Speaker verification using a novel set of dynamic features

Author/Contributor:
Nosratighods, Mohaddeseh; Ambikairajah, Eliathamby; Epps, Julien

Publication details:
18th International Conference on Pattern Recognition
pp. 266-269
076952521 (ISBN)

Event details:
18th International Conference on Pattern Recognition
Hong Kong

Publication Date:
2006

Publisher DOI:
http://dx.doi.org/10.1109/ICPR.2006.1071

SPEAKER VERIFICATION USING A NOVEL SET OF DYNAMIC FEATURES

Mohaddeseh Nosratighods[1], Eliathamby Ambikairajah[1,2], and Julien Epps[2,1]

[1]*School of Electrical Engineering and Telecommunications,*
The University of New South Wales, Sydney, NSW 2052, Australia
[2]*National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia*
m.nosratighods@student.unsw.edu.au, ambi@ee.unsw.edu.au, julien.epps@nicta.com.au

Abstract

Dynamic cepstral features such as delta and delta-delta cepstra have been shown to play an essential role in capturing the transitional characteristics of the speech signal. In this paper, a set of new dynamic features for speaker verification system are introduced. These new features, known as Delta Cepstral Energy (DCE) and Delta-Delta Cepstral Energy (DDCE), can compactly represent the information in the delta and delta-delta cepstra. Further, it is shown theoretically that DCE carries the same information as the delta cepstrum using an entropy criterion. Experimental speaker verification results on the TIMIT database support the theoretical result, showing a significant improvement in terms of equal error rate compared with conventional feature extraction methods using delta and delta-delta cepstra.

1. Introduction

Extracting suitable dynamic, speaker dependent features has a significant effect on the recognition accuracy of an automatic speaker verification system. Several studies have investigated the usage of short-term dynamic features such as delta and delta-delta cepstra. However, they can increase the complexity of the speaker verification system by doubling or tripling the dimensionality of the feature vector. This extra dimensionality imposes more computational load and consequently more processing time to the back-end system comprising the pattern matching and decision making [1]. Furui [2,3] first utilized a combination of cepstral coefficients based on the LPC cepstra and their regression coefficients for speaker-independent isolated word recognition, and established the effectiveness of combination of temporal and dynamic features in reducing the frequency of large individual error rates.

An early promising study of the application of Maximum Entropy Discrimination (MED) for feature selection in speech recognition can be found in [4]. The MED algorithm was compared with a classical wrapper feature selection and a hybrid wrapper/MED algorithm was proposed. Recently, a new approach for speaker verification based on segmental dynamic information was introduced [5]. This method used the speaker information (entropy) included in the Gaussian component strings.

This paper introduces a novel front-end processing method using a new set of dynamic cepstral features, named Delta Cepstral Energy (DCE). DCE not only greatly reduces the dimensionality of the feature vector compared to delta and delta-delta cepstra, but also provide the same performance. The concept of speaker entropy, which conveys the information contained in one speaker's speech data based on the extracted features, is also introduced to facilitate comparative evaluation of the proposed methods.

2. Delta Cepstral Energy

Short-term dynamic features such as delta and delta-delta coefficients can be used for improving speech and speaker verification system by modelling the short-term transitional information in the speech. However, they can increase the size of the feature vector by up to 24 dimensions. Here we introduce the Delta Cepstral Energy (DCE) and the Delta-Delta Cepstral Energy (DDCE), which each compactly characterize the delta and delta-delta cepstral information in a one dimensional feature. The DCE and DDCE for a single frame are calculated as follows:

$$DCE = E_{\Delta MFCC} = \sum_{l=1}^{L} (\Delta MFCC_l)^2 , \quad (1)$$

$$DDCE = E_{\Delta^2 MFCC} = \sum_{l=1}^{L} (\Delta^2 MFCC_l)^2 , \quad (2)$$

where $\Delta MFCC_l$ and $\Delta^2 MFCC_l$ are the l^{th} delta and delta-delta cepstral coefficients respectively , and L is the number of MFCCs. The proposed dynamic features, given in (1) and (2), can track temporal information in the speech signal at significantly reduced complexity relative to delta and delta-delta cepstrum.

3. Entropy Computation

In this section, we derive expressions for the entropies of $\Delta MFCC$ and $\Delta^2 MFCC$, and for the DCE and DDCE, before employing this entropy measurement to compare the information content of these features.

3.1. Entropy of the delta cepstrum and delta-delta cepstrum

Consider a K dimensional random vector X. The entropy of X, $H(X)$ is the information content associated with the random vector X. On the other hand, $H(X)$ can also be considered as the information we are expecting to receive when outcome X occurs. The entropy of the random vector X is computed by estimating its probability density function (PDF), which can be calculated either from its histogram or parameterized distribution.

Consider the difference Z of two random vectors X and Y. The entropy of random vector Z is:

$$H(Z) = H(Z_1) + H(Z_2) + ... + H(Z_K) \quad (3)$$

if and only if $Z_1, Z_2, ..., Z_K$ are all independent random variables, for which [6]:

$$H(Z_i) = -\int f_{z_i}(z_i) \ln(f_{z_i}(z_i)) dz_i \quad (4)$$

Assume the PDF of each dimension of X follows a Gaussian distribution,

$$f_{x_i}(X_i) = \frac{1}{\sqrt{2\pi\sigma_{x_i}}} e^{-\frac{1}{2}(\frac{X_i - \mu_{x_i}}{\sigma_{x_i}})^2}, i = 1,...,K \quad (5)$$

where μ_{x_i} and $\sigma_{x_i}^2$ are the mean and variance of the i^{th} dimension. If Y also follows the Gaussian distribution with mean μ_y and σ_y^2, it can be shown that Z is itself a Gaussian random vector with mean $\mu_z = \mu_x - \mu_y$ and variance $\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2Cov(x, y)$, where $Cov(x, y)$ is the cross-covariance between random vectors X and Y. The entropy of each dimension of Z can be obtained by replacing the Gaussian PDF in equation (4):

$$H(Z_i) = \ln(\sigma_i \sqrt{2\pi e}) \quad (6)$$

If the MFCC features are assumed to follow a Gaussian distribution with a diagonal covariance matrix[1], the entropy of $\Delta MFCC$ can be expressed as

$$H(Z) = \frac{K}{2} + K \ln(\sqrt{2\pi}) + \sum_{j=1}^{K} \ln(\sigma_{\Delta MFCC_j}) \quad (7)$$

Similarly, the entropy of $\Delta^2 MFCC$ can be obtained via equation (5),

$$H(Z) = \frac{K}{2} + K \ln(\sqrt{2\pi}) + \sum_{j=1}^{K} \ln(\sigma_{\Delta^2 MFCC_j}) \quad (8)$$

3.2. Entropy of the DCE and DDCE

It can be proved [8] that for K independent Gaussian random variables Z_i with mean μ_{z_i} and variance $\sigma_{z_i}^2$, the sum of squares

$$W = \sum_{i=1}^{K} \frac{(Z_i - \mu_{z_i})^2}{\sigma_{z_i}^2} \quad (9)$$

follows a χ^2 distribution with K degrees of freedom ,

$$f(W) = \frac{1}{2^{K/2} * \Gamma(K/2)} W^{\frac{K}{2}-1} * e^{-\frac{W}{2}}, K = 1,2,... \quad (10)$$

where Γ is the gamma function. The entropy of the random variable W, which can be considered as the DCE, can then be found via equations (4) and (10),

$$H(W) = \frac{K}{2} + \ln(2\Gamma(\frac{K}{2})) + (1 - \frac{K}{2})\Psi(\frac{K}{2}) \quad (11)$$

where $\Psi(n)$ is the digamma function and is defined as,

$$\Psi(n) = \frac{d}{dx} \ln \Gamma(x) = H_{n-1} + \gamma \quad (12)$$

Here H_n and γ denote the harmonic number and Euler-Mascheroni constant, respectively.

$$H_n = \sum_{l=1}^{n} \frac{1}{l}, \quad \gamma = 0.57721 \quad (13)$$

The only difference between the entropy calculation of DCE and DDCE is in the random variable W, which is shifted and scaled by the mean and variance of $\Delta^2 MFCC$ in DDCE.

As an example, the histograms of $\Delta MFCC$, $\Delta^2 MFCC$, DCE and DDCE for a female speaker are

[1] In practice this can be achieved using cumulative distribution function matching [7].

illustrated in Figure 1, where $K = 12$. It can be clearly seen (Fig.1) that $\Delta MFCC$ and $\Delta^2 MFCC$ follow a similar Gaussian distribution, but the variance of $\Delta^2 MFCC$ is smaller than that of $\Delta MFCC$. Also, DCE and DDCE show a χ^2 distribution with $K = 12$ degree freedom with mean $K = 12$ and variance $2K = 24$.

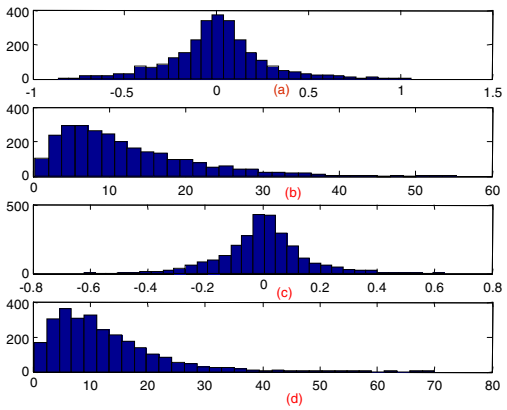

Figure1. Histogram of (a) *ΔMFCC*; (b) *DCE;* (c) *Δ²MFCC*; (d) *DDCE* for a female speaker from the TIMIT database.

3.3. Entropy Comparison of Delta Cepstrum and DCE

Comparing equations (7) and (11), let

$$P = \ln(2\Gamma(\frac{K}{2})) + (1 - \frac{K}{2})\Psi(\frac{K}{2}) - K\ln(\sqrt{2\pi}) - \sum_{j=1}^{K}\ln(\sigma_{\Delta MFCC_j}) \quad (14)$$

where P is close to zero if and only if H_{DCE} is comparable to $H_{\Delta MFCC}$. H_{DCE} can not be more than $H_{\Delta MFCC}$ in theory since DCE is completely obtained from $\Delta MFCC$, however it can be shown empirically that $P > 0\ \forall K$, in the TIMIT database. When K is increased, the first three terms in equation (11) do not increase as rapidly as the last term decreases, so H_{DCE} is very close to $H_{\Delta MFCC}$ for a larger number of Mel cepstral coefficients.

4. EXPERIMENTS

4.1. Database

Speaker verification experiments were conducted using the TIMIT database, including 168 speakers (112 male, 56 females) from the "test" portion of the database. As in [9], speaker models with 32 Gaussian mixtures were trained using eight utterances of approximately 24 seconds total duration. Again, similarly to Reynolds [9], each speaker was considered to be a claimant while the remaining speakers acted as impostors, resulting in 334 true and 52538 imposter tests.

4.2. Evaluation Measures

The evaluation of the speaker verification system is based on Detection Error Trade-off (DET) curves, which show the trade-off between false alarms (FA) and false rejection (FR) errors. The Equal Error Rate, is also computed, the threshold in which FA and FR error rates are equal.

4.3. Speaker Verification System

Twelve-dimension Mel-frequency cepstral coefficients (MFCC) were first extracted from the speech at a frame length of 20 ms, overlapped by 10 ms. Cumulative distribution function (CDF) matching [7] was applied to MFCC coefficients to warp their distributions into Gaussian distributions. DCE and DDCE were calculated based on the warped delta MFCCs and appended to form a 14-dimensional feature vector.

The pattern recognition in the speaker verification back-end was based on Gaussian mixture models (GMMs), which are widely used in text-independent speaker recognition. GMMs are an implementation of Bayesian Decision Theory, which is used to derive a two-class decision classifier in a verification system. GMMs are parametric models used to estimate continuous probability density functions from a set of multi-dimensional components. During training a background model set (BMS) for each speaker, comprising 5 maximally-spread close and 5 maximally-spread far speakers, was developed from training data [9]. Thirty-two Gaussian mixtures were used for each BMS and GMM. During testing, the score of each trial was obtained from the average likelihood ratios between the claimed target and its corresponding BMS [9].

4.4. Results

The efficacy of the DCE and DDCE was compared against the MFCC+ΔMFCC (24 dimensions) for the TIMIT database under clean conditions. The DDCE was also appended to the MFCC+DCE (14 dimensions), and this was compared with MFCC+ΔMFCC+Δ²MFCC (36 dimensions). MFCC+DCE shows a 3% improvement over MFCC+ΔMFCC, and MFCC+DCE+DDCE shows a 3% improvement over MFCC+ΔMFCC+Δ²MFCC, in terms of equal error rate, under clean conditions. Thus

the DCE can reduce the dimensionality by nearly two or three times while providing the same performance, as seen in Figure 2.

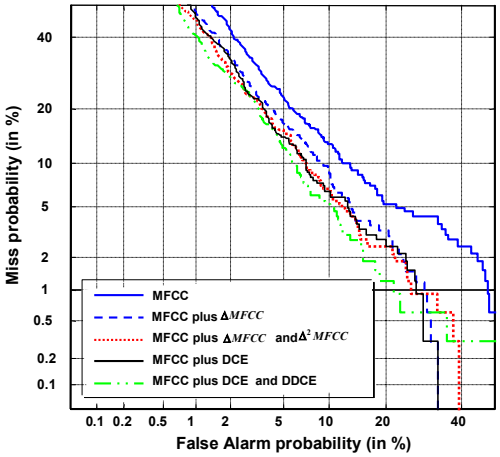

Figure 2. DET curves comparing DCE and DDCE-based front-ends with the more traditional MFCC and ΔMFCCs under clean conditions *with CDF matching*.

If CDF matching was not applied to MFCC, the results were slightly different (Figure 3). DCE and DDCE still improve performance 2% over MFCC+ΔMFCC and 3% over MFCC+ΔMFCC+Δ^2MFCC. Empirical observations show that DCE and DDCE nearly obey the Chi-Square distribution even without applying CDF matching. Hence, even in this case DCE and DDCE have greater entropy than traditional ΔMFCCs, resulting in better performance.

5. Conclusion

We introduced the DCE and DDCE, novel dynamic features for speaker verification that are substitutes for the traditional delta and delta-delta cepstra. DCE and DDCE not only reduce the feature set dimension by nearly two or three times, but also show a 3% improvement in accuracy over MFCC+ΔMFCC and MFCC+ΔMFCC+Δ^2MFCC respectively in terms of EER. This paper has shown that DCE and DDCE carry the same information as delta and delta-delta cepstrum in terms of entropy. Experimental results on the TIMIT database showed that DCE as a feature gave an improvement over the baseline MFCC+ΔMFCC. Our future work will concentrate on investigating these proposed features under noisy conditions, using NIST database.

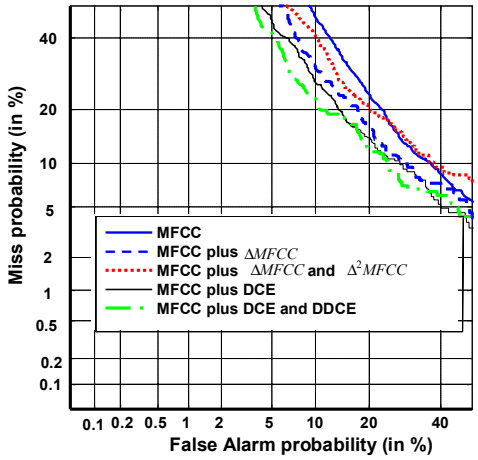

Figure 3. DET curves comparing DCE and DDCE-based front-ends with the more traditional MFCC and ΔMFCCs for clean conditions *without CDF matching*.

7. References

[1] Campbell, J.P., "Speaker Recognition:A Tutorial", Proc. of the IEEE, vol.85, no.9, 1997, pp.1437-1461.

[2] Furui, S., "Comparison of speaker recognition methods using statistical features and dynamic features", *Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 3, 1981, pp. 342-350.

[3] Furui, S., "Comparison of speaker recognition methods using statistical features and dynamic features", *Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-34, no. 1, 1986, pp. 52-5.

[4] Valente, F., Wellekens, C., "Maximun entropy discrimination (MED) feature selection for speech recognition", in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 327 – 332.

[5] Xiang, B., "Text-Independent speaker verification with dynamic trajectory model", *IEEE Signal Processing Letters*, vol. 10, no. 5, 2005, pp.141-143.

[6] Sayood, K., *Introduction to data compression,* Morgan Kaufmann Publishers, 2000.

[7] Pelecanos, J., Sridharan, S., "Feature warping for robust speaker verification", in *Proc. of a Speaker Odyssey* , 2001, pp. 213-218.

[8] Scharf, L., *Statistical Signal Processing: Detection, Estimation and Time Series Analysis,* Addison-Wesley, New York, 1991.

[9] D. A. Reynolds, R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech & Audio Processing,* 1995, pp. 91-108.