 measures with the 3 major principal components for the parameter F_0. Combined the three PCs account for 84.4% of all variance in the original data.

Measures	*Principal Components*		
	1	2	3
DTW Dist	0.26	-0.07	0.03
Weighted DTW Dist.	0.25	-0.16	0.05
Border DTW Dist.	0.26	-0.07	0.03
Warp Len./Contour Len.	-0.26	0.03	-0.19
Vert. Trans.	-0.26	0.03	-0.19
Horiz. Trans.	-0.21	*-0.36*	0.07
Diag. Trans.	0.25	0.19	0.06
Num. Vert. Excn	-0.25	0.04	-0.19
Num. Horiz. Excn	-0.22	-0.21	0.02
Num. Diag. Excn	-0.25	-0.11	-0.07
Max. Len. Vert. Excn	-0.25	0.03	-0.15
Max. Len. Horiz. Excn	-0.04	*-0.42*	0.28
Max. Len. Diag. Excn	0.16	0.07	-0.09
Off Diag. Wpath Dist.	-0.15	*-0.32*	0.1
Mean	-0.27	0.06	-0.06
Std. Dev.	-0.16	0.26	*0.46*
Min.	-0.24	-0.04	-0.22
Max.	-0.26	0.12	0.1
Range	-0.18	0.25	*0.44*
Mean Rate Change	-0.16	0.22	*0.45*
Length	0	*0.5*	-0.29

Table E.2: Principal Component loadings of the 3 *major* principal components of F_0 with the 21 original measures of F_0. Combined the 3 presented principal components account for 84.4% of all variance in the original measures of F_0.

PC-1 accounts for a surprising 61.9% of all variance in the original data. Examining the correlation values it appears hard to label this dimension as most measures have medium-strong correlations with none predominating. However, based on knowledge of F_0 and the speaker set this would appear to be differences in mean value. PC-2 accounts for 14.2% of the entire variance of F_0. Examining the correlation values it appears to be a quantification of temporal differences. PC-3 accounts for 8.3% of all variance. Based on correlation values it appears to be related to the distribution (range) of the original contours.

E.3 Voicing

Table E.3 shows the correlations (loadings) of the 18 measures (min, max, and range excluded) with the 5 major PCs of the parameter voicing. Combined the 5 PCs account for 72.7% of all variance in the original data.

Measures	*Principal Components*				
	1	2	3	4	5
DTW Dist	0.23	-0.13	*0.41*	-0.17	0.23
Weighted DTW Dist.	0.28	-0.16	*0.31*	-0.23	0.09
Border DTW Dist.	0.24	-0.15	*0.4*	-0.2	0.18
Warp Len./Contour Len.	0.04	*-0.46*	-0.18	0.08	0.08
Vert. Trans.	0.03	*-0.47*	-0.18	0.07	0.07
Horiz. Trans.	*0.34*	0.17	-0.12	0.24	0.1
Diag. Trans.	*-0.35*	0	0.19	-0.26	-0.13
Num. Vert. Excn	-0.08	*-0.34*	-0.17	0.09	0.12
Num. Horiz. Excn	0.27	0.12	-0.2	*-0.41*	*-0.34*
Num. Diag. Excn	0.25	0.04	-0.27	*-0.43*	*-0.3*
Max. Len. Vert. Excn	0.03	*-0.43*	-0.17	-0.07	-0.05
Max. Len. Horiz. Excn	*0.31*	0.1	-0.06	*0.34*	0.24
Max. Len. Diag. Excn	-0.28	0.02	*0.31*	0.13	-0.08
Off Diag. Wpath Dist.	*0.32*	-0.11	-0.04	-0.08	0.01
Mean	-0.19	0.2	-0.23	-0.26	*0.44*
Std. Dev.	-0.13	0.11	-0.25	*-0.36*	*0.6*
Mean Rate Change	-0.07	-0.03	-0.28	0.09	-0.12
Length	*-0.3*	-0.29	0.06	-0.2	-0.07

Table E.3: Principal Component loadings of the 5 *major* principal components of voicing with the 18 original measures of voicing. Combined the 5 presented principal components account for 72.7%% of all variance in the original measures of voicing.

PC-1 accounts for 29.4% of all variance in measures of voicing. Based on the correlation values the 'dimension' of PC-1 appears unclear, though temporal related. PC-2 accounts for 19.1% of variance and appears a quantification of differences in duration. PC-3 accounts for 12.0% of variance and is most strongly correlated to the DTW distance measures. PC-4 accounts for 6.9% of all variance, although its 'quality' or 'dimension' is unclear. Finally, PC-5 accounts for 5.2% of variance and appears a measure of the variance of voicing.

E.4 Zero Crossing Rate

Table E.4 shows the correlations (loadings) of the 21 measures with the 6 major PCs of the parameter zero crossing rate. Combined the 6 PCs account for 83.1% of all variance in the original data.

Measures	*Principal Components*					
	1	2	3	4	5	6
DTW Dist	*0.36*	-0.07	0.05	-0.02	0.02	-0.19
Weighted DTW Dist.	*0.36*	-0.01	0.06	-0.09	0.08	-0.25
Border DTW Dist.	*0.36*	-0.06	0.05	-0.02	0.03	-0.21
Warp Len./Contour Len.	0.01	*-0.39*	*-0.3*	-0.21	0.04	-0.02
Vert. Trans.	0	*-0.41*	-0.29	-0.19	0.03	-0.01
Horiz. Trans.	0.04	*0.36*	*-0.37*	-0.07	-0.05	0
Diag. Trans.	-0.04	-0.08	*0.51*	0.18	0.03	0
Num. Vert. Excn	-0.13	*-0.33*	-0.11	0.19	*-0.39*	-0.21
Num. Horiz. Excn	-0.13	0.22	-0.18	0.29	*0.53*	-0.17
Num. Diag. Excn	-0.22	-0.08	-0.21	*0.39*	0.11	-0.26
Max. Len. Vert. Excn	0.07	-0.21	-0.17	*-0.32*	*0.44*	0.23
Max. Len. Horiz. Excn	0.11	0.2	-0.2	-0.22	*-0.53*	0.11
Max. Len. Diag. Excn	0.13	-0.02	*0.43*	-0.1	-0.01	0.07
Off Diag. Wpath Dist.	0.21	0.17	-0.01	-0.28	0.18	-0.21
Mean	*-0.31*	-0.08	-0.02	0.07	-0.07	0.04
Std. Dev.	*-0.3*	0.08	0.11	*-0.3*	0.03	-0.1
Min.	0.01	0	-0.01	-0.05	-0.12	*-0.73*
Max.	*-0.31*	0.08	0.1	*-0.31*	0.04	-0.13
Range	-0.29	0.08	0.11	*-0.33*	0.02	-0.17
Mean Rate Change	-0.27	0.03	0.11	-0.25	0.01	-0.12
Length	-0.03	*-0.48*	0.17	-0.03	0.06	-0.01

Table E.4: Principal Component loadings of the 6 *major* principal components of zero crossing rate with the 21 original measures of zero crossing rate. Combined the 6 presented principal components account for 72.7%% of all variance in the original measures of zero crossing rate.

PC-1 accounts for 18.7% of the total variance and is most strongly correlated to measures of the mean and the DTW distances. PC-2 accounts for 17.6% of variance and appears a measure of differences in duration. PC-3 accounts for 15.8% of the variance of the original data and appears a measure of how well the two contours matched (diagonal transitions on warp path). PC-4 accounts for 10.3% of variance and appears a quantification of distribution. PC-5 accounts for 5.7% of variance and appears a contrast of micro timing adjustments with a single adjustment (horizontal excursions versus max. length horizontal excursion). Finally, PC-6 accounts for 5.0% of the total variance and is strongly correlated with the minimum measure.

Appendix F

Listener Instructions

The following list of instructions were presented verbally to all listeners prior to conduct of the perception experiments. Listeners were invited to question anything that was unclear and experiments were only initiated when all listeners affirmed their understanding.

1. You will be taking part in three separate listening experiments; speaker identification; sex identification; and dialect assignment.

2. For each experiment you will hear a presented utterance or set of utterances and must respond by indicating your decision upon the response sheet provided.

3. In all cases you must provide a response; do not skip any as "too hard" to decide.

4. There is no right or wrong answer, and this is not a trial or test of the listener.

5. Do not expect that your responses should be balanced amongst the choices.

6. Each response is numbered. Please work down the columns then over the page.

7. Each of the three listening experiments will take of the order of 10 minutes. Between each experiments we will take a break of approximately 5 minutes. Please feel free to relax during this period.

8. The utterances you will hear have been machine processed. Some may sound 'artificial' or synthetic. Do not deliberately listen for this effect.

9. If at any stage in an experiment you lose sequence then notify the experimenter. A new response sheet will be provided and you may restart the experiment.

10. For the speaker identification experiment you will hear three utterances for each decision. The first two utterances are always the same and are samples of the speakers Alan and Peter respectively. The third utterance is unknown. You must decide whether it is Alan or Peter by indicating in one of the two boxes your choice.

11. For the speaker sex identification experiment a single utterance is presented for each decision. You must decide whether the speaker is male or female and indicate your decision by marking one of the two boxes provided.

12. Australian dialect is a spectrum running from broad or 'thick' accent (a heavy Australian accent) through to cultivated or 'narrow' (often no detectable 'Australian accent').

13. For the speaker dialect experiment a single utterance is presented for each decision. You must decide where on the dialect spectrum that utterance lies and indicate your choice by marking a point somewhere along the line, marked 'cultivated' to 'broad', provided.

14. THANK YOU.

Bibliography

[Abr73] Albert S. Abrams. *Minimal Auditory Cues for Distinguishing Black from White Talkers.* PhD thesis, City University of New York, 1973.

[ACS77] Moya L. Andrews, W. Miles Cose, and Raymond G. Smith. Effects of alcohol on the speech of non alcoholics. *Central States Speech Journal*, 28:140–143, 1977.

[Ada71] Corinne Adams. English speech rhythm and the foreign learner. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 824–832, 1971.

[Ano68] Anon. On the track of voiceprints. *Nature*, 218:513–514, May 1968.

[ANS85] Åsa Nilsonne and Johan Sundberg. Differences in ability of musicians and non musicians to judge emotional state from the fundamental frequency of voice samples. *Music Perception*, 2(4):507–516, 1985.

[Ata72] Bishnu S. Atal. Automatic speaker recognition based on pitch. *J. Acoust. Soc. Am.*, 52(6):1687–1697, 1972.

[Ata74] Bishnu S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Am.*, 55(6):1304–1312, June 1974.

[Ata76] Bishnu S. Atal. Automatic recognition of speakers from their voices. *Proc. IEEE*, 64(4):460–475, April 1976.

[BCJ+69] Richard H. Bolt, Franklin S. Cooper, Edward E. David Jr., Peter B. Denes, James M. Pickett, and Kenneth N. Stevens. Identification of a speaker by speech spectrograms. *Nature*, 166:338–343, October 1969.

[BCJ+73] Richard H. Bolt, Franklin S. Cooper, Edward E. David Jr., Peter B. Denes, James M. Pickett, and Kenneth N. Stevens. Speaker identification by spectrograms: Some further observations. *J. Acoust. Soc. Am.*, 54(2):531–534, 1973.

[BCW88] Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language.* Wadsworth & Brooks/Cole, Pacific Grove California, 1988.

[Ber67] John R. L-B. Bernard. *Some Measurements of Some Sounds of Australian English.* PhD thesis, Sydney University, 1967.

[Ber90] C. Bernasconi. On instantaneous and transitional spectral information for text-dependent speaker verification. *Speech Communication*, 9(2):129–139, April 1990.

[BHN89] W. J. Barry, C. E. Hoequist, and F. J. Nolan. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, 3(4):355–366, October 1989.

[BLN+73] John W. Black, William Lashbrook, Ernest Nash, Herbert J. Oyer, Charles Pedrey, Oscar I. Tosi, and Henry Truby. Reply to "speaker identification by speech spectrograms: Some further observations". *J. Acoust. Soc. Am.*, 54(2):535–537, 1973.

[BP66] Peter D. Bricker and Sandra Pruzansky. Effects of stimulus content and duration on talker identification. *J. Acoust. Soc. Am.*, 40(6):1441–1449, 1966.

[BRD75] Eileen M. Brennan, Ellen B. Ryan, and William E. Dawson. Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research*, 4(1):27–36, 1975.

[Bre71] Ruth M. Brend. Male-female intonation patterns in American English. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 866–870, 1971.

[BSG90] Younes Bennani, Francoise Fogelman Soulie, and Patrick Gallinari. A connectionist approach for automatic speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, pages 265–268, 1990.

[BSR73] Bruce L. Brown, William J. Strong, and Alvin C. Rencher. Perceptions of personality from speech: Effects of manipulations of acoustical parameters. *J. Acoust. Soc. Am.*, 54(1):29–35, 1973.

[Bus71] Clara N. Bush. Temporal ratios of sound segments and the perception of English dialect differences. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 666–673, 1971.

[BW88] Michael Barlow and Michael Wagner. Prosody as a basis for determining speaker characteristics. In Michael Wagner, editor, *Proc. Second Australian Int. Conf. on Speech Science and Technology*, pages 80–85, Sydney, November 1988. Australian Speech Science and Technology Association.

[CB69] Frank R. Clarke and Richard W Becker. Comparison of techniques for discriminating among talkers. *Journal of Speech and Hearing Research*, 12:747–761, 1969.

[CF89] Arnon Cohen and Igal Froind. On text independent speaker identification using a quadratic classifier with optimal features. *Speech Communication*, 8(1):35–44, 1989.

[Che78] Ronald S. Cheung. Feature selection via dynamic programming for text-independent speaker identification. *IEEE TRANS. ASSP*, ASSP-26(5):397–403, 1978.

[CL87] Sin-Horng Chen and Min-Tau Lin. On the use of pitch contour of Mandarin speech in text-independent speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Proc.*, pages 1418–1421, 1987.

[CM72] S. Cort and T. Murray. Aural identification of children's voices. *J. Acoust. Soc. Am.*, 51:131(A), 1972.

[Col71] Ralph O. Coleman. The perception of maleness and femaleness in the voice and its relationship to vowel formant frequencies. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 471–478. IEEE, 1971.

[Col76] Ralph O. Coleman. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech and Hearing Research*, 19:168–180, 1976.

[Col77] A. M. Collins. Acquisition of an Australian speech data base. *Proceedings of the Digital Equipment Users Society*, 3(5):1575–1577, 1977.

[Cos83] Leda Cosmides. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(6):864–881, 1983.

[Cow36] M. Cowan. Pitch and intensity characteristics of stage speech. *Arch. Speech*, Suppl. 1, 1936.

[CWH87a] D. G. Chidlers, Ke Wu, and D. M. Hicks. Voice conversion: A model for studying voice quality and speaker normalisation. In *Proc. European Conf. on Speech Technology*, pages 488–491, 1987.

[CWH87b] D. G. Childers, Ke Wu, and D. M. Hicks. Factors in voice quality: Acoustic features related to gender. In *Proc. ICASSP 87*, pages 293–296, 1987.

[CWHY89] D. G. Childers, Ke Wu, D. M. Hicks, and B. Yegnanarayana. Voice conversion. *Speech Communication*, 8(2):147–158, June 1989.

[CYW85] D. G. Childers, B. Yegnanarayana, and Ke Wu. Voice conversion: Factors responsible for quality. In *Proc. ICASSP 85*, pages 748–751, 1985.

[DK38] Delwin Dusenbury and Franklin H. Knower. Experimental studies of the symbolism of action and voice-I: A study of the specificity of meaning in facial expression. *The Quarterly Journal of Speech*, 24(3):425–435, 1938.

[Dod71a] George R. Doddington. A method of speaker verification. *J. Acoust. Soc. Am.*, 49(1):139(A), 1971.

[Dod71b] George R. Doddington. *A Method of Speaker Verification.* PhD thesis, The University of Wisconsin, 1971.

[Dod85] George R. Doddington. Speaker recognition - identifying people by their voices. *Proceedings of the IEEE*, 73(11):1651–1664, Nov 1985.

[Dom90] Wim A. Van. Dommelen. Acoustic parameters in human speaker recognition. *Language and Speech*, 33(3):259–272, July 1990.

[DS71] S. K. Das and S. L. Saleeby. Speaker verification experiments. *J. Acoust. Soc. Am.*, 49(1):138(A), 1971.

[Dun84] George H. Dunteman. *Introduction to Multivariate Analysis.* Sage Publications, Beverly Hills, 1984.

[EFS76] Paul Ekman, Wallace V. Freisen, and Klaus R. Scherer. Body movement and voice pitch in deceptive interaction. *Semiotica*, 16(1):23–27, 1976.

[End71] W. Endres. Changes of human voice caused by age, disguise, and simulation. *J. Acoust. Soc. Am.*, 49(1):138(A), 1971.

[Eng71] Walburga Von Raffler Engel. Intonational and vowel correlates in contrasting dialects: A suggestion for further research. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 768–773, 1971.

[FBG77] C. Abraham Fenster, Lily Klebranoff Blake, and Alan M. Goldstein. Accuracy of vocal emotional communications among children and adults and the power of negative emotions. *Journal of Communication Disorders*, 10:301–314, 1977.

[FBS84] J. Fokes, Z. S. Bond, and M. Steinberg. Patterns of English word stress by native and non-native speakers. In *Proceedings of the Tenth International Congress of Phonetic Sciences*, pages 682–686, 1984.

[FH41] Grant Fairbank and LeMar W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Speech Monographs*, 8:85–90, 1941.

[FH82] Hiroya Fujisaki and Keikichi Hirose. Modeling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation. In *Proceedings of the XIII th International Congress of Linguistics*, pages 57–70, Tokyo, 1982.

[FH84] James Emile Flege and James Hillenbrand. Limits on phonetic accuracy in foreign language speech production. *J. Acoust. Soc. Am.*, 76(3):708–721, 1984.

[FIS72] Sadaoki Furui, Fumitada Itakura, and Shuzo Saito. Talker recognition by long-time averaged speech spectrum. *Electronics and Communications in Japan*, 55-A(10):54–61, 1972.

[Fle84] James Emil Flege. The detection of French accent by American listeners. *J. Acoust. Soc. Am.*, 76(3):692–707, 1984.

[FP39] Grant Fairbanks and Wilbert Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, 6:87–104, 1939.

[Fuj88] Hiroya Fujisaki. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Osamu Fujimura, editor, *Vocal Physiology: Voice Production, Mechanisms and Functions*, chapter 30, pages 347–355. Raven Press Ltd., New York, 1988.

[Fur78] Sadaoki Furui. Effects of long-term spectral variability on speaker recognition. *J. Acoust. Soc. Am.*, 64(Suppl. 1):S183, 1978.

[Fur81a] Sadaoki Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-29(2):254–272, April 1981.

[Fur81b] Sadaoki Furui. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-24(3):342–350, 1981.

[FW80] John E. Freund and Ronald E. Walpole. *Mathematical Statistics.* Prentice-Hall International, London, 1980.

[GK68] James W. Glenn and Norbert Kleiner. Speaker identification based on nasal phonation. *J. Acoust. Soc. Am.*, 43(2):368–372, 1968.

[GMH89] Marylou Pausewang Gelfer, Karen Parker Massey, and Harry Hollien. The effects of sample duration and timing on speaker identification accuracy by means of long-term spectra. *Journal of Phonetics*, 17(4):327–338, October 1989.

[Gol76] Ursula G. Goldstein. Speaker identifying features based on formant tracks. *J. Acoust. Soc. Am.*, 59(1):176–182, 1976.

[GR69] B. Gold and L. R. Rabiner. Parallel processing techniques for estimating pitch periods of speech in the time domain. *J. Acoust. Soc. Am.*, 46:442–448, 1969.

[Har79] David E. Hartman. The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders*, 12:53–61, 1979.

[Haz73] Barry Hazen. Effects of differing phonetic contexts on spectrographic speaker identification. *J. Acoust. Soc. Am.*, 54(3):650–660, 1973.

[HD76] David E. Hartman and Jeffrey L. Danhauer. Perceptual features of speech for males in four perceived age decades. *J. Acoust. Soc. Am.*, 59(3):713–715, 1976.

[Hes83] Wolfgang Hess. *Pitch Determination of Speech Signals.* Springer-Verlag, 1983.

[HHP88] Eva B. Holmberg, Robert E. Hillman, and Joseph S. Perkell. Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal and loud voice. *J. Acoust. Soc. Am.*, 84(2):511–529, 1988.

[HL88] Bernard Harmegnies and Albert Landercy. Intra-speaker variability of the long term speech. *Speech Communication*, 7(1):81–86, 1988.

[HM77] Harry Hollien and Wojciech Majewski. Speaker identification by long-term spectra under normal and distorted speech conditions. *J. Acoust. Soc. Am.*, 62(4):975–980, 1977.

[HMD82] Harry Hollien, Wojciech Majewski, and E. Thomas Doherty. Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, 10:139–148, 1982.

[HP69] Harry Hollien and Patricia Paul. A second evaluation of the speaking fundamental frequency characteristics of post-adolescent girls. *Language and Speech*, 12:119–124, 1969.

[HR72] G. D. Hair and T. W. Rekeita. Automatic speaker verification using phoneme spectra. *J. Acoust. Soc. Am.*, 51(1):131(A), 1972.

[HSBW68] Michael H. L. Hecker, Kenneth N. Stevens, Gottfried Von Bismarck, and Carl E. Williams. Manifestation of task induced stress in the acoustic speech signal. *J. Acoust. Soc. Am.*, 44(4):993–1001, 1968.

[HT78] Harry Hollien and Gilbert C. Tolhurst. The aging voice. In *Transcripts of the Seventh Symposium Care of the Professional Voice Part II: Life-Span Changes in the Human Voice*, pages 67–73. The Voice Foundation, 1978.

[HW71] William L. Hays and Robert L. Winkler. *Statistics: Probability, Inference, and Decision.* Holt, Rinehart and Winston, Inc., New York, 1971.

[Ing68] Frances Ingemann. Identification of the speaker's sex from voiceless fricatives. *J. Acoust. Soc. Am.*, 44(4):1142–1144, 1968.

[INN78] A. Ichikawa, A. Nakajima, and K. Nakata. Speaker verification from actual telephone voice. *J. Acoust. Soc. Am.*, 64(Suppl. 1):S182, 1978.

[IP86] John Ingram and Jeff Pittam. Acoustic correlates of accent change: Vietnamese schoolchildren acquiring Australian English. In *Proc. First Australian Conference on Speech Science and Technology*, pages 246–251, Canberra, November 1986.

[JHH84] Charles C. Johnson, Harry Hollien, and James W. Hicks. Speaker identification utilising selected temporal speech features. *Journal of Phonetics*, 12:319–326, 1984.

[Joh90] Keith Johnson. The role of perceived speaker identity in F_0 normalisation of vowels. *J. Acoust. Soc. Am.*, 88(2):642–654, August 1990.

[JPL+85] A. Jimenez, F. Del Pozo, M. Laraña, E. Vázquez, and J. L. Zorenda. Analysis of emotional markers in speech. In *IEEE/Seventh Annual Conference of the Engineering in Medicine and Biology Society*, pages 901–904. IEEE, 1985.

[JR90] Richard D. Jacques and Michael P. Rastatter. Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners. *Folia Phoniatrica*, 42(3):118–124, May 1990.

[KDVRKK87] Bruce E. Koenig, Jr Donald V. Ritenour, Barbara A. Kohus, and Artese Savoy Kelly. Reply to "some fundamental considerations regarding voice print identification" [j. acoust. soc. am. 82, 687-688(1987)]. *J. Acoust. Soc. Am.*, 82(2):688–689, August 1987.

[Ker62] L. G. Kersta. Voiceprint identification. *Nature*, 196(4861):1253–1257, December 1962.

[KK90] Dennis H. Klatt and Laura C. Klatt. Analysis, resynthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87(2):820–857, February 1990.

[Kno41] Franklin H. Knower. Analysis of some experimental variations of simulated vocal expressions of the emotions. *The Journal of Social Psychology*, 14:369–372, 1941.

[KO84] H. Kuwabara and K. Ohgushi. Acoustic characteristics of professional male announcers' speech sounds. *Acustica*, 55:233–240, 1984.

[Koe80a] Bruce E. Koenig. Speaker identification part 1. *FBI Law Enforcement Bulletin*, 49(1):1–4, January 1980.

[Koe80b] Bruce E. Koenig. Speaker identification part 2. *FBI Law Enforcement Bulletin*, 49(2):20–22, February 1980.

[Koe86] Bruce E. Koenig. Spectrographic voice identification: A forensic survey. *J. Acoust. Soc. Am.*, 79(6):2088–2090, 1986.

[KPL88] F. Klingholz, R. Penning, and E. Liebhart. Recognition of low-level alcohol intoxication from speech signal. *J. Acoust. Soc. Am.*, 84(3):929–935, 1988.

[LAJ80] Norman J. Lass, Celest A. Almerino, and Laurie F. Jordan. The effect of filtered speech on speaker race and sex identification. *Journal of Phonetic*, 3:101–112, 1980.

[Lar71] Conrad Lariviere. Some acoustic and perceptual correlates of speaker identification. In *Proceedings of the Seventh International Congress of Phonetic Sciences*, pages 558–564. IEEE, 1971.

[LBNS78] Norman J. Lass, Amy S. Beverly, Debra K. Nicosia, and Laurel A. Simpson. An investigation of speaker height and weight identification by means of direct estimation. *Journal of Phonetics*, 6:69–76, 1978.

[LF85a] Sue Ellen Linville and Hilda B. Fisher. Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *J. Acoust. Soc. Am.*, 78(1):40–48, 1985.

[LF85b] Sue Ellen Linville and Hilda B. Fisher. Acoustic characteristics of women's voices with advancing age. *Journal of Gerontology*, 40(3):324–330, 1985.

[LGP84] Gordon E. Legge, Carla Grosmann, and Christina M. Pieper. Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2):298–303, 1984.

[LH74] K. P. Li and G. W. Hughes. Talker differences as they appear in correlation matrices of continuous speech spectra. *J. Acoust. Soc. Am.*, 55(4):833–837, April 1974.

[LHB+76] Norman J. Lass, Karen R. Hughes, Melanie D. Bowyer, Lucille T. Waters, and Victoria T. Bourne. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *J. Acoust. Soc. Am.*, 59(3):675–678, 1976.

[LHH66] K. P. Li, G. W. Hughes, and A. S. House. Correlation characteristics and dimensionality of speech spectra. *J. Acoust. Soc. Am.*, 46(4):1019–1026, 1966.

[Li87] K. P. Li. Separating phonetic and speaker features of vowels in formant space. In *Proc. ICASSP 87*, pages 1469–1472, 1987.

[Lin73] C. E. Linke. A study of pitch characteristics of female voices and their relationship to vocal effectiveness. *Folia Phoniatrica*, 25:173–185, 1973.

[Lin88] Sue Ellen Linville. Intra speaker variability in fundamental frequency stability: an age related phenomenon? *J. Acoust. Soc. Am.*, 83(2):741–745, 1988.

[Lip87] Richard P. Lippman. An introduction to computing with neural nets. *IEEE ASSP Magazine*, pages 4–22, April 1987.

[LK86] Sue Ellen Linville and Edward W. Korabic. Elderly listeners' estimates of vocal age in adult females. *J. Acoust. Soc. Am.*, 80(2):692–694, 1986.

[LKE85] Diana Van Lancker, Jody Kreiman, and Karen Emmorey. Familiar voice recognition: Patterns and parameters part I: Recognition of backward voice. *Journal of Phonetics*, 13(1):19–33, 1985.

[LKW85] Diana Van Lancker, Jody Kreiman, and Thomas D. Wickers. Familiar voice recognition: Patterns and parameters part II: Recognition of rate-altered voices. *Journal of Phonetics*, 13(1):39–52, 1985.

[LMK78] Norman J. Lass, Pamela J. Mertz, and Karen L. Kimmel. The effect of temporal speech alterations on speaker race and sex identifications. *Language and Speech*, 21(3):279–290, 1978.

[LR71] H. Levitt and L. R. Rabiner. Analysis of fundamental frequency contours in speech. *J. Acoust. Soc. Am.*, 49(2):569–582, 1971.

[LST+85] D. Robert Ladd, Kim E.A. Silverman, Frank Tolkmitt, Günther Bergmann, and Klaus R. Scherer. Evidence for the independent function of intonation contour type, voice quality and F_0 range in signaling speaker affect. *J. Acoust. Soc. Am.*, 78(2):435–444, 1985.

[LT90] Brian C. Lovell and Ah Chung Tsoi. Speaker verification using artificial neural networks. In *Proceedings of the Third Australian International Conference on Speech Science and Technology*, pages 298–303, Melbourne, November 1990.

[LTMB79] Norman J. Lass, John E. Tecca, Robert A. Mancuso, and Wanda I. Black. The effects of phonetic complexity on speaker race and sex identification. *Journal of Phonetics*, 7:105–118, 1979.

[Luc69] James E. Luck. Automatic speaker verification using cepstral measurements. *J. Acoust. Soc. Am.*, 46(4):1026–1032, 1969.

[Lum72] Robert C. Lummis. Test of an automatic speaker verification method with intensively trained professional mimics. *J. Acoust. Soc. Am.*, 51:131(A)–132(A), 1972.

[Lum73] Robert C. Lummis. Speaker verification by computer using speech intensity for temporal registration. *IEEE Trans. Audio and Electroacoustics*, AU-21(2):80–89, April 1973.

[MAHG76] J.D. Markel and Jr. A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, 1976.

[Mat89] Hiroshi Matsumoto. Text-independent speaker identification from short utterances based on a piecewise discriminant analysis. *Computer Speech and Language*, 3(2):133–150, 1989.

[MD78] John D. Markel and Steven B. Davis. Text-independent speaker identification from a large linguistically unconstrained time-spaced data base. In *Proc. ICASSP 78*, pages 287–290. IEEE, 1978.

[MD79] John D. Markel and Steven B. Davis. Text-independent speaker identification from a large linguistically unconstrained time-spaced data base. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-24(1):74–82, 1979.

[MH74] Wojciech Majewski and Harry Hollien. Euclidean distance between long-term speech spectra as a criterion for speaker identification. In *Speech Communication Seminar*, pages 301–310, Stockholm, August 1974.

[MHAL84] Jerald B. Moon and Jr Herbert A. Leeper. Speaker race identification of selected Adult North American Indians. *Folia Phoniatrica*, 36:174–181, 1984.

[MHSN73] Hiroshi Matsumoto, Shizuo Hiki, Toshio Sone, and Tadamoto Nimura. Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Trans. Audio and Electroacoustics*, AU-21(5):428–436, 1973.

[Mil86] J. Bruce Millar. Quantification of speaker variability. In *Proc. First Australian Conf. Speech Science and Technology*, pages 228–233, Canberra, November 1986.

[Mil88] J. Bruce Millar. Stability of long term acoustic features. In *Proc. Second Australian International Conf. Speech Science and Technology*, pages 314–319, Sydney, November 1988.

[Mil91] J. Bruce Millar. Private communication, May 1991.

[MML88] E. Mencel, J.B. Moon, and H.A. Leeper. Speaker race identification of North American Indian children. *Folia Phoniatrica*, 40(4):175–182, 1988.

[MOJ77] John D. Markel, Beatrice T. Oshika, and Augustine H. Gray Jr. Long term feature averaging for speaker recognition. *IEEE Trans. ASSP*, ASSP-24(4):330–337, 1977.

[MT34] E. Murray and J. Tiffin. An analysis of some basic aspects of effective speech. *Arch. Speech*, 1:61–83, 1934.

[NA90] G. S. Neiman and J. A. Applegate. Accuracy of listener judgements of perceived age relative to chronological age in adults. *Folia Phoniatrica*, 42(6):327–330, November 1990.

[NND89] Jayant M. Naik, Lorin P. Netsch, and George R. Doddington. Speaker verification over long distance telephone lines. In *Proc. ICASSP 89*, pages 524–527. IEEE, 1989.

[Nod89] Hideki Noda. On the use of the information on individual speaker's position in the parameter space for speaker recognition. In *Proc. ICASSP 89*, pages 516–519. IEEE, 1989.

[OM90] J. Oglesby and J. S. Mason. Optimisation of neural models for speaker identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 261–264, 1990.

[O'S86] Douglas O'Shaughnessy. Speaker recognition. *IEEE ASSP Magazine*, pages 4–17, October 1986.

[PB52] Gordon E. Peterson and Harold L. Barney. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.*, 24(2):174–184, 1952.

[Pfe78] Larry L. Pfeifer. New techniques for text-independent speaker identification. In *Proc. ICASSP 78*, pages 283–286. IEEE, 1978.

[Pfo54] Paul L. Pfoff. An experimental study of the communication of feeling without contextual material. *Speech Monographs*, 21(2):155–156, 1954.

[PGC88] Jeff Pittam, Cynthia Gallois, and Victor Callan. The long-term acoustic characteristics of emotion. In *Proc. Second Australian Internation Conf. Speech Science and Technology*, pages 320–325, Sydney, November 1988.

[PGC90] Jeff Pittam, Cynthia Gallios, and Victor Callan. The long-term spectrum of perceived emotion. *Speech Communication*, 9(3):177–187, June 1990.

[PI90] Jeff Pittam and John Ingram. Connected speech processes in Vietnamese-Australian. In *Proc. Third Australian International Conf. Speech Science and Technology*, pages 84–89, Melbourne, November 1990.

[PM64] Sandra Pruzansky and M. V. Mathews. Talker recognition procedure based on analysis of variance. *J. Acoust. Soc. Am.*, 36(11):2041–2047, 1964.

[PPS54] I. Pollack, J.M. Pickett, and W.H. Sumby. On the identification of speakers by voice. *J. Acoust. Soc. Am.*, 26(3):403–406, 1954.

[Pri89] P. J. Price. Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, 8(3):261–277, September 1989.

[Pru63] Sandra Pruzansky. Pattern matching procedure for automatic talker recognition. *J. Acoust. Soc. Am.*, 35(3):354–358, 1963.

[PSMJ66] Paul H. Ptacek, Eric K. Sander, Walter H. Maloney, and C. C. Roe Jackson. Phonatry and related changes with advanced age. *Journal of Speech and Hearing Research*, 9:353–360, 1966.

[RB74] W. J. Ryan and K. W. Burk. Perceptual and acoustic correlates of aging in the speech of males. *Journal of Communication Disorders*, 7:181–192, 1974.

[RC78] Ellen Bouchard Ryan and Harry L. Capadano. Age perceptions and evaluative reactions toward adult speakers. *Journal of Gerontology*, 33(1):98–102, 1978.

[RD79] Alan R. Reich and James E. Duke. Effects of selected vocal disguises upon speaker identification by listening. *J. Acoust. Soc. Am.*, 66(4):1023–1028, 1979.

[RDCS73] Mark Ross, Robert J. Duffy, Harry S. Cooker, and Russell L. Sargeant. Contributions of the lower audible frequencies to the recognition of emotions. *American Annals of the Deaf*, 118:37–42, 1973.

[Rei81] Alan R. Reich. Detecting the presence of vocal disguise in the male voice. *J. Acoust. Soc. Am.*, 69(5):1458–1461, 1981.

[RH72] T. W. Rekieta and G. D. Hair. Mimic resistance of speaker verification using phoneme spectra. *J. Acoust. Soc. Am.*, 51:131(A), 1972.

[RJ90] Michael P. Rastatter and Richard D. Jacques. Formant frequency structure of aging male and female vocal tract. *Folia Phoniatrica*, 42(6):312–319, November 1990.

[RLS90] Aaron E. Rosenberg, Chin-Hui Lee, and Frank K. Soong. Sub-word unit talker verification using hidden markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 269–272, 1990.

[RMC76] Alan R. Reich, Kenneth L. Moll, and James F. Curtis. Effects of selected vocal disguises upon spectrographic speaker identification. *J. Acoust. Soc. Am.*, 60:919–925, 1976.

[Ros71] Aaron E. Rosenberg. Listener performance in speaker verification tasks. *J. Acoust. Soc. Am.*, 50(1):106(A), 1971.

[Ros72] Aaron E. Rosenberg. Listener performance in a speaker-verification task with deliberate imposters. *J. Acoust. Soc. Am.*, 51:132(A), 1972.

[Ros76] Aaron E. Rosenberg. Automatic speaker verification: A review. *Proc. IEEE*, 64(4):475–487, April 1976.

[RR83] Lorraine A. Ramig and Robert L. Ringel. Effects of physiological aging on selected acoustic characteristics of the voice. *Journal of Speech and Hearing Research*, 26:22–30, 1983.

[SAKG83] L.A. Streeter, W. Apple, R.M. Krauss, and K.M. Galotti. Acoustic and perceptual indicators of emotional stress. *J. Acoust. Soc. Am.*, 73(4):1354–1360, 1983.

[Sam75] Marvin R. Sambur. Selection of acoustic features for speaker identification. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-23(2):176–182, 1975.

[Sam76] Marvin R. Sambur. Speaker identification using orthogonal linear prediction. *IEEE Trans. Acoust., Speech, Signal Proc.*, ASSP-24(4):283–289, 1976.

[SAS85] SAS Institute Inc. *SAS User's Guide: Basic*, version 5 edition edition, 1985.

[Sch68] Martin F. Schwartz. Identification of speaker sex from isolated, voiceless fricatives. *J. Acoust. Soc. Am.*, 43(5):1178–1179, 1968.

[SDH87] Thomas Shipp, E. Thomas Doherty, and Harry Hollien. Some fundamental considerations regarding voice identification. *J. Acoust. Soc. Am.*, 82(2):687–688, August 1987.

[SF78] Shuzo Saito and Sadaoki Furui. Personal information in dynamic characteristics of speech spectra. In *Proc. 4th Int. Joint. Conf. on Pattern Recognition 78*, pages 1014–1018. IEEE, 1978.

[SG90] Michael Savic and Sunil K. Gupta. Variable parameter speaker verification system based on hidden markov modelling. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 281–284, 1990.

[SH69] Thomas Shipp and Harry Hollien. Perception of the aging male voice. *Journal of Speech and Hearing Research*, 12:703–710, 1969.

[SH87] P. M. Spinks and S. M. Hiller. *Audlab User's Guide*, February 1987.

[SKG+77] Lynn A. Streeter, Robert M. Krauss, Valerie Geller, Christopher Olson, and William Apple. Pitch changes during attempted deception. *Journal of Personality and Social Psychology*, 35(5):345–350, 1977.

[Ski35] E. Ray Skinner. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness. *Speech Monographs*, 2:81–137, 1935.

[SLF74] Lo-Soun Su, K.-P. Li, and K. S. Fu. Identification of speakers by use of nasal coarticulation. *J. Acoust. Soc. Am.*, 56(6):1876–1882, December 1974.

[SMB81] M. Shridar, N. Mohankrishnan, and M. Baraniecki. Text-independent speaker recognition using orthogonal linear prediction. In *Proc. ICASSP 81*, pages 197–200. IEEE, 1981.

[Smi62] J. E. Keith Smith. Decision theoretic speaker recognizer. *J. Acoust. Soc. Am.*, 34:1988(A), 1962.

[SNS85] A. Schmidt-Nielsen and Karen R. Stern. Identification of known voices as a function of familiarity and narrow-band coding. *J. Acoust. Soc. Am.*, 77(2):658–663, 1985.

[SNS86] A. Schmidt-Nielsen and Karen R. Stern. Recognition of previously unfamiliar speakers as a function of narrow-banded processing and speaker selection. *J. Acoust. Soc. Am.*, 79(4):1174–1177, 1986.

[Spe88] Linda E. Spencer. Speech characteristics of male-to-female transsexuals: A perceptual and acoustic study. *Folia Phoniatrica*, 40(1):31–42, January 1988.

[SR68] Martin F. Schwartz and Helen E. Rine. Identification of speaker sex from isolated, whispered vowels. *J. Acoust. Soc. Am.*, 44(6):1736–1737, 1968.

[SR88] Frank K. Soong and Aaron E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-36(6):871–879, June 1988.

[SRB82] R. Schwartz, S. Roucos, and M. Berouti. The application of probability density estimation to text-independent speaker identification. In *Proc. ICASSP 82*, pages 1649–1652. IEEE, 1982.

[SS72] Linda C. Sobell and Mark B. Sobell. Effects of alcohol on the speech of alcoholics. *Journal of Speech and Hearing Research*, 15:852–860, 1972.

[SSC82] Linda C. Sobell, Mark B. Sobell, and Robert F. Coleman. Alcohol-induced dysfluency in nonalcoholics. *Folia Phoniatrica*, 34:316–323, 1982.

[Sto81] Margaret L. Stoicheff. Speaking fundamental frequency characteristics of non-smoking female adults. *Journal of Speech and Hearing Research*, 24:437–441, 1981.

[SWCW68] K.N. Stevens, C.E. Williams, J.R. Carbonell, and Barbara Woods. Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *J. Acoust. Soc. Am.*, 44(6):1596–1607, 1968.

[TG90] P. D. Templeton and B. J. Guillemin. Speaker identification based on vowel sounds using neural networks. In *Proceedings of the Third Australian International Conference on Speech Science and Technology*, pages 280–285, Melbourne, November 1990.

[Tit89] Ingo R. Titze. Physiological and acoustic differences between male and female voices. *J. Acoust. Soc. Am.*, 85(4):1699–1707, April 1989.

[TK86] Tohru Takagi and Hisao Kuwabara. Contributions of pitch, formant frequency and bandwidth to the perception of voice-personality. In *Proceedings of ICASSP, 1986*, pages 889–892. IEEE, 1986.

[TKE68] F. Trojan and K. Kryspin-Exner. The decay of articulation under the influence of alcohol and paraldehyde. *Folia Phoniatrica*, 20:217–238, 1968.

[TOL+72] Oscar Tosi, Herbert Oyer, William Lashbrook, Charles Pedrey, Julie Nicole, and Ernest Nash. Experiment on voice identification. *J. Acoust. Soc. Am.*, 51(6):2030–2043, 1972.

[TON72] Oscar Tosi, Herbert Oyer, and Ernest Nash. Latest developments in voice identification. *J. Acoust. Soc. Am.*, 51:132(A), 1972.

[TOP+71] Oscar Tosi, Herbert Oyers, Charles Pedrey, William Lashbrook, and Julie Nicol. An experiment on voice identification by visual inspection of spectrograms. *J. Acoust. Soc. Am.*, 49(1):138(A), 1971.

[Tos79] Oscar Tosi. *Voice Identification: Theory and Legal Applications.* University Park Press, Baltimore, 1979.

[Voi64] William D. Voiers. Perceptual bases of speaker identity. *J. Acoust. Soc. Am.*, 36(6):1065–1073, 1964.

[Voi79] William D. Voiers. Toward the development of practical methods of evaluating speaker recognition. In *Proc. ICASSP 1979*, pages 793–796. IEEE, 1979.

[Wag78] Michael Wagner. *The Application of a Learning Technique for the Identification of Speaker Characteristics in Continuous Speech.* PhD thesis, Australian National University, Canberra, September 1978.

[WD75] D. A. Wasson and R. W. Donaldson. Speech amplitude and zero crossings for automatic identification of human speakers. *IEEE Trans. Acoust., Speech, Signal Proc.*, pages 390–392, 1975.

[Wea24] A. T. Weaver. Experimental studies in vocal expressions. *Journal of Applied Psychology*, 8:23–56, 1924.

[WH80] Kim A. Wilcox and Yoshiyuki Horii. Age and changes in vocal jitter. *Journal of Gerontology*, 35(2):194–198, 1980.

[WKK+83] J. Wolf, M. Krasner, K. Karnofsky, R. Schwartz, and S. Roucos. Further investigation of probablistic methods for text-independent speaker identification. In *Proc. ICASSP 83*, pages 551–554. IEEE, 1983.

[WM72] Ronald E. Walpole and Raymond H. Myers. *Probability and Statistics for Engineers and Scientists.* Collier-Macmillan, New York, New York, 1972.

[Wol72] Jared J. Wolf. Efficient acoustic parameters for speaker recognition. *J. Acoust. Soc. Am.*, 51(6):2044–2056, 1972.

[Woo78] C. A. Wood. Speaker identification by analysis of sound islands. *J. Acoust. Soc. Am.*, 64(Suppl. 1):S183, 1978.

[WS72] Carl E. Williams and Kenneth E. Stevens. Emotions and speech: Some acoustic correlates. *J. Acoust. Soc. Am.*, 52(4):1238–1250, 1972.

[XOM89] L. Xu, J. Oglesby, and J. S. Mason. The optimisation of perceptually-based features for speaker identification. In *Proc. ICASSP 89*, pages 520–523. IEEE, 1989.

[YI84] M. Yokoyama and K. Inoue. Extraction of information on personal perception based on the sense of voice quality. *Japanese Journal of Ergonomics*, 20(1):41–48, 1984.

Thesis Corrections

The following pages comprise thesis corrections requested by the examiners. The corrections are in four parts; an errata section; an addition (section 4.1) to Chapter 4; an addition (section 5.1.1) to Chapter 5; and a new conclusion chapter.

The errata section covers typographical errors, errors of expression, and other corrections possible for such a format. The additional sections (4.1 and 5.1.1) deal with the issues of linguistic/phonetic normalisation and sentence choice for the speech database collected. The new conclusion chapter (Chapter 11) replaces the old chapter and ties the results of the thesis more closely to other work in the area.

Errata

Page 1, Paragraph 3, Line 5 *Change* "Speaker recognition systems..." *to* "Automatic speaker recognition systems...".

Page 4, Paragraph 3, Line 3 *Change* "...answer such questions as..." *to* "...address such issues as...".

page 5, Paragraph 2, Lines 3-4 *Change* "Stimulus material, consisting of a sentence from one of 40 speakers, was presented to a group of 10 listeners in pairs." *to* "Stimulus material, consisting of pairs of sentences from a speaker population of 40, were presented to a group of 10 listeners."

Page 5, Paragraph 4, Line 3 *Change* "(15 male, 6 female)" *to* "(15 male, 9 female)".

Page 9, Paragraph 2, Line 8 *Change* "For bandwidth alterations accurate speaker recognition was lost at a scale of..." *to* "Speaker recognition rates fell to chance level for bandwidth alterations of...".

Page 9, Paragraph 3, Line 5 *Change* "...the energy termed based..." *to* "...the energy term based...".

Page 10, Paragraph 6, Line 1 *Change* "Streeter, Apple, Draus and Coalatti..." *to* "Streeter, Apple, Krauss, and Galotti...".

Page 22, Paragraph 3, Line 2 *Change* "...on the criteria of Σ..." *to* "...on the criteria of maximum Σ...".

Page 23, Paragraph 6, Line 7 *Change* "...dropped rapidly showing..." *to* "...dropped rapidly, with respect to increasing averaging time, showing...".

Page 25, Paragraph 4, Line 5 *Change* "...an equal error rate of..." *to* "...an equal error rate (false acceptance rate equals false rejection rate) of...".

Page 27, Paragraph 4, Line 1 *Change* "...approach to examing encoded..." *to* "... approach to examining encoded...".

Page 28, Paragraph 4, Line 11 *Change* "...of Log Area Ration..." *to* "...Log Area Ratio...".

Page 33, Paragraph 7, Line 2 *Change* "...three dialects of Australian English:..." *to* "...three varieties of Australian English:...".

Page 35, Paragraph 1, Line 5 *Change* "...in laymens' speech." *to* "...in laymen's speech."

Page 36, Paragraph 4, Line 6 *Change* "...more centralised." *to* "...more centralised within the range."

Page 37, Paragraph 2, Line 6 *Change* "...and open trails,..." *to* "...and open trials,...".

Page 45, Paragraph 4, Line 7 *Change* "...examine prosodic or suprasegmental..." *to* "...examine time varying properties of prosodic or suprasegmental parameters for...".

Page 54, Paragraph 1, Line 8 *Change* "This ranking system..." *to* "This hierarchical comparison and ranking system...".

Page 50, Paragraph 3, Line 4 *Change* "..., in th form..." *to* "..., in the form...".

Page 73, Paragraph 5, Line 1 *Change* "...discriminate analysis..." *to* "...discriminant analysis...".

Page 58, Paragraph 5, Line 1 *Change* "...the median-3 filter was:" *to* "...the mean-3 filter was:"

Page 58, Paragraph 6, Line 8 *Add* "No form of smoothing was performed across voicing boundaries for the F_0 contours."

Page 59, Paragraph 3, Line 1 *Change* "...through experimental runs with..." *to* "...through experimental trials with...".

Page 59, Paragraph 4, Line 3-4 *Change* "...ensures the separation of voicing from F_0..." *to* "...ensures the separation of the glottal source into voicing and F_0...".

Page 62, Paragraph 1, Line 5 *Change* "...therefore desirous to..." *to* "...therefore desirable to...".

Page 62, Paragraph 2, Line 3 *Change* "...properties, of two contours." *to* "properties, though also incorporating aspects of the time-invariant nature (see section 5.2.2) of two contours."

Page 63, Paragraph 1, Line 2 *Change* "...two contours. There are..." *to* "...two contours. Hence the DTW derived measures, will be referred to as dynamic measures while the previously defined static measures will be use to examine time-invariant encoding. In order to measure time-dependent encoding as accurately as possible normalisation (see section 5.2.2) is also applied in order to eliminate time-dependent encoding information from the DTW (dynamic) measures. There are...".

Page 64, Paragraph 6, Line 2 *Change* "...a theoretical optional diagonal..." *to* "...a theoretical optimal diagonal (simple macro-linear time normalisation)...".

Page 66, Paragraph 1, Line 5 *Change* "...respect to to the..." *to* "...respect to the...".

Page 66, Paragraph 2, Line 6 *Change* "...its a good match." *to* "...it is a good warp (time) match."

Page 72, Paragraph 2, Lines 2-4 *Change* "The warp path...to use it." *to* "Dynamic time warping performs both macro and micro time normalisation on two contours in order to minimise the total difference between the two contours. The DTW distance, conventionally used in most DTW approaches, is a measure of this total difference. However, the warp path, calculated in the process of the time normalisation and distance calculation, encodes the macro and micro time normalisation performed, and hence the relative dynamics of the two contours. This information encoded in the warp path is not used by conventional DTW based systems."

Page 86, Paragraph 1, Line 2 *Change* "...and Ingram [PGC90] or..." *to* "...and Ingram [PI90] or...".

Page 86, Paragraph 5, Line 1 *Change* "Based on all four plots..." *to* "Based on all three plots...".

Page 92, Paragraph 1, Line 4 *Change* "...and contrasts with Furui..." *to* "...and contrasts to a degree with Furui...".

Page 95, Paragraph 3, Line 1 *Change* "...for all sentences dynamic..." *to* "for three of the four sentences dynamic...".

Page 95, Paragraph 5, Line 2 *Change* "This is not surprising as dialect..." *to* "One possible explanation for this is that dialect..."

Page 98, Table 6.8, Caption *Add* "Results are for all four sentences combined."

Page 101, Paragraph 3, Line 5 *Change* "Surprisingly there is a drop..." *to* "There is a drop...".

Page 116, Paragraph 2, Line 5 *Change* "...dialect differences exits across the..." *to* "...dialect differences exist and may be detected across the...".

Page 119(122), Paragraph 5, Lines 5-6(1-2) *Change* "Fujisaki [FH82,Fuj88] considers... based on a linear F_0." *to* "One area of continuing debate regarding F_0 is the best scale, whether on the basis of production mechanism or automated recognition results, to represent F_0. Proponents such as Fujisaki [FH82,Fuj88] cite a log scale for F_0 on the basis that the vocal fold movement under syntactic and lexical constraints is best modelled logarithmically. On the basis of these results a log scale model appears superior to a linear scale model in terms of speaker recognition performance."

Page 130, Paragraph 5, Line 5 *Change* 88.2% *to* 80.3%.

Page 130, Paragraph 5, Line 8 *Change* 3.1% *to* 2.8%.

Page 135, Paragraph 4, Line 2 *Change* "...process, and generally found to..." *to* "...process, and in 2 of the 3 cases was found to...".

Page 135, Paragraph 5, Lines 3-6 *Change* "In section 5.4...were proposed." *to* "In section 5.4 two variants on the DTW distance were proposed. These were the Weighted DTW distance, which sought to include warp path derived information in the distance; and the Border DTW distance, which eliminated two leading and trailing values from the interval over which the distance is calculated."

Page 143, Paragraph 1, Line 2 *Change* "...(level), by up to a factor of 2, worse..." *to* "...(level) worse...".

Page 144, Paragraph 1, Line 1 *Change* "...of of..." *to* "...of...".

Page 148, Paragraph 4, Line 4 *Change* "...'consistently' strong correlations, markedly in excess of those of other measures." *to* "...correlation values markedly in excess of those of other measures."

Page 155, Paragraph 3, Line 1 *Change* "...be garnered from the plot." *to* "...be gathered from the plot."

Page 155, Paragraph 3, Line 6 *Change* "...if such a policy were..." *to* "if such a policy of using speaker dependent threshold values for automatic speaker recognition were...".

Page 162, Paragraph 4, Line 6 *Change* "Therefore it could reasonably expected..." *to* "Therefore it could reasonably be expected...".

Page 163, Paragraph 1, Line 3 *Change* "Possibly there..." *to* "The reason for this phenomenon is unclear though possibly there...".

Page 163, Paragraph 3, Line 4 *Change* "...other two. Thus, while..." *to* "...other two. This may be attributable to the fact that the warp path and the warp distance are extracting two different and separate items of data and that seeking to incorporate them into a single measure in such a manner leads to a loss, rather than gain, of information. Thus, while...".

Page 164, Paragraph 5, Line 2 *Change* "Surprisingly, while..." *to* "While...".

Page 164, Paragraph 6, Line 2 *Change* "...superior parameter with..." *to* "...superior parameter (as has been suggested by previous research [Wea24,HHP88]) with...".

Page 164, Paragraph 6, Lines 3-4 *Change* "...with voicing at a surprising 71.8%..." *to* "...with voicing yielding a discrimination rate of 71.8%...".

Page 165, Paragraph 3, Line 4 *Change* "...single sentence DTW distance discriminated sex with a mean of 88.2% as opposed..." *to* "...single sentence, DTW distance discriminated sex with a mean of 88.2%, as opposed...".

Page 165, Paragraph 4, Line 3 *Change* "...discrimination rates. There appears..." *to* "...discrimination rates. These may be attributable to the fact that the DTW distance and the warp path measures extract different sources of information:– combining them into a single measure leads to information loss. There appears...".

Page 165, Paragraph 6 *Add* "The application of such an individual thresholding system in order to determine speaker sex pre-supposes a knowledge of speaker identity:- a finer categorisation than speaker sex. However analysis of the variations in threshold between speakers may lead to new categorisations of speakers and hence a realistic pre-processing stage that may be applied to determine threshold levels for a speaker sex decision mechanism."

Page 166, Paragraph 2, Line 8 *Change* "...degree of accuracy based..." *to* "...degree of accuracy (though the maximum result achievable remains unclear) based...".

Page 169, Paragraph 3, Line 1 *Change* "Ideally, a..." *to* "Extending the analysis of the four sentences, ideally a...".

Page 179, Paragraph 1, Lines 3-6 *Change* "Parameters of...characteristic examined." *to* "Parameters (timing, F_0, voicing, and energy) of the utterances will be systematically altered to derive new utterances. The new utterances will be played to listeners and the results will be used to gauge the correlation between listener perception of the speaker characteristic and the parameter altered."

Page 180, Paragraph 1 *Add* "For both the initial exploratory experiment and the following reported experiments listeners were given an initial familiarisation period in which they heard the utterances of A and B repeatedly. The presentation of each unknown utterance to which the listeners were asked to respond was preceded by the same utterance from A and B, in order that the listener had a ready reference for the two speakers."

Page 180, Paragraph 2 *Add* "Figures 9.4 to 9.17 and tables 9.1 to 9.14 present the results of the experiments."

Page 183, Paragraph 6, Line 1 *Change* "...parameter it maybe seen..." *to* "...parameter it may be seen...".

Page 183, Paragraph 6, Line 2 *Change* "..., and in all cases bar 1 higher, shifts..." *to* "..., and in all cases bar 1, higher shifts...".

Page 189, Paragraph 1, Line 4 *Change* "...like' A. In general..." *to* "...like' A. Another possible explanation is the strong double peak in B's original energy contour between 15 and 30 time frames. Warping B's energy tends to destroy this double peak while warping A's energy can not create the strong double peak. In general...".

Page 198, Paragraph 4, Line 8 *Change* "...comparisons the was..." *to* "...comparisons there was...".

Page 199, Paragraph 4, Line 4 *Change* "...and the hence the..." *to* "...and hence the...".

Page 208, Paragraph 2 *Add* "The results based on the warped energy serve as a reminder of the limitations on warping. When there are marked differences in the two contours being compared (such as the extreme double peaks in B's energy contour between 15 and 30 time frames), particularly in terms of range, dynamic time warping will only perform with limited success and some form of normalisation is required if greater success is required."

Page 208, Paragraph 5 *Add* "More analysis of this phenomenon is required."

Additional Sections

4.1 Linguistic-Phonetic Control

One possible source of variability between different instances of the same utterance is that corresponding to the communicative intent. Speakers repeating the same sentence on different occasions may use different intonation patterns (e.g., falling/rising) or place stress at different points within the sentence. This linguistically/phonetically induced variability is beyond the scope of the current work but its influence must be minimised, if not eliminated, in order that it does not affect the results obtained (due to the influence of these linguistic/phonetic factors on the acoustic parameters being examined).

Following each recording session by each speaker; all material recorded during that session was listened to by the investigator. If linguistic/phonetic variations were detected the speaker was asked to re-record the session or individual sentences. In practice a small number of speakers fell into this category of being asked to re-record a session for linguistic/phonetic reasons. By the third recording session all speakers were found to be adept at maintaining consistency.

A second stage of listening occurred during the digitisation process during which each sentence was manually extracted from the tape. At this stage if linguistic/phonetic variability was detected by the investigator the particular utterance was marked as an "error". The number of errors for each of the 15 sentences in the recording set was then used to select the four sentences (due to time, storage, and processing constraints) with the least errors that also met other criteria; such as including different sentence types in the final set of four sentences. If any instances of the four selected sentences were marked as errors they were discarded and not used in experimentation.

It is worth remarking that linguistic/phonetic derived variability is unlikely to neatly parallel the speaker characteristics dialect or sex and is also highly likely (if present) to be inconsistent for a single speaker. In effect any linguistic/phonetic variability not eliminated will most likely adversely affect any results obtained; *not* lead to artificially inflated results.

5.1.1 Material Inadequacies

Unfortunately the four sentences selected for analysis do not comprise an ideal set for examination. The sentences taken together are not phonetically balanced and do not represent the full range of typical Australian English sentence types. Constraints of time, processing capability

and data availability meant that a compromise in terms of volume and content of speech material used in experiments was required. While these compromises must of necessity impinge upon the generality of results they do not invalidate the results obtained.

At the time of data collection it was envisaged that all fifteen sentences would be used in the analysis experiments. In many ways speech research is data driven:- the fifteen sentences were selected for historical reasons; both due to previous well known studies in the area of speaker recognition [Dod71b, Wol72], and due to the availability of already recorded data from a number of North American and Australian speakers.

However, due to the detailed nature of the analysis conducted and time constraints of the thesis it was necessary to select a subset of the sentences for analysis. A subset size of four sentences was decided upon as a compromise between the objectives of completing the thesis while also attempting to adequately capture some degree of acoustic and phonetic diversity. All utterances of each of the fifteen possible sentences were analysed as to possible linguistic/phonetic variation and the number of such occurrences and plain pronunciation errors marked for each sentence. The four sentences were then selected so as to effectively minimise the errors or linguistic/phonetic variation in the subset selected. Other factors including mean sentence duration, sentence "type" (declarative, question etc.), and phonemic content (phonetic balance of list) were also taken into account. These later factors were subordinate to the goal of minimising pronunciation errors and linguistic/phonetic variation in the subset members in order that sufficient speech data (after the elimination of utterances showing such errors or variation) be employed to allow meaningful statistical analysis of any results obtained. The set of four sentences chosen is the result of such a compromise.

A phonemic analysis of the four selected sentences shows that a few phonemes of Australian English are missing. In particular there are no voiced fricatives (two voiceless), no laterals, and no back monopthongs. The phonetic composition of the four sentences taken together is the result of a number of decision factors, only one of which was phonetic balance. It is reasonable to expect that higher performance than that reported here would be obtainable with a more balanced sentence set.

It is worth noting that *were* the sentence set artificially designed (which was not the case) so as to yield optimal performance the results obtained on such data would still (regardless) have application in such areas as automatic speaker recognition. For such tasks implementors attempt to obtain maximum recognition performance and often design of an 'optimal' utterance set is one way to obtain improvements.

Therefore, the four sentences used for analysis are derived from a series of decisions, constraints, goals, and compromises including historical precedence, time limitations, and attempting to minimise linguistic/phonetic variability. The resulting sentences are not phonetically balanced, and a larger more balanced set of sentences is required to adequately examine and address all the issues raised relating to data-dependence of results. However, given the constraints and goals imposed, the sentences selected do allow meaningful analysis of the encoding of speaker characteristics in prosodic parameters.

Chapter 11

Conclusion

A database of sentence-long utterances from nineteen adult speakers of Australian English was collected. Four prosodic parameters— energy, fundamental frequency, voicing, and zero crossing rate were extracted and analytical and perceptually based investigations of the parameters' correlations to the speaker characteristics identity, sex, and dialect were carried out. Via this relatively uncommon approach (Lass, Linville, Pruzansky and Xu [LMK78, LF85a, Pru63, XOM89]) of combining perceptual and analytic methods a number of previously known results were confirmed, and new results regarding the relationship of the parameters to the speaker characteristics were discovered.

Analytical experiments were conducted using four of the fifteen different recorded sentences. Discriminant analysis was applied to examinations of the characteristics identity and sex, and least-squares-fit analysis for speaker dialect. Twenty-one measures of the properties of each of the parameters were examined, the measures being logically split into two groups:- dynamic—measures of the time varying properties of the parameter contours, and static—measures of the time invariant properties of the parameters.

It was found that identity, sex, and dialect could be detected to significant degrees based on the parameters and sentences used:- identity and sex discrimination at 75% and 96% respectively, and dialect correlated at 0.58. These results show the high degree of speaker characteristic encoding in the prosodic parameters and further the research of other scientists who have examined prosodics; Doddington [Dod71b], Lummis [Lum73], and Wasson and Donaldson [WD75] for speaker identity; Weaver and others [Wea24, HHP88, Mil88] for speaker sex; and Fokes and others [FBS84, Ada71] for speaker dialect. Further, the results illustrate the utility of prosodic acoustic parameters as inputs to automatic systems for the determination of speaker characteristics; for example automatic speaker verification systems.

In order to examine the form of identity, sex, and dialect encoding in the prosodic parameters, time varying and time invariant properties of the parameters were investigated separately. Several researchers:- Furui [Fur81b], Soong and Rosenberg [SR88], and Bernasconi [Ber90] have performed similarly motivated investigations for identity alone, examining only spectral acoustic parameters. The current thesis expands on previous work by employing a new technique, investigating prosodic parameters, and examining multiple speaker characteristics rather than

speaker identity alone.

Comparison of dynamic measure based, and static measure based discrimination and correlation rates showed that for all three speaker characteristics the dynamic measure set performance was equal to or superior to that of the static set, though combined performance exceeded that of either alone. Clearly the dynamic measures of the prosodic parameters extract more speaker-related information than the static measures, though dynamic measures do not encapsulate all of the information extracted by the static measures.

Normalisation, that is linear shifting of parameter contours into the range 0–1, was used in order to 'distill' the dynamic properties of the parameters. This technique does not appear to have been previously examined. Besides enhancing the separation of time varying and time invariant properties of parameter contours, the technique appears to have applications to such methods as Artificial Neural Networks which often need normalisation of inputs to ensure stability. Discrimination rates for identity, and correlation for dialect, dropped little following normalisation, while sex discrimination rates dropped sharply. Discrimination or correlation rates for each characteristic were contrasted between static measures of the non-normalised parameters and dynamic measures of the normalised parameters. Significant differences were found in all cases such that speaker identity and dialect were found to be more strongly encoded in the time varying properties of the contours, while speaker sex was more strongly encoded in the time invariant properties.

A novel extension of the dynamic time warp algorithm was employed based on significant enhancements to early work by Saito and Furui [SF78]:- measures of the calculated warp path between two contours were examined for speaker characteristic encoded information. Most DTW based schemes implicitly calculate the warp path and discard it after deriving the DTW distance. It was found that both speaker identity and dialect were strongly encoded in the warp path measures, to such an extent that they were significantly 'better' than the DTW distance. For speaker sex the warp path measures were marginally inferior to the DTW distance. Clearly the DTW warp path—a measure of the relative dynamics of two contours—encodes temporal related speaker characteristic information to a high degree, which may be used to discriminate the speaker characteristics.

The illustrated improvements in discrimination for a DTW based scheme which incorporates measures of the warp-path over a conventional DTW based scheme has several implications. Firstly re-evaluations of DTW based schemes when compared to other decision mechanisms such as Hidden Markov Models; e.g., Naik, Netsch, and Doddington [NND89]; should be carried out such that warp-path measures are incorporated in the DTW scheme. Secondly, it appears reasonable to expect that the performance of systems based on DTW, such as the speaker verification at Texas Instruments [Dod85], could be markedly improved via the incorporation of warp path measures.

The four basic parameters— energy, fundamental frequency, zero crossing rate, and voicing were individually investigated for encoded speaker characteristic information. *All* four parameters were found to have encoded information relating to *each* of the three characteristics. Fundamental frequency was the 'best' (highest degree of encoding) parameter for all three speaker

characteristics, though degree of encoding, and in general the relative 'worth' of parameters was speaker characteristic dependent. These results contribute to the body of knowledge regarding the encoding of speaker characteristics in individual prosodic parameters:- while considerable work has been done on fundamental frequency and identity [Dod71b, Ata72] or sex [Wea24, HHP88] the other three parameters have received little or no attention as to their encodings of identity [JHH84], sex [Mil88] or dialect.

Four variant representations of fundamental frequency, based on log versus linear scale and interpolation across unvoiced or concatenated voiced only, were examined. For both speaker identity and dialect the log representation of F_0 was found to be markedly superior to the linear scale. In all cases interpolation was equal or superior to concatenation. Though these results are inconclusive as evidence supporting or denying a representation of fundamental frequency based on physical articulator motion constraints [FH82, Fuj88] the results show the utility of using a log scale for fundamental frequency, when F_0 is used as an input to an automatic recognition system; e.g. Doddington [Dod71b], Atal [Ata72] or Chen and Lin [CL87].

Discriminant or correlation performance for the three speaker characteristics was adequately modelled as a growth function of the amount of speech material used. These results continue early work by Pollack et. al. [PPS54] who modelled listener performance for judgement of identity tasks as a function of utterance duration. However, factors other than amount of speech material appear to influence discriminant/correlation levels so that accurate estimates of optimal performance were not possible, nor were general guidelines regarding choice of utterance for text-dependent recognition systems. Clearly these issues of utterance duration and content have major implications for the design and implementation of automatic systems (e.g., automatic speaker verification) yet little research has been conducted in the area and further work is required.

The twenty-one measures were individually compared and contrasted. It was found that all measures extracted some encoded speaker-related information and that no single measure stood out as being consistently strong for all combinations of characteristic-parameter. Clearly, the form or nature of encoding of speaker characteristics is parameter and speaker characteristic dependent.

Discriminant and correlation results were analysed on the basis of the individual speakers in the speaker population. For all three speaker characteristics results were found to be variable between individual speakers, and in particular highly variable for speaker dialect (showing that prosodic correlates of dialect are general population 'trends', not firm constraints). That is, that for all three speaker characteristics their encoding within the parameters was speaker dependent. Very little concerted research has been conducted in this area of speaker dependency of results, or the uniqueness of the speakers in the speaker population [Nod89]. These issues of speaker population homogeneity/heterogeneity on a number of scales have major implications for such important areas of speech technology as database design, collection and evaluation, and the performance evaluation of automatic speech recognition and automatic speaker recognition systems. Clearly there is considerable scope for more research to be conducted in this area.

Perceptual experiments were conducted using a single sentence and a different subset of the

speaker population for each of the three speaker characteristics. A novel method of analysis-resynthesis using linear prediction to construct a composite utterance from a number of utterances, and allowing the individual manipulation and alteration of energy, fundamental frequency, voicing, and timing was devised and used to evaluate listener utilisation of acoustic cues to speaker characteristics. A number of other researchers have examined speaker characteristic encoding via the manipulation of acoustic parameters with different degrees of sophistication:- Lass et. al. [LMK78, LAJ80], Van Lancker et. al. [LKE85, LKW85], Childers et. al. [CYW85, CWH87a, CWH87b, CWHY89], Takagi and Kuwabara [TK86] and Dommelen [Dom90]. The analysis-resynthesis technique of the thesis ranks among the most sophisticated approaches to parameter alteration and is unique in the use of multiple (more than 2) speakers to build composite utterances.

Identity perception experiments revealed that listeners used prosodic parameters to identify speakers with a high degree of accuracy based on a combination of the four examined parameters. These results further the work of other researchers, such as Dommelen [Dom90] and Johnson [Joh90] who have chiefly examined fundamental frequency, as the prosodic cue for speaker identification.

Weighting of parameters as perceptual cues to identity was found to vary between the parameters and be dependent upon speaker. That is, listener cue utilisation was speaker dependent. In a parallel of the analytical examination of the warp path (dynamic differences) parameter contours from speakers were warped to match that of the different speakers' contours for the same sentence. In some parameter-speaker combinations listener perception was altered significantly based on the warping— showing listener utilisation of the dynamics of a contour for identity perception—while in others little or no alteration occurred. Clearly in some cases at least listeners use the dynamic (time varying) properties of a parameter more than its static (time invariant) properties to form judgements of speaker identity. This new method of contour warping and perceptual trials requires considerable further investigation both of the technique itself and the results.

Sex perception experiments showed the significance of mean fundamental frequency in listener perception of sex. This result confirms the well known pre-eminence of mean fundamental frequency in listener perception of speaker sex [Col76, LHB+76, Joh90]. Different dynamics of F_0 and the parameters energy, and voicing were found to have no significant influence upon listener perception of sex.

Dialect perception experiments were conducted using utterances of speakers from either end of the dialect spectrum. Naive listener response was generally found to be consistent, though under certain conditions of parameter alteration it became highly variable. This result tends to confirm that of Brennan et. al. [BRD75], that naive listeners are capable of judging degree of accent accurately and consistently. Encoding of the parameters energy, fundamental frequency, and voicing showed no significant shift in listener perception consistent with the dialect of the originator of the parameter. Alterations in duration of utterance were found to significantly influence results such that shorter utterances were perceived as more cultivated while lengthened utterances were perceived as broader. Whether this listener perception is an externally imposed

stereotype (e.g., media influence) and not a true representation of prosodic correlates of dialect for the Australian population, or a model drawn from listener experience is unclear. However, results of the analysis section did show the significance of duration for speaker dialect, while researchers such as Lass et. al. [LMK78, LAJ80] have shown that temporal cues do play a part in listener perceptions of accent or dialect.

The results highlight several areas in which further work may be carried out. Both analytical and perceptual techniques may be applied to other speaker characteristics, such as emotion, and to other speech parameters, such as spectral parameters. The perceptual experiments were limited in scope, the addition of more speakers and listeners would greatly strengthen and add to the results already achieved. Further examination of the phenomenon of listener perception of warped parameters is required. The assumption of the dialect difference as a linear scale may be an over-simplification and non-linear transformations may yield better results. Further investigation is required in order to determine what constitutes a good utterance for speaker recognition systems, and build accurate mathematical models of recognition performance based on parameters of the experiment— utterance, acoustic parameters, number of speakers etc. Following on from this, individual speaker variance in encoding of speaker characteristics requires further examination both so that databases may be accurately quantified and compared, and so that existing recognition systems may be 'fine tuned' by targeting 'trouble' speakers. A means of quantifying the dynamics of a contour (DTW warp path) that did not require the comparison of two contours would be advantageous. A less restricted data set, where utterances were sampled under normal conversational conditions, and hence are more variable and dynamic appears desirable. Finally, implementing measures of the warp path in an existing recognition system would allow their evaluation under practical, application conditions.