Speech based Continuous Emotion Prediction: An investigation of Speaker Variability and Emotion Uncertainty

Ting Dang

Supervisor: Dr. Vidhyasaharan Sethu

Joint Supervisor: Prof. Eliathamby Ambikairajah

A thesis submitted in fulfilment

of the requirement for the degree of

Doctor of Philosophy

School of Electrical Engineering and Telecommunications

Faculty of Engineering

June 2018

UNSW SYDNEY | Australia's Global University

Thesis/Dissertation Sheet

Surname/Family Name	:	**Dang**
Given Name/s	:	**Ting**
Abbreviation for degree as give in the University calendar	:	**PhD**
Faculty	:	**Engineering**
School	:	**School of Electrical Engineering and Telecommunications**
Thesis Title	:	**Speech based Continuous Emotion Prediction: An investigation of Speaker Variability and Emotion Uncertainty**

Abstract 350 words maximum: (PLEASE TYPE)

Understanding and describing human emotional state is important for many applications such as interactive human-computer interface design and clinical diagnosis tools. Speech based emotion prediction is generally viewed as a regression problem, where speech waveforms are labelled in terms of affective attributes such as arousal and valence, with numerical values indicating the short-term emotion intensity. Current research on continuous emotion prediction has primarily focused on improving the backend, developing novel features or improving feature selection techniques. However, emotion expressions or perceptions are in general heterogeneous across individuals, owing to a wide range of factors, such as cultural background and speaker's gender. The impact of these sources of variations on the continuous emotion prediction systems has not been fully explored yet and is the focus of this thesis.

Speaker variability, i.e., differences in emotion expression among speakers, has been shown to be one of the most confounding factors in categorical emotion recognition system, but there is limited literature that analyses the effect on continuous emotion prediction systems. In this thesis, a probabilistic framework is proposed to quantify speaker variability in continuous emotion systems in both the feature and the model domains. Furthermore, three compensation techniques for speaker variability are developed and in-depth analyses in both the feature and model spaces are carried out.

Another confounding factor is the inter-rater variability, i.e., difference in emotion perception among raters, which is ignored in current approaches by taking the average rating across multiple raters as the 'true' representation of the emotion states. However, differences in perception among raters suggest that prediction certainty varies with time. A novel approach for the prediction of emotion uncertainty is proposed and implemented by including the inter-rater variability as a representation of the uncertainty information in a probabilistic model. In addition, Kalman filters are incorporated into this framework to take into account the temporal dependencies of the emotion uncertainty, as well as providing the flexibility to relax the Gaussianity assumption on the emotion distribution that reflects the uncertainty.

The proposed frameworks and methods have been extensively evaluated on multiple state-of-the-art databases and the results have suggested the effectiveness.

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

ORIGINALITY STATEMENT

'I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously publ:shed or wri e:1 by an'Jther person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other ec: 1ca'.:o'.":al institution, except where due acknowledgement is made in the thes·s. P.Jny contribution made to the research by others, with v,hom I rave worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also cec!;,re that the intellectual content of this thesis is the product of my own w0r!<. exce;;t to the extent that assistance from others in the proj cfs <iAsisri a'1d concep!ion or in style, presentation and linguistic expression is acknow!edged.'

Signed

Date

ABSTRACT

Understanding and describing human emotional state is important for many applications such as interactive human-computer interface design and clinical diagnosis tools. Speech based emotion prediction is generally viewed as a regression problem, where speech waveforms are labelled in terms of affective attributes such as arousal and valence, with numerical values indicating the short-term emotion intensity. Current research on continuous emotion prediction has primarily focused on improving the backend, developing novel features or improving feature selection techniques. However, emotion expressions or perceptions are in general heterogeneous across individuals, depending on a wide range of factors, such as cultural background and speaker's gender. The impact of these sources of variations on the continuous emotion prediction systems has not been fully explored yet and is the focus of this thesis.

Speaker variability, i.e., differences in emotion expression among speakers, has been shown to be one of the most confounding factors in categorical emotion recognition system, but there is limited literature that analyses the effect on continuous emotion prediction systems. In this thesis, a probabilistic framework is proposed to quantify speaker variability in continuous emotion systems in both the feature and the model domains. Furthermore, three compensation techniques for speaker variability are developed and in-depth analyses in both the feature and model spaces are carried out.

Another confounding factor is the inter-rater variability, i.e., difference in emotion perception among raters, which is ignored in current approaches by taking the average rating across multiple raters as the 'true' representation of the emotion states. However, differences in perception among raters suggest that prediction certainty varies with time. A novel approach for the prediction of emotion uncertainty is proposed and implemented by including the inter-rater variability as a representation of the uncertainty information in a probabilistic model. In addition, Kalman filters are incorporated into this framework to take into account the temporal dependencies of the emotion

uncertainty, as well as providing the flexibility to relax the Gaussianity assumption on the emotion distribution that reflects the uncertainty.

The proposed frameworks and methods have been extensively evaluated on multiple state-of-the-art databases and the results have demonstrated the potential of the proposed solutions.

ACKNOWLEDGEMENTS

I could not have completed this thesis without the guidance and support of many people. Foremost, I would like to express my sincere gratitude to my supervisor Dr Vidhyasaharan Sethu for all of his support, encouragement and advice. Special thanks must also go to my joint-supervisor Professor Eliathamby Ambikairajah for his always helpful insights, suggestions and enthusiasm. Together they have guided and challenged me, enabling me to get the best of my PhD. I could not have asked for better supervision.

I would also like to thank Associate Professor Julien Epps for his helpful suggestions and constant encouragements. I am grateful to Professor Roland Goecke and Dr. Munawar Hayat for their collaboration and help with the Audio/Visual Emotion Challenge and Workshop 2017.

I would like to thank my colleagues at the UNSW Signal Processing research group: Kalani Wataraka Gamage and Kaavya Sriskandaraja for speech processing discussions and suggestions; Stefanie Brown for paper writing suggestions; Saad Irtza, Jianbo Ma, Hang Li, Sarith Fernando, Zhaocheng Huang, Brian Stasak, Gajan Suthokumar, Namalka Kananke, Dr Phu Ngoc Le, Tharshini Gunendradasan for fun and moral support.

I acknowledge the following sources of funding which have enabled me to pursue my PhD; UNSW Sydney for an Australian Postgraduate Award, and Data 61 Csiro for a Research Postgraduate Award. I thank the School of Electrical Engineering and Telecommunications UNSW for supporting me throughout my studies. I would like thank my friends and family for their ongoing love and support.

LIST OF ACRONYMS AND ABBREVIATIONS

AVEC	**A**udio/**V**isual **E**motion **C**hallenge
BLSTM	**B**idirectional **L**ong **S**hort-**T**erm **M**emory
BoTW	**B**ag-**o**f-**T**ext-**W**ords
CC	**C**orrelation **C**oefficient
CCC	**C**oncordance **C**orrelation **C**oefficient
CNNs	**C**onvolutional **N**eural **N**etworks
eGeMAPS	**E**xtended **G**eneva **M**inimalistic **A**coustic **P**arameter **S**et
EM	**E**xpectation **M**aximisation
EWE	**E**valuator **W**eighted **E**stimator
FA	**F**actor **A**nalysis
FW	**F**eature **W**arping
GMM	**G**aussian **M**ixture **M**odel
GMR	**G**aussian **M**ixture **R**egression
GP	**G**aussian **P**rocess
HM	**H**eterogeneous **M**apping
IEMOCAP	**I**nteractive **E**motional Dyadic **Mo**tion **Ca**pture **D**atabase (IEMOCAP)
IFN	**I**terative **F**eature **N**ormalisation
LLDs	**L**ow **L**evel **D**escriptors

LPCs	**L**inear **P**redictive **C**oefficients
LSTM	**L**ong **S**hort-**T**erm **M**emory
MAP	**M**ax **A** **P**osterior
MFCCs	**M**el **F**requency **C**epstral **C**oefficients
MP	**M**ost **P**robable
MSE	**M**ean **S**quared **E**rror
OA	**O**utput-**A**ssociate
PAV	**P**robabilistic **A**coustic **V**olume
PCA	**P**rinciple **C**omponent **A**nalysis
PLLR	**P**hone **L**og-**L**ikelihood **R**atios
PLS2	**P**artial **L**east **S**quare **2**
PLSDR	**P**artial **L**east **S**quare **D**imension **R**eduction
RNN	**R**ecurrent **N**eural **N**etwork
RVM	**R**elevance **V**ector **M**achines
SAL	**S**ensitive **A**rtificial **L**istener
SDC	**S**hifted **D**elta **C**oefficients
SVR	**S**upport **V**ector **R**egression
UBM	**U**niversal **B**ackground **M**odel
VAD	**V**oice **A**ctivity **D**etection
VAM	The **V**era **A**m **M**ittag German Audio-Visual Spontaneous Speech Database

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Speech based continuous emotion prediction

Human behaviour is a complex process which manifests an intricate interplay among the human brain and the body [3]. Understanding and describing human behaviour is crucial to many fields including analysis of affective states, underlying cognitive load, impaired social behaviours, etc. Affect, refers to the feelings we experience as part of everyday life in the form of moods and emotions [4]. Emotions represent the mental feeling, such as happiness and anger. They can lead us to engage appropriately in a given situation. For instance, a good teacher who is capable of observing student's emotion states can adjust his/her teaching plan, since emotion states can affect students' concentration and task solving ability [5]. Similarly, a machine that automatically recognises students' emotion states can be useful. With the rapid development of technology, it makes the automatic emotion recognition a possible task which facilitates many aspects of everyday life. For example, in the customer service application, a machine detecting the specific customer's emotion state as extremely angry can immediately pass him/her to professional assistants, which will dramatically reduce aftermarket complaints. Similarly, an automated personal assistant able to detect a speaker's emotion state can take the right action to interact with the speaker, such as playing comforting music while they are sad, or playing jokes when them are happy. While humans are easily able to recognise the emotion state of a speaker, it is still a challenge for machines to automatically recognise humans' emotion states.

Emotion is generally represented in two ways. One of the representations is in the form of a small number of categorical classes, where six basic emotion states that are universal for all human beings are shown in Figure 1.1 [6], including anger, fear, disgust, joy, sadness and surprise. However, it is argued that human also exhibit more complex and subtle states, such as thinking and embarrassment, which are not reflected by the six basic categorical emotion representations [7], and are often not adequately considered when modelling human emotions. Therefore, the dimensional description of

Figure 1.1: Six basic emotion categories: anger, fear, disgust, joy, sadness and surprise [1].

human affect is advocated [8], where the most widely adopted dimensions are arousal (ranging from deactivated to activate) and valence (ranging from unpleasant to pleasant). Numerical values are used to indicate the type and degree of the emotions. These dimensions of arousal and valence are related to one another in a systematic manner [9]. Each basic emotion can be represented as a combination of the type and degree of the emotional continuum, which is able to cover almost all of the complex and

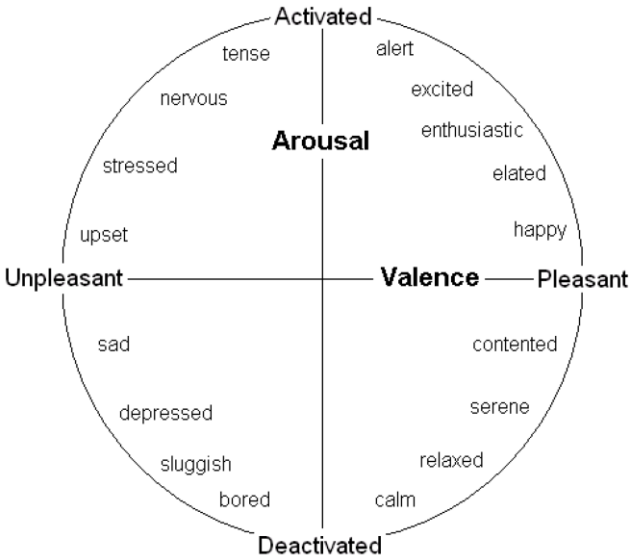


Figure 1.2: A graphical representation of the circumplex model of affect: the horizontal axis represents the valence dimension (pleasant vs. unpleasant) and the vertical axis represents the arousal dimension (activated vs. deactivated).

subtle human emotion states, shown in Figure 1.2. The emotion intensity in terms of arousal and valence is indicated by numerical values such as range within [-1,1], with a larger value representing high arousal or valence emotion states. For instance, happiness is an emotion state with high arousal and high valence, while depression is an emotion state with low arousal and low valence values. Fontaine., et al. [10] have argued against the notion that a two dimensional description of emotion is sufficient for characterizing humans' emotional experiences. For example, anger and fear are both negative valence and high activation, but they are differentiated based on a third dimension: dominance, which represents controlling and dominant versus controlled or submissive one feels (see Figure 2.2 for details). Due to the advantages of dimensional representation that maps all human emotion states to a three dimensional arousal-valence-dominance space, therefore, our research mainly focuses on the dimensional representation of emotion owing to those advantages.

Affect behaviour can be communicated via multiple modalities including the speech, face, body language, etc., among which speech is one of the most significant human behavioural signal, containing a rich variety of information. It is also easy and nonintrusive to collect, making it an excellent advantage for speech based recognition system design. Linguistic information, referring to the words spoken, is able to reflect a speakers' emotion state. On the other hand, paralinguistic information (acoustic cues) has been reported to be effective in predicting the emotion states of speakers, and is harder to disguise compared to the linguistic information [11]. In certain circumstances, one might exhibit more informative emotional information than the other. For instance, language is often more informative for valence, but may not show significant advantages in arousal predictions compared to speech acoustic cues. For example, a speaker's pitch (fundamental frequency of speech) and speech energy tends to increase with increasing positive emotion states [12]. Owing to these advantages, this thesis will focus on the speech and the inference of emotional states (dimensional representation) from speech.

As mentioned above, the dimensional emotion labels are represented by numerical values indicating the emotion intensity in terms of arousal and valence. These values are generally obtained for each small interval, such as per 20 or 40 million seconds. The aim of a speech based automatic

detection system is to predict these numerical values from the speech segment within each small interval, referred as the speech-based continuous emotion prediction system, shown in Figure 1.3. Speech waveforms are labelled with a specific numerical value for valence attribute indicating the short-term emotion intensity. The numerical labels (solid line) of the speech frames are generally achieved by averaging multiple raters' evaluations (dash lines) as perceived by several raters listening to the speech (and watching associated videos if available). The speech-based continuous emotion prediction system aims to capture the relationship between the speech and the corresponding emotion intensity, which is generally a regression model, and the system is expected to output the emotion predictions continuously (in small time intervals, i.e. 20ms) for unknown speech.

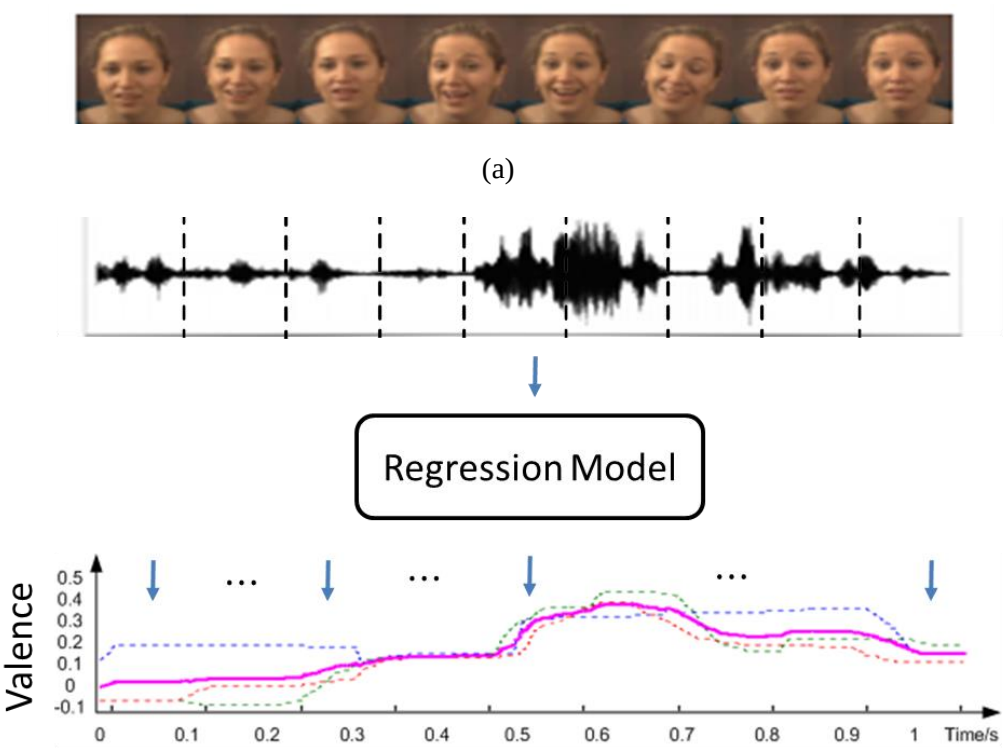


Figure 1.3: A speech based continuous emotion prediction system. (a) one video clip with facial expression (not used in system)[2]; (b) Speech waveform are segmented to small chunks with each chunk annotated with valence intensity (solid line), which are averaged among three ratings(dash lines). The speech based continuous emotion prediction system is developed as a regression model that captures the relationship between speech and the valence intensity.

Current research on continuous emotion prediction has primarily focused on either improving the backend, developing novel features or improving feature selection techniques for choosing the most discriminative feature set from a large pool of (generally statistical) features. These systems are typically built on the implicit assumption that the only source of variability in the model is the emotional content, but not other factors unrelated to emotion. However, emotion expressions or perceptions are in general heterogeneous across individuals, which can be affected by a wide array of factors ranging from culture background, speaker's age and gender, to speaker's health conditions [3]. This heterogeneity introduced from different sources in the continuous emotion prediction systems has not been fully explored yet.

Previous studies have reported a negative influence of the additional variability not related to emotion on both categorical emotion recognition and continuous emotion prediction systems, which generally leads to unreliable predictions [13, 14]. For example, speaker variability was shown to reduce the ability to distinguish between different emotional classes such as happiness, anger etc. [14], while variability in phoneme level has been observed as changes in magnitude and direction of formants in different emotional speech across vowels [15]. Among the different sources of variability unrelated to emotion that are typically present in speech, speaker variability (differences between speakers) has been shown to be one of the most confounding factors in categorical emotion recognition systems [13-16], but only limited literature has considered speaker variability in continuous emotion prediction systems [17-20]. Z-normalisation [18, 19] and i-vector normalisation [17, 20] are generally utilised; these being the standard variability compensation techniques employed in the field of speech processing, and specifically speaker verification systems. Speaker-dependent systems have also been proposed in conjunction with score level fusion to improve prediction accuracy [21]. While the majority of speaker variability compensation techniques applied in continuous emotion prediction systems have been based on those applied to the emotion classification problem, the fundamental premise behind speaker variability compensation for classification and regression systems is quite different. In the case of emotion classification systems, the aim is generally to maximise inter-class variability while minimising intra-class variability. However, in a continuous emotion prediction system that is cast as a regression problem, the problem of compensating for speaker variability is typically framed as that of reducing inter-class variability when treating each speaker as a distinct class, while trying to preserve the information that is more emotion specific. Consequently, the direct application of compensation methods developed for classification problem to continuous emotion prediction systems may be suboptimal and not appropriate. The speaker variability in continuous emotion prediction will be further discussed in Section 2.4.2.

Another confounding issue is the variability introduced by multi-raters, which is generally neglected by taking the average or weighted average among multi-raters as the representation of the

emotion states. Taking the average of these individual ratings to produce a 'gold standard' representing the emotion intensity smooths out discrepancies between raters. However, perception differences among raters and other possible sources of variability suggest that the certainty of emotion intensity may not be consistent. For instance, happiness is easy to recognise while frustration can be more ambiguous, which will be indicated by low and high inter-rater variability respectively. An optimal emotion prediction system is expected to consider the varying prediction certainty with time. Several studies [22-25] have showed the importance of taking information from multiple raters into account for both categorical emotion recognition and continuous emotion prediction systems, or have argued that emotion attributes should be ranked instead of trying to predict absolute values. However, investigation on the inter-rater variability in continuous emotion prediction systems is still lacking, which will be discussed in details in Section 2.4.3.

In addition, temporal dependencies between the acoustic observations have been shown to be critical for continuous emotion prediction tasks since affect evolves over time [26, 27]. The inclusion of both historical and future emotional information is generally achieved by a bidirectional long short-term memory recurrent neural network (BLSTM-RNN), prevailing over the standard statistic modelling techniques. However, LSTM-RNN easily falls into over fitting when the size of the training data is small [28]. The output-associate (OA) framework was proposed to capture temporal dependencies of a single affect dimension as well as the dependencies between the dimensional affect attributes of arousal and valence [29-31]. While these studies investigated emotions' evolving nature in terms of numerical values of emotion attributes referred as hard labels (i.e. mean ratings among multiple raters of arousal and valence), only a limited number of researchers have taken the temporal dependencies of the uncertainty of emotion prediction into account. Current techniques cannot be directly used to explore the temporal dependencies of such emotion uncertainty, since they target on the hard labels only. Thus by exploring the temporal dependencies of emotion uncertainty, it aims to reveal the evolving process of emotion label distributions.

1.2 Thesis objectives

Given the limitations raised in the previous section, the three principle objectives of this thesis are presented in this section. Firstly, the speaker variability in terms of probability distributions is systematically formulated, and to compensate speaker variability in continuous emotion prediction tasks. In meeting this objective, this thesis aims to:

- Characterise speaker variability in feature space and in model space.
- Develop compensation techniques for speaker variability in continuous emotion prediction based on the analysis of the difference between speaker-dependent probability distributions in feature space.

Secondly, I intend to determine how the existing representation of emotion states as hard labels can be improved to a more appropriate representation as distributions which include uncertainty of emotion prediction, by explicitly accounting for multi-rater variability in the system. To achieve this objective we:

- Include inter-rater variability as a representation of the information of uncertainty in a probabilistic framework.
- Analyse the effect of inter-rater variability on the conventional emotion prediction system (utilising mean rating as ground truth) based on the information of predicted uncertainty of emotion labels.

Finally, I want to explore how the temporal dependencies of emotion attributes can be incorporated to improve the existing prediction systems, in terms of both hard labels and label distributions by:

- Investigating different approaches in the front-end and back-end to incorporate temporal dependencies of hard labels for arousal and valence.
- Developing techniques to capture the temporal dependencies of emotion label distributions.

1.3 Organisation of thesis

The remainder of this thesis is organised as follows:

Chapter 2 describes emotion representations and provides an overview of the current techniques for predicting emotion intensity in terms of three affective attributes; namely, arousal, valence and dominance. The emotion databases currently in common usage are introduced, focusing on the three databases used to generate all of the experimental results presented in this thesis. It also highlights the challenges of the variability introduced from speakers and multi-raters and the importance of temporal dependencies.

Chapter 3 characterises speaker variability in terms of a probability distribution in continuous emotion prediction systems. It explores the differences in distributions of features between speakers in feature and model spaces. Two metrics of inter-speaker and intra-speaker difference are adopted to quantify the confounding effect of speaker variability in continuous emotion prediction systems.

Chapter 4 proposes three techniques to compensate for speaker variability in continuous emotion prediction systems, which are carried out in either the feature or model spaces. In-depth analysis reports and compares the compensation effect of these techniques. It highlights that the way speakers express their preferred emotion states is a key difference and compensating it in the feature space shows significant potential.

Chapter 5 characterises the inter-rater variability in continuous emotion prediction systems. In addition, the delay caused by sensing and judgment between an annotator's perceptual observations and their decision-making is also compensated in the ground truth and individual raters respectively. The analyses reveal low agreement among raters and inconsistency within individual rater, which leads a deeper consideration of the inter-rater variability.

Chapter 6 proposes a novel framework that is able to incorporate the uncertainty information of speech frames by explicitly accounting for multi-rater variability in the system. In addition, the

correlation between the uncertainty and the performance of a conventional emotion prediction system utilising average rating as the ground truth is investigated, by comparing the prediction performance in the low and high uncertainty regions.

Chapter 7 considers the temporal dependencies that occur in the evolution of emotion represented as hard labels. The incorporation of temporal dependencies for hard labels is explored in a wide range from feature extraction to the design of system structures.

Chapter 8 analyses the incorporation of temporal dependencies into the emotion uncertainty prediction, i.e. label distributions. A distribution assumption is made on the label distribution and two new measurements were adopted to evaluate the system performance in terms of the similarity between predicted and ground truth distributions.

Chapter 9 concludes the thesis with a summary of the research contributions and presents potential future research directions to follow up from this thesis.

1.4 Major contributions

The research presented in this thesis provides original contributions to the automatic assessment of emotion intensity in terms of arousal, valence and dominance using paralinguistic speech cues. The major contributions can be summarised as follows:

- **A probabilistic framework is developed to quantify speaker variability in continuous emotion systems in both the feature space and model space**. Two measures for quantifying the speaker variability have been formulated, which are based on the exploration of the difference in probability distribution of features between the speaker-dependent models in both the feature and model spaces. Results suggest that speaker variability is a confounding factor for continuous emotion prediction in both spaces, and a significant difference was observed in terms of the inter-speaker variability, while only minor differences were observed in terms of the intra-speaker variability.

- **A factor analysis based speaker normalisation technique is formulated** for continuous emotion prediction. The proposed technique operates directly on the feature space and decomposes it into speaker and emotion specific sub-spaces. Speaker-specific information is removed from the original feature space and only the emotion-specific information is kept. The improvements achieved by employing the proposed method suggest the effectiveness of the proposed scheme.

- **A novel Partial least square dimension reduction (PLSDR) based compensation technique for speaker variability is introduced**. It aims to project the original feature space to a lower-dimensional latent space, which simultaneously minimises speaker variability, and maximises the mutual-information between the projected features and the underlying ground truth. A key advantage of the proposed method is its capability to achieve feature dimension reduction and speaker normalisation in the same step. Experimental analyses indicate that the proposed technique is able to compensate for speaker variability in both spaces.

- **A feature mapping based compensation technique for speaker variability is proposed**. It aims to compensate for the speaker variability appearing in the different ways that speakers express similar emotion states. Speaker-dependent models are developed and compared to the speaker-independent model, to determine the speaker-dependent shifting and scaling parameters of compensation from a model point of view. The advantage of feature mapping based normalisation is the additional compensation effect on the local variability of speaker-dependent distributions, while the other two compensation techniques mentioned above only minimise the distance between speaker-dependent distributions.

- **A novel approach for predictions of uncertainty in the lables is developed by utilizing the inter-rater variability, which relaxes the assumption of the ground truth as the mean ratings**. This is implemented by including the inter-rater variability as a representation of the information of uncertainty in a probabilistic Gaussian mixture regression (GMR) model. In addition, the correlation between the uncertainty and the performance of typical emotion

prediction systems that utilise average rating as the ground truth is investigated, by comparing the prediction performance in the low uncertainty regions and the high uncertainty regions. Result suggests the effectiveness of predicting the uncertainty in the labels using inter-rater variability in terms of distribution estimation. They also indicate the correlation between the proposed framework and the conventional emotion prediction.

- **Temporal dependencies regarding the emotion mean ratings among multi-raters are investigated by extending output-associate (OA) structure within a multimodal fusion framework, and by further incorporating a measure of uncertainty associated with each prediction within this framework.** OA takes into account the contextual and temporal dependencies that exist within and between predicted arousal and valence values when performing multimodal fusion. Superior performance in arousal and valence predictions suggest the effectiveness of the method. In addition, the predicted uncertainties firstly obtained from different sub-systems are combined with the predictions, and serves as additional information in the second stage OA fusion framework. Consistent improvements were observed when incorporating prediction uncertainty across various system configurations for predicting arousal and valence, suggesting the importance of taking into consideration prediction uncertainty for fusion and more broadly the advantages of probabilistic predictions.

- **A dynamic multi-rater GMR is proposed, aiming to obtain the predictions of uncertainty in the emotion labels by taking into account their temporal dependencies**. This framework is achieved by incorporating feedforward and backward Kalman filters into a GMR to estimate the time-dependent label distribution that reflects the emotion uncertainty. It also provides the benefits of relaxing the label distribution from Gaussian assumption to that of a Gaussian mixture model (GMM). In addition, two new measurements to estimate emotion uncertainty from the GMM are adopted. This study is the first attempt to incorporate temporal dependencies in distribution predictions in continuous emotion prediction.

1.5 List of publications

Journal Papers

Dang, T., Sethu, V., Epps, J., & Ambikairajah, E. (under review). Compensation techniques for speaker variability in continuous emotion prediction, IEEE Transaction on Affective Computing.

Conference Papers

Dang, T., Sethu, V., Epps, J., & Ambikairajah, E. (2018). Dynamic multi-rater Gaussian Mixture Regression incorporating temporal dependencies of emotion uncertainty using Kalman filters. Proc. ICCASP 2018.

Dang, T., Stasak, B., Huang, Z., Jayawardena, S., Atcheson, M., Hayat, M., Le, P., Sethu, V., Goecke, R. and Epps, J.. Investigating Word affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017. In Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge. ACM

Dang, T., Sethu, V., Epps, J., & Ambikairajah, E. (2017). An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression. Proc. Interspeech 2017, 1248-1252.

Dang, T., Sethu, V., & Ambikairajah, E. (2016). Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction. In INTERSPEECH (pp. 913-917).

Huang, Z., **Dang, T.**, Cummins, N., Stasak, B., Le, P., Sethu, V., & Epps, J. (2015, October). An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (pp. 41-48). ACM.

Huang, Z., Stasak, B., **Dang, T.**, Wataraka Gamage, K., Le, P., Sethu, V., & Epps, J. (2016, October). Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016.

In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (pp. 19-26). ACM.

2 SPEECH BASED EMOTION PREDICTION: A REVIEW

2.1 Emotion Representations

Psychologists have pursued three major approaches towards emotion modelling that can be distinguished: categorical, dimensional and appraisal-based approaches [32]. The categorical approach is one of the most commonly adopted approaches as it describes human emotion as a small number of emotions that are universally recognised, with the six basic emotions including anger, disgust, fear, happiness, sadness and surprise [7]. However, it is also argued that people exhibit more complex, subtle and non-basic emotions in their everyday social life, such as embarrassment or thinking, which are not covered in the categorical framework. Therefore, the dimensional approach was proposed, attempting to conceptualise human emotions by defining several affective dimensions. The dimensional model of emotion suggests that a common and interconnected neurophysiological system is responsible for all affective states, which contrasts the theory of basic emotion that different emotions arise from separate neural systems [9]. Several dimensional models of emotion have been developed, though there are just a few that remain as the dominant, currently accepted models [33]. The dimensions of arousal, valence and dominance are the most commonly used in these dimensional models [8].

One typical model is the circumplex model of emotion. This model assumes that each emotion can be represented as a combination of two dimensions, namely, arousal and valence. Arousal indicates the degree of activated emotions while valence indicates the degree of pleasant emotions. The varying degree of both arousal and valence can be combined to map all complex and subtle emotion states, as shown in Figure 2.1 (same as Figure 1.2). The other model is the PAD emotional state model which uses three numerical dimensions ,namely, pleasure (valence), arousal and dominance, to represent all emotions [34]. Apart from arousal and valence, the third dimension of dominance is added, which reflects the degree of controlling and dominant nature of the emotion. The PAD model [34] is shown in Figure 2.2.

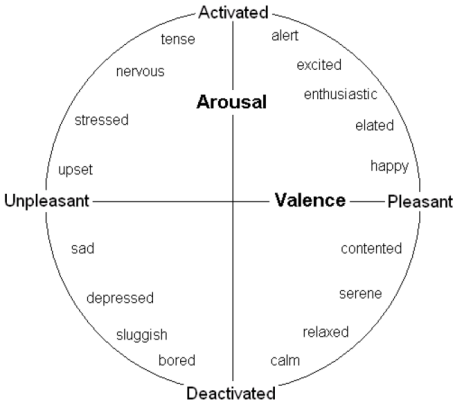


Figure 2.1: A graphical representation of the circumplex model of affect: the horizontal axis represents the valence dimension (pleasant vs. unpleasant) and the vertical axis represents the arousal dimension (activated vs. deactivated).

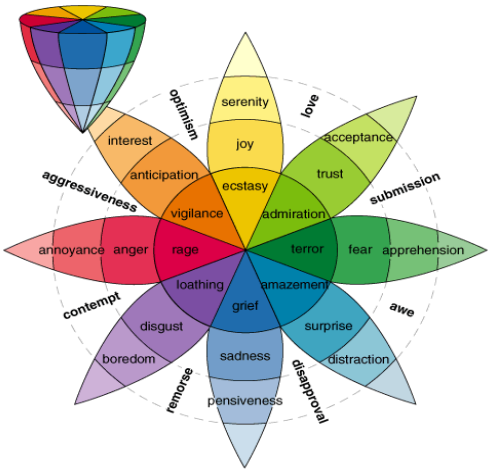


Figure 2.2: The PAD model of affect. The cone's vertical dimension represents intensity and the circle represents degrees of similarity among the emotions. The eight sectors are designed to indicate that there are eight primary emotion dimensions defined by the theory arranged as four pairs of opposites. In the exploded model the emotions in the blank spaces are the primary dyads or dyadic emotions (mixtures of two primary emotions) [27].

The third set of psychological models is componential models which are based on the appraisal theory. These assume that emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world [35]. A computational tractable model proposed by Ortony, Clore and Collins commonly referred as OCC model [36] based on appraisal theory is now established as a standard model for emotions and has mostly been used in affect synthesis [7]. However, only limited literature has investigated appraisal theory from an engineering point of view, instead focusing on the theory.

A variety of automatic systems have been developed based on three emotion representations, especially with focus on the first two sets. An emotion recognition system aims to identify emotion categories form multi-modal behaviour signals, which is generally achieved by developing classifiers to different emotion categories, or finding the hard margins between different emotion categories. It is referred as a classification system. A dimensional emotion prediction system aims to predict the emotion intensity (i.e. arousal and valence) in terms of continuous numerical values, which targets on a regression model to capture the relationship between the multi-modal behaviour signals and the continuous numerical values.

While categorical emotion representations are probably more actionable and interpretable, dimensional models of emotion are better representations in terms of providing more tractable analysis. Therefore, I focus on the dimensional emotion representation throughout the thesis with more emphasis on the two dimensional circumplex model, since previous literature has reported that arousal and dominance are highly correlated [37]. A similar performance was also generally observed for arousal and dominance prediction, suggesting that the prediction framework for arousal and dominance prediction may be similar.

2.2 Dimensional Emotion Prediction

A speech-based dimensional emotion prediction system generally aims to develop a regression model, often referred to as the back-end, which captures the relationship of the emotional speech and emotional state in terms of attributes such as arousal and valence. Emotional speech is generally represented by a set of features that are extracted by a suitable front-end, and the emotion intensity such as arousal and valence is represented by a set of time-varying numerical values, herein referred to as labels.

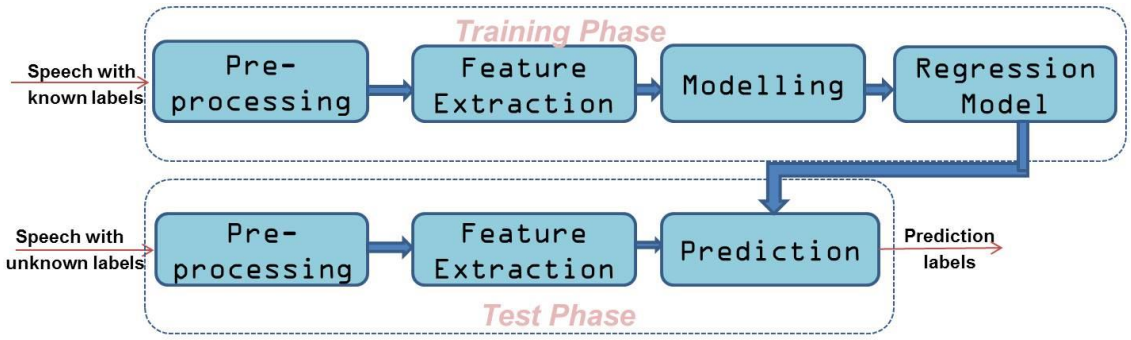


Figure 2.3: Block diagram of a speech-based dimensional emotion prediction system, comprising of the training and test phases.

2.2.1 System Overview

A continuous emotion prediction system is generally comprised of two phases, namely the training phase and the test phase as shown in Figure 2.3. During the training phase, speech samples with known labels are pre-processed and representative features are extracted to indicate the speaker's personal emotional information. Regression modelling techniques are used to develop a regression model. In the test phase, the same feature sets are extracted from speech with unknown labels, referred as test features. Predictions are then estimated using the regression model based on the test features.

Details of the dimensional emotion prediction systems are described in Figure 2.4. Features are extracted from small time durations segments referred to as frames, which generally range from 20ms to 40ms in length. The frame-wise features are referred to as low level descriptors (LLDs). Several statistical descriptions (functionals) are then applied to the LLDs within a longer window of several seconds to calculate the statistic features, e.g. the mean, standard deviation, skewness, etc. The statistical features capture the statistical characteristics within the window, since emotion is a slowly varying process. Different regression modelling techniques have been developed and applied in this field. An overview of these techniques is presented in Section 2.2.2 and 2.2.3.

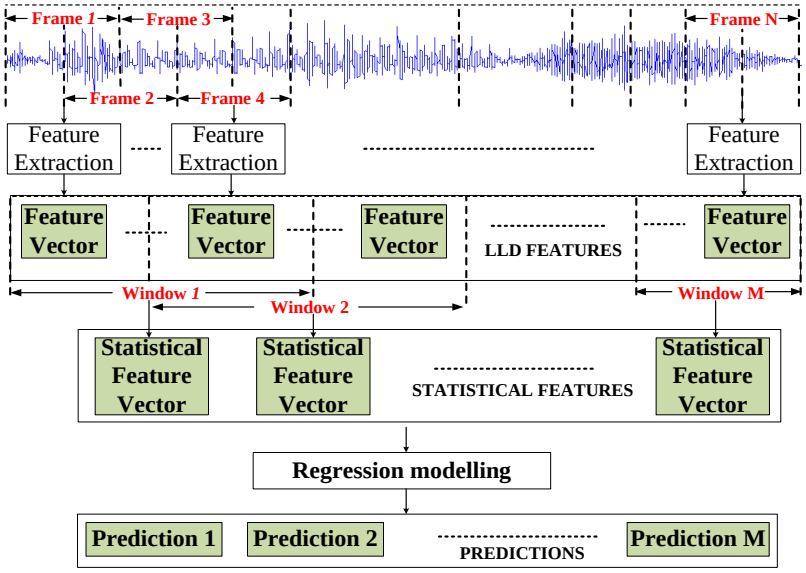


Figure 2.4: Speech based continuous emotion prediction system description.

2.2.2 Features and Feature Selection Techniques

Features, a way of extracting and suitably representing relevant information form the raw signals, can be used as a higher parametric representation of speech waveforms. These are generally extracted at the first step in a continuous emotion prediction system, since the raw speech signals contain large amount of data which is computationally expensive to process, and it is also hard to distinguish the desired emotion information directly from raw speech. An important step in speech-based continuous emotion prediction systems is the extraction of the suitable features that characterise the emotion-specific information. The current analysis on speech-based features can be divided into three categories: acoustic features, linguistic features and feature embedding.

2.2.2.1 Acoustic and prosodic features and feature selection techniques

A suitable set of acoustic features, which reflect the characteristics of speech sound, is the most widely adopted feature set in continuous emotion prediction systems. Spectral features and prosodic features are two feature sub-sets that have proven effective in capturing the emotion-specific information [11]. Spectral features indicating the distribution of the spectral energy across the range of frequency of speech were shown to be effective in distinguishing between emotion categories [38]. For instance, happiness was proven to contain high energy at high frequencies, while sadness had low

energy at the same range [39, 40]. Mel-frequency cepstral coefficients (MFCCs) developed based on the properties of the human auditory system [41] and linear predictive coefficients (LPCs) that model the characteristic of the human vocal tract [42] are the two most widely adopted spectral feature. Prosodic features, e.g. pitch, jitter and shimmer, reflecting the auditory quality of sound have been shown to be indicative of emotional state [11]. In particular, the concept of pitch, referring to the fundamental frequency (f0) of the vibration of the vocal folds when excited by air from the lungs passing over the vocal folds, is of great importance [43-46]. Formants showing the resonant frequencies of the vocal tract, and voice quality reflected in signal amplitude, energy and duration are another two commonly adopted prosodic features [11].

Both spectral features and prosodic features are extracted on a frame basis referred as LLDs as discussed in Section 2.2.1. Since LLDs only capture the information in a short duration that ignores the statistic property over a longer window, the statistical features are proposed by applying several functions to the LLDs over a longer duration. A large number of researchers have shown that statistical features are superior to LLDs in terms of system performance [47-50], and most existing literature that uses the statistical feature sets have demonstrated good performance in emotion prediction [51-54]. However, the statistical features are generally of high dimensions due to the large number of functions applied, and it is common to use dimensionality reduction techniques in speech emotion recognition applications in order to reduce the feature dimensions.

There are generally two approaches for dimensionality reduction: feature selection and feature transformation. In feature selection, the main objective is to find the feature subset that achieves the best possible system performance. This subset is usually characterised by an easy to calculate function, called the feature selection criterion. Such a criterion searches for the optimal subset of features by minimising the cross validation error [55], or maximising the mutual information between the features and the class labels[56]. On the other hand, feature transform techniques aim to find a suitable linear or nonlinear mapping from the original feature space to another space with reduced dimensionality while preserving as much relevant classification/regression information as possible. The reader may refer to [57, 58] for excellent reviews on dimensionality reduction techniques. In particular, PCA is

one of the most widely adopted dimension reduction technique in continuous emotion prediction systems [59, 60], aiming to preserve enough variability in feature space, as well as reducing the feature dimension for the followed modelling techniques.

2.2.2.2 Linguistic features

The lexical content of speech is important and a straightforward way to convey emotions. It is generally extracted at the phoneme level or word level, thus the first step is to recognise the phoneme or word sequences. However, phoneme or word recognisers were shown to have poor performance, making linguistic feature extraction unreliable [61, 62], which has slowed down research into continuous emotion prediction development using linguistic features. With improvements in data collection in many databases, transcripts that enable a more accurate analysis on the linguistic features have started to be provided.

Phone Log-Likelihood Ratios (PLLR) features have been investigated for continuous emotion prediction. They show promising performance in predicting arousal and valence, which motivates feature analysis at the phoneme level and opens new research possibilities in this field [63]. A bag-of-words feature representation based on the transcripts has also been proposed and generated with OpenXBOW [64] in AVEC 2017, which uses 521 unigrams, i.e. 521 single word. It is referred as bag-of-text-words (BoTW). The histograms of 521 unigrams are created over a segment of 6 seconds in time, and the logarithm is taken from the term frequencies. Thus the BoTW features contain 512 dimensional features in total. Significant improvements in arousal, valence and likability (how much a subject expresses a positive or a negative attitude while speaking) predictions have been observed using the BoTW features over the standard acoustic feature set of *Extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [65]. Apart from phoneme- and word-level features, nonverbal vocal gestures such as laughter or filler have also been analysed. Information extracted from these segments has been shown to be complimentary to the acoustic and linguistic features [66]. Though only limited body of literature has attempted to use these linguistic features, or nonverbal vocal gestures, in continuous emotion prediction systems, the promising potential has opened a new research avenue in this field.

2.2.2.3 Feature embedding

With the great computation power of deep learning, some research has focused on learning deep features with neural networks. Convolutional neural networks (CNNs) and LSTM methods have been combined as a feature extractor that is applied to the raw speech waveforms, with the aim of automatically learning the best representation of the raw speech signal directly. This technique was shown to outperform traditional approaches using acoustic features significantly [67]. Additionally, CNNs combined with auto-encoders were also used to learn affect-salient features for speech based emotion recognition systems. An auto-encoder is a neural network used for unsupervised learning, with the aim of learning a representation (encoding) for a set of data. A variant of auto-encoders has been adopted at the first stage to learn the local invariant features, and convolution layers were connected to the output of auto-encoders to form a series of feature maps. Then these feature maps are subsampled and stacked into one feature vector as input feature for the regression modelling techniques [68]. The experimental results showed that the proposed method lead to stable and robust recognition performance in complex scenarios, e.g. with variations of speaker and environmental distortions. In [69], CNNs have been directly applied to LLDs to extract the emotion salient feature vector without need to apply utterance-level statistics. CNNs and auto-encoders have been the most widely used neural network structures for feature extraction [70-72]. CNNs learn the visual patterns in a three dimensional space that works as filtering processing and well captures the salient speech segments, and auto-encoders are capable of performing dimension reduction and reconstruction that fit well in the situations where variability from other sources affect the main task. Though such features have achieved promising results, analysis of what emotion-related information the network captures is still lacking.

2.2.3 Regression modelling techniques

Regression models aim to capture the relationship between speech-based features and the affective dimensions of arousal and valence. The regression model generally takes one of three approaches: (a) it defines a hard margin represented by a set of parameters which is obtained by optimising a specific objective function, e.g. support vector regression or neural networks; (b) it models the joint

probability distribution of features and labels, e.g. Gaussian mixture regression models;or (c) it represents a probabilistic mapping function from the features to the labels, e.g. RVM. All three approaches have shown promising results. The general idea behind each regression modelling technique will be discussed in this section, while details of the specific modelling techniques utilised in this thesis will be presented in Section 2.3.

2.2.3.1 Support vector regression

Support vector regression (SVR) [73] is one of the most popular regression techniques in continuous emotion prediction systems. It is used extensively in the speech-based emotion recognition and prediction. owing to one of the main advantages that SVR generalises well in many applications [74-76]. The basic idea behind SVR is to find a linear mapping function between the feature and labels as shown in Figure 2.5. The '$\varepsilon - tube$' shown within the black solid lines is modelled using all the feature vectors represented by red dots. SVR is a sparse approach that only a subset of vectors are adopted to characterise the model, referred as support vectors appearing in the edge of the '$\varepsilon - tube$'. The slack variable ξ is introduced to cope with the infeasible constraints of the optimisation problem in SVR. To deal with the non-linear mapping with SVR, kernels are adopted which maps the original feature space to a high dimensional space where the non-linear mapping can be converted to a linear problem.

One of the main advantages of SVR is the good generalisation that fits well in many applications, since the $'\varepsilon - tube$' aims to yield the most fitted region around a regression line. Another benefit of SVR is the efficiency of modelling computation, since it only adopts the support vectors close to the tube, and the number of parameters tuned during the training phase is limited. In addition, the kernel functions can easily provide a path of dealing non-linear mapping problems. However, the error permitted and the kernel functions utilised has to be properly identified since SVR depends greatly on these parameters.

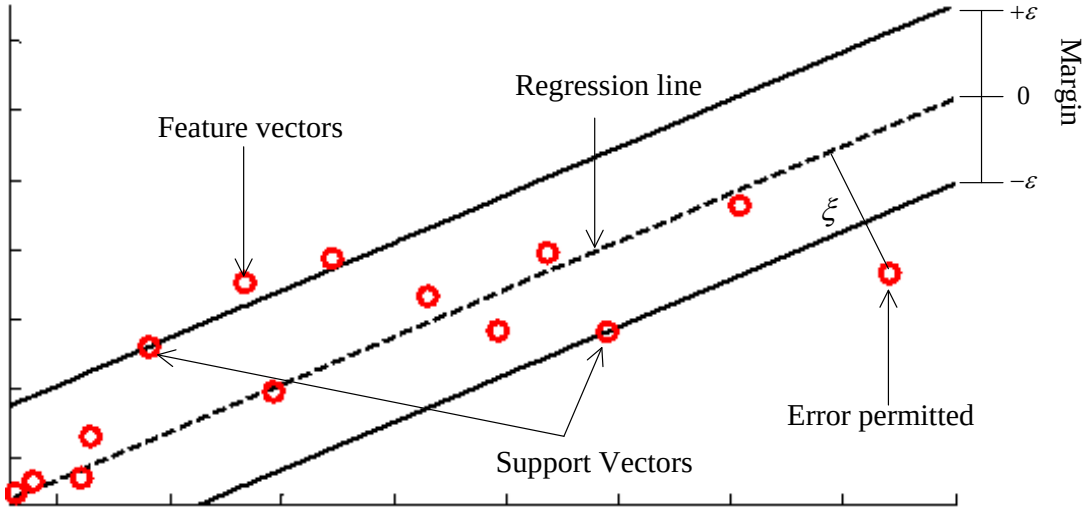


Figure 2.5: Concept of Support vector regression highlighted by a two-dimensional example. Feature vectors are represented by the red dots .A tube is fitted around the regression dash line, where small deviations from the feature vectors are permitted. Those red dots lie in the margin of tube are the support vectors.

2.2.3.2 Long short term memory recurrent neural network

Deep learning structures, especially recurrent neural networks (RNNs) and LSTM-RNNs, have attracted more and more attention in recent years for use in continuous emotion prediction systems, since they are able to automatically capture the long-term temporal dependencies of emotion's evolving nature. They take advantage of the sequential information with the output being dependent on the previous computations, while 'memory cells' within the RNN store the information calculated so far. A typical RNN is shown in Figure 2.6.

In Figure 2.6 $X_t = [X_{1t}, X_{2t}, \cdots X_{Nt}]^T$ represents the N dimensional feature vector at time t, and Y_t represents the corresponding prediction. W_{in} and W_{out} represent the weight matrix connecting two conservative layers, and W_r is the weight matrix that connect the previous time step to current time step, which are able to memorise the past information as shown in the unfolded structure. However, researchers have found that RNNs can only remember short-term temporal dependencies [77, 78]. With increasing length of windows, RNNs become unable to connect past information to the current state, thus LSTM-RNN is proposed to capture long-term information. LSTM-RNNs [24, 52, 79, 80] as a special kind of RNN is capable to model the long-term dependencies, since it introduces gates in the network that are able to control how much of information to let through.

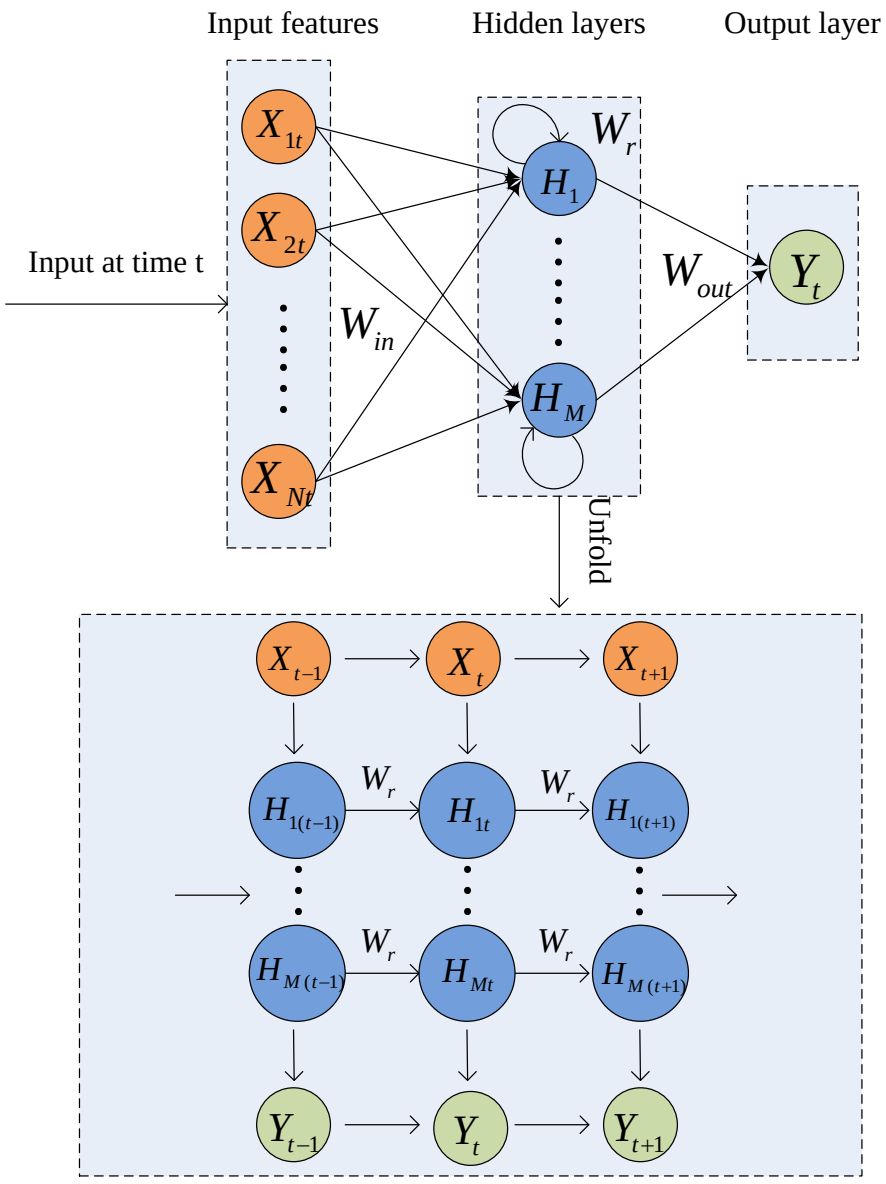


Figure 2.6: A recurrent neural network and the unfolding in time of the computation involved in its forward computation. X_t represents input feature vector at time t, and N represents the feature dimensions; H_{Mt} represents the M_{th} hidden unit at time t; Y_t represents prediction at time t. W_{in} represents the input weight matrix, W_r represents the recurrent weight matrix and W_{out} represents the output weight matrix.

2.2.3.3 Gaussian mixture regression

GMR [81-83] aims to construct a Gaussian mixture model (GMM) for the joint density of the features and the affective dimensions referred as labels during the training phase, and then derives the conditional density and regression functions from the joint GMM during the testing phase. It is developed for multivariate nonlinear regression modelling and provides the flexibility of allwoing

multi-dimensional predictions. Researches have employed it in continuous emotion predictions in recent years [81] and superior performance is observed compared to neural networks[81].

2.2.3.4 Relevance vector machines

The relevance vector machine (RVM) [84] is a probabilistic model that adopts the Bayesian framework for obtaining sparse solutions to regression tasks. RVM aims to find a set of weights associated with each feature dimensions during the training process, and then predictions are made using these trained weights during the test phase. The Bayesian framework is introduced by assigning a Gaussian prior to the weight of each dimension and a Gaussian assumption to the bias term. One of the key advantages of RVM is the sparse representation of the weights, which ensures that the model will be characterised by a subset of the feature dimensions, thus achieving feature selection during the training phrase. Another advantage of RVM is that the probabilistic output provides the prediction uncertainty as a Gaussian distribution, which could be useful in clinical applications [85].

Apart from these regression modelling techniques, a variety of other regression approaches have also been explored with regards to continuous emotion prediction. Distance-based fuzzy k-nearest neighbours and rule-based fuzzy-logic estimators [86] have been investigated, but were found to be less effective than SVR systems. Gaussian Process (GP) as a nonparametric model has been tested in continuous emotion prediction systems and showed superior advantages in preference learning, in which the goal is to learn a predictive preference model and to predict the label ranking, i.e. training instance $\boldsymbol{X}_1$ displaying higher arousal intensity than instance $\boldsymbol{X}_2$[87-89]. This thesis mainly focuses on GMR and RVM, since they have shown robust and superior performance in continuous emotion prediction systems and also provide the flexibility with different requirements, which are further discussed in Section 2.3.

2.2.4 Evaluation metrics

Most existing evaluation metrics in continuous emotion prediction focus on either the absolute value difference, or the trend of value changing between the time series of system prediction and the ground truth provided by the databases, e.g. generally the mean ratings among individual annotators.

Pearson's correlation coefficient (CC) [90], and mean squared error (MSE) are the two most commonly used metrics. Pearson's correlation coefficient indicates the strength and direction of the linear relationship between the affective ground truth $\boldsymbol{y}$ and the affective predictions $\boldsymbol{x}$ by the trained model, and is calculated as:

$$\rho_{cc} = \frac{cov(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}} \tag{2.1}$$

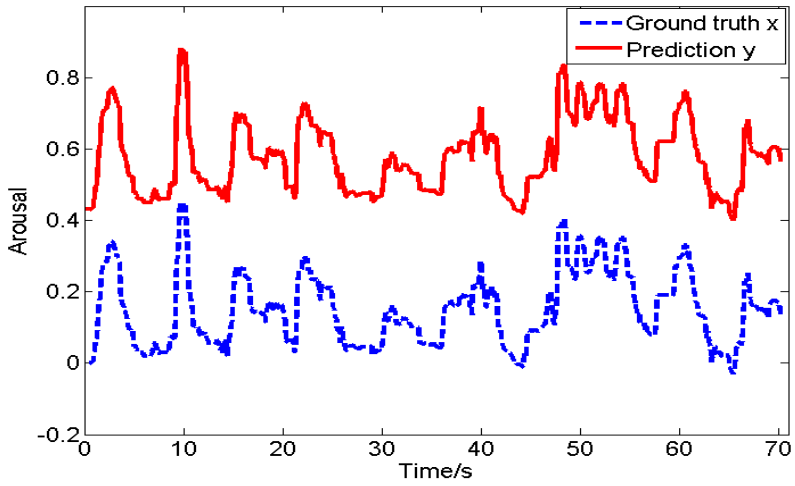
where cov represents the covariance of two variables, and $\sigma_{\boldsymbol{x}}$ and $\sigma_{\boldsymbol{y}}$ are the standard deviations of $\boldsymbol{x}$ and $\boldsymbol{y}$. A higher value of ρ_{cc} indicates a strong correlation between two variables, thus a better performance of the emotion prediction system.

Mean squared error measures [91] the average of the squares of the error terms, i.e. the difference between the ground truth and the predictions, and is calculated as

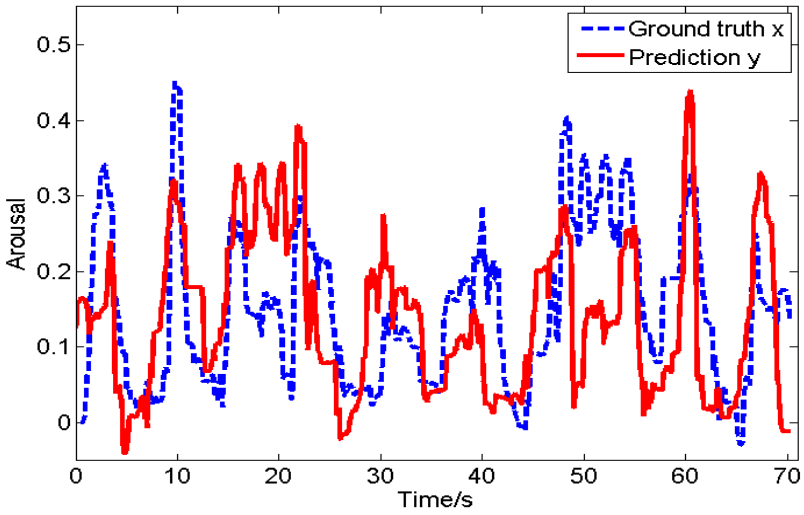
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{2.2}$$

where i represents the frame number and N is the total number of frames in sequences $\boldsymbol{x}$ and $\boldsymbol{y}$. A smaller value of MSE represents a better performance of the prediction systems.

However, ρ_{cc} and MSE can only partially represent the emotion prediction system performance. For example, a high ρ_{cc} may not always represent good predictions due to the high MSE, as shown in Figure 2.7(a), and a low MSE may not always indicate good predictions either due to the low ρ_{cc}, as seen in Figure 2.7(b). Finding a trade-off between these two metrics may be a better representation the system performance, but it is not straightforward to employ. Therefore, an evaluation metric that integrate both the correlation and the MSE in continuous emotion prediction systems was proposed[76], namely, the concordance correlation coefficient (CCC) [92]. The CCC measures the agreement between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ as:



(a)



(b)

Figure 2.7: Plot of predictions x and ground truth y within a time segment, showing examples where CC and MSE measures fail to indicate poor prediction performance. In (a) I observe a high CC coupled with a high MSE; (b) shows a low CC and low MSE.

$$\rho_{ccc} = \frac{2\rho_{cc}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (u_x - u_y)^2} \tag{2.3}$$

where u_x and u_y are the mean value of prediction sequence $\boldsymbol{x}$ and ground truth $\boldsymbol{y}$. CCC is able to indicate the overall system performance, and thus has been commonly adopted in continuous emotion prediction systems since AVEC 2016.

Even though CC, MSE and CCC have all been adopted in continuous emotion prediction systems for years, there is still not adequate support on the optimal evaluation metric. More importantly, as

discussed in Section 1.4, I have investigated emotion prediction as a distribution prediction using inter-rater variability, instead of treating emotion prediction as a point estimation that uses the mean rating as the ground truth. In this case, the evaluation metrics of the distribution predictions should be able to deal with a time sequence of distributions, where these three evaluation metrics cannot be directly employed since they are only applicable for time sequence of point estimations.

2.3 Two regression models

This section will introduce two back-ends mainly adopted in this thesis: RVM and GMR. These are two relatively new approaches to be used in emotion prediction systems, and they show great potential for emotion prediction tasks, as RVM performing well for high dimensional feature sets and GMR modelling features from a probabilistic point of view.

2.3.1 Relevance Vector Machines (RVM)

RVMs are a relatively new approach to multi-dimensional regression which is gaining in popularity in the field of continuous emotion prediction [19, 29, 93]. RVM can be considered as a sparse Bayesian method analogous to support vector regression (SVR) [84, 94]. A key advantage of RVM over SVR in the context of multi-modal learning is its heterogeneous mapping (HM) property, which allows any arbitrary kernel function to be used in conjunction with a RVM. HM allows not only the mappings of contextual temporal information, but also a convenient multimodal fusion technique, which negates the need to train and heuristically combine multiple predictors [85].

RVM forms the regression function as:

$$y(\boldsymbol{x}_n, \boldsymbol{w}) = \boldsymbol{w}\Phi(\boldsymbol{x}_n) + \epsilon_n = \sum_{p=1}^{P} w_p \Phi_p(\boldsymbol{x}_n) + \epsilon_n \tag{2.4}$$

where $\boldsymbol{w} = [w_1, \cdots, w_P]^T$ is an estimated set of sparse weights, also known as regression parameters, and $\boldsymbol{x}_n$ denotes an M-dimensional feature vector at frame n. $\Phi = [\Phi_1(\boldsymbol{x}_*), \cdots, \Phi_P(\boldsymbol{x}_n)]^T$ is a set of potentially non-linear transforms performed on $\boldsymbol{x}_n$ and ϵ_n is the training noise vector. In the Bayesian approach used in RVMs all noise terms are assumed to have a Gaussian distribution, such that

$$\epsilon_n \sim N(0, \sigma^2) \tag{2.5}$$

Given the assumption of $\boldsymbol{w}$ a zero-mean Gaussian prior, it ensures that RVM learns a sparse representation of $\boldsymbol{w}$, where the majority of elements in $\boldsymbol{w}$ are zero. This encourages sparsity by declaring smaller weights as more probable [84]:

$$P(\boldsymbol{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{P} N(0, \alpha_i^{-1}) \tag{2.6}$$

where $\boldsymbol{\alpha} = [\alpha_1, \cdots \alpha_P]^T$ is the inverse variance hyperparameter, and is analogous to regularisation terms in SVR or ridge regression. The training phase of the RVM regression model searches for the most probable (MP) values of $\boldsymbol{\alpha}$ and σ^2, called $\boldsymbol{\alpha}_{MP}$ and $\boldsymbol{\sigma}_{MP}^2$, using an iterative Bayesian inference procedure. In the testing phase, $\boldsymbol{\alpha}_{MP}$ and $\boldsymbol{\sigma}_{MP}^2$ are used to make a prediction and to estimate the level of uncertainty associated with that prediction.

2.3.1.1 RVM training

Assuming the independence of data points $\boldsymbol{y}_t$ at each time t, RVM aims to maximise the likelihood of the complete data set $P(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)$ as:

$$P(\boldsymbol{y}|\boldsymbol{w}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp(-\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{w}\boldsymbol{\Phi}\|^2) \tag{2.7}$$

where $\boldsymbol{\Phi} = [\Phi(\boldsymbol{x}_1), \Phi(\boldsymbol{x}_2), \cdots \Phi(\boldsymbol{x}_N)]^T$, $\boldsymbol{y} = [y_1, y_2, \cdots y_N]^T$ and N represents the feature dimensionality of $\boldsymbol{\Phi}$. Within $\boldsymbol{\Phi}$, $\Phi(\boldsymbol{x}_N) = [1, K(\boldsymbol{x}_n, \boldsymbol{x}_1), K(\boldsymbol{x}_n, \boldsymbol{x}_2), \cdots, K(\boldsymbol{x}_n, \boldsymbol{x}_N)]$, where $K(\cdot)$ represents the kernel functions. Thus the frame-wise label $\boldsymbol{y}_t$ is distributed as a Gaussian variable $P(y_t|\boldsymbol{x}_*) = N(y_t|\boldsymbol{w}\emptyset(\boldsymbol{x}_*), \sigma^2)$.

With the Gaussian prior $\boldsymbol{\alpha}$ defined, the objective function expands to $P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2)$, which can be estimated as:

$$P(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2)P(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2)}{P(\boldsymbol{y})} \tag{2.8}$$

This posterior $P(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})$ cannot be analytically computed in full, but equation (2.8) can instead be decomposed as:

$$P(\boldsymbol{w}, \boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) = P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2)P(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) \tag{2.9}$$

The first term in the right side of equation (2.9) can be further expanded as:

$$\begin{aligned} P(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\alpha}, \sigma^2) &= \frac{P(\boldsymbol{y}|\boldsymbol{w}, \sigma^2)P(\boldsymbol{w}|\boldsymbol{\alpha})}{P(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2)} \\ &= (2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{u})^T\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{w}-\boldsymbol{u})) \end{aligned} \tag{2.10}$$

where the posterior covariance $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = (\sigma^{-2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{A}) \tag{2.11}$$

and the posterior mean $\boldsymbol{u}$ is

$$\boldsymbol{u} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{y}, \tag{2.12}$$

where $\boldsymbol{A} = diag(\alpha_0, \alpha_1, \cdots \alpha_N)$, and σ^{-2} represents the inverse variance of the noise term in equation (2.5).

In order to estimate the hyperparameter posterior $P(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})$ on the right hand side in equation (2.9), a delta function has been adopted to approximate $P(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})$, as explained in [84]:

$$\int P(\boldsymbol{y}|\,\boldsymbol{\alpha}, \sigma^2)\,\delta(\boldsymbol{\alpha}_{MP}, \sigma^2_{MP})d\boldsymbol{\alpha}d\sigma^2 = \int P(\boldsymbol{y}|\,\boldsymbol{\alpha}, \sigma^2)\,P(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y})d\boldsymbol{\alpha}d\sigma^2 \tag{2.13}$$

Therefore, the learning process becomes the search for the hyperparameter posterior mode:

$$argmax_{\boldsymbol{\alpha},\sigma^2}\,P(\boldsymbol{\alpha}, \sigma^2|\boldsymbol{y}) \propto P(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2)P(\boldsymbol{\alpha})P(\sigma^2) \tag{2.14}$$

which aims to search for $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP}.

The estimation of $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP} cannot be obtained in close form, and are instead estimated iteratively. The updating of $\boldsymbol{\alpha}$ follows [95] as:

$$\alpha_i^{new} = \frac{\gamma_i}{u_i^2} \tag{2.15}$$

$$\gamma_i = 1 - \alpha_i\boldsymbol{\Sigma}_{ii} \tag{2.16}$$

where $\boldsymbol{\Sigma}_{ii}$ is the i^{th}diagonal element of the posterior weight covariance from equation (2.11), and γ_i, ranging from 0 to 1, indicates how 'well-determined' the corresponding weight w_i is by the dataset.

In terms of the hyperparameter σ^2, differentiation gives the updated variance ${\sigma^2}^{new}$ as:

$$ {\sigma^2}^{new} = \frac{\|\boldsymbol{y} - \boldsymbol{\Phi u}\|}{N - \boldsymbol{\Sigma}_i \gamma_i} \tag{2.17} $$

where N represents the number of data samples and $\boldsymbol{u}$ is shown in equation (2.12). The iteratively estimated most probable hyperparameters $\boldsymbol{\alpha}_{MP}$ and σ^2_{MP} from equation (2.15) and (2.17) are then used for the predictions.

2.3.1.2 RVM prediction

For a new data point $\boldsymbol{x_n}$, the prediction y_n of the test feature vector is made as:

$$ P(y_n|\boldsymbol{y}, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) = N(y_n|u_n, \sigma_n^2) \tag{2.18} $$

where

$$ u_n = \boldsymbol{u}^T \Phi(\boldsymbol{x}_n) \tag{2.19} $$

$$ \sigma_n^2 = \sigma^2_{MP} + \Phi(\boldsymbol{x}_n)^T \boldsymbol{\Sigma} \Phi(\boldsymbol{x}_n) \tag{2.20} $$

In practice, $\boldsymbol{w}$ is set to fixed values $\boldsymbol{u}$ in equation (2.19) for the purpose of point prediction. Note that the variance σ_n^2 in equation (2.20) is comprised of the estimated noise σ^2_{MP} and a term$\Phi(\boldsymbol{x}_n)^T \boldsymbol{\Sigma} \Phi(\boldsymbol{x}_n)$ representing the uncertainty in the prediction.

RVM presents the learnt regression model as the most relevant set of extracted feature dimensions, meaning the technique explicitly performs both dimensionality reduction and feature selection without the need for holding out a subset of validation data. This is a desirable quality as it helps to minimise the chances of over-fitting during system development. More importantly, the RVM output is probabilistic given that the σ_n^2 contains the uncertainty in the prediction as shown in equation (2.20). The probabilistic output could be useful in many applications, i.e. clinical practice.

2.3.2 Gaussian Mixture Model (GMM)

GMMs are probabilistic models for representing data within an overall population, and have been used extensively in speech processing. Real-world data can follow a multimodal distribution, thus fitting them to a unimodal model generally gives a poor fit. GMMs aim to model the data distribution

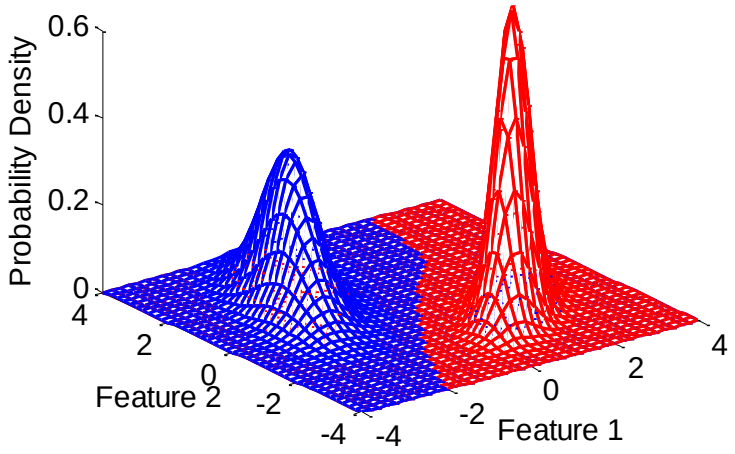


Figure 2.8: Data fitting to a 2-mixture Gaussian Mixture Model. Feature 1 and Feature 2 are the 2 dimensional features.

as a mixture of multiple unimodal Gaussian distributions as shown in Figure 2.8. Furthermore, GMMs maintain many of the theoretical and computational benefits of Gaussian models, making them practical for efficiently modelling very large datasets.

GMMs are parameterised by two types of values, the mixture component weights and the component means and covariance. For a GMM with M components, the model can be represented as:

$$\lambda = P(\boldsymbol{x}) = \sum_{i=1}^{M} w_i N(\boldsymbol{x}|\boldsymbol{u}_i, \boldsymbol{\Sigma}_i) \tag{2.21}$$

where i represents the mixture number and M represents the total mixture number, and w_i is the mixture component weight. $N(\boldsymbol{x}|\boldsymbol{u}_i, \boldsymbol{\Sigma}_i)$ is the multivariate normal distribution of mixture i, with mean $\boldsymbol{u}_i$ and covariance $\boldsymbol{\Sigma}_i$. The varying number of mixture components ensures that GMM is able to model any real distribution as a combination of multiple Gaussians. The i_{th} mixture component $N(\boldsymbol{x}|\boldsymbol{u}_i, \boldsymbol{\Sigma}_i)$ in equation (3.9) can be further expanded as:

$$N(\boldsymbol{x}|\boldsymbol{u}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{u}_i)^T {\boldsymbol{\Sigma}_i}^{-1} (\boldsymbol{x} - \boldsymbol{u}_i)\right) \tag{2.22}$$

And a valid $P(\boldsymbol{x})$ requires w_i satisfy:

$$\sum_{i=1}^{M} w_i = 1 \tag{2.23}$$

The GMM parameters $\theta = [w, \boldsymbol{u}, \boldsymbol{\Sigma}]$ are estimated based on the maximum likelihood from a set of training data. Assuming the data points are independent, the likelihood of a GMM model λ is:

$$P(\theta|\boldsymbol{x}) \propto P(\boldsymbol{x}|\theta) = \prod_{n=1}^{N} P(\boldsymbol{x}_n|\theta) \tag{2.24}$$

where n represents the frame number. The GMM parameters θ are estimated by maximising the likelihood function $P(\boldsymbol{x}|\theta)$ in equation (2.24). Generally the expectation maximisation (EM) [96] algorithm is utilised to estimate the GMM parameters θ by maximising the loglikelihood $lnP(\boldsymbol{x}|\theta)$.

During the training phase, the GMM parameters θ are first randomly initialised, and then estimated iteratively. The posterior probability $P(i|\boldsymbol{x}_n)$ of each i_{th} mixture component given data point $\boldsymbol{x}_n$ is updated as:

$$P(i|\boldsymbol{x}_n) = \frac{w_i N(\boldsymbol{x}_n|\boldsymbol{u}_i, \boldsymbol{\Sigma}_i)}{\sum_{k=1}^{M} w_k N(\boldsymbol{x}_n|\boldsymbol{u}_k, \boldsymbol{\Sigma}_k)} \tag{2.25}$$

Then the mean $\boldsymbol{u}_i$ is updated over the entire training set as:

$$\boldsymbol{u}_i = \frac{1}{N_i} \sum_{n=1}^{N} P(i|\boldsymbol{x}_n)\, \boldsymbol{x}_n \tag{2.26}$$

$$N_i = \sum_{n=1}^{N} P(i|\boldsymbol{x}_n) \tag{2.27}$$

where N_i can be interpreted as the effective number of points assigned to mixture i. Similarly, the covariance $\boldsymbol{\Sigma}_i$ can be updated as:

$$\boldsymbol{\Sigma}_i = \frac{1}{N_i} \sum_{n=1}^{N} P(i|\boldsymbol{x}_n)\, (\boldsymbol{x}_n - \boldsymbol{u}_i)(\boldsymbol{x}_n - \boldsymbol{u}_i)^T \tag{2.28}$$

The weights w_i are then estimated as:

$$w_i = \frac{N_i}{N} \tag{2.29}$$

Another method for GMM training is adapting the GMM parameters from a pre-trained universal background model (UBM), referred as GMM-UBM. It especially suits the situation where only a small set of training data is available. The main premise behind GMM-UBM training is to use the prior information of the potential model represented by UBM (λ_{UBM}) . Maximum a Posterior (MAP) adaptation is utilised to adapt the GMM parameters from UBM.

The UBM is firstly trained as in (2.26) – (2.29). Given another set of adaptation data $\boldsymbol{x}$**,** the new

statistics of $\boldsymbol{u}_i' = E_i(\boldsymbol{x})$, $\boldsymbol{\Sigma}_i' = E_i(\boldsymbol{x}^2) - E_i^2(\boldsymbol{x})$ and w_i' are similarly estimated as in (2.26) – (2.29). The new parameters are estimated as a combination of the UBM and the new statistics as:

$$\widehat{\boldsymbol{u}}_i = (1 - \alpha_i)\boldsymbol{u}_i + \alpha_i \boldsymbol{u}_i^{'} \tag{2.30}$$

$$\widehat{\boldsymbol{\Sigma}}_i = (1 - \alpha_i)(\boldsymbol{\Sigma}_i + \boldsymbol{u}_i) + \alpha_i E_i(\boldsymbol{x}^2) - \widehat{\boldsymbol{u}}_i^2 \tag{2.31}$$

$$\widehat{w}_i = [(1 - \alpha_i)w_i + \alpha_i * w_i'] * \varphi \tag{2.32}$$

where φ is a scalar that guarantees the weight components sum to unity. α_i is the adaptation coefficient for each mixture i, ensuring the adaptation conducted only when the training data is reliable . α_i is represented as:

$$\alpha_i = \frac{N_i}{N_i + r} \tag{2.33}$$

where r is the relevance factor that controls the degree of adaptation. The GMM-UBM adaptation scheme is utilised in this Chapter since the emotion database is small.

2.3.3 Gaussian Mixture Regression (GMR)

GMR aims to develop a joint density function over features and labels during the training phase, and then derives the conditional probability and regression functions for the test features.

Let $\boldsymbol{X}_n = [\boldsymbol{x}_n^{\mathrm{T}}, \Delta\boldsymbol{x}_n^{\mathrm{T}}]^{\mathrm{T}}$ represent the features consisting of the static information (low level descriptors) and dynamic information (generally delta features) and $\boldsymbol{Y}_n = \boldsymbol{y_n}$ represent labels at frame n, where the delta values are calculated as in [97] as

$$\Delta\boldsymbol{x_n} = \frac{\boldsymbol{x_{n+1}} - \boldsymbol{x_{n-1}}}{\boldsymbol{2}} \tag{2.34}$$

The training features and labels are represented as $\boldsymbol{X} = [\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, \cdots \boldsymbol{X}_N^T]^{\mathrm{T}}$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^T, \boldsymbol{Y}_2^T, \cdots \boldsymbol{Y}_N^T]^T$, where N represents the total number of frames. The GMM $\lambda^{[\boldsymbol{Z}]}$ of the joint probability distribution of features and labels is trained using all the joint features $\boldsymbol{Z}_n = [\boldsymbol{X}_n^T, \boldsymbol{Y}_n^T]^T$ by the EM algorithm as [83]:

$$\lambda^{[\boldsymbol{Z}]} = \sum_{m=1}^{M} w_m N\left([\boldsymbol{X}, \boldsymbol{Y}]; \begin{bmatrix} \boldsymbol{u}_m^{(\boldsymbol{X})} \\ \boldsymbol{u}_m^{(\boldsymbol{Y})} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \\ \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} & \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} \end{bmatrix}\right) \tag{2.35}$$

where m is the mixture number, M is the total number of mixtures, and w_m is the weight for each mixture. $\boldsymbol{u}_m^{(X)}$ and $\boldsymbol{u}_m^{(Y)}$represent the mean vectors of the m^{th} mixture component for the features and labels respectively. The matrices $\boldsymbol{\Sigma}_m^{(XX)}$and $\boldsymbol{\Sigma}_m^{(YY)}$ represent the covariance of the m^{th} mixture for the features and labels. $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$ are the cross-covariance matrices of the m^{th} mixture for the features and labels. Full covariance matrices are employed to better capture statistical properties of the features and labels. The reader may refer to Section 3.2.3 of this thesis for the full details of the GMM.

In order to find label $\boldsymbol{Y}_n$ for each frame n, the conditional probability of label $\boldsymbol{Y}_n$ given $\boldsymbol{X}_n$ is estimated, as shown in Figure 2.9. A three mixture GMM is modelled as $\lambda^{(\boldsymbol{Z})}$. Given the test feature vector $\boldsymbol{X}_n$, the conditional probability $P\left(\boldsymbol{Y}_n\middle|\boldsymbol{X}_n,\lambda^{(\boldsymbol{Z})}\right)$ is estimated as the cross-section. Assuming the independence of data points $\boldsymbol{Y}_n$, the overall conditional probability $P\left(\boldsymbol{Y}\middle|\boldsymbol{X},\lambda^{(\boldsymbol{Z})}\right)$ is represented as [83]:

$$P\left(\boldsymbol{Y}\middle|\boldsymbol{X},\lambda^{(\boldsymbol{Z})}\right)=\prod_{n=1}^{N}\sum_{m=1}^{M}P\left(\boldsymbol{Y}_n\middle|\boldsymbol{X}_n,\lambda^{(Z)}\right)=\prod_{n=1}^{N}\sum_{m=1}^{M}P(m|\boldsymbol{X}_n,\lambda^{(\boldsymbol{Z})})P\left(\boldsymbol{Y}_n\middle|\boldsymbol{X}_n,m,\lambda^{(\boldsymbol{Z})}\right) \quad (2.36)$$

where $P(m|\boldsymbol{X}_n,\lambda^{(\boldsymbol{Z})})$ represents the probability of $\boldsymbol{X}_n$ belonging to the m^{th} mixture as:

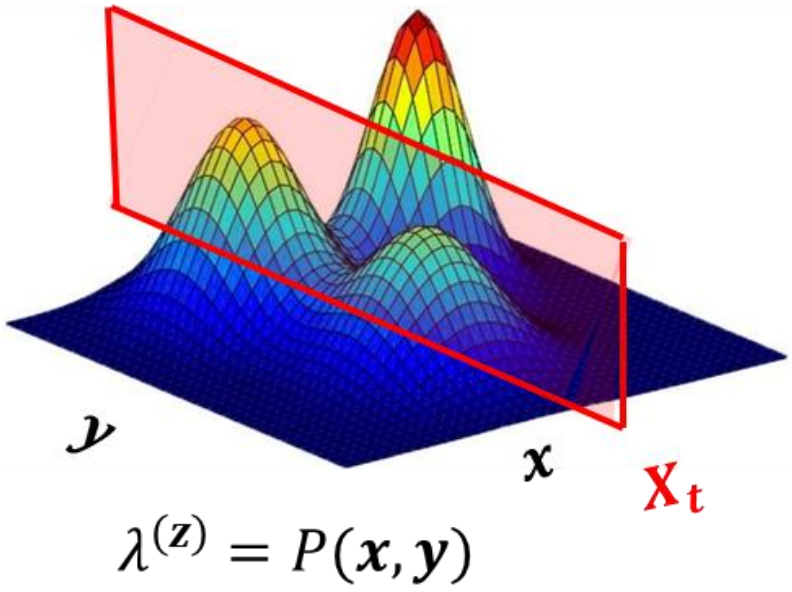


Figure 2.9: An example of a two dimensional GMM $\lambda^{(\boldsymbol{Z})}$with three mixture components. x and y represent the variables of feature and label; $\boldsymbol{X}_t$ represents the test features at frame t.

$$P\left(m\middle|\boldsymbol{X}_n,\lambda^{(\boldsymbol{Z})}\right)=\frac{w_m N(\boldsymbol{X}_n;\boldsymbol{u}_m^{(\boldsymbol{X})},\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})})}{\sum_{k=1}^{M} w_k N(\boldsymbol{X}_n;\boldsymbol{u}_k^{(\boldsymbol{X})},\boldsymbol{\Sigma}_k^{(\boldsymbol{XX})})} \tag{2.37}$$

In equation (2.36) the posterior probability $P(\boldsymbol{Y}_n|\boldsymbol{X}_n,m,\lambda^{(\boldsymbol{Z})})$ is a Gaussian distribution with the mean $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and covariance $\boldsymbol{D}_m^{(\boldsymbol{Y})}$

$$P\left(\boldsymbol{Y}_n\middle|\boldsymbol{X}_n,m,\lambda^{(\boldsymbol{Z})}\right)=N(\boldsymbol{Y}_n;\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})},\boldsymbol{D}_m^{(\boldsymbol{Y})}) \tag{2.38}$$

where the mean $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and covariance $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ of the Gaussian distribution are

$$\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}=\boldsymbol{u}_m^{(\boldsymbol{Y})}+\boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-\mathbf{1}}(\boldsymbol{X}_n-\boldsymbol{u}_m^{(\boldsymbol{X})}) \tag{2.39}$$

$$\boldsymbol{D}_m^{(\boldsymbol{Y})}=\boldsymbol{\Sigma}_m^{(\boldsymbol{YY})}-\boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-\mathbf{1}}\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \tag{2.40}$$

It can be seen that $P\left(\boldsymbol{Y}_n\middle|\boldsymbol{X}_n,\lambda^{(Z)}\right)$ (from equation (2.36)) is also a GMM for each frame n [83]. The time sequence $\widehat{\boldsymbol{Y}}=[\boldsymbol{Y}_1^T,\boldsymbol{Y}_2^T,\cdots\boldsymbol{Y}_N^T]^T$ is estimated based on maximising the function in equation (2.36) over consecutive frames and the EM algorithm is generally applied [83].

As in [83], the auxiliary function is iteratively maximised with respect to $\widehat{\boldsymbol{Y}}$:

$$Q(\widehat{\boldsymbol{Y}},\boldsymbol{Y})=\sum_{m=1}^{M} P(m|\boldsymbol{X},\boldsymbol{Y},\lambda^{(\boldsymbol{Z})})log\,P\left(\widehat{\boldsymbol{Y}},m\middle|\boldsymbol{X},\lambda^{(\boldsymbol{Z})}\right) \tag{2.41}$$

The final estimated label vector $\widehat{\boldsymbol{Y}}$ is obtained:

$$\widehat{\boldsymbol{Y}}=(\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}})^{-1}\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}^{(\boldsymbol{Y})}} \tag{2.42}$$

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}}=diag[\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}},\overline{\boldsymbol{D}_2^{(\boldsymbol{Y})^{-1}}},\cdots,\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}},\cdots,\overline{\boldsymbol{D}_N^{(\boldsymbol{Y})^{-1}}}] \tag{2.43}$$

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}^{(\boldsymbol{Y})}}=[\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_1^{(\boldsymbol{Y})}}^T,\overline{\boldsymbol{D}_2^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_2^{(\boldsymbol{Y})}}^T,\cdots,\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_n^{(\boldsymbol{Y})}}^T,\cdots\overline{\boldsymbol{D}_N^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_N^{(\boldsymbol{Y})}}^T] \tag{2.44}$$

$$\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}}=\sum_{m=1}^{M}\gamma_{m,n}\,\boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}} \tag{2.45}$$

$$\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_n^{(\boldsymbol{Y})}}^T=\sum_{m=1}^{M}\gamma_{m,n}\,\boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})} \tag{2.46}$$

$$\gamma_{m,n}=P(m|\boldsymbol{X}_n,\boldsymbol{Y}_n,\lambda^{(\boldsymbol{Z})}) \tag{2.47}$$

The derivation of (2.41) is given in the Appendix.

A good approximation algorithm to the EM algorithm [83, 98] with the dominant mixture sequence $\widehat{\boldsymbol{m}}$ has been shown to be effective in voice conversion systems [83]. The likelihood in equation (2.36) can be approximated with a mixture component sequence for each frame as:

$$\widehat{\boldsymbol{m}} = argmax_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})}) \tag{2.48}$$

where $\widehat{\boldsymbol{m}} = [\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_n, \cdots \hat{m}_N]$ with $\hat{m}_n$ indicating the dominant mixture component at frame n:

$$\hat{m}_n = arg \max_{1 \le m \le M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) \tag{2.49}$$

Then the approximated estimated label $\widehat{\mathbf{Y}}$ can be estimated based on the dominant mixture component sequence shown as:

$$\hat{\boldsymbol{y}} = argmax_{\boldsymbol{y}} P(\widehat{\boldsymbol{m}}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})})\, P(\boldsymbol{Y}|\boldsymbol{X}, \widehat{\boldsymbol{m}}, \lambda^{(\boldsymbol{Z})}) \tag{2.50}$$

Instead of considering the conditional probability distribution $P\left(\boldsymbol{Y}_n \middle| \boldsymbol{X}_n, \lambda^{(Z)}\right)$ for each frame to be a GMM, the approximate algorithm takes the approach of adopting the dominant Gaussian mixture component as the posterior probability. Our preliminary experimental results indicate that this approximate algorithm gives comparable results to the EM algorithm in continuous emotion prediction, and was thus utilised throughout this chapter. A convenient means to estimate the uncertainty as the standard deviation of the dominant mixture component for each frame n is provided by replacing the GMM with a single Gaussian distribution.

2.4 Databases

A large number of early emotion databases are comprised of recordings of acted behaviours, where several speakers read scripts with different emotion states. The annotations of such databases mainly focused on emotion categories. SUSAS (Speech Under Simulated and Actual Stress), collected by J. Hansen at the University of Colorado Boulder in 1999, contains 32 speakers and 16000 utterances in total [99]. Burkhardt, F. collected a German database of acted emotional speech [100], containing ten sentences performed in 6 target emotions by ten actors. The Serbian Emotional Speech Database [101]

contains six actors' recordings with five acted emotion states. However, scientists have found that acted emotion in speech is different from spontaneous emotional speech, and the optimal features and methods for analysing these two kinds of speech are different [102, 103]. With the increasing attention drawn to continuous emotion prediction systems, databases with continuous annotations of arousal and valence became an urgent need. Researchers then started to collect spontaneous speech databases, closer to realistic speech, with both categorical and continuous emotion annotations.

The FAU Aibo Emotion Corpus [104] consists of 9 hours of German speech of 51 children from the age 10-13 years, spontaneously interacting with the Sony's pet robot Aibo. This database is annotated with 11 emotion categories for each small chunk. The length of chunks is in between word level and turn level, which is generated by annotating the whole database to generate turns in terms of free phrases, dislocations, vocatives, etc. by a coarse syntactic boundary system, and then applying the similar rule to the turns to create the chunks. The Vera am Mittag German Audio-Visual Spontaneous Speech Database (VAM) [105] is an audio-visual database containing 12 hours of recordings of a German TV talk-show, and is annotated in terms of three emotion dimensions: arousal, valence and dominance. The most popular database for emotion recognition and prediction before 2015 was the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [106], which was collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). IEMOCAP [106] is an audio-visual database collected in dyadic interactions between professional actors. It is one of the early databases that recorded dyadic interactions between male-female professional actors. 10 subjects were recorded in 5 dyadic sessions, amounting to approximately 12 hours of recordings. Each utterance and video clip is annotated by three annotators in terms of the categorical emotion labels and the continuous emotion intensity on an utterance basis. The categorical emotion labels include specific types of emotions (happiness, anger, sadness, frustration, neutral state, etc.), while the continuous emotion intensity includes three dimensions: arousal, valence and dominance with values ranging from 1 to 5 and a step size of 1. The final ground truth is the average of the multiple annotations.

The most recent released and commonly used databases for continuous emotion prediction systems since 2015 are the USC CreativeIT database [107] , the RECOLA database [108], the SEMAINE database [109] and the SEWA database [65], which can be found in Section 2.4.1 The limitations of current databases is presented in Section 2.4.2.

In general, annotations of the databases are gathered by asking several evaluators to listen to the emotional recordings, and watch the video if available, and move a cursor within a 1D or 2D emotional space to give their perception about the emotion intensity of the speech samples. The ground truth of the speech sample is obtained as the weighted or unweighted average among several evaluators [107-109].

2.4.1 Databases with Frame-level Annotation

The four databases described in this section have all been utilised in this thesis. Most experimental results are carried out using multiple databases to prove their effectiveness in different conditions, i.e. acted and spontaneous, English and French, etc. The details of the usage of each database including the partitions of training and development sets and the cross validation settings will be specifically discussed in the experimental section in each chapter.

2.4.1.1 The CreativeIT database

The USC CreativeIT multimodal database created by Metallinou et al. [107] provides a novel bridge between the study of theatrical improvisation and human expressive behaviour in dyadic interaction. The theatrical improvisation technique of Active Analysis is able to provide naturally induced affective and goal-driven interaction, and ensures a more spontaneous dynamic interaction between two actors.

The USC CreativeIT database consists of two different theatrical techniques; the two-sentence exercise, where each actor is restricted to saying one predefined sentence with a given verb driving their emotions and actions; and the paraphrase exercise, where actors are asked to act out a given script with their own words and interpretation. The database contains multimodal behaviour signals including speech and video.

In total it contains 8 sessions of 90 sentences recorded by 16 speakers in English. The attributes that are annotated for each session include arousal, valence and dominance at both frame-level and utterance-level. All continuous annotations at frame-level are performed by watching the session videos and using Feeltrace software [110] that enables the user to continuously move a mouse around a computer screen so as to indicate the attribute value in a range from -1 to 1. For the discrete annotations at the utterance-level, annotators are asked to provide a label ranging from 1 to 5. Each recording is evaluated by 2 to 4 evaluators. The final continuous attribute values are obtained by averaging all individual annotations.

2.4.1.2 The RECOLA database

The RECOLA database [108] consists of multimodal spontaneous speech in French collected in a remote collaborative framework. 23 dyadic interactions of 46 participants including 27 females and 19 males are recorded in terms of their audio, video and physiology behaviour signals. The two participants are placed in separate rooms and engage in a remote discussion about a simple task paradigm. The annotation tool developed for and used in this database is ANNEMO [111], which enables one-dimensional continuous affective annotation instead of joint two-dimensional annotations, in order to reduce the cognitive load ensuring high quality annotation. The two dimensions of arousal and valence were annotated separately every 40 milliseconds with values ranging from -1 to +1 and a step size of 0.01. Post processing was performed to reduce the unwanted variability of each annotation, and finally the ground truth of an utterance is estimated by taking the mean of the annotations provided by all six annotators.

The RECOLA database is used for AVEC 2015 and AVEC 2016. Speech data from 27 speakers was equally divided into training, development and test partitions. The affect labels of the designated test partition were not released for the challenge purpose. This partition of RECOLA database is most commonly used in the existing literatures for a direct comparison with the challenge results. Further information about RECOLA can be found in [18].

2.4.1.3 The SEMAINE databases

The SEMAINE database [109] is collected using the Sensitive Artificial Listener (SAL) paradigm, where the subject naturally interacts with an operator (role-played by a subject). The database consists of three basic scenarios: i) solid SAL, where human operators play the roles of the SAL characters; ii) semi-automatic SAL, where the system speaks to participant using phrases selected by a human operator from a pre-defined list; and iii) automatic SAL where an automated system chooses what to say. The participant interacts with four characters, or operators, with different 'personality'. 141 conversation sessions were recorded in total, and only the speech data from 20 speakers recorded over 26 sessions using the solid SAL scenario is used in this dissertation owing to its inclusion of spontaneous speech.

Five affective dimensions including valence, activation (arousal), power (dominance), expectation and intensity, and four emotion categories are annotated for each solid SAL conversation. Final annotations were the average of the individual ratings from a number of 2-8 raters for arousal, valence and dominance (power), which were obtained using the FEELTRACE tool [110].

2.4.1.4 The SEWA database

The SEWA database [109] is an audio-visual multimodal database in German which is collected 'in-the-wild', i.e. using the webcams and microphones from computers in the participants' homes or offices. Participants are given a task of discussing a commercial they have just viewed in pairs, without any scripts or constraints in regards to emotion provisions; thus it is a database with spontaneous and natural behaviour signals. These dynamic conversations were limited to 3 minutes.

The SEWA database is annotated in terms of arousal, valence and likability, i.e. how much a subject expresses a positive or a negative attitude while speaking. Each speech utterance (clip) is annotated by 6 annotators aged between 20 and 24. Post processing is applied to each individual annotation to normalise them to the same range and to remove any bias. Then the ground truth is calcuclated as a weighted average over the six normalised annotations based on the evaluator weighted estimator (EWE) approach [112].

The SEWA database was released in 2017 and has been used for AVEC 2017. In total, the 32 pairs of 64 participants aging from 18 to 60 were divided into three partations: training, development and test, which contain 34, 14 and 16 utterances respectively. It should be noted that the labels of the test partition were not provided since AVEC 2017 challenge organisers hold them to validate the system performance of participants' submissions.

One crucial improvement of the SEWA database over other databases is the provided transcripts, which have been transcribed from the video chats manully. Timestamps have been used to indicate which subject is speaking. This has motivated more investigations on the lingustic analysis and experimental results have shown superior performance [65, 113].

2.4.2 Limitations of current databases

Most of the existing emotion databases have some limitations with regards to assessing the performance of the emotion recognition and prediction systems. Some of these limitations [11] can be briefly summarised as:

(a) Most emotion databases do not simulate emotions in a natural and clear way. They have used different techniques such as the sensitive artificial listener (SAL) paradigm or the theatrical improvisation technique of active analysis, but there has been no analysis as to which method is the most appropriate to generate natural expressions of emotion.

(b) Transcripts were not provided in most early databases, which makes the analysis of linguistic content difficult.

(c) Many databases are recorded in a controlled environment, where the recording quality may not exactly match a realistic scenario.

(d) The 'ground truth' of emotion recognition and prediction is obtained by averaging the human ratings based on the perception of emotion. This may not be the optimal way to define the ground truth.

2.5 Challenges

2.5.1 Overview

Continuous emotion prediction has attracted much attention in the last few decades, and has been mainly focused on finding representative acoustic and linguistic features, or on developing more advanced regression modelling techniques for prediction. Most of the literature has focused on the acoustic features, and the eGeMAPS feature set [114] proposed in 2016 has been one of the most widely adopted acoustic feature set. This is used as a standard minimalistic set of voice parameters, to ensure compliance with state-of-the-art methods allowing appropriate comparison of results across studies. More and more research has focused on linguistic features owing to the increased availability of speech database transcripts. In addition, the nonverbal vocal gestures that extracted from the segments, such as 'laughter', 'pause' or 'filler', have also been proven to provide complimentary information in predicting emotion categories. This has motivated further analysis of continuous emotion prediction [66].

Regarding regression modelling techniques, there has been an increasing interest in deep learning structures in emotion recognition and prediction, especially recurrent neural networks. One of its key advantages is the ability to model the short-term and long-term temporal dependencies since these aspects play a crucial role in human emotion expression and prediction. Other regression techniques including GMR, RVM, and SVR also show great potential in different system configurations.

However, many challenges still exist in continuous emotion prediction systems, which are listed as follows:

- While the investigation of acoustic features has solidified, the analysis of linguistic features has just started owing to the transcripts provided. Despite the success of the initial analysis using PLLR features [63] and the bag-of-words representation of acoustic LLDs and text-based features [65], there still lacks knowledge regarding the appropriate use of transcripts.

- The back-end of deep learning structures has proven promising as an area of research in speech processing, especially LSTM in continuous emotion prediction. However, there is still limited literature that analyses LSTM, giving insight into what and how it captures emotion-related information.
- As mentioned in Section 2.4.3, a standard principle of data collection has not been determined yet, which will limit the development on the speech-based automatic systems.
- Another great challenge in continuous emotion prediction is the emotion-unrelated factors introduced into the systems. Emotion expression and perception vary among different speakers due to cultural background, gender or even religion [115, 116]. This results in speaker variability and annotation variability respectively. Speaker variability refers to the difference in emotion expression occurring between different speakers, while annotation variability refers to the difference in emotion perception occurring different evaluators, further discussed in Section 2.5.3.

Speaker variability has been shown to be one of the most confounding factors in continuous emotion prediction systems. Different speakers sound different and their corresponding speech characteristics in the feature space may be different, and some speakers may be more expressive while others are more introverted. All these differences among speakers will introduce variability into continuous emotion prediction systems, which may lead to less accurate models and predictions. The inter-rater variability, indicating the agreement among annotations, is generally neglected in current systems since the typical approach is to use some weighted or unweighted average among annotators as the ground truth. However, some literature argues the importance of this variability and this has motivated serious consideration in recent systems [117]. This thesis focuses on these two categories of variability in continuous emotion prediction systems, and provides insights into how the two sources of variability affect continuous emotion predictions systems. Furthermore, compensation techniques and a novel framework to compensate for or adopt this variability have been proposed to address the problem. In addition, the long-term temporal dependencies of emotion have also been considered and incorporated into the proposed framework.

2.5.2 Speaker Variability

The expression of emotion varies among different speakers due to many factors, including cultural background, gender and religion [115, 116], thus the features extracted from different speakers not only contain emotion-specific information, but also speaker-specific information, as shown in Figures 2.10 and 2.11. Figure 2.10 represents three speakers with different emotion expression ranges. It can be seen that speaker 3 is more expressive than speaker 2, as indicated by the wider distribution represented by the black dashed lines. The varying feature distribution of the three speakers is what

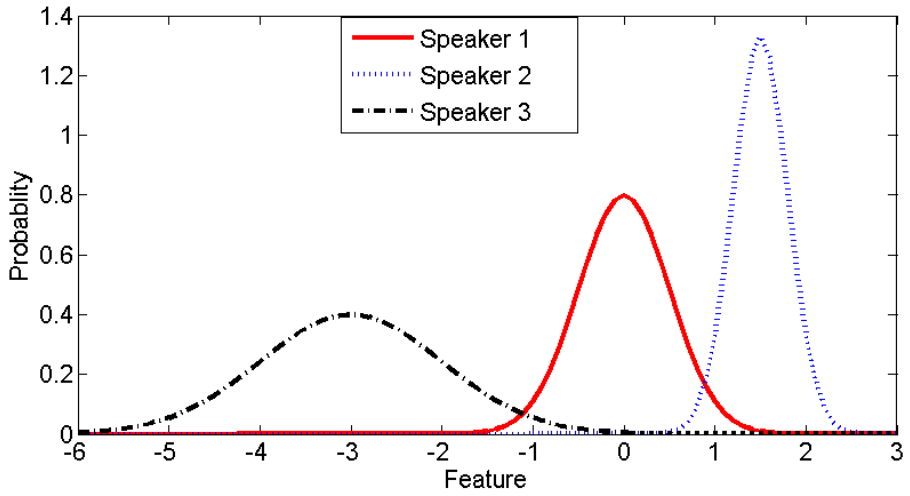


Figure 2.10: An example of 1-dimension feature distribution of three speakers. The distributions for three speakers are quite different and it introduces the speaker variability in the regression models.

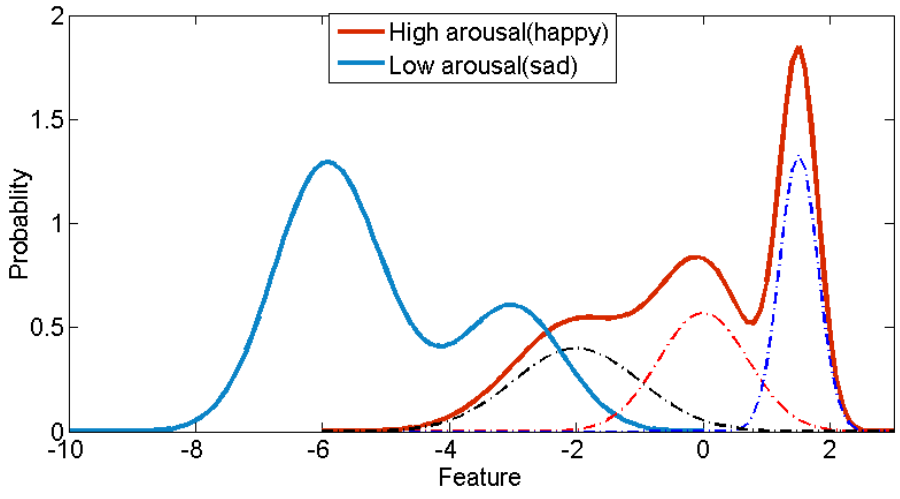


Figure 2.11: An example of 1-dimensional feature distribution of high arousal and low arousal emotion states. The red solid line represents the overall distribution of high arousal (happy) state of three speakers, indicated by the coloured dash lines; The blue line indicates the overall distribution of low arousal (sad) that is also generated by the same three speakers.

introduces speaker variability into the regression model. Regarding making predictions about an unknown speaker whose speaker-specific information was not included in the regression model, the system may produce a less reliable prediction. Figure 2.11 shows two feature distributions of high and low arousal states, i.e., happy and sad. The overall distribution of high arousal (happy) state is modelled by all the speakers, where the dashed lines shown in Figure 2.10 indicate three speaker-dependent distributions. Similarly, the overall distribution of low arousal (sad) represented by the blue solid line is also generated using all the speaker information. As observed, the distributions of high and low arousal overlap significantly, due to that the way to express similar emotion state varies among speakers. It reduces the discrimination between emotion states, and makes emotion recognition and prediction a harder task.

Some of the literature has considered speaker variability in emotion recognition systems, but only a small number of studies have taken it into account in continuous emotion prediction systems. A general idea is to borrow the variability compensation techniques used in other fields like speaker verification, however, since addressing speaker variability in continuous emotion prediction is carried out in a regression diagram while other speech processing fields solve classification problems, it is expected that the compensation on speaker variability in continuous emotion prediction will work differently.

This thesis mainly analyses the effect of speaker variability in continuous emotion prediction systems from a probabilistic view in Chapter 3, and proposes compensation techniques for speaker variability in Chapter 4. Analyses on the compensation effect of the proposed techniques are further discussed in Chapter 4, which provides some insights and motivated further investigation in this field.

2.5.3 Inter-rater variability

Apart from human expression difference in emotion, the perception difference also plays a crucial role in continuous emotion predictions. Annotation schemes are used to evaluate the emotional information present in speech data, significantly depending on individual perception. However, emotion perception differs among evaluators, which can result in a very low inter-rater agreement

level of different raters [7]. Taking the average of annotations to force a ground truth makes the labels less reliable. The valence annotation of one speech utterance and video clip is shown in Figure 2.12 [81]. Three coloured lines indicate three individual ratings. As observed, inter-rater agreement varies in different segments. For instance, a large variance appears in the green segment at the start of the speech and a high agreement is achieved in the yellow segment where the speaker tends to be more positive. The existing systems treat the inter-rater variability equally over the entire utterance, but the inter-rater variability is supposed to convey the useful information of emotion certainty varying over time. Hence, how to properly take advantage of the inter-rater variability in continuous emotion prediction systems is still a challenge. More importantly, the typical framework treating emotion prediction as a point estimation which ignores the inter-rater variability warrants deeper consideration.

In Chapter 5 and 6, the impact of inter-rater variability is analysed by observing the raters' reaction

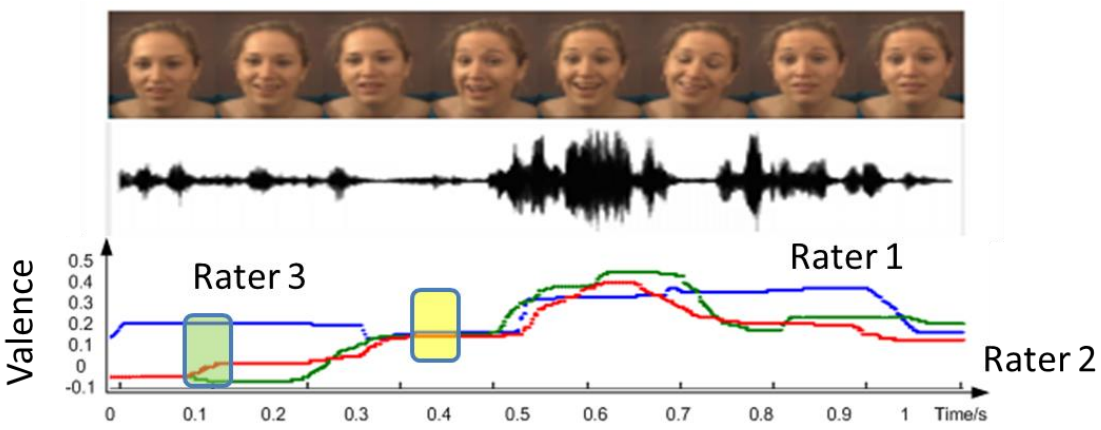


Figure 2.12: An example of 1-dimensional inter-rater variability. A video clip and the corresponding speech segment is shown on top of this figure. Three raters evaluate the video clip and speech segment simultaneously, with the individual ratings as coloured lines. The regions rectified in green and yellow regions indicate the different inter-rater variability, displayed as low and high agreement among three raters.

lag and quantifying the inter-rater variability, as well as proposing a framework to incorporate the inter-rater variability in continuous emotion predictions systems.

2.5.4 Temporal Dependencies

The incorporation of long-term dependencies is critical for continuous emotion prediction tasks [118]. Most statistical models such as SVR, RVM and GMR are not able to take temporal dependency into account on their own, but RVM and GMR can be used with additional techniques, which have the

flexibility to incorporate long-term dependencies. These are achieved by an output-associate structure for RVM (OA-RVM) and extracting dynamic features that capture information change among multiple frames for GMR [31, 83, 93]. In addition, the LSTM and RNNs with the memory cells have shown great success in continuous emotion prediction systems in automatically capturing long-term knowledge [80], and CNNs were also further explored to incorporate the past and future information for current states in continuous emotion prediction systems [118].

To capture the long-term knowledge in continuous emotion prediction systems, this thesis has expanded OA-RVM to a multimodal fusion framework, and developed GMR based system with shifted delta cepstral (SDC) features in Section 7. Furthermore, based on the great potential our novel framework that predicts emotion distribution using inter-rater variability shows, incorporating temporal dependencies in this paradigm is also investigated, which will be further discussed in Section 8.

2.6 Summary

This section has briefly explained the overview of speech-based continuous emotion prediction systems, and illustrated the current developments of feature extraction, regression modelling techniques and the commonly used evaluation matrices in this field. A variety of emotion databases were introduced and specially emphasised the four popular databases which are utilised in this thesis. One key limitation in the current database is the way to generate the emotion labels, by averaging the ratings of multiple raters, which motivates our continued research on the analyses of inter-rater variability as one target in this thesis. The main challenges I focus on in this thesis mainly are the human expression and perception difference, i.e. speaker variability and inter-rater variability respectively, which show a negative effect in continuous emotion prediction systems, and the temporal dependencies that is critical for continuous emotion prediction tasks. The in-depth analyses and proposed techniques will be further explained in details in the following chapters.

3 CHARACTERISATION OF SPEAKER VARIABILITY

3.1 Introduction

Speech production is a complex process involving the lungs, glottis, vocal tract, etc., all of which vary across different speakers [119]. This will result in differences in speech among speakers. Moreover, the linguistic context (i.e. words, syllables and phonemes) used to express emotional states also varies across speakers, due to cultural, gender and age differences [120, 121]. Speaker variability, referring to the differences in expression of emotional states among speakers, generally manifests as: (a) differences in how speakers express their gamut of emotional states; and (b) differences in how the same emotional state is expressed by different speakers. Additionally it has been shown to be one of the most confounding factor in categorical emotion recognition and continuous emotion prediction systems, which results in negative effect in emotion prediction systems. For example, the system developed using those speakers who tend to be positive will generate a less accurate prediction to the speaker who tends to be negative. Thus, compensating speaker variability in the continuous emotion prediction systems is necessary and important.

Most of the compensation techniques for speaker variability are proposed in categorical emotion recognition systems [14, 16, 122], while only limited studies have concerned with speaker variability in the continuous emotion predictions, and most of them directly borrowed the techniques from other fields, such as speaker verification and categorical emotion prediction that cast as classification problems. However, as stated in Section 2.4.2, the fundamental premise behind the compensation techniques for classification and regression systems may be quite different. Consequently, compensating speaker variability in a regression problem could not directly borrow the concepts in a classification problem.

Most existing literatures adopt z-normalisation [18, 123] and compensation in the i-vector domain [17, 20] for speaker variability in continuous emotion prediction systems. In [18], Valstar et al. examined speaker-dependent z-normalisation where the normalisation parameters were calculated and applied individually for each speaker, assuming that the feature distribution for each speaker was

different. On the other hand, the methods based on i-vectors [17, 20] operate on the basis that the total variability model can be viewed as a projection of a model of the distribution of the feature space to a more informative low-dimensional space. As previously stated, given that these two approaches were originally proposed for classification models, they can be adopted in continuous emotion prediction systems, but may not be an ideal solution for a regression problem.

Other approaches for speaker variability compensation have included the use of speaker-dependent systems followed by score level fusion strategies [21] [124]. Mencattini et al. [21] proposed the dynamic cooperative speaker models which developed single-speaker-regression-models for each speaker and obtained the predictions by combining a subset of model outputs, which were dynamically selected as the most concordant among them within a time period. Their proposed method assumed that the relationships between features and labels were different for different speakers and that the cooperative strategy of merging the predictions that exhibit a common consensus minimised the speaker variability by eliminating those speakers in which system prediction was shown less correlated with other speakers based on the concordance correlation coefficient criterion. In summary, compensating for speaker variability has led to benefits for continuous emotion prediction systems. However, these compensation methods either aim to obtain predictions from speaker-dependent systems and fuse them, or simply utilise methods developed for classification problems, such as z-normalisation and i-vector compensation. To the best of the author's knowledge, there is a dearth of analyses of how speaker variability affects continuous emotion prediction systems, and how to compensate for speaker variability in their specific case.

Speaker variability can manifest as differences in terms of the feature distribution, e.g. one speaker expresses more positive emotions while another only shows negative emotions, which leads to a wider overall feature distribution due to the distinct speaker-dependent distributions as depicted in Figure 2.10. It can also appear as differences in the relationship between individual features and labels, e.g. the differences between female and male speech expressing the same emotion, which results in less discriminative distributions between different emotion states as in Figure 2.11. In this Chapter, the aim is to analyse how these speaker-dependent feature distributions and speaker-dependent regression models differ in continuous emotion prediction systems.

A probabilistic framework to quantify speaker variability in continuous emotion systems in both the feature space and the model space, i.e. the learnt relationship between features and continuous attribute labels, is proposed. As discussed in Section 2.2.3, the regression model generally takes one of the three approaches: either (a) it models the joint probability distribution of features and labels (e.g. Gaussian mixture regression models); (b) it represents a probabilistic mapping function from the features to the labels (e.g. RVM); or (c) it defines a hard margin represented by a set of parameters which is obtained by optimising a specific objective function (e.g. support vector regression or neural networks). All three approaches are affected by speaker variability, but the first approach involves a generative model of the joint distribution over the features and labels, which lends itself to quantitative analyses of the effect of speaker variability in the feature space. Thus, this is the approach taken throughout this chapter.

The rest of this Chapter is organised as follow: a novel approach to quantify speaker variability is presented in Section 3.2. Experimental settings and results are presented in Sections 3.3 and 3.4 respectively; and the findings will be summarised in Section 3.5.

3.2 Formulation of speaker variability

As discussed in Section 3.1, speaker variability is characterised by studying: (a) differences in how speakers express their gamut of emotional states; and (b) differences in how the same emotional state is expressed by different speakers. In order to quantify these things in terms of generative modelling of the joint distribution over feature and the affective attribute label spaces of arousal, valence and dominance, they are respectively analysed as: (a) differences in the marginal distributions over the feature space (marginalised across the label space), which captures the differences in the acoustic characteristics of different speakers; and (b) differences between conditional distributions of the features given the emotional state, which captures the differences in how different speakers express similar emotional states.

3.2.1 Quantifying speaker variability

Given the joint distribution $P(\boldsymbol{x}, \boldsymbol{y}|s)$ of features $\boldsymbol{x}$ and multi-dimensional attribute labels $\boldsymbol{y}$ for each speaker s, it can be expected that the differences in the speech characteristics between two speakers, $s = i$ and $s = j$ will be reflected in the differences between the marginal distributions $P(\boldsymbol{x}|s = i)$ and $P(\boldsymbol{x}|s = j)$. Specifically, if speaker variability is a significant confounding factor I would expect the differences between the speaker specific marginal distributions to be greater than the difference between speaker specific marginal distributions and a speaker independent distribution $P(\boldsymbol{x})$. Similarly, I would also expect the differences between speaker specific conditional distributions of the features given affect labels, $P(\boldsymbol{x}|\boldsymbol{y}, s = i)$ and $P(\boldsymbol{x}|\boldsymbol{y}, s = j)$, to be greater than the differences between speaker specific conditional distributions $P(\boldsymbol{x}|\boldsymbol{y}, s)$ and the speaker independent conditional distribution $P(\boldsymbol{x}|\boldsymbol{y})$.

Furthermore, if speaker variability was a significant confounding factor, I would also expect speaker independent distributions of features from all speakers, both $P(\boldsymbol{x})$ and $P(\boldsymbol{x}|\boldsymbol{y})$, to be 'broader' distributions when compared to speaker specific feature distributions $P(\boldsymbol{x}|s)$ and $P(\boldsymbol{x}|\boldsymbol{y}, s)$. Consequently, the broadness of these marginal and conditional feature distributions are estimated, called 'widths', to test this hypothesis.

3.2.2 Proposed distribution based measurements

In this work the difference between distributions in terms of the symmetric KL divergence is estimated (sometimes simply referred to as KL divergence in this Chapter for ease of reading). This is a measure of dissimilarity between two probability distributions, with a larger KL divergence indicating a greater separation between them. The symmetric KL divergence I_{SKL} between two distributions $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$ is given by [14]:

$$I_{SKL}(P_1, P_2) = \frac{1}{2}\left|\int_x P_1(\boldsymbol{x}) In \frac{P_1(\boldsymbol{x})}{P_2(\boldsymbol{x})} d\boldsymbol{x} + \int_x P_2(\boldsymbol{x}) In \frac{P_2(\boldsymbol{x})}{P_1(\boldsymbol{x})} d\boldsymbol{x}\right| \quad (3.1)$$

Specifically, a Monte-Carlo estimate of the symmetric KL divergence proposed in [14] is utilised to quantify the separation between two distributions. For details on how this Monte-Carlo approximation

is implemented, the reader is referred to [14]. The separation between speaker specific marginal feature distributions, $P(\boldsymbol{x}|s)$, are then estimated as the average KL divergence between one speaker specific model and all other speaker models as follows:

$$S_{\boldsymbol{x}}(i) = \frac{1}{N_s - 1} \sum_{\forall j, j \neq i} I_{SKL}\big(P(\boldsymbol{x}|s = i), P(\boldsymbol{x}|s = j)\big) \tag{3.2}$$

where $S_{\boldsymbol{x}}(i)$ denotes the average KL divergence between the marginal distribution of features for the i^{th} speaker and speaker specific marginal distributions of features for all other speakers, and N_s is the total number of speakers.

The KL divergence between a speaker specific marginal feature distribution and the speaker independent feature distribution is then given by:

$$U_{\boldsymbol{x}}(i) = I_{SKL}\big(P(\boldsymbol{x}|s = i), P(\boldsymbol{x})\big) \tag{3.3}$$

where $U_{\boldsymbol{x}}(i)$ denotes the KL divergence between the marginal distribution of features for the i^{th} speaker and the speaker independent distribution of features (from all speakers).

In order to estimate the width of a distribution, the probabilistic acoustic volume (PAV) [125] is employed, depicted in Figure 3.1. PAV can be viewed as the hyper-volume corresponding to the cross-

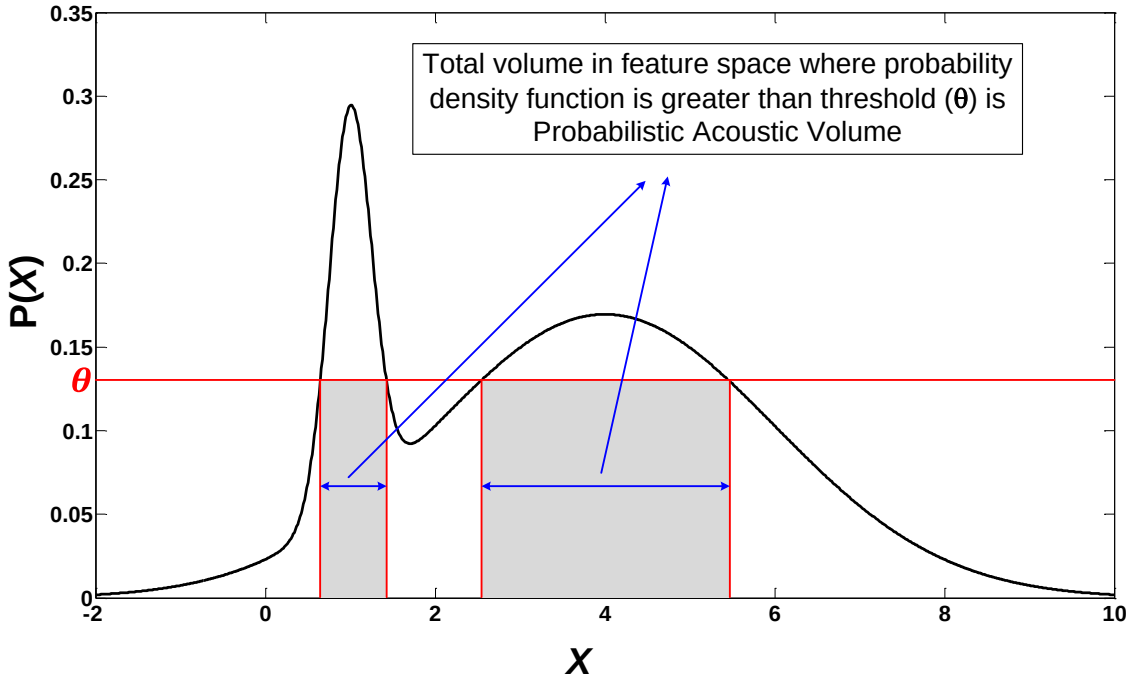
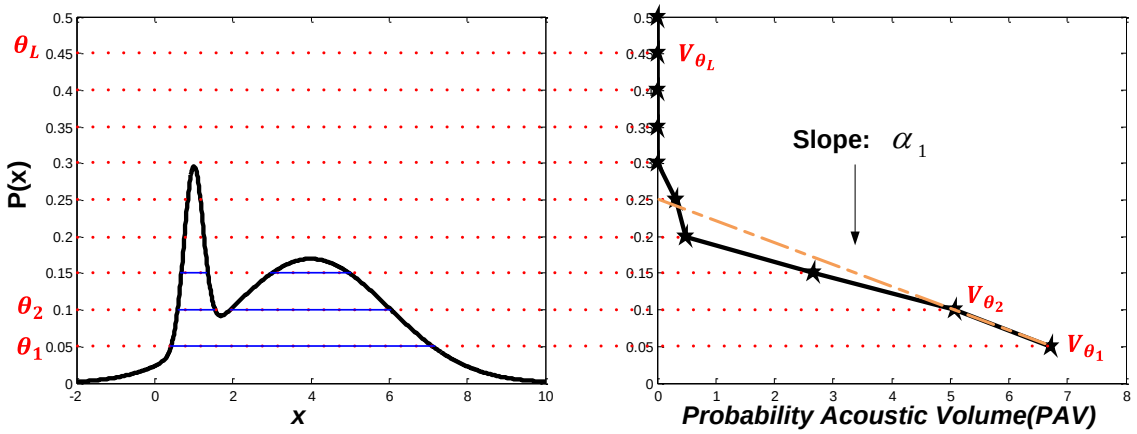
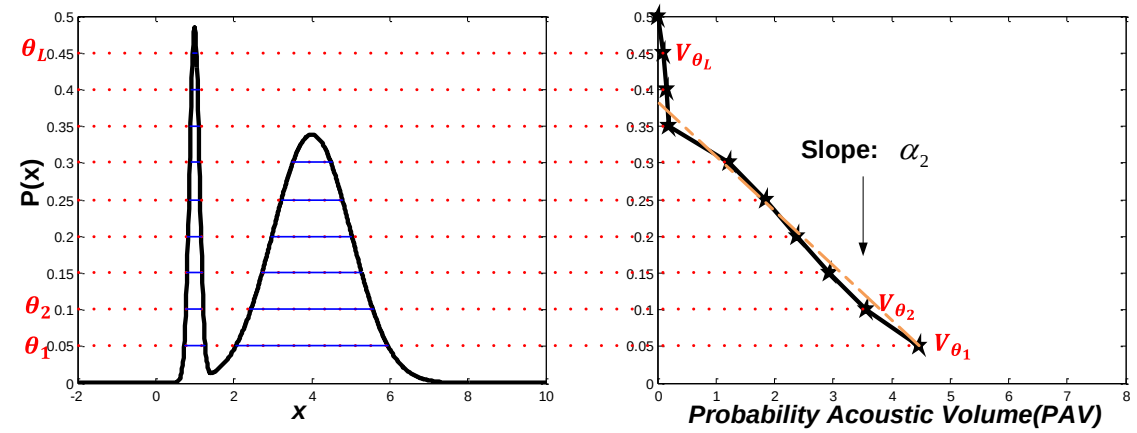


Figure 3.1: Estimation of the probabilistic acoustic volume of 1-dimensiaonl distribution P(x). θ is the threshold; The probablistic acoustic volume is indicated by the grey area.



(a)



(b)

Figure 3.2: (a) Estimate of probability acoustic volume (PAV) profile for a sample 1-dimensional feature distribution reflecting two small feature clusters; (b) estimate of probability acoustic volume profile for an example 1-dimensional feature distribution reflecting two wide low density feature clusters exhibiting a 'low concentration' of features.

section of a multivariate distribution at a specified threshold. Given one threshold θ, the total acoustic volume of a distribution $P(\boldsymbol{x})$ can be estimated as the grey area, where the probability $P(\boldsymbol{x})$ larger than θ is obtained as:

$$V_\theta = \int f(x)dx, \;\; where, f(x) = \begin{cases} 1, P(\boldsymbol{x}) > \theta \\ 0, P(\boldsymbol{x}) < \theta \end{cases} \tag{3.4}$$

A Monte Carlo approach is utilised to compute V_θ as in [126].

Further, by defining a series of thresholds $\theta = [\theta_1, \theta_2, \cdots \theta_L]$, the corresponding PAV can be computed as $V = [V_{\theta_1}, V_{\theta_2}, \cdots V_{\theta_L}]$ as shown in Figure 3.2(a). The rate at which the corresponding PAVs vary (referred to as PAV profiles) is indicative of the width of the distribution. For instance, a

wide probability distribution shown in the left graph of Figure 3.2(a) gives rise to a shallow PAV profile, as shown in the right graph of Figure 3.2(a), while a narrower probability distribution shown in the left half of Figure 3.2(b) is indicated by a steep PAV profile as in the right graph of Figure 3.2(b). Therefore, the best linear fit to a PAV profile is estimated and the slope of the line is utilised as a measure of the steepness of the profile and consequently the width of the distribution. Noting that the slopes will be negative, a lower value indicates a steeper PAV profile, and therefore a narrower probability distribution. For example, $\alpha_1 > \alpha_2$ in Figure 3.2.

Similar to estimating these measures from the marginal feature distributions, the symmetric KL divergence and PAV based measures on the conditional distributions of the features given attribute labels $P(\boldsymbol{x}|\boldsymbol{y}, s)$ are also estimated. However, it is less straightforward to estimate $P(\boldsymbol{x}|\boldsymbol{y}, s)$ since it is the product of $P(\boldsymbol{x}_f|\boldsymbol{y}_f, s)$ over all frames f, and this does not always have a closed form. Instead, the conditional probability $P(\boldsymbol{x}|\boldsymbol{y}, s)$ can be rewritten as follows:

$$P(\boldsymbol{x}|\boldsymbol{y}, s) = \frac{P(\boldsymbol{x}, \boldsymbol{y}|s)}{P(\boldsymbol{y}|s)} \tag{3.5}$$

It can be observed that $P(\boldsymbol{x}|\boldsymbol{y}, s)$ is proportional to $P(\boldsymbol{x}, \boldsymbol{y}|s)$, which can be estimated as a GMM if $P(\boldsymbol{y}|s)$ is consistent for all speakers. Consequently, a speaker specific transformation of the affect labels, $\boldsymbol{y}$, to a normalised label, $\tilde{\boldsymbol{y}}$ is estimated such that:

$$P(\tilde{\boldsymbol{y}}|s) = P(\tilde{\boldsymbol{y}}), \qquad \forall s \tag{3.6}$$

This speaker specific label normalisation is based on feature mapping [127] and implemented by modelling speaker specific and speaker independent label distributions $P(\boldsymbol{y}|s)$ and $P(\boldsymbol{y})$ as GMMs. Following this normalisation, the conditional feature distributions of interest are approximated as:

$$P(\boldsymbol{x}|\tilde{\boldsymbol{y}}, s) = \frac{1}{P(\tilde{\boldsymbol{y}})} P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s) \tag{3.7}$$

Noting that $P(\tilde{\boldsymbol{y}})$ in equation (3.7) is identical for all speakers, the KL divergences between the conditional feature distributions of interest were estimated as:

$$S_{\boldsymbol{x}|\boldsymbol{y}}(i) = \frac{1}{N_s - 1} \sum_{\forall j, j \neq i} I_{SKL}\big(P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s = i), P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s = j)\big) \tag{3.8}$$

where, $S_{\boldsymbol{x}|\boldsymbol{y}}(i)$ denotes the average KL divergence between the conditional distribution of features given the affect label for the i^{th} speaker and the corresponding speaker specific conditional distributions for all other speaker. Again, N_s is the total number of speakers.

Similarly, the symmetric KL divergence as in equation (3.1) between a speaker specific conditional feature distribution given the labels and the speaker independent conditional feature distribution given the labels is then given by:

$$U_{\boldsymbol{x}|\boldsymbol{y}}(i) = I_{SKL}\big(P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s = i), P(\boldsymbol{x}, \tilde{\boldsymbol{y}})\big) \tag{3.9}$$

Finally, the widths of the conditional distributions $P(\boldsymbol{x}|\boldsymbol{y}, s)$ and $P(\boldsymbol{x}|\boldsymbol{y})$ are estimated as the widths of $P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s)$ and $P(\boldsymbol{x}, \tilde{\boldsymbol{y}})$ in terms of the slope of the linear fit their PAV profiles, as shown in Figure 3.2. Note that these measurements are also used to quantify the reduction effect in speaker variability after employing the proposed compensation techniques for a direct comparison in Chapter 4.

3.2.3 Gaussian Mixture Model

GMMs are utilised to model the marginal distribution $P(\boldsymbol{x}|s = i)$ and the conditional distribution $P(\boldsymbol{x}|\boldsymbol{y}, s = i)$, since it is in concordance with the GMR back-end, which models the joint GMM of feature and label distributions (further discussed in Chapter 4.3.2). Additionally, GMMs also provides the framework for a range of subsequent transforms for variability reduction in the supervector and i-vector domain [128, 129]. Details of GMMs can be referred to Section 2.4.2.

3.3 Experimental settings

The proposed analysis of speaker variability, as described in Section 3.2, was carried out on the USC CreativeIT database. The front-end employed 65 low-level descriptors (LLDs) and their first order derivatives were extracted using OpenSMILE [130], to match those used in the Computational Paralinguistics Challenge 2013 (ComParE 2013) audio feature set [131]. Three second windows with

a one second shift between windows were used to compute statistical features by applying five functionals (maximum, minimum, mean, standard deviation, and range) of each of the LLDs. PCA was used to reduce the feature dimensionality to 40, while aiming to preserve approximately 70-85% of the data variability [16, 17]. Dynamic features and labels were calculated as in [16], and concatenated with the original features and labels in order to capture the information of emotion change that has been shown to improve emotion prediction [19]. The final feature dimension was 80, including dynamic features. This feature set was utilised in the front-end for all the emotion prediction systems and in all analyses reported in this Chapter.

Experimentally GMMs with four full covariance components and 10 iterations of the EM algorithm were found to be the optimal settings for the speaker-dependent and speaker-independent models. Due to data scarcity, training individual speaker models from scratch will lead to unreliable models, consequently a universal background model (UBM) was first trained using half of each speaker's data, and speaker-dependent models were developed via Maximum a Posterior (MAP) adaptation [132] using the other half of each speaker's data.

The Monte-Carlo approach that was used to estimate the KL divergence and PAV profiles made use of 100000 samples. In terms of the PAV profiles, the overall probability of all test utterances is computed, by estimating the probability of each frame-wise test feature vector fitting to the trained GMMs. Then 37 thresholds ranging from 25 to 75 percentiles in 2 percentile steps were applied to the overall probability as in [125] to estimate the PAV profiles. Thresholds less than the 25^{th} percentile and higher than the 75^{th} percentile are not considered as in [125], since the PAV calculated using less than 25^{th} percentiles approximately equals to 1 and that higher than 75^{th} percentile is around 0.

3.4 Experimental results

3.4.1 Marginal Probability Distribution

The average KL divergence between the marginal feature distributions of a speaker and all other speakers $S_x(i)$ for each of the 16 speakers is shown in Figure 3.3 in light blue. These are compared to the KL divergences between the distributions for each speaker and a speaker independent UBM,

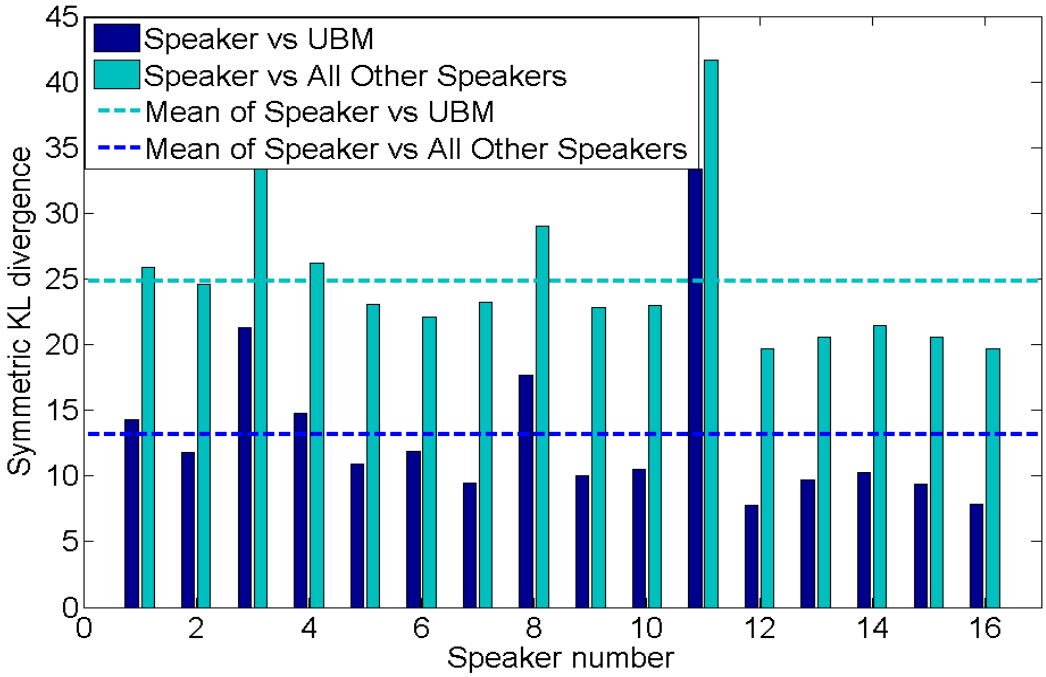


Figure 3.3: Symmetric KL divergence for speaker-dependent models and UBM of marginal probability distributions for all speakers. The average KL divergence calculated between one speaker and all other speakers (light blue) is smaller than that between that speaker and UBM (dark blue), indicating the separation between speaker-dependent distributions in terms of the marginal probability distributions.

$U_x(i)$, shown in dark blue. Finally, the mean values of both measures across all 16 speakers are indicated by the dark and light blue dotted lines. It can be seen that $S_x(i)$ is consistently greater than $U_x(i)$, indicating that the separation between speaker specific models is greater than the separation between a speaker model and the speaker independent model (UBM).

Table 3.1 shows the mean and median values of the slopes of the linear fits to the PAV profiles of the speaker specific marginal distributions and compares them to the slope of the linear fit to the PAV profile of the speaker independent UBM. It should be noted that the slope of PAV profile of Speaker 1 achieves extremely low value, i.e. -9.67*10^55, indicating that feature distribution of Speaker 1 is quite concentrated compared to other speakers (slope values ranging within [-3.45 -2.60]). Speaker 1

Table 3.1: Slope of PAV profiles for marginal distributions

	Speaker-independent	Speaker-dependent		
		Mean	Median	Standard deviation
Slope of PAV α	-2.79	-2.95	-2.91	0.26

was a clear outlier and consequently not included. Consequently, the mean and standard deviation were calculated only for Speakers 2 to 16. It was observed that the mean and median slopes for the speaker specific distributions are lower than that of the speaker-independent UBM, indicating that the speaker models are 'narrower' than the speaker independent UBM, which is as expected.

Taken together, the results shown in Figure 3.3 and Table 3.1 suggest that marginal feature distributions for individual speakers are distinct from each other and 'narrower' than the marginal distribution over features from all speakers. Therefore, pooling data from multiple speakers to train a single model will lead to a broader model which in turn are likely to lead to less accurate predictions, showing that speaker variability is a significant confounding factor.

3.4.2 Conditional Probability Distribution

In order to estimate both separation between the conditional distributions of features given affect labels and their widths, this section proposes normalising the affect labels $\boldsymbol{y}$ to obtain normalised labels $\tilde{\boldsymbol{y}}$, such that the distribution of the normalised labels is consistent across all speakers, as described in Section 3.1.2. Also, as mentioned therein, feature mapping [19] was used to map the speaker-dependent labels to a consistent distribution. However, feature mapping adapts the mean and covariance, but not the weights of a GMM. Thus the mapped label distributions $P(\tilde{\boldsymbol{y}}|s)$ are similar across all speakers s, but not identical. Consequently, it should be noted that the comparison of $P(\boldsymbol{x}|\boldsymbol{y}, s)$ values are only indicative and not definitive.

To verify that the label normalisation technique works as required, the average KL divergence I_{SKL} is estimated as in equation (3.1) between label distribution for one speaker and all other speakers $S_{\boldsymbol{y}}(i)$ as:

$$S_{\boldsymbol{y}}(i) = \frac{1}{N_s - 1} \sum_{\forall j, j \neq i} I_{SKL}\big(P(\boldsymbol{y}|s = i), P(\boldsymbol{y}|s = j)\big) \tag{3.10}$$

where N_s is the total number of speaker. The KL divergence between speaker specific label distributions and speaker independent label distributions $U_{\boldsymbol{y}}(i)$ for the un-normalised labels is also estimated as:

$$U_{\boldsymbol{y}}(i) = I_{SKL}\big(P(\boldsymbol{y}|s=i), P(\boldsymbol{y})\big) \tag{3.11}$$

Then these are compared to the equivalent measures estimated on the normalised labels $S_{\tilde{\boldsymbol{y}}}(i)$ in yellow and $U_{\tilde{\boldsymbol{y}}}(i)$ in red in Figure 3.4. From this comparison it can be seen that the KL divergences are dramatically reduced after normalisation, suggesting the normalisation was effective in making all the label distributions similar to each other as intended.

Following this, as in the case of the marginal distributions described in Section 3.3.1 the average KL divergence between one speaker and all other speakers $S_{\boldsymbol{x}|\boldsymbol{y}}(i)$, shown in dark blue for each of the 16 speakers, are compared with the average KL divergence computed between one speaker and UBM as in light blue $U_{\boldsymbol{x}|\boldsymbol{y}}(i)$ in Figure 3.5. The widths of these conditional distributions in terms of slopes of linear fits to their PAV profiles are compared in Table 3.2. These comparisons agree with the observations made in the case of the marginal distributions and lend further support to the suggestion

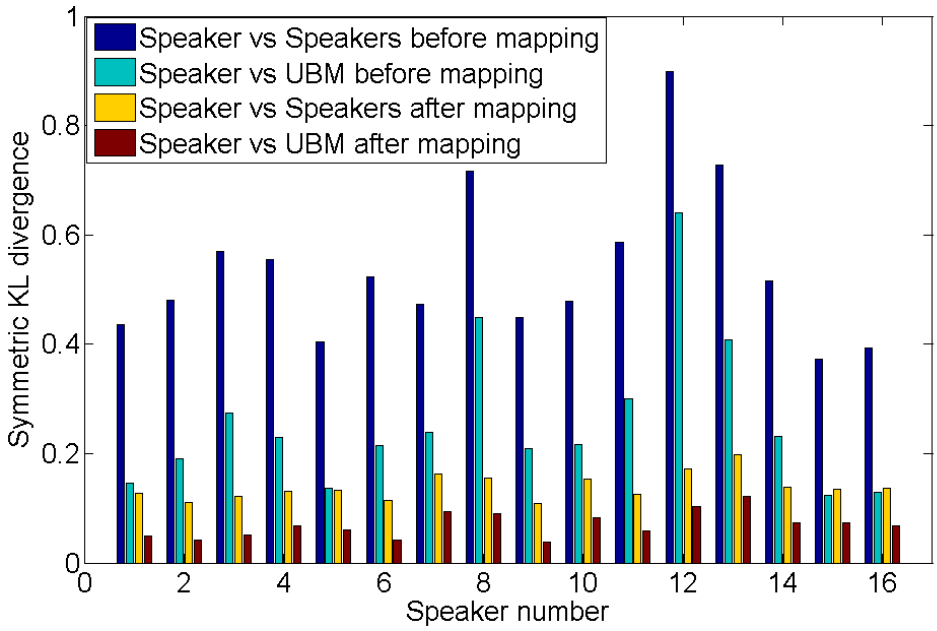


Figure 3.4: Symmetric KL divergence before and after mapping of $P(\boldsymbol{y}|s)$ for all speakers. The average KL divergence between one speaker and all other speakers $S_y(i)$ before label normalisation is represented in dark blue; The average KL divergence between one speaker and UBM $U_y(i)$ before label normalisation is represented in light blue; The average KL divergence $S_{\tilde{y}}(i)$ after label normalisation is represented in yellow; The average KL divergence $U_{\tilde{y}}(i)$ after label normalisation is represented in red. The speaker-dependent label distribution after normalisation is significantly smaller than that before normalisation, indicating that the mapped speaker-dependent distributions are similar. This ensures a effective comparison in terms of the conditional probability distributions (referring to equations (3.6) and (3.7)).

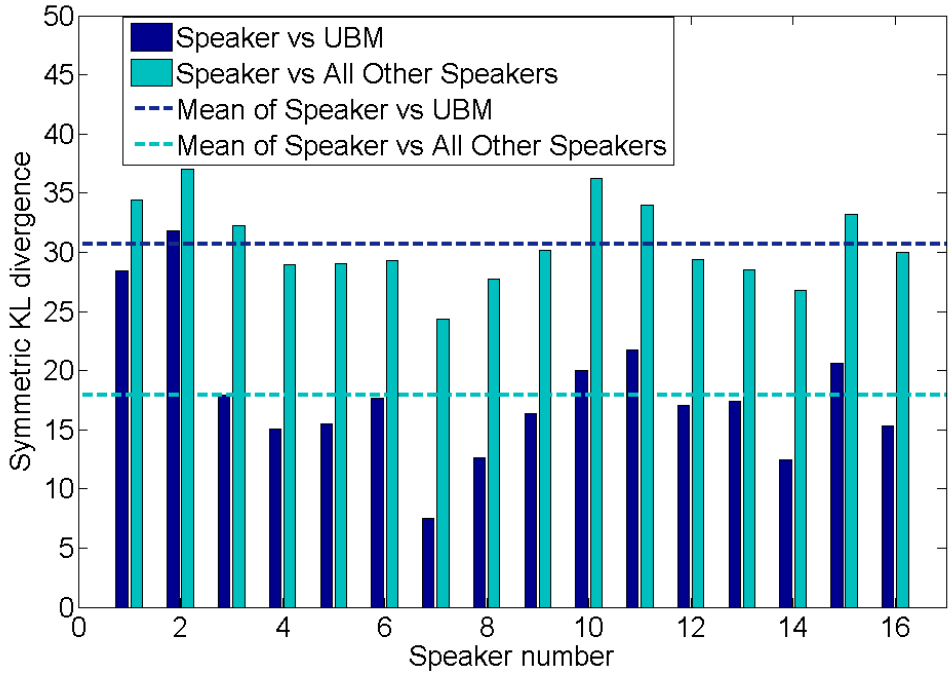


Figure 3.5: Symmetric KL divergence for speaker-dependent models and the UBM of conditional probability distribution. The average KL divergence calculated between one speaker and all other speakers (light blue) is smaller than that between that speaker and UBM (dark blue), indicating the separation between speaker-dependent distributions in terms of the conditional probability distributions.

that speaker variability is a significant confounding factor and compensating for speaker variability prior to training emotion prediction systems may be beneficial.

3.5 Summary

This chapter has analysed speaker variability in continuous emotion prediction systems in terms of marginal and conditional feature distributions in order to gain insight into how speaker variability affects emotion prediction systems. Measures of inter- and intra-speaker variability in terms of the symmetric KL divergence and the probabilistic acoustic volume profiles respectively, were adopted to quantify speaker variability based on comparisons of speaker-dependent Gaussian mixture models of the feature space. It was found that speaker variability showed negative effect in inter-speaker variability indicated by the distinct speaker-dependent feature distributions, but only altered the intra-speaker variability slightly, for both marginal and conditional probability distributions. In addition, the intra-speaker variability was compared to the local variability in a speaker independent universal

background model and it was found that the speaker independent model was broader than that of the speaker-dependent models, suggesting that differences between speakers increased the widths of the speaker-independent feature distributions, which is not expected since the variability within the broader model may generate less reliable predictions. This analysis suggests that speaker variability is indeed a significant confounding factor and that compensating for speaker variability prior to training emotion prediction systems may be beneficial. This motivates our research on compensation techniques for speaker variability in continuous emotion prediction systems, which appear in Chapter 4.

4 COMPENSATION TECHNIQUES FOR SPEAKER VARIABILITY

4.1 Motivation and Introduction

Speaker variability has been shown to be a significant confounding factor in speech based emotion prediction systems in Chapter 3. However, most current continuous emotion prediction systems either ignore speaker variability or adopt normalisation techniques that were originally proposed for classification problems. Two of the most widely adopted techniques are speaker-wise z-normalisation [17, 18] and i-vectors [19, 20].

Some of the other compensation methods for speaker variability in classification problems include: (a) normalisation techniques such as joint factor analysis based normalisation methods [9], iterative feature normalisation [12], and an auto-encoder based transfer learning method [11]; and (b) model compensation techniques, which improve the model representation to decrease the variability [13-14]. These are less employed in continuous emotion prediction systems owing to certain constraints that that prevent them from being applied directly to regression systems.

As previously stated in Chapter 3, the fundamental premise behind speaker variability compensation in classification and regression systems is expected to be quite different. In terms of emotion classification systems, inter-class variability is supposed to be maximised while intra-class variability is minimised. However, as a regression problem of continuous emotion prediction, the aim of compensating for speaker variability is to reduce inter-class variability, where the class refers to speaker. Therefore, the methods utilised for classification problems cannot be directly applied to continuous emotion prediction systems.

Based on the analysis in Section 3.4, this chapter proposes three compensation methods based on factor analysis, partial least square dimension reduction (PLSDR) [23] and feature mapping [24], and further compares these against a number of the state-of-the-art techniques [13, 21, 25]. Factor analysis based normalisation aims to decompose the feature space into emotion-specific and speaker-specific

spaces, and reduce the speaker information contained in in the speaker-specific space. PLSDR, a technique widely used in chemometrics [26] that projects the observed data to its latent structure, assumes that the observed data is generated by a process driven by only a small number of latent variables. Generally high dimensional feature vectors with dimensionality ranging from hundreds to thousands, which are computed by stacking a number of functionals of LLDs, are utilised to represent the emotional content in speech signals, and to develop the prediction systems. However, they might be redundant and can benefit from a low dimensional representations in the latent space. The proposed PLSDR based normalisation technique projects the original feature space to a latent space that minimises the speaker variability. Feature mapping for GMMs was shown to be promising for channel compensation in speech based speaker verification systems [24]. This technique maps the channel-dependent feature space to a channel-independent feature space. A similar concept is adopted and expanded to compensate for speaker variability in continuous emotion prediction systems, by similarly mapping the speaker-dependent space to a speaker-independent space.

The proposed compensation techniques are described in Section 4.2. Then the key experiment settings and evaluation techniques are explained in Section 4.3, and the performance comparison between the proposed and the state-of-the-art compensation techniques for speaker variability follow in Sections 4.4 and 4.5. The summarisation is given in Section 4.6.

4.2 Proposed compensation techniques

4.2.1 Factor analysis based normalisation

The proposed speaker normalisation technique views speaker identity as an underlying factor that affects speech features within a factor analysis framework. Specifically, it assumes features extracted from speech are comprised of a common vector, a speaker identity component and a residual vector that contains mainly emotion-related features as given below,

$$\boldsymbol{x}_{ij} = \boldsymbol{u} + \boldsymbol{F}\boldsymbol{y}_i + \boldsymbol{\varepsilon}_{ij} \tag{4.1}$$

where $\boldsymbol{x}_{ij}$ represents the feature vector estimated from the j^{th} frame of speech from the i^{th} speaker, u is the independent mean over all speakers, $\boldsymbol{y}_i$ is the vector of speaker factors, $\boldsymbol{F}$ is the factor loading

matrix that captures the speaker variability, and $\boldsymbol{\varepsilon}_{ij}$ is the residual component that contains emotion specific information. This is mathematically similar to the PLDA model [21].

Speaker normalisation is then accomplished by subtracting the speaker identity component $\boldsymbol{F}\boldsymbol{y}_i$ from the raw features $\boldsymbol{x}_{ij}$ to give the normalised features $\widetilde{\boldsymbol{x}}_{ij}$, so that

$$\widetilde{\boldsymbol{x}}_{ij} = \boldsymbol{x}_{ij} - \boldsymbol{F}\boldsymbol{y}_i - \boldsymbol{u} \tag{4.2}$$

Note that it might not be an ideal solution to linearly separate emotion-specific and speaker-specific information in feature space. However, based on the model assumption, we conducted an initial analysis by assuming speaker information as an additive factor to the emotion content.

In this model, the speaker factors $\boldsymbol{y}_i$ are assumed to follow a standard normal distribution and the residuals $\boldsymbol{\varepsilon}_{ij}$ are assumed to follow a zero-mean normal distribution with a covariance $\boldsymbol{\Sigma}$, i.e.

$$\boldsymbol{y}_i \sim \mathcal{N}(0, I) \tag{4.3}$$

$$\boldsymbol{\varepsilon}_{ij} \sim \mathcal{N}(0, \boldsymbol{\Sigma}) \tag{4.4}$$

Parameters $\boldsymbol{\theta} = [\boldsymbol{u}, \boldsymbol{F}, \boldsymbol{\Sigma}]$ and $\boldsymbol{y}_i$ of the model should be estimated during the training phase using training data from all speakers. The training procedure is discussed in Section 4.2.1.1 which is identical to that given in [133], and the test procedure is presented in Section 4.2.1.2.

4.2.1.1 Model Parameter Estimation

Let $\boldsymbol{X}_i = [{\boldsymbol{x}_{i1}}^T, {\boldsymbol{x}_{i2}}^T \cdots {\boldsymbol{x}_{iM_i}}^T]^T$ represent the concatenated statistical features of the i^{th} speaker of $M_i \times D$ dimensions, where N is the number of speakers, M_i represents the number of frame of feature vectors from the i^{th} speaker and D represents the feature dimension. In the training phase, the aim is to find the optimal parameter set $\boldsymbol{\theta}$ that maximises the model likelihood of $P(\boldsymbol{X}|\boldsymbol{\theta})$, given some training data $\boldsymbol{X} = [{\boldsymbol{X}_1}^T, {\boldsymbol{X}_2}^T \cdots, {\boldsymbol{X}_N}^T]^T$. Here, the EM algorithm as mentioned in Section 3.2.3 is used to solve the problem as follows:

Firstly, equation (4.1) can be rewritten as:

$$\boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{u} \\ \vdots \\ \boldsymbol{u} \end{bmatrix} + \begin{bmatrix} \boldsymbol{F} \\ \vdots \\ \boldsymbol{F} \end{bmatrix} \boldsymbol{y}_i + \begin{bmatrix} \boldsymbol{\varepsilon}_{i1} \\ \vdots \\ \boldsymbol{\varepsilon}_{iM_i} \end{bmatrix} \tag{4.5}$$

At this point, it is helpful to introduce the notation, $\boldsymbol{A} = [\boldsymbol{F}^T, \boldsymbol{F}^T, \dots, \boldsymbol{F}^T]^T$, $\boldsymbol{m} = [\boldsymbol{u}^T, \boldsymbol{u}^T, \dots, \boldsymbol{u}^T]^T$and $\boldsymbol{\varepsilon}_i = \left[\boldsymbol{\varepsilon}_{i1}^T, \boldsymbol{\varepsilon}_{i2}^T, \dots, \boldsymbol{\varepsilon}_{iM_i}^T\right]^T$, where $\boldsymbol{\varepsilon}_i$ is of mean zero and covariance matrix $\boldsymbol{\Sigma'}$ as shown in equation (4.6).

$$\boldsymbol{\Sigma'} = \begin{bmatrix} \boldsymbol{\Sigma} & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Sigma} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Sigma} \end{bmatrix} \tag{4.6}$$

For each speaker i, the posterior of the speaker identity $\boldsymbol{y}_i$can be rewritten using Bayes theorem as:

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\theta}) = \frac{P(\boldsymbol{X}_i|\boldsymbol{y}_i, \boldsymbol{\theta})P(\boldsymbol{y}_i)}{P(\boldsymbol{X}_i)} \tag{4.7}$$

Since $P(\boldsymbol{X}_i)$ is consistent, the $P(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\theta})$ is proportional as:

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\theta}) \propto P(\boldsymbol{X}_i|\boldsymbol{y}_i, \boldsymbol{\theta})P(\boldsymbol{y}_i) \tag{4.8}$$

where, the posterior probability $P(X_i|y_i, \theta)$ is a Gaussian distribution as below:

$$P(\boldsymbol{X}_i|\boldsymbol{y}_i, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m} + \boldsymbol{A}\boldsymbol{y}_i, \boldsymbol{\Sigma'}) \tag{4.9}$$

Since $P(\boldsymbol{X}_i|\boldsymbol{y}_i, \boldsymbol{\theta})$ and $P(\boldsymbol{y}_i)$ in equation (4.8) are both Gaussian distributions, the posterior distribution $P(\boldsymbol{y}_i|\boldsymbol{X}_i, \theta)$ is also a Gaussian distribution given by

$$P(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\theta}) = \mathcal{N}(E[\boldsymbol{y}_i], cov(\boldsymbol{y}_i)) \tag{4.10}$$

where

$$E[\boldsymbol{y}_i] = (\boldsymbol{A}^T\boldsymbol{\Sigma'^{-1}}\boldsymbol{A} + \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{\Sigma'^{-1}}(\boldsymbol{X}_i - \boldsymbol{m}) \tag{4.11}$$

and

$$cov[\boldsymbol{y}_i] = (\boldsymbol{A}^T\boldsymbol{\Sigma'^{-1}}\boldsymbol{A} + \boldsymbol{I})^{-1} \tag{4.12}$$

The model parameters, $\boldsymbol{\theta} = [\boldsymbol{u}, \boldsymbol{F}, \boldsymbol{\Sigma}]$, are optimised using EM algorithm which aims to maximise the auxiliary function $Q(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$

$$Q(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t) = \sum_{i=1}^{N}\sum_{j=1}^{M_i}\int P(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{\theta}_{t-1})\log[\,P(\boldsymbol{X}_{ij}|\boldsymbol{y}_i, \boldsymbol{\theta}_t)P(\boldsymbol{y}_i)]d\boldsymbol{y}_i \tag{4.13}$$

where t indicates the iteration number.

The updated parameters $\boldsymbol{\theta}$ can be obtained by calculating the derivatives of $Q(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t)$ and are given as:

$$\boldsymbol{u} = \frac{1}{N \cdot M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} \boldsymbol{x}_{ij} \tag{4.14}$$

$$\boldsymbol{F} = \left(\sum_{i=1}^{N} \sum_{j=1}^{M_i} (\boldsymbol{x}_{ij} - \boldsymbol{u}) E[\boldsymbol{y}_i]^T \right) \left(\sum_{i=1}^{N} E[\boldsymbol{y}_i \boldsymbol{y}_i^T] \right)^{-1} \tag{4.15}$$

$$\boldsymbol{\Sigma} = \frac{1}{N * M_i} \sum_{i=1}^{N} \sum_{j=1}^{M_i} diag \left[(\boldsymbol{x}_{ij} - \boldsymbol{u}) \left(\boldsymbol{x}_{ij} - \boldsymbol{u} \right)^T - \boldsymbol{F} E[\boldsymbol{y}_i] \left(\boldsymbol{x}_{ij} - \boldsymbol{u} \right)^T \right] \tag{4.16}$$

4.2.1.1 Speaker Normalisation for Test Utterances

During the testing phase, the speaker factors $\boldsymbol{y}_t$ are estimated from $P(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{\theta})$, where $\boldsymbol{z}_t$ represents the test data. As the posterior probability of test speaker factor $P(\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{\theta})$ is a Gaussian distribution as in equation (4.8), the expectation value $E[\boldsymbol{y}_t | \boldsymbol{z}_t, \boldsymbol{\theta}]$ is adopted as the final estimation for the test speaker factor $\boldsymbol{y}_t$, where $\boldsymbol{A} = \boldsymbol{F}$ and $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$, since the normalisation is carried out at the frame-level. The normalised feature vectors $\tilde{\boldsymbol{z}}_t$ are calculated as given by

$$\tilde{\boldsymbol{z}}_t = \boldsymbol{z}_t - \boldsymbol{F}\boldsymbol{E}[\boldsymbol{y}_t] - \boldsymbol{u} \tag{4.17}$$

where

$$E[\boldsymbol{y_t}] = (\boldsymbol{F}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{F} + \boldsymbol{I})^{-1} \boldsymbol{F}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}_t - \boldsymbol{u}) \tag{4.18}$$

4.2.2 PLSDR based normalisation

4.2.2.1 Conventional PLSDR

PLSDR assumes that a few underlying factors can account for the data variability [134]. It aims to project the original feature space to a lower-dimension latent variable space, which maximises the covariance between the latent factors and the underlying ground truth. Let $\boldsymbol{X} = [\boldsymbol{X}_1{}^T, \boldsymbol{X}_2{}^T \cdots, \boldsymbol{X}_N{}^T]^T$ represents the feature vectors where N denotes the total number of frame. PLSDR decomposes the feature matrix $\boldsymbol{X}$ as:

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} \tag{4.19}$$

where $\boldsymbol{X}$ is an $N \times D_1$ matrix, N represents total number of frames, and D_1 represents the original feature dimensions; $\boldsymbol{T}$ is an $N \times D_2$ matrix comprising of latent components (a low-dimensional representation of $\boldsymbol{X}$) and D_2 denotes the dimensionality of the latent components ($D_2 < D_1$); $\boldsymbol{P}$ is a $D_1 \times D_2$ matrix; and $\boldsymbol{E}$ is an $N \times D_1$ matrix representing residual factors. Conversely, given the data, $\boldsymbol{X}$, the latent components, $\boldsymbol{T} = [\boldsymbol{t}^{1^T}, \boldsymbol{t}^{2^T}, \cdots \boldsymbol{t}^{D_2{}^T}]^T$ can be calculated by estimating each column vector $\boldsymbol{t}^i$ sequentially as:

$$\boldsymbol{t}^i = \boldsymbol{X}\boldsymbol{w}^i \tag{4.20}$$

where $\boldsymbol{w}^i$ ($for\ 1 \leq i \leq D_2$) is a weight vector corresponding to elements of the weight matrix $\boldsymbol{W} = [\boldsymbol{w}^{1^T}, \boldsymbol{w}^{2^T}, \cdots \boldsymbol{w}^{D_2{}^T}]^T$, a $D_2 \times D_1$ matrix that projects the original feature space $\boldsymbol{X}$ to the lower dimensional space $\boldsymbol{T}$. For computational purposes, the label $\boldsymbol{Y}$ is similarly decomposed as for $\boldsymbol{X}$ in equation (4.19) as:

$$\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}^T + \boldsymbol{F} \tag{4.21}$$

where $\boldsymbol{Y}$ is an $N \times D_3$ matrix, D_3 denotes the label dimensionality. $\boldsymbol{U} = [\boldsymbol{u}^{1^T}, \boldsymbol{u}^{2^T}, \cdots \boldsymbol{u}^{D_3{}^T}]^T$ is a $N \times D_4$ projected matrix of $\boldsymbol{Y}$, similar as T to X in (4.19). $\boldsymbol{Q}$ is a $D_4 \times D_3$ matrix, and $\boldsymbol{F}$ is the residual matrix similar as $\boldsymbol{E}$ in (4.19). PLSDR aims to find $\boldsymbol{w}^i$ from equation (4.20) by maximising the covariance between the new projected latent feature vector $\boldsymbol{X}\boldsymbol{w}^i$ and labels $\boldsymbol{u}^i$ as:

$$\begin{aligned} \boldsymbol{w}^i &= \underset{\boldsymbol{w}^i}{argmax} \left(cov(\boldsymbol{X}\boldsymbol{w}^i, \boldsymbol{u}^i) \right) \\ &= \underset{\boldsymbol{w}^i}{argmax} \frac{1}{N-1} \left(\boldsymbol{X}\boldsymbol{w}^i - \overline{\boldsymbol{X}\boldsymbol{w}^i} \right)^T (\boldsymbol{u}^i - \bar{\boldsymbol{u}}^i) \end{aligned} \tag{4.22}$$

where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are first mean-centred. This means that the covariance calculation simplifies to

$$\boldsymbol{w}^i = \underset{\boldsymbol{w}^i}{argmax} \left(\frac{1}{N-1} (\boldsymbol{X}\boldsymbol{w}^i)^T \boldsymbol{u}^i \right) \tag{4.23}$$

A Lagrange multiplier is added to solve the maximisation problems under the constraint $\boldsymbol{w}^{i^T}\boldsymbol{w}^i = 1$. It can be proven that $\boldsymbol{w}^i$ corresponds to the first eigenvector of $\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$ [135], which takes into account the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Finally, $\boldsymbol{W}$ and $\boldsymbol{T}$ are obtained. For further details of the algorithm, the reader is referred to [59, 136-138].

4.2.2.2 Proposed PLSDR based normalisation

Instead of only considering the covariance between the features and labels, a method that aims to project the original features to a latent space that additionally minimises the speaker variability is developed. This method focuses on manipulating the speaker-specific covariance between the features and labels. Firstly, features and labels from each speaker s, denoted here as $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$, are decomposed individually, similarly to equations (4.19) and (4.21), as:

$$\boldsymbol{X}_s = \boldsymbol{T}_s \boldsymbol{P}_s^T + \boldsymbol{E}_s \tag{4.24}$$

$$\boldsymbol{Y}_s = \boldsymbol{U}_s \boldsymbol{Q}_s^T + \boldsymbol{F}_s \tag{4.25}$$

where $\boldsymbol{T}_s = \boldsymbol{X}_s \boldsymbol{W}$. $\boldsymbol{T}_s$, $\boldsymbol{P}_s$, $\boldsymbol{E}_s$, $\boldsymbol{U}_s$, $\boldsymbol{Q}_s$ and $\boldsymbol{F}_s$ all keep the same denotation as conventional PLSDR in 4.2.2.1, but specific to each speaker s here. The projection matrix $\boldsymbol{W}$ maximises the summation of speaker-specific covariance between the new features $\boldsymbol{X}_s \boldsymbol{W}$ and $\boldsymbol{Y}_s$ for each speaker s. It can be estimated as

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{argmax} \left(\sum_{s=1}^{S} cov(\boldsymbol{X}_s \boldsymbol{W}, \boldsymbol{Y}_s) \right) \tag{4.26}$$

This projection matrix introduces the speaker-specific information, and the projected feature space is expected to minimise the speaker variability according to (4.26). Similar to equation (4.23), $\boldsymbol{X}_s$ and $\boldsymbol{Y}_s$ for each speaker s can be mean-centred individually and the simplified criterion can be estimated as

$$\boldsymbol{w}^i = \underset{\boldsymbol{w}^i}{argmax} \sum_{\boldsymbol{s}=1}^{S} \frac{1}{N_s - 1} \left(\boldsymbol{X}_s \boldsymbol{w}^i \right)^T \boldsymbol{u}_s^i \tag{4.27}$$

where N_s represents the total number of frames of each speaker s. The Lagrange multiplier λ is generally introduced to solve equation (4.26) under the constraint $\left\| \boldsymbol{w}^i \right\| = 1$ as:

$$\boldsymbol{L}(\boldsymbol{w}^i) = \sum_{\boldsymbol{n}=1}^{s_n} \frac{\left(\boldsymbol{X}_s \boldsymbol{w}^i \right)^T \boldsymbol{u}_s^i}{N_s - 1} - \lambda (\boldsymbol{w}^{iT} \boldsymbol{w}^i - 1) \tag{4.28}$$

By taking the derivative of this equation with respect to $\boldsymbol{w}^i$, it comes to:

$$\frac{\delta \boldsymbol{L}(\boldsymbol{w}^i)}{\delta \boldsymbol{w}^i} = \sum_{s=1}^{S} \frac{\boldsymbol{X}_s{}^T \boldsymbol{u}_s^i}{N_s - 1} - 2\lambda \boldsymbol{w}^i = 0 \tag{4.29}$$

The PLS2 algorithm [59, 136-138] is used to estimate the parameters $\boldsymbol{W}$, and $\boldsymbol{\theta} = [\boldsymbol{T}_s, \boldsymbol{P}_s, \boldsymbol{Q}_s, \boldsymbol{U}_s]$ for the proposed method that can accommodate multi-dimensional labels $\boldsymbol{Y}$. In our case, $\boldsymbol{Y}$ is a matrix of either two or three columns representing the affective attributes of interest. The use of multi-dimensional labels, as opposed to employing a different decomposition for each affective attribute, aims to take the dependencies between the affective attributes into account.

The algorithm is developed by estimating the column vectors of $\boldsymbol{T}_s$ and $\boldsymbol{W}$ sequentially. Let $[\boldsymbol{t}_s^i, \boldsymbol{p}_s^i, \boldsymbol{q}_s^i, \boldsymbol{u}_s^i]$ represent the i^{th} column vectors of matrices $[\boldsymbol{T}_s, \boldsymbol{P}_s, \boldsymbol{Q}_s, \boldsymbol{U}_s]$ for each speaker s respectively. Based on equation (4.29), the i^{th} column vector $\boldsymbol{w}^i$ of weight matrix $\boldsymbol{W}$ can be calculated under the constraint $\|\boldsymbol{w}\| = 1$ as

$$\boldsymbol{w}^i = \sum_{s=1}^{S} \frac{\boldsymbol{X}_s^{i\,T} \boldsymbol{u}_s^i}{N_s - 1} \Bigg/ \left\| \sum_{s=1}^{S} \frac{\boldsymbol{X}_s^{i\,T} \boldsymbol{u}_s^i}{N_s - 1} \right\| \tag{4.30}$$

where N_s represents the total frame number of speaker s. $\boldsymbol{X}_s^i$ denotes the residual information of the original feature vectors of speaker s, since the information in $\boldsymbol{X}_s$ will be reduced after each column estimation and $\boldsymbol{w}^i$ is then computed using the information left as in equation (4.35). $\boldsymbol{u}_s^i$ is the i^{th} column vector of $\boldsymbol{U}_s$. Note that the first column vector u_s^1 should be initialised for estimation of $\boldsymbol{w}^1$ in (4.30), which generally takes the first column of $\boldsymbol{Y}_s$ as in [59, 136-138].

Then the i^{th} column vectors $[\boldsymbol{t}_s^i, \boldsymbol{p}_s^i, \boldsymbol{q}_s^i, \boldsymbol{u}_s^i]$ can then be estimated and updated iteratively until they converge, the details of which can be found in [139]. These parameters are:

$$\boldsymbol{t}_s^i = \boldsymbol{X}_s^i \boldsymbol{w}^i \tag{4.31}$$

$$\boldsymbol{q}_s^i = \boldsymbol{Y}_s^{i\,T} \boldsymbol{t}_s^i \Big/ \left\| \boldsymbol{Y}_s^{i\,T} \boldsymbol{t}_s^i \right\| \tag{4.32}$$

$$\boldsymbol{p}_s^i = \boldsymbol{X}_s^{i\,T} \boldsymbol{t}_s^i / \boldsymbol{t}_s^{i\,T} \boldsymbol{t}_s^i \tag{4.33}$$

$$\boldsymbol{u}_s^i = \boldsymbol{Y}_s^i \boldsymbol{q}_s^i \tag{4.34}$$

It can be seen that $\boldsymbol{u}_s^i$ was updated as a linear combination of the columns of $\boldsymbol{Y}_s^i$ with weights $\boldsymbol{q}_s^i$ as in equation (4.34). This allows information about multiple affective attributes from $\boldsymbol{Y}_s$ to be introduced into the estimation. $\boldsymbol{X}_s^i$ and $\boldsymbol{Y}_s^i$ represent the residual information as:

$$\boldsymbol{X}_s^i = \boldsymbol{X}_s^{i-1} - \boldsymbol{t}_s^{i-1}\boldsymbol{p}_s^{i-1^T} \tag{4.35}$$

$$\boldsymbol{Y}_s^i = \boldsymbol{Y}_s^{i-1} - \boldsymbol{u}_s^{i-1}\boldsymbol{q}_s^{i-1^T} \tag{4.36}$$

where $1 \leq i \leq D_2$. Thus every i^{th} column vector of $\boldsymbol{W}$ and $[\boldsymbol{T}_s, \boldsymbol{P}_s, \boldsymbol{Q}_s, \boldsymbol{U}_s]$ can be estimated sequentially, until a predefined number of columns (dimensions), D_2, are achieved.

During the estimation of each column vectors $[\boldsymbol{t}_s^i, \boldsymbol{p}_s^i, \boldsymbol{q}_s^i, \boldsymbol{u}_s^i]$ in equations (4.31)-(4.34), they are also optimised iteratively. The estimation of the optimal column vectors $\boldsymbol{w}^i$ and $[\boldsymbol{t}_s^i, \boldsymbol{p}_s^i, \boldsymbol{q}_s^i, \boldsymbol{u}_s^i]$ continues until they reach the converge condition, $\sum_{s=1}^{S}(\boldsymbol{u}_s^j - \boldsymbol{u}_s^{j-1}) \leq \varepsilon$, where j is the iteration number for the optimisation and ε is a predefined threshold.

The normalised features $\boldsymbol{T}_s$ can be obtained individually as per equation (4.31) for each speaker, and the regression model for emotion prediction is developed based on $\boldsymbol{T}_s$ over the entire training partition. During the testing phase, features $\boldsymbol{X}_t$ extracted form test speech (as opposed to features $\boldsymbol{X}_s$ from training speech from speaker s) are mean-centred and the trained weight matrix $\boldsymbol{W}$ is used to project the test features to the normalised feature space to obtain $\boldsymbol{T}_t$ as follows:

$$\boldsymbol{T}_t = \boldsymbol{X}_t\boldsymbol{W} \tag{4.37}$$

The regression model is trained using normalised features $\boldsymbol{T}_s$, and predictions are made based on each normalised new test features $\boldsymbol{T}_t$.

The proposed PLSDR based speaker normalisation aims to achieving both feature dimensionality reduction and speaker normalisation simultaneously. Since the proposed technique is applied directly on the features prior to model training, it can be easily used with other types of models and in other regression problems in other fields.

4.2.3 Feature mapping based normalisation

This section introduces a speaker normalisation method based on feature mapping [140] that directly transforms the feature space to reduce mismatch between marginal feature distributions $P(\boldsymbol{x}|s)$ for different speakers, which were originally laid out in Section 3.2.

Feature mapping was first introduced within the field of speaker verification for a system that aimed to minimise the channel variability from model space [127]. It mapped the channel-dependent

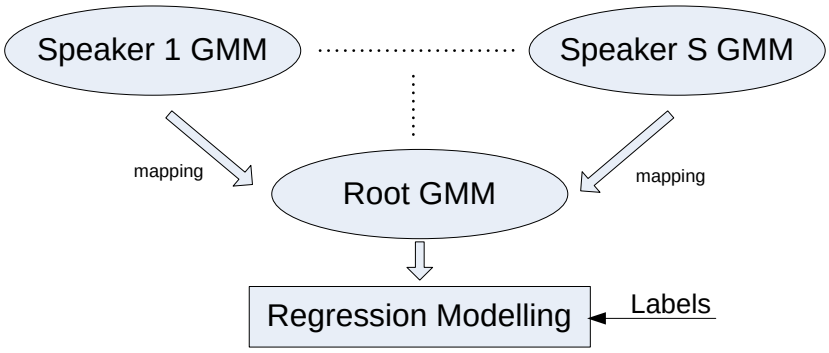


Figure 4.1: Feature mapping structure. The speaker-dependent GMMs are mapped to a root GMM (speaker-independent GMM). The mapping rule is speaker-dependent. The normalised features are used for the regression modelling techniques.

feature space to a channel-independent space, and developed speaker-dependent mapping rules. The technique proposed in this section aims to reduce the speaker variability in continuous emotion prediction systems by adapting the speaker-dependent GMMs of the feature space to a speaker-independent GMM, where the root GMM can be viewed as a speaker-independent model as shown in Figure 4.1.

In this chapter, the root GMM is chosen to be a UBM trained on data from multiple speakers in order to make it speaker-independent. During the test phase, the distribution of features of the test speaker $P(\boldsymbol{x}|t)$ is similarly mapped to the UBM $P(\boldsymbol{x})$. It should be noted that the root GMM does not have to be a UBM and a speaker adaptation scheme can be envisioned. In this way, a model of the test speaker's feature distribution is chosen as the root GMM [141]. However, there appears to be inadequate data from the target test speakers in the databases employed herein to estimate a reliable model for use as the root GMM, and the performance of the speaker adaptation approach was poor. Henceforth, the main focus is only the speaker normalisation approach in this section, a description of which follows.

Firstly, speaker-dependent GMMs of $P(\boldsymbol{x}|s)$ are developed in the feature space. Then speaker-dependent mapping rules are developed by comparing the speaker-dependent GMMs and the UBM individually. This is achieved by a linear affine transformation, shifting the mean and adapting the covariance matrix of the speaker-dependent GMMs towards that of the UBM. The speaker-dependent mapping is carried out by the transformation obtained for each Gaussian mixture component instead of for the overall distribution of the GMM. After mapping, the normalised joint model of $P(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ is

developed using normalised features $\tilde{\boldsymbol{x}}$ from all speakers.

In terms of the mapping principle for each speaker *s*, the dominant mixture *k* for the feature vector at frame i was first calculated as:

$$k = arg \max_{1 \le m \le M} w_m P_m(\boldsymbol{x}_i|s) \tag{4.38}$$

where

$$P_m(\boldsymbol{x_i}|s) = N\left(\boldsymbol{x}_i; \boldsymbol{u}_{sm}^{\boldsymbol{x}}, \boldsymbol{\Sigma}_{sm}^{(xx)}\right), \tag{4.39}$$

$\boldsymbol{x_i}$ represents the feature vector at frame i and $P_m(\boldsymbol{x_i}|s)$ represents the probability of feature vector $\boldsymbol{x_i}$ belonging to m^{th} mixture component of the speaker-dependent GMM $P(\boldsymbol{x}|s)$. $\boldsymbol{u}_{sm}^{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{sm}^{(xx)}$ represent the mean and covariance matrix of the m^{th} mixture component of the speaker-dependent GMMs for speaker *s*. The dominant mixture component modelled by $P_k(\boldsymbol{x_i}|s)$, to which $\boldsymbol{x}_i$ is most likely to belong, is determined for each speaker s as in equation (4.38). Then the feature vector $\boldsymbol{x}_i$ is mapped to a speaker-independent space, by following the affine transformation from this dominant mixture component k to the corresponding k^{th}mixture component of the UBM from a distribution view. It should be noted that covariance matrices $\boldsymbol{\Sigma}_{k}^{(xx)}$and $\boldsymbol{\Sigma}_{sk}^{(xx)}$of k^{th} mixture of the UBM and speaker-dependent GMM are full covariance matrices, though the approach presented in [127] only dealt with diagonal covariance matrices.

When using full covariance matrices for each mixture component, the mapping of $\boldsymbol{x}_i$ to $\tilde{\boldsymbol{x}}_i$ for speaker s is given as:

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{u}_{sk}^{\boldsymbol{x}}) + \boldsymbol{u}_k^{\boldsymbol{x}} \tag{4.40}$$

where $\boldsymbol{W}$ represents the transformation matrix for scaling and rotation, in order to match the speaker-dependent mean and covariance $\boldsymbol{\Sigma}_{sk}^{(xx)}$ to the speaker-independent mean and covariance $\boldsymbol{\Sigma}_{m}^{(xx)}$. The means $\boldsymbol{u}_{sk}^{\boldsymbol{x}}$ and $\boldsymbol{u}_k^{\boldsymbol{x}}$ of the k^{th}dominant mixture for speaker dependent GMM and UBM can be directly found from models of the joint distributions $P(\boldsymbol{x}, \boldsymbol{y}|s)$ and $P(\boldsymbol{x}, \boldsymbol{y})$ respectively. To estimate $\boldsymbol{W}$, first the covariance matrices corresponding to both sides of equation (4.40) are found as:

$$cov(\tilde{\boldsymbol{x}}_i) = cov(\boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{u}_{sk}^{\boldsymbol{x}}) + \boldsymbol{u}_k^{\boldsymbol{x}}) \tag{4.41}$$

$$= cov(\boldsymbol{W}\boldsymbol{x}_i)$$

$$\therefore cov(\tilde{\boldsymbol{x}}_i) = \boldsymbol{W}cov(\boldsymbol{x}_i)\boldsymbol{W}^T$$

Since $\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} = cov(\tilde{\boldsymbol{x}}_i)$ and $\boldsymbol{\Sigma}_{sm}^{(\boldsymbol{xx})} = cov(\boldsymbol{x}_i)$ are already obtained in the joint model $P(\boldsymbol{x}, \boldsymbol{y}|s)$ and $P(\boldsymbol{x}, \boldsymbol{y})$ respectively, equation (4.41) can be rewritten as:

$$\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} = \boldsymbol{W}\boldsymbol{\Sigma}_{sm}^{(\boldsymbol{xx})}\boldsymbol{W}^{\boldsymbol{T}} \tag{4.42}$$

Cholesky decomposition can be used to further decompose the covariance matrices as:

$$\boldsymbol{\Sigma}_{sm}^{(\boldsymbol{xx})} = \boldsymbol{Q}\boldsymbol{Q}^{\boldsymbol{T}} \tag{4.43}$$

$$\boldsymbol{\Sigma}_m^{(\boldsymbol{xx})} = \boldsymbol{R}\boldsymbol{R}^{\boldsymbol{T}} \tag{4.44}$$

Substituting equations (4.43) and (4.44) into (4.42), it is noted that

$$\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{W}\boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{W}^T = \boldsymbol{W}\boldsymbol{Q}(\boldsymbol{W}\boldsymbol{Q})^T \tag{4.45}$$

The transformation matrix $\boldsymbol{W}$ is finally obtained as

$$\boldsymbol{W} = \boldsymbol{R}\boldsymbol{Q}^{-1} = \boldsymbol{R}\boldsymbol{Q}^T \tag{4.46}$$

The mapping is performed on a vector basis. The normalised joint model $P(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ can be developed using normalised features, and predictions are made based on $P(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ and the normalised test features.

One of the obstacles to feature mapping is that it only maps features by shifting the mean and transforming the covariance, not taking the weights of each mixture into consideration. It can therefore reduce the variability between speakers, but may not map speaker specific feature distributions $P(\boldsymbol{x}|s)$ to a common distribution, since the target models could differ in terms of their mixture weights.

4.3 Experimental settings

In this section the experimental settings are seperated into two parts: one for factor analysis based normalisation technique, and the other for the PLSDR and feature mapping based normalisation techniques, since different back-ends were utilised for each. Factor analysis based normalisation was

carried out in the original high dimensional feature space, thus RVM was used as the back-end due to its inherent feature selection property. GMR was utilised as the regression modelling technique for the PLSDR and feature mapping based normalisation techniques, aligned with the analysis in Chapter 3.

4.3.1 Factor analysis based normalisation

The USC CreativeIT database [107] and the SEMAINE database [142] (referring back to Section 2.4.2) were utilised to evaluate the factor analysis based normalisation method. The same feature set described in Section 3.3.1 was utilised and RVM was adopted as the back-end. The proposed speaker normalisation was applied with 13-dimensional and 12-dimensional speaker factor vectors with the USC CreativeIT and the SEMAINE databases respectively. The dimensionality of the speaker factor vectors $\boldsymbol{y}_i$ was chosen based on the number of speakers in the training dataset. The speaker normalisation model parameters were estimated with 10 iterations of the EM algorithms in both cases. Person's correlation coefficient is adopted as the evaluation metric.

The experiments on the USC CreativeIT database were conducted in a leave-one-session-out cross validation manner to avoid the model starvation owing to the limited size of the database[81]. The SEMAINE database on the other hand was split into a distinct training set comprising of speech data from 12 randomly selected speakers and a distinct test comprising of speech from the remaining 6 speakers. Apart from the system performance, analyses of the compensation effect of the proposed technique are also reported on the USC CreativeIT database in Section 4.5.1.

4.3.2 PLSDR and feature mapping based normalisation

Three databases were used to evaluate the proposed PLSDR and feature mapping based normalisation methods: the USC CreativeIT database [107], the SEMAINE database [142] and the RECOLA database in Section 2.3.2, which featured in AVEC 2016 [18]. Similarly to Section 4.4.1, further analyses of the proposed techniques are carried out on the USC CreativeIT database for a direct comparison in Section 4.5.2.

An overview of the system is shown in Figure 4.2. Statistical features are extracted and GMR [81] is used as the regression technique. All the parameters are identical to those described in Section 3.3.

PLSDR and feature mapping based speaker normalisation, shown as orange blocks in Figure 4.2, are two proposed compensation techniques for speaker variability and are evaluated independently here.

4.3.2.1 Front-end

In order to ensure that the systems are compared fairly, the feature dimension after PLSDR is set to 80, to be consistent with the system described in Section 3.3. It should be noted that the PLSDR based method carries out feature normalisation after incorporating dynamic information, while the baseline approach with PCA reduces the dimension to 40 and then incorporates the dynamic information (in the form of deltas) since this was found to exhibit better performance [81].

4.3.2.2 Back-end

GMR allows for the possibility of jointly modelling all the affective dimension of interest, i.e. where arousal, valence and dominance are jointly predicted. In this case, the label vector $\boldsymbol{Y}_s$ used for GMR is a six-dimension vector comprising of the three affective attributes (arousal, valence and dominance) and their temporal derivatives. Inclusion of the derivatives takes the dependencies among emotion attributes into account as well. For details about GMR the reader is referred to [81, 83].

GMMs with four mixture components and full covariance matrices are used. Twenty iterations of the EM algorithm are used to train the GMMs. Finally, post-processing of the prediction is implemented using a binomial filter to smooth them, and then mean and variance normalisation for scaling purposes [29].

4.3.2.3 Evaluation metrics

All systems evaluated on a database used identical training, development and test partitions to allow direct comparison of accuracy on that database. The experiments on the USC CreativeIT database were conducted in a leave-one-session-out cross validation manner [81]. The SEMAINE database was split into a distinct training set comprising of speech data from 12 randomly selected speakers and a distinct test set of speech from the remaining 6 speakers with 28 utterances. In the RECOLA database, the experiments were carried out using the training set and the development set (serving as test set) as partitioned by AVEC 2016 [92].

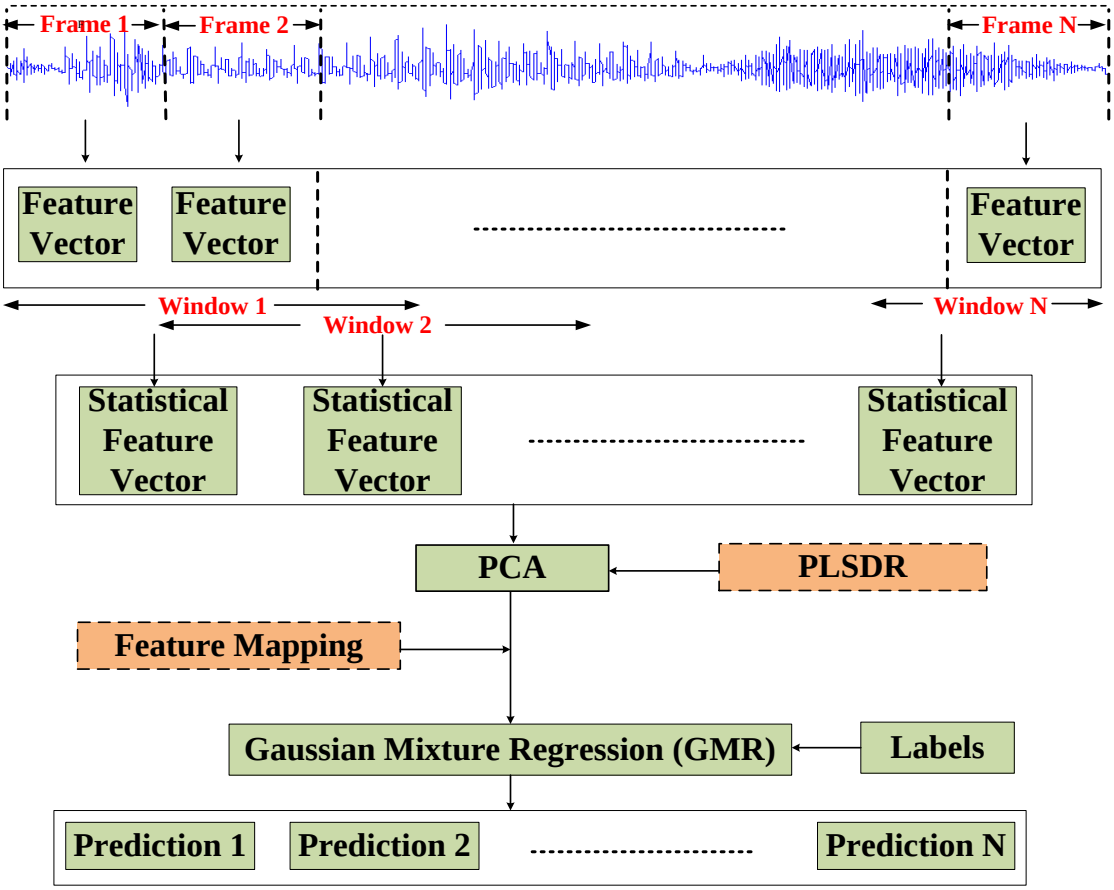


Figure 4.2: System overview of speech-based emotion prediction systems with proposed compensation techniques for speaker variability. Orange boxes indicate the proposed techniques in the system flow, either of the proposed technique is utilised in the system.

The performance of these continuous emotion prediction systems was evaluated using Pearson's correlation coefficients averaged across the test utterances as the performance metric. In addition, inter-speaker variability and intra-speaker variability, as outlined in Section 3.2.2, were also compared before and after the compensation techniques on the USC CreativeIT database.

4.4 Experimental results

This section compares the effect of speaker variability before and after the proposed speaker normalisation techniques using the CreativeIT database. The performance of emotion prediction systems, employing the proposed techniques on the three databases, are reported and compared with the state-of-the-art techniques. The experimental results are separated into two sections: factor analysis based normalisation, and the PLSDR and feature mapping based normalisation.

4.4.1 Factor analysis based normalisation

4.4.1.1 Analysis of the speaker variability

In order to determine how speaker variability negatively affects the performance of continuous emotion prediction systems, the performance of a speaker independent emotion prediction system was compared to that of speaker-specific emotion prediction systems on the USC CreativeIT database. Speaker-specific emotion prediction systems refer to those that are trained and tested on data from the same speaker. For this experiment, speaker-specific systems were trained on 2/3 of the data and tested on the remaining 1/3 of the data from 14 of the 16 speakers in the database, as there was insufficient data from the remaining two speakers to train and test a speaker-specific system. The performance of the speaker independent system is estimated on data from all 8 sessions in the database in a leave-one-session-out cross-fold validation. Both systems use only voiced speech for training and no feature normalisation.

The results of the experiment are shown in Figure 4.3, where the performance of the 14 speaker-specific system as well as the average speaker-specific performance is compared to the performance of the speaker independent system in terms of mean correlation coefficient between predicted attribute values and ground truth labels based on human annotators (included in the database).The consistently superior performance of the speaker-specific systems suggests that speaker variability degrades the performance of speech based continuous emotion prediction systems.

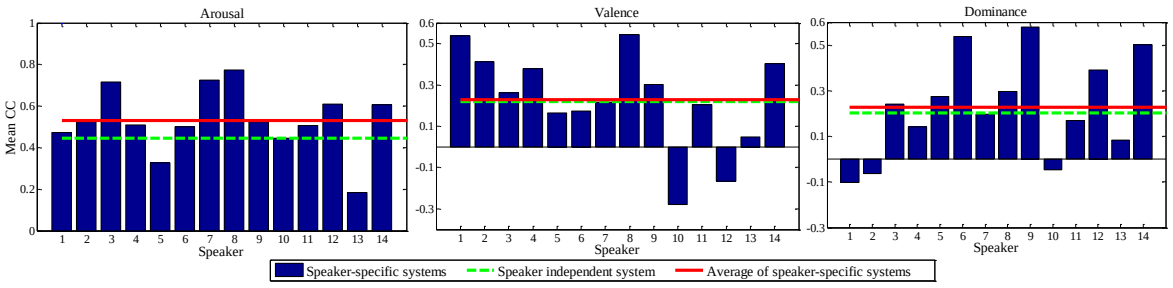


Figure 4.3: Speaker independent vs. speaker-specific systems - correlation coefficient evaluated on the USC CreativeIT database. Blue bars represent the speaker-specific systems, the red line represents the average of speaker-specific systems, and the dotted green line represents the speaker independent system.

In addition to this, the F-ratio is used as a measure of dissimilarity between speaker classes in order to investigate the effect of the normalisation on raw features [143]. This is the ratio of inter-class variability over intra-class variability given by:

$$F-ratio = \frac{\frac{1}{N}\sum_{i=1}^{N}(u_i - u)}{\frac{1}{N \cdot M_i}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(x_{ij} - u_i)} \tag{4.47}$$

where u_i represents the mean of features estimated from the i^{th} speaker and other notations are the same as in Section 4.2.1.

In this experiment, each speaker is treated as a distinct class and adopt the average F-ratio of speaker classes as a measure of feature dissimilarity between speakers per feature dimension. A larger F-ratio value indicates a more separated feature and therefore greater speaker variability. F-ratios of the first 50 dimensions (out of 650) of the un-normalised feature vector are compared to the F-ratios of corresponding 50 dimensions of the normalised feature vector in Figure 4.3. In addition the F-ratios of the same 50 dimensions of the speaker identity component ($\boldsymbol{FE}[\boldsymbol{y}_i]$) as in equation (4.1) are also shown in Figure 4.4. From this figure it can be seen that consistently the largest F-ratios correspond to the speaker identity component and the smallest F-ratios to the normalised feature vectors which suggests that the proposed speaker normalisation method is operating as expected and is able to decompose the feature space into a speaker subspace and a residual subspace, which includes emotion information.

4.4.1.2 Performance with factor analysis based speaker normalisation

The final validation of the proposed speaker normalisation technique was carried out on both the USC CreativeIT database and the SEMAINE database by comparing the performances of the basic emotion prediction systems with and without speaker normalisation.

The performance is shown in Table 4.1. It can be seen the proposed speaker normalisation consistently improves the relative performance of emotion prediction systems that use voiced speech on both databases by 8.2%, 11.7% and 7% in USC CreativeIT and 11.0%, 95.7% and 1.9% in SEMAINE for arousal, valence and dominance respectively. However, no improvement is shown on

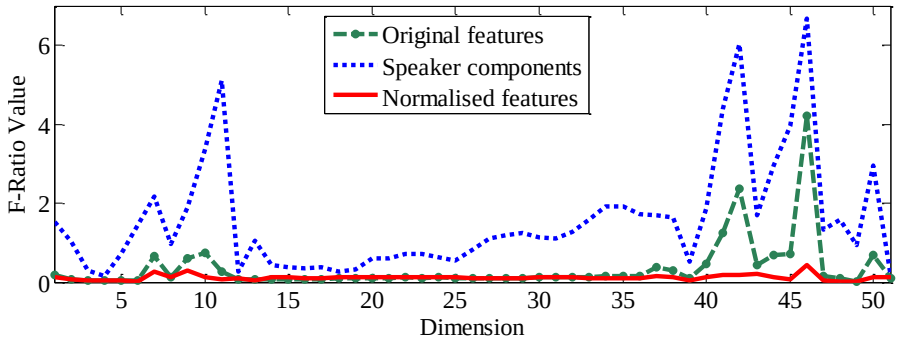


Figure 4.4: F-ratio comparison among original features, speaker component and normalised features

Table 4.1: Performance on two databases.

A means arousal, V means valence and D means dominance.

		Mean Correlation Coefficient(CC)		
		A	**V**	**D**
USC CreativeIT Database	Model with VAD	0.447	0.220	0.201
	Model with VAD + Normalisation	***0.483***	***0.246***	***0.215***
	Model without VAD	0.527	0.238	0.237
	Model without VAD +Normalisation	0.526	0.231	0.220
SEMAINE Database	Model with VAD	0.453	0.106	0.623
	Model with VAD + Normalisation	***0.503***	***0.208***	***0.635***
	Model without VAD	0.429	0.116	0.611
	Model without VAD +Normalisation	***0.521***	***0.211***	***0.643***

the system that uses all frames in USC CreativeIT database, possibly owing to the lack of voiced speech in USC CreativeIT database. It also doesn't show a large difference between the two systems with and without VAD on the SEMAINE database.

4.4.2 PLSDR and feature mapping based normalisation

This section analyses the PLSDR and feature mapping based compensation techniques. The analyses of KL divergence before and after PLSDR and feature mapping based normalisation are presented in Sections 4.5.2.1 and 4.5.2.2, in terms of marginal and conditional probability distributions (as discussed in Section 3.2.2). The experimental results conducted on the three databases from the previous section are reported in Section 4.5.2.3.

4.4.2.1 *Marginal probability distribution*

The average KL divergence between the feature distributions of the 16 speakers and all other speakers $S_x(i)$ (as defined in Section 3.2.2) after the application of the two proposed compensation techniques are shown in Figure 4.5 and compared to the KL divergences prior to speaker normalisation.

It can be seen from the results that the proposed PLSDR based speaker normalisation has the greatest effect on the features from speakers who are the furthest from the other speakers (speakers 3 and 11 in Figure 4.5) and brings them in line with the other speakers. On the other hand, feature mapping based speaker normalisation reduces the differences between feature distributions across all speakers, but in the case of speakers whose distribution was significantly different in the first place (speaker 11), the difference after normalisation is still fairly high unlike in the case of PLSDR. Finally, it should also be noted that the variability in the average KL divergence between speakers for each speaker $S_x(i)$ after normalisation was lower for PLSDR (standard deviation of $S_x(i)$ was 3) compared to feature mapping (standard deviation of $S_x(i)$ was 5.7).

The slopes of linear fits to the PAV profiles discussed in Section 3.2.2 before and after compensation were also estimated as measures of the effect of the proposed speaker normalisation methods on the 'widths' of feature distributions and reported in Table 4.2. As in Section 3.2.2, the

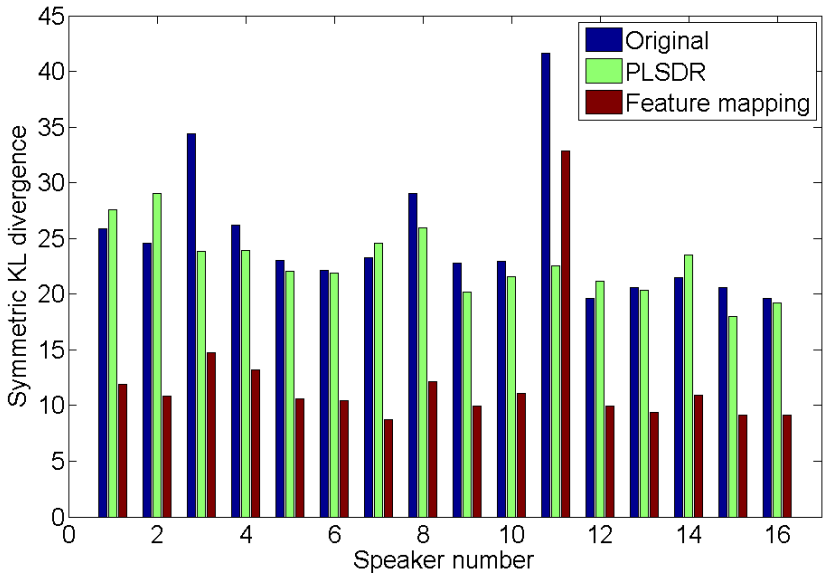


Figure 4.5: Symmetric KL divergence of marginal probability P(x|s) before and after compensation.

Table 4.2: Slope of PAV profiles for marginal distributions

	Speaker-independent	Speaker-dependent		
		Mean	Median	Standard deviation
Original	-2.79	-2.95	-2.91	0.26
PLSDR	-3.54	-3.52	-3.52	0.18
Mapping	-2.64	-2.78	-2.79	0.14

slope of speaker 1 (an outlier) is not included in the calculation of the mean or the standard deviation. These results show that in the original feature space (prior to any speaker normalisation), the speaker independent feature distribution was significantly 'broader' than the speaker specific distributions. However, after speaker normalisation using either proposed method, the 'widths' of the speaker independent and speaker dependent feature distributions were much more similar, suggesting that the differences between speakers were reduced. It is also interesting to note that the PLSDR based normalisation method reduced the variability in the feature space, namely the 'widths' of the speaker dependent and speaker independent distributions, compared to the un-normalised feature space. This is indicated by a smaller mean value of slopes of speaker-dependent models compared to the slope value of speaker-independent model. However, the feature mapping based approach increased the variability (the widths) in the feature space slightly.

4.4.2.2 Conditional probability distribution

As in Section 4.5.2.1, the analysis carried out on the marginal feature distributions is repeated, on the conditional distributions of the features $\boldsymbol{x}$ given the affect labels $\boldsymbol{y}$ as well by approximating the true conditional distribution $P(\boldsymbol{x}|\boldsymbol{y}, s)$, by the label normalised joint distribution, $P(\boldsymbol{x}, \tilde{\boldsymbol{y}}|s)$ (refer to Section 3.2.2). These results are reported in Figure 4.6 and Table 4.3 and broadly concur with the results corresponding to the marginal feature distributions in Figure 4.5 and Table 4.2. However, compared to $S_{\boldsymbol{x}}(i)$ shown in Figure 4.6, speakers 3 and 11 do not appear to be significantly different from the other speakers in terms of the conditional feature distributions, and feature mapping appears to perform worse for speaker 11. Finally the effect of the PLSDR based approach on the variability ('widths') of the conditional distributions appears to be greater than that on the marginal distributions while the effect of feature mapping on the variability of the conditional distributions is much more

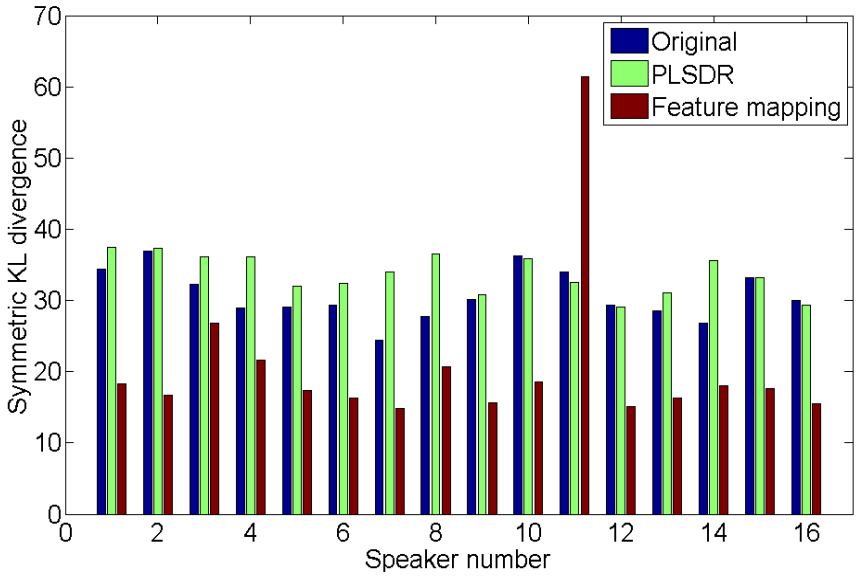


Figure 4.6: Symmetric KL divergence of conditional probability $P(x|y,s)$ before and after compensation.

Table 4.3: Slope of PAV profiles of conditional probability distribution

	Speaker-independent	Speaker-dependent		
		Mean	Median	Standard deviation
Original	-2.85	-2.98	-2.96	0.26
PLSDR	-3.73	-3.80	-3.83	0.24
Mapping	-2.96	-3.10	-3.21	0.22

subdued. On the whole, these results suggest that the PLSDR based normalisation approach may be somewhat superior to the feature mapping based approach when viewed through the lens of their effects on the conditional distributions of the features given affect labels.

4.4.2.3 Performance on three databases

In this section, the performance of GMR based continuous emotion prediction systems using PLSDR (GMR-PLSDR) and feature mapping (GMR-FM) based speaker normalisation techniques on three well-established databases are reported, and compared to the commonly employed speaker-wise Z-normalisation (GMR-Z-N). These systems were also compared to a PCA based system without any speaker normalisation (GMR-PCA). In addition, two additional systems employing speaker normalisation techniques that have been shown to be effective in emotion classification systems, namely feature warping (FW) [122] and iterative feature normalisation (IFN) [144], were also implemented and evaluated on the USC CreativeIT database for comparison. The reported

performances of other recent speaker normalisation techniques, including i-vector [17] and factor analysis (FA) based normalisation [145], are also shown here for reference. The performances of these systems on the three databases are reported in Tables 4.4 to 4.6. The evaluation metric is mean Pearson's correlation coefficient averaged across the utterance. It should be noted that the systems without 'GMR' in their label used different feature sets or back-ends, and this should be taken in to account when comparing performances.

As can be seen from Tables 4.4 to 4.6, the proposed GMR-PLSDR and GMR-FM systems are shaded in grey, and the best performance achieved is highlighted in red. The GMR-PLSDR system outperformed the commonly used speaker wise z-normalisation (GMR-Z-N) and showed consistent improvements on all three databases. Notably, the improvement in valence was more significant than for arousal for all databases. Generally, valence cannot be predicted well using speech alone, consequently the improvements achieved by compensating for speaker variability suggests that large differences may exist between speakers when expressing positive or negative emotions. The improvement shown by GMR-PLSDR also indicates that taking into consideration the mutual information between features and labels during speaker normalisation may be beneficial for valence prediction. The proposed GMR-FM provided a small improvement over GMR-PCA (with no speaker normalisation) on all three databases. However, it did not outperform GMR-PLSDR, concurring with the results in Section 4.5.1.

Finally, the results in Table 4.4 reveal that speaker normalisation techniques developed for emotion classification systems, feature warping [122] and iterative feature normalisation [144], do not perform as well in continuous emotion prediction, possibly due to the fact that feature warping assumes the feature dimensions are independent and iterative feature normalisation relies on a definition of a neutral state that is difficult to properly articulate in the dimensional emotion representation. The i-vector and factor analysis based normalisation techniques did not perform as well as the proposed techniques either.

The results in Table 4.5 obtained using the SEMAINE database are compared to the state-of-the-art systems presented in AVEC 2012, including the baseline and winner papers for both the frame-level

Table 4.4: Performance on the CreativeIT database

A-arousal, V-valence, D-dominance; PCC- Pearson's correlation coefficients; CCC- Concordance CC

		PCC			CCC		
		A	V	D	A	V	D
USC CreativeIT	GMR-PCA	0.562	0.229	0.229	0.358	0.127	0.124
	GMR-Z-N	0.580	0.251	0.223	0.370	0.144	0.110
	GMR-PLSDR	***0.617***	***0.266***	***0.280***	***0.405***	***0.155***	***0.137***
	GMR-FM	***0.573***	***0.246***	***0.248***	***0.330***	***0.080***	***0.093***
	GMR-FW [20]	0.533	0.229	0.196	0.248	0.043	0.052
	IFN [13]	0.373	0.176	0.214	-	-	-
	I-vector [26]	0.492	-	0.145	-	-	-
	FA [17]	0.526	0.231	0.220	-	-	-

Table 4.5: Performance on the SEMAINE database

		PCC			CCC		
		A	V	D	A	V	D
SEMAINE	GMR-PCA	0.479	0.288	0.268	0.340	0.167	0.100
	GMR-Z-N	0.441	0.249	0.332	0.304	0.147	0.161
	GMR-PLSDR	***0.594***	***0.400***	***0.433***	***0.389***	***0.228***	***0.217***
	GMR-FM	***0.552***	***0.354***	***0.289***	***0.387***	***0.168***	0.082
	AVEC12 baseline[146]	0.054	0.062	0.019	-	-	-
	AVEC12 winner 1[147]	0.445	0.017	0.380	-	-	-
	AVEC12 winner 2[148]	0.257	0.270	0.147	-	-	-

Table 4.6: Performance on the RECOLA database

			PCC		CCC	
			A	V	A	V
RECOLA	GMR-PCA		0.766	0.427	0.629	0.281
	GMR-Z-N		0.813	0.471	0.667	0.208
	GMR-PLSDR		***0.820***	***0.513***	***0.670***	0.248
	GMR-FM		***0.781***	***0.443***	***0.631***	0.202
	State-of-the-art [21]	Proposed	-	-	~0.7	~0.32
		Average	-	-	~0.6	~0.21
		Global	-	-	~0.58	~0.18

* Note that '~' indicates the approximated values since these results were reported as a boxplot in [19].

and word-level sub-challenges. Though the data split for training, development and test are not identical between our experiments and AVEC 2012 challenges, the overall data size utilised in these experiments is similar, with 87 sessions in our experiments and approximately 96 sessions for AVEC 2012 challenge. It can be observed that the performance with the proposed compensation techniques outperformed those systems significantly.

The performance on the RECOLA database was also compared to a state-of-the-art system employing dynamic cooperative speaker models. It can be seen that our proposed systems outperformed the systems which utilises the global predictions (i.e. one speaker-independent regression system) and average predictions (i.e. mean of the predictions from all speaker-dependent systems). Though our GMR-PLSDR could not outperform the system with cooperative regression models, it is still a simpler system that does not require developing the speaker-dependent regression systems.

4.5 Comparison of compensation techniques

It is of interest to compare the compensation techniques in the feature space and model space. This is carried out by comparing the average relative improvements over the GMR-PCA baseline of feature normalisation and model adaptation schemes respectively in terms of PCC. Since valence is better predicted with a video signal and dominance is highly correlated with arousal [19, 145], the comparison of speech-based arousal prediction is more indicative and will now be discussed in this section. The analysis on valence will be considered as the future work.

Speaker-wise z-normalisation, factor analysis and feature mapping based normalisation techniques were carried out at the feature level, which can be characterised as compensation methods in the feature space. The PLSDR based techniques target model adaptation, since they alter the relationship between features and labels, directly interfering with the speaker-dependent conditional probability distribution. Thus the relative improvements of the GMR-Z-N and GMR-FM systems over the baseline GMR-PCA are calculated as the average overall improvement achieved by compensation techniques in the feature space. Similarly, the mean of the relative improvements of the systems

Table 4.7: Comparison of normalisation and adaptation for arousal

	USC CreativeIT	SEMAINE	RECOLA
Feature space	4.7%	12.6%	6.7%
Model space	24.2%	41.5%	13.6%

GMR-PLSDR over the baseline is computed, as the overall improvement of compensation techniques in the model space. The performance was calculated on the same three databases as in Section 4.5, and shown in Table 4.7. It was observed that the compensation techniques conducted in model space outperformed those in feature space on all three databases, suggesting that model adaptation performs better than feature normalisation for speaker variability compensation in continuous emotion prediction. Despite the different experimental settings, the comparison of relative improvement may still provide some useful insights.

4.6 Summary

Three speaker normalisation techniques based on factor analysis, partial least square dimension reduction (PLSDR) and feature mapping were proposed and tested over three databases. Improvements were observed in arousal, valence, and dominance prediction evaluated over three databases, validating the effectiveness of the proposed techniques. Follow up analysis of the feature space distributions was also conducted to verify the effect of the proposed techniques, in terms of the F-ratio, inter-speaker and intra-speaker variability before and after compensation.

Specifically in terms of the PLSDR and feature mapping based normalisation techniques, the results of the analyses and the validation on the three databases in terms of affect prediction systems complemented each other and indicated that both speaker normalisation techniques were effective in reducing speaker variability with the PLSDR based method being a little superior. In particular, the analyses showed that the PLSDR based method had a greater effect on features from those speakers that were more different from the other speakers. Finally, the experimental results also showed that the proposed methods outperformed current approaches to speaker normalisation. It can also be extended to explore other confounding factors apart from speaker variability.

Only speaker normalisation techniques are explored in this section and the adaptation of the trained models to the target speaker is not explored, though this may be a promising avenue for future work. Another limitation of the work reported in this chapter is that the GMR employed in this chapter did not take temporal information into account, and ignores the evolving nature of emotions.

5 CHARACTERISATION OF INTER-RATER VARIABILITY

5.1 Introduction

As mentioned in Chapter 2, continuous emotion prediction is generally viewed as a regression problem, where a speech waveform is labelled with a specific numerical value for each affective attribute indicating the short-term emotion intensity. The numerical affective attribute labels corresponding to each speech frame are generally obtained by averaging labels from multiple raters as perceived by them when listening to the speech (and watching associated videos if available). In current continuous emotion prediction systems, the back-end regression models are trained using these average ratings as targets, which neglects the possible information indicated by multiple raters, such as the disagreement among multiple raters. Thus, averaging these individual ratings to produce a 'gold standard' may not be the optimal strategy to generate the underlying emotion labels, since it forces the conflicting information between raters to be de-emphasised during system design.

One of the key information neglected in current continuous emotion systems is the inter-rater variability, referring to the disagreement among the multiple raters. The conventional system using mean ratings only considers the average but did not take into account the uncertainty information of the average rating: a high inter-rater variability may indicate a high uncertainty of the emotion state and vice versa as discussed in Section 2.5.3. Therefore, using the information of multiple raters instead of the mean rating to represent the underlying emotion attributes may be beneficial since they can provide more comprehensive information, i.e. the uncertainty of the emotion states.

Only a very limited number of studies have thus far considered inter-rater variability in continuous emotion prediction systems [23-25, 149-151]. These studies either presented methods to improve the inter-rater reliability, or adopted the inter-rater variability as additional information in the continuous emotion prediction system designs. However, a systematic analysis on the relationship between inter-rater variability and emotion states/categories has not yet been fully investigated.

The other important clue that multiple raters can provide is the inherent reaction lags between the individual annotations and the underlying emotional content, which is introduced by the time delay of the evaluator first sensing the stimulus and then defining his/her judgements. Previous studies [2] have only reported that compensating for the reaction lag between the mean rating and the underlying emotional content improves the system performance dramatically, but they assumed that difference in reaction lags between individual annotators are negligible, which may not hold true.

In this chapter I primarily focus on the inter-rater variability and the individual reaction lags. To evaluate the inter-rater variability, two measurements are adopted in terms of the pair-wise Pearson's correlation coefficient and Cronbach's alpha; the former estimating the average inter-rater variability of the pair-wise correlation, and the latter estimating the inter-correlation among multiple raters. Good surveys of different inter-rater reliability measures can be found in [152] and [153]. Additionally, a probabilistic framework is proposed to quantify the inter-rater variability in terms of different emotion categories which are clustered based on the two dimensional arousal and valence ratings, aiming to reveal the correlation between them and to provide a path to utilise the information from multiple raters for emotion prediction systems. The RECOLA database is utilised for the analyses in terms of the inter-rater variability since it contains the highest number of annotators for all the utterances, i.e. six annotators, amongst all publically available emotion databases.

Regarding the individual reaction lags, first the compensation effect is investigated in the mean ratings, and then proposed a compensation technique for the individual annotation delay based on maximising their inter-correlation.

The rest of this chapter is organised as follows: the delay compensation techniques for both the mean rating and the individual annotators are first described in Section 5.2, since the individual ratings after compensation for the reaction lags can support a more accurate analysis of the inter-rater variability. The inter-rater variability estimated by the two measurements mentioned above and the proposed probabilistic framework is discussed in Section 5.3. The key experimental settings and evaluation techniques are explained in Section 5.4. The chapter is summarised in Section 5.5.

5.2 Delay effect and compensation techniques

Section 5.2.1 describes the compensation method of the mean ratings, and Section 5.2.2 expands this concept to individual raters.

5.2.1 Delay Compensation for the mean rating

Generally the annotators listen to the speech, sensing and judging the underlying emotion states, and finally make their decisions and move the joysticks. This will introduce a delay between the speech and the annotations, referred as the reaction lag, as shown in Figure 5.1. The perceived emotion states are delayed than the underlying emotion states owing to the reaction lag. This in turn affects the reliability of the ground truth utilised for the emotion prediction systems, since the regression models are developed using the wrong labels will output wrong emotion predictions. Therefore, the

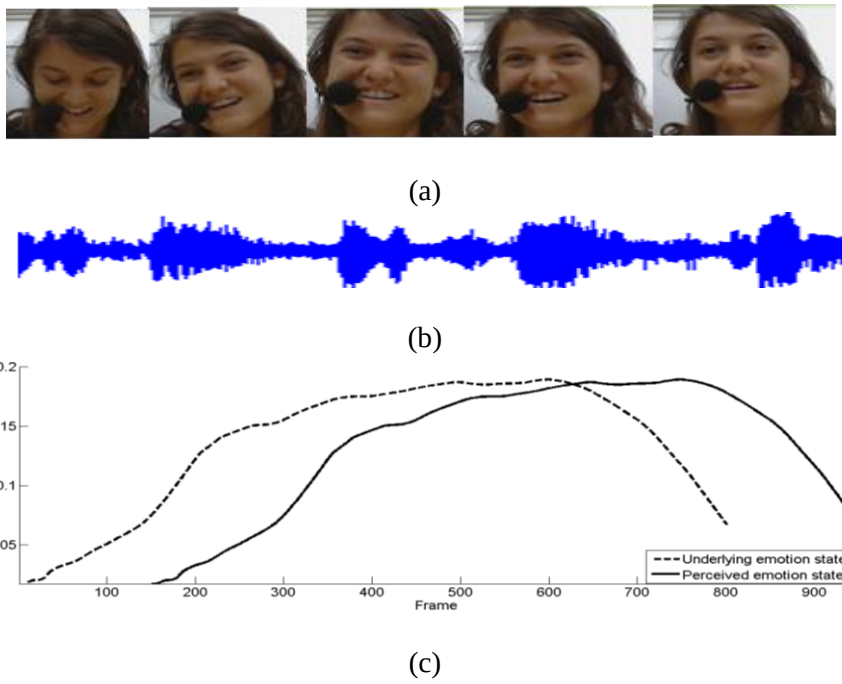


Figure 5.1: An example of reaction lag betteen the annotation and the speech. (a) Facial expression of one speaker that is shown for understanding but not used in this thesis; (b) speech segments of the corresponding speaker; (c) the perceived emotion states (mean rating among multiple raters) represented by solid line and the underlying emotion states represented by dash line. The perceived emotion states are delayed than the underlying emotion states.

annotation delay is generally compensated before training the regression models.

Similar to the work in [2], temporal shifts were applied to speech files in order to realign the features with the ground truth. The frame shift was achieved by dropping first N ground truth scores and last N input feature frames before regression training. This is applied to each file in the training set, and these realigned features and labels are then utilised for the regression modelling. This method is generally applied to most of emotion prediction systems [18, 19, 29].

However, the compensation in the training process will lead to the need to shift the prediction back by the same amount. As shown in Figure 5.2(a), the predictions and the ground truth are displayed without compensation for annotation delay. With the compensation for annotation delay in Figure 5.2(b), the predictions were produced by the regressor which is trained with the shifted training files. They are shifted forward in time by N frames, when compared to the ground-truth labels. In order to realign the predictions with the labels, the predictions are redshifted back N frames. This can be either by directing shift or by a smoothing filter that introduces the time shift. Filtering has been shown to be effective for smoothing output predictions, helping to minimise adverse effects due to noisy predictions and offer rough estimations for undetected frames in facial features [154]. Filtering also introduces an output-delay proportional to the filter length; a FIR filter length of $2N + 1$ introduces a delay of N, where N remains the number of scores/frames dropped as per the previous paragraph. Hence, post-processing filters are used for resolving the synchronisation issue in predicted outputs caused by the introduction of a delay in the training phase (Figure 5.2c), and unless stated this is applied to all systems reported herein.

In this chapter, a smoothing filter is used not only to help remove high frequency noise present in predictions, but also to realign predictions generated by a system trained on frame shifted features. As this filter will be applied over longer timescales (2s to 4s), the commonly used mean filter [155] that applies equal weights to all samples could be an unsuitable choice of filter. I therefore apply a binomial filter, which is a Gaussian shaped filter that gives greater weight to predictions adjacent to the prediction and less weight to the predictions further away. The binomial filter coefficients are obtained by the successive convolutions of $(1,1)$. For example, with order $n = 3$ applications the

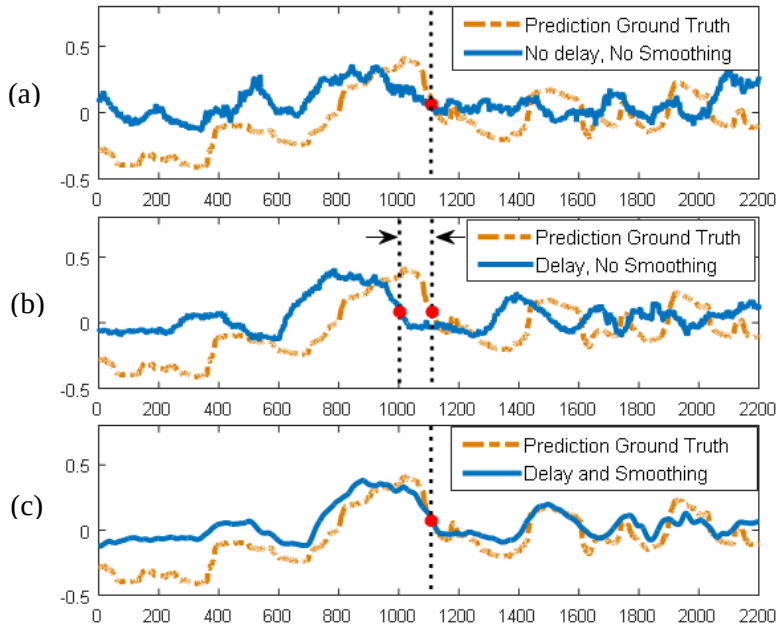


Figure 5.2: Effect of annotation delay compensation on a set of predicted arousal ratings. (a) Predictions without delay compensation and smoothing are noisy and not well matched with the ground truth labels. (b) Applying temporal shifts to the training data improves system performance but results in predictions that are advanced in time compared to their ground truth. (c) Applying a binomial filter to these predictions not only smooths the output but resolves the synchronisation issue [29].

binomial filter weights are given by $[1,1] * [1,1] = [1,2,1]$, and with order $n = 4$ the weights are given by $[1,1] * [1,1] * [1,1] = [1,3,3,1]$. They tail off smoothly towards zero near the edges. For large numbers of applications, the weights become Gaussian and the filtering approximates Gaussian kernel smoothing. In this case, the order $2N + 1$ is used and the coefficients $h(n)$ are formed as:

$$h(n) = \underbrace{[1,1] * \cdots * [1,1]}_{2N} \tag{5.1}$$

The coefficients are finally normalised by dividing $\sum_{n=1}^{2N+1} h(n)$ to make the summation of all coefficients to 1. The predictions after applying binomial filter will be realigned as in [29]. It should be noted that the length of binomial filter can be flexible. The reason we adopted the length related to annotation delay is to achieve the smoothing and realignment simultaneously. This framework has been tested within a wide range of delays targeting the optimal delay compensation for arousal and valence respectively, and experimental results are discussed in Section 5.4.1.

5.2.2 Delay Compensation for individual annotators

The method of compensating reaction lag in mean ratings based on validation performance cannot be directly applied to individual annotators, since reaction delays between each individual annotator may be different, which will increase the computational requirements dramatically. Let N_r represents the individual rater's frame shift for the r^{th}rater. For R raters in total, if using the first rater as a reference, a set of frame shifts $[N_2, N_3 \cdots N_R]$ for all the other raters $(2 \leq r \leq R)$ needs to be verified. N_r is validated over a wide range of values, assuming relative delays from -5 to 5 seconds with a step of 0.2 second (51 total steps), this will lead to 51^{R-1} combinations to be tested, and increasing number of raters and delay frames tested will further increase the computational load. Note that the frame shifts are verified from negative delays to positive delays to overcome the problem that the first rater responds slower than other raters. A more effective strategy is proposed to compensate for the reaction lags in individual annotators based on the inter-correlation coefficient.

A speech segment annotated by three raters is shown in Figure 5.3. Taking Rater 1 as a reference, it is observed that Rater 3 (red) responds to the decreasing change in affect level much more slowly compared to Rater 1 indicated by the orange arrow in Figure 5.3, while Rater 2 (green) responds faster

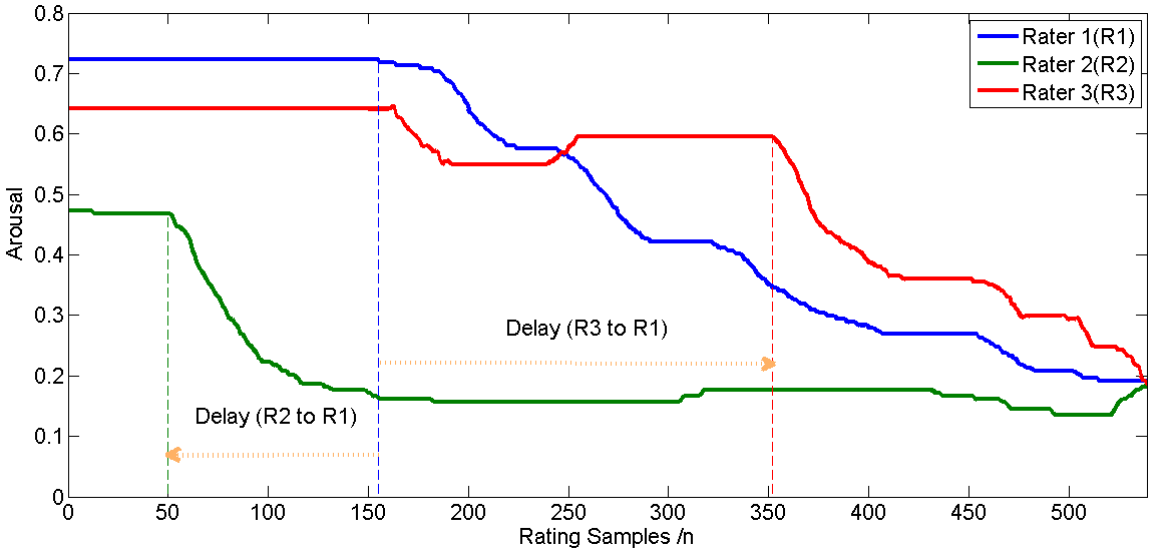


Figure 5.3: Three individual ratings of a speech segment. The delay is observed as the difference among the starting point of a decreasing trend in the ratings.

than Rater 1 (blue). A compensation method is proposed that aims to maximise the inter-rater 'correlation coefficient' among three raters based on the Pearson's correlation coefficient (CC). Pearson's CC is only capable of calculating the correlation between two annotators, so first the idea must be expanded to three annotators.

Pearson's CC between two variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is [90]:

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{cov(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}} \tag{5.2}$$

where the covariance of $\boldsymbol{x}$ and $\boldsymbol{y}$ *is*

$$cov(\boldsymbol{x}, \boldsymbol{y}) = E[\boldsymbol{xy}] - E[\boldsymbol{x}]E[\boldsymbol{y}] \tag{5.3}$$

To expand this concept to three variables, the covariance between three variables $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{z}$ is calculated as:

$$cov(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = E[\boldsymbol{xyz}] - E[\boldsymbol{x}]E[\boldsymbol{yz}] - E[\boldsymbol{y}]E[\boldsymbol{xz}] - E[\boldsymbol{z}]E[\boldsymbol{xy}] + 2E[\boldsymbol{x}]E[\boldsymbol{y}]E[\boldsymbol{z}] \tag{5.4}$$

The 'correlation' between three variables can then be calculated as:

$$\rho(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \frac{cov(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})}{\sigma_{\boldsymbol{x}}\sigma_{\boldsymbol{y}}\sigma_{\boldsymbol{z}}} \tag{5.5}$$

Similar to Section 5.2.1, different frame shifts are investigated. One rater, randomly chosen as Rater 1, is utilised as the reference without any frame shift, then Rater 2 and Rater 3 are shifted with regards to Rater 1. The optimal delay for each rater is finally chosen as the time shift values that achieve the maximum $\rho(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$, and the individual ratings are then realigned based on these optimal delays. In order to validate the effectiveness of the realignment of the individual raters, a system which utilises the mean realigned ratings is compared to the system with the original mean ratings. All other system configurations are kept same for a direct comparison.

5.3 Inter-rater reliability

Recall from Section 5.1, our definition of inter-rater variability, it indicates the disagreement among multiple raters. A high inter-rater variability represents a high level disagreement among raters, which has been paid far less attention to, when defining the emotion intensity as the mean of multiple ratings.

A systematic understanding of how inter-rater variability relates to the emotion categories is still lacking. This section first proposes two measurements of inter-rater variability in Section 5.3.1, and analyses the correlation between the inter-rater variability and emotion categories/clusters in Section 5.3.2.

5.3.1 Measurements of inter-rater variability

The most widely adopted metrics used to measure the inter-rater variability are the mean of pair-wise Pearson's correlation coefficient and Cronbach's alpha [156]. The mean of pair-wise Pearson's correlation coefficient is given as:

$$u(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N) = \frac{1}{\sum_{k=1}^{R-1} k} \sum_{i=1}^{R} \sum_{j=i+1}^{R} \rho(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{5.6}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ represents the i^{th} and the j^{th} individual annotations, and R represents the total rater number. Cronbach's alpha α is able to estimate the reliability between multiple raters and is given as:

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \sigma_{\boldsymbol{x}_i}^2}{\sigma_{\boldsymbol{y}}^2}\right) \tag{5.7}$$

where $\sigma_{\boldsymbol{x}_i}^2$ is the variance of the i^{th} rating, $\boldsymbol{x}_i$ represents the i^{th} individual annotations, and $\sigma_{\boldsymbol{y}}^2$ represents the variance of the observed total ratings given that $\boldsymbol{y} = \boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_R$. These two measurements reflect the reliability of the ratings, and a high value of $\rho(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_N)$ or α indicates a low inter-rater variability.

5.3.2 Correlation between inter-rater variability and emotion clusters

The correlation between the inter-rater variability and emotion clusters/categories is analysed in a probabilistic framework, similar to that proposed in chapter 3. I aim to gain some insights into the inter-rater variability, i.e. is inter-rater variability different for different emotion categories? The first challenge in this framework is the definition of emotion categories in terms of dimensional labels of arousal and valence. Here the *K*-means clustering method [157] is adopted to cluster the speech frames into *K* different emotion categories in the arousal-valence space. This is directly applied to

mean ratings. *K*-means clustering aims to partition the observations into *K* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype for each cluster. It is used to allocate speech frames to different clusters that represent different emotional states.

Let $\boldsymbol{x}_n$ represent the two-dimensional mean ratings of arousal and valence at frame n, and $\boldsymbol{X} = [\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots \boldsymbol{x}_N^T]^T$ represents the entire training data of *N* frames. *K*-means clustering partitions *N* frames into *K* sets $\boldsymbol{S} = [S_1, S_2, \cdots S_K]$ with the aim of minimising the sum of distances between the training data and their corresponding centroids. The Euclidean distance is the metric generally adopted, given by:

$$\underset{\boldsymbol{S}}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{\boldsymbol{x} \sim S_i} \|\boldsymbol{x} - \boldsymbol{u}_i\|^2 \tag{5.8}$$

where $\boldsymbol{u}_i$ is the mean of the data points in S_i. Readers can refer to [157] for further details.

The second key point in the analysis of the relationship between the inter-rater variability and the emotion categories/clusters is the representation of the inter-rater variability, which is computed as the standard deviation among multiple raterss. Let $\boldsymbol{\sigma}_n$ represent the standard deviation among *R* raters for each frame :

$$\boldsymbol{\sigma}_n = \sqrt{\frac{1}{R} \sum_{r=1}^{R} (\boldsymbol{x}_{n,r} - \overline{\boldsymbol{x}}_n)^2} \tag{5.9}$$

where $\overline{\boldsymbol{x}}_n$ represents the average of $\boldsymbol{x}_{n,r}$ among *R* raters. These inter-rater variability $\boldsymbol{\sigma}_n$ will be assigned to the sets $\boldsymbol{S} = [S_1, S_2, \cdots S_k]$ of each cluster according to the *K*-means clusters. GMMs are then estimated to model the distributions of inter-rater variability within each emotions cluster, referred as the emotion-dependent inter-rater distributions, $P(\boldsymbol{\sigma}|S_i)$. Similar to Section 3.2.1, GMMs $P(\boldsymbol{\sigma}|S_i)$ aim to model the probabilistic distribution of the inter-rater variability $\boldsymbol{\sigma}_n$ for each emotion cluster S_i. Note that GMMs were developed for each speaker in Section 3.2.1, while here they are created for each emotion clusters S_i.

In order to quantify the inter-rater variability between different emotion clusters, the symmetric KL divergence is adopted which calculates the similarity between the emotion-dependent GMMs $P(\boldsymbol{\sigma}|S_i)$ and $P(\boldsymbol{\sigma}|S_j)$, similar to equation (3.1):

$$I_{SKL}\left(P(\boldsymbol{\sigma}|S_i), P(\boldsymbol{\sigma}|S_j)\right) = \frac{1}{2}\left|\int_x P(\boldsymbol{\sigma}|S_i) In \frac{P(\boldsymbol{\sigma}|S_i)}{P(\boldsymbol{\sigma}|S_j)} dx + \int_{\boldsymbol{x}} P(\boldsymbol{\sigma}|S_j) In \frac{P(\boldsymbol{\sigma}|S_j)}{P(\boldsymbol{\sigma}|S_i)} d\boldsymbol{x}\right| \tag{5.10}$$

A Monte-Carlo sample method is utilised to computationally estimate KL divergence in (5.10). Readers can refer to [14] more details. The average KL divergence between one emotion-dependent distribution $P(\boldsymbol{\sigma}|S_i)$ and all other emotion-dependent distributions is finally estimated as follows:

$$A(i) = \frac{1}{K-1}\sum_{\forall j, j\neq i} I_{SKL}\left(P(\boldsymbol{\sigma}|S_i), P(\boldsymbol{\sigma}|S_j)\right) \tag{5.11}$$

It should be noted that the absolute value of the symmetric KL divergence $A(i)$ in (5.11) is not straightforward to determine the similarity between two distributions. Thus two reference symmetric KL divergence are adopted, i.e. the similarity between the emotion-dependent distributions from the same clusters, and the similarity between the emotion-dependent distribution $P(\boldsymbol{\sigma}|S_i)$ and a emotion-independent distribution $P(\boldsymbol{\sigma})$ (similar as Section 3.2.2). In terms of the similarity between emotion-dependent distributions from the same clusters, the inter-rater variability $\boldsymbol{\sigma}$ are divided equally to two parts for the same emotion cluster, represented as $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$**,** then the emotion dependent distributions are developed as $P(\boldsymbol{\sigma}_1|S_i)$ and $P(\boldsymbol{\sigma}_2|S_i)$. The KL divergence between the same emotion clusters are

$$R(i) = I_{SKL}\left(P(\boldsymbol{\sigma}_1|S_i), P(\boldsymbol{\sigma}_2|S_i)\right) \tag{5.12}$$

Regarding the similarity between the emotion-dependent distribution $P(\boldsymbol{\sigma}|S_i)$ and a emotion-independent distribution $P(\boldsymbol{\sigma})$ (referring to UBM) , the KL divergence is calculated as:

$$U(i) = I_{SKL}(P(\boldsymbol{\sigma}|S_i)\,, P(\boldsymbol{\sigma})) \tag{5.13}$$

Under the assumption that inter-rater variability manifest different for different emotions, it is expected that $A(i)$ is significantly larger than $R(i)$ and $U(i)$, with $U(i)$ larger than $R(i)$.

One key challenge in this framework is the cluster number *K* which cannot be specifically defined, thus a range of different values are tested. In order to investigate the effect of the total cluster number

K in our proposed framework, the mean values of $A(i)$ and $U(i)$ averaged over all emotion clusters are estimated and compared for each specific K value as:

$$A_f(K) = \sum_{i=1}^{K} A(i) \tag{5.14}$$

$$R_f(K) = \sum_{i=1}^{K} R(i) \tag{5.15}$$

$$U_f(K) = \sum_{i=1}^{K} U(i) \tag{5.16}$$

where K represents the cluster number, and is tested from 4 to 10 clusters. Finally $A_f(k)$, $R_f(K)$ and $U_f(k)$ are compared in Section 5.4.2. One of the limitations in this work is that the standard deviation among multiple raters as a global representation of inter-rater variability does not capture all the potential biases in individual annotators.

5.4 Experimental settings and results

5.4.1 Delay Compensation

5.4.1.1 Delay Compensation in mean ratings

A range of different time delays ranging from 0 to 10 seconds with a step of 1 second for arousal and valence respectively was tested in the RECOLA database. The eGeMAPS feature set was adopted as the front-end, and RVM is utilised as the back-end. The concordance correlation coefficient (CCC) is utilised as the evaluation metric, as shown in equation (2.3). The results depicted in Figure 5.4 show that a substantial improvement in CCC was seen with the introduction of delay compensation, increasing from around 0.276 to 0.691 for arousal and from 0.140 to around 0.5 for valence (Figure 5.4). Also it shows that valence rating responds more rapidly than arousal, which is consistent with previous studies [2].

Our analysis reveals that compensating the annotation delay improves the system performance substantially. Based on these results, a 4-second delay for arousal and a 2-second delay for valence was selected as the optimal delay value. Unless otherwise stated, these delay values were used in all subsequent systems.

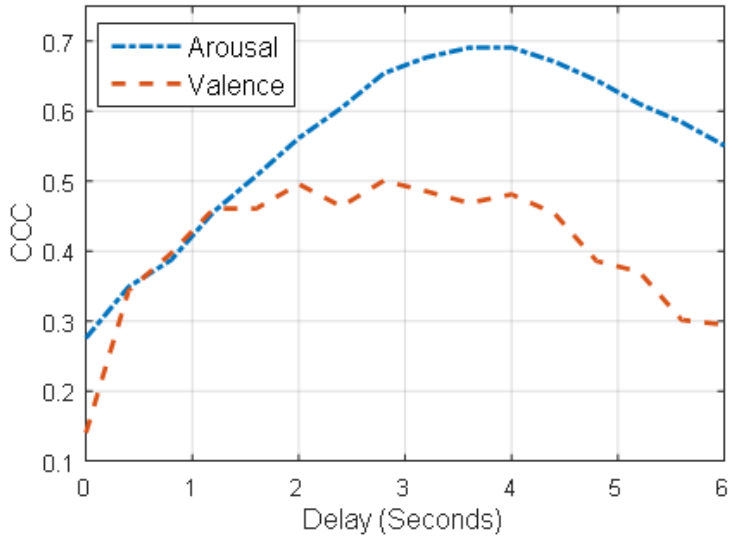


Figure 5.4: Delay compensation using frame shift and smoothing. The best delay for arousal was found to be 4s and the best delay value for valence was 2s.

5.4.1.2 Delay Compensation in individual ratings

Here, the USC CreativeIT database (annotated by 2 to 4 evaluators) is adopted since it makes the initial analysis a simpler task. Similar to Section 5.4.1.1, a range of different time delays ranging from [-10s 10s] seconds with a step of 0.2 seconds was tested for each rater. Note that the number of annotators for each utterance in the USC CreativeIT database is not fixed, which could be 2, 3, or 4 annotators, and for a given utterance they would be all different. Thus the compensation is carried out on an utterance basis rather than on the database as a whole. The inter-rate 'correlation coefficients' among three raters achieved with different individual delays for one utterance are shown in Figure 5.5. It can be seen that the inter-correlation changes with varying delay, and the optimal delay is chosen as the values achieving the maximum CC.

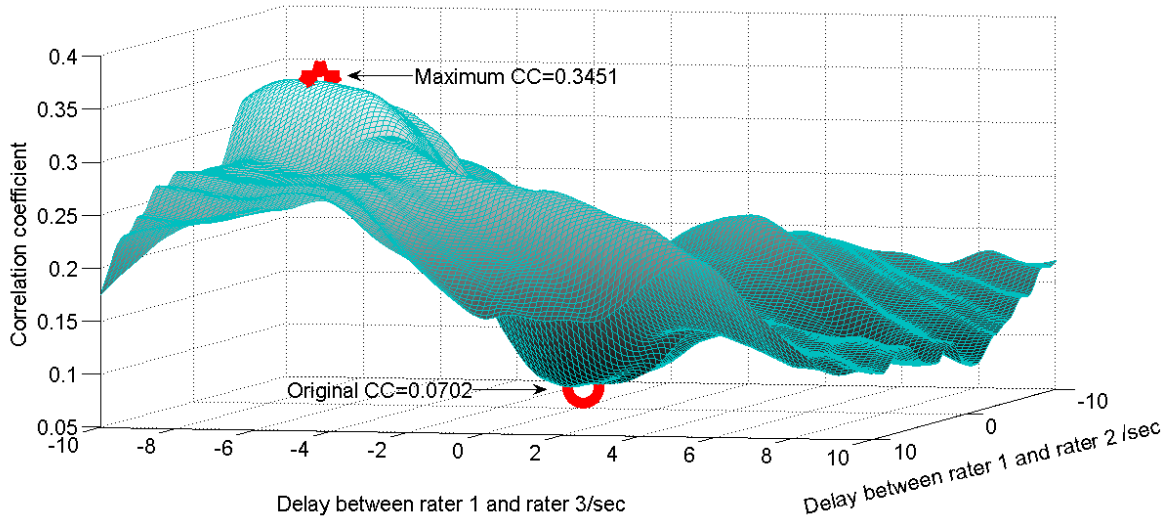


Figure 5.5: Delay compensation for individual annotators using frame shift. The original CC without realignment (at (0,0) in the x-y plane) and maximum CC, where the individual ratings have been realigned, are indicated by a red circle and a red star, respectively.

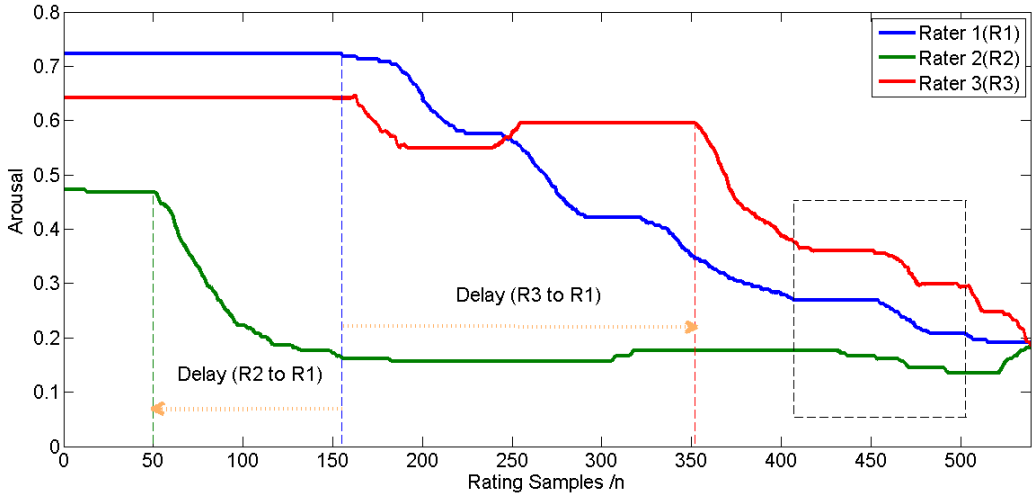
The mean rating averaged among the realigned individual ratings are utilised for the regression system design, and compared with the system using the original mean ratings. In addition to arousal and valence, the system performance for dominance is also developed. As stated in [145], the feature set discussed in Section 4.4.2 performs well in the USC CreativeIT database, and was found slightly better than eGeMAPS in our analysis. Thus the 650 dimensional feature set was utilised and PCA was applied to reduce the feature dimension. Similarly, GMR was found to perform well in these system configurations and was utilised in this section. The system performance is evaluated in terms of the Pearson's correlation coefficient and the results are shown in Table 5.1.

The system with delay compensation could not outperform the baseline without delay compensation between individual raters, possible owing to the delay between individual raters not

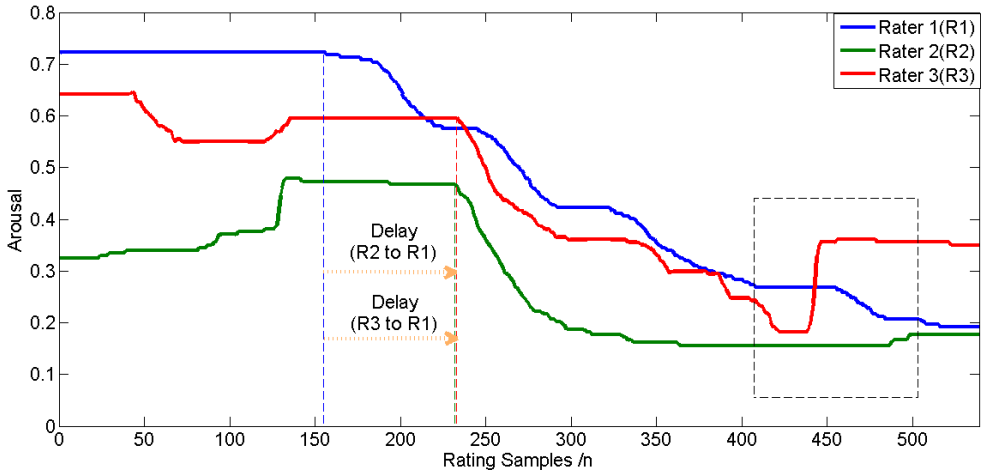
Table 5.1: Comparison of System performance with and without delay compensation in individual rating in terms of correlation coefficients

	Arousal	**Valence**	**Dominance**
with delay compensation	0.4868	0.2472	0.2240
without delay compensation	0.5539	0.2096	0.2412

being consistent over the entire utterance. Figure 5.6 shows that Rater 2 and Rater 3 were realigned with approximately the same starting point of the decreasing trend of arousal intensity, and their delays with respect to R1 were significantly reduced. However, the three ratings in the black dash box in Figure 5.6(b) becomes dislocation comparing to that in Figure 5.6(a). Consequently, it may be that the realignment of individual ratings may only work for some specific segments. One possible solution can be to implement the proposed method for each small segment instead of the entire



(a)



(b)

Figure 5.6: Delay compensation for individual ratings based on maximizing CC. (a) original individual ratings without delay compensation, where R2 and R3 were observed significant delay regarding to R1 at the starting point of a decreasing trend; However, the individual ratings within the black dash box were well aligned; (a) individual ratings with delay compensation, where R2 and R3 were realigned and the delay regarding R1 was reduced ; however, the individual ratings within the black dash box becomes misaligned.

utterance, which is future work as discussed in Section 9.2. Due to the worse performance with compensation in the individual annotators, the following analysis on inter-rater variability will focus on the original ratings without compensation.

5.4.2 Analysis on inter-rater variability

5.4.2.1 Measurements of inter-rater variability

As discussed in Section 5.1, the RECOLA database is used owing to its uniform use of a relatively large number (6) of raters. The pair-wise Pearson's correlation coefficient for arousal and valence measuring inter-rater variability is shown in Tables 5.2 to 5.5 (next page). R_i represents Rater i in these tables. The mean Pearson's CC of each rater was calculated by averaging the five pair-wise Pearson's CC respectively. A significant difference was observed among the six raters in terms of the mean Pearson's CC, suggesting a high inter-rater variability. The inter-rater variability as lower in valence compared to that of arousal, though speech-based valence prediction systems generally

Table 5.2: Pair-wise Pearson's Correlation Coefficients (CC) for arousal in training partition. R1 to R6 represents Rater 1 to Rater 6 respectively; Mean represents the average among the five pair-wise CC.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**
R1	1	0.459	0.594	0.407	0.457	0.423
R2	0.459	1	0.521	0.479	0.346	0.334
R3	0.594	0.521	1	0.382	0.400	0.412
R4	0.407	0.479	0.382	1	0.262	0.323
R5	0.457	0.346	0.400	0.262	1	0.241
R6	0.423	0.334	0.412	0.323	0.241	1
Mean	***0.468***	***0.428***	***0.462***	***0.371***	***0.341***	***0.347***

Table 5.3: Pair-wise CC for arousal in validation partition.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**
R1	1	0.493	0.732	0.434	0.523	0.429
R2	0.493	1	0.602	0.451	0.353	0.383
R3	0.732	0.602	1	0.578	0.543	0.536
R4	0.434	0.451	0.578	1	0.447	0.568
R5	0.523	0.353	0.543	0.447	1	0.310
R6	0.429	0.383	0.536	0.568	0.310	1
Mean	***0.522***	***0.456***	***0.598***	***0.496***	***0.435***	***0.445***

Table 5.4: Pair-wise CC for valence in training partition.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**
R1	1	0.620	0.545	0.609	0.285	0.615
R2	0.620	1	0.539	0.556	0.155	0.549
R3	0.545	0.539	1	0.511	0.178	0.502
R4	0.609	0.556	0.511	1	0.239	0.548
R5	0.285	0.155	0.178	0.239	1	0.309
R6	0.615	0.549	0.502	0.548	0.309	1
Mean	***0.535***	***0.484***	***0.455***	***0.493***	***0.233***	***0.505***

Table 5.5: Pair-wise CC for valence in validation partition.

	R1	**R2**	**R3**	**R4**	**R5**	**R6**
R1	1	0.601	0.512	0.610	0.359	0.512
R2	0.601	1	0.595	0.573	0.294	0.528
R3	0.512	0.595	1	0.478	0.182	0.505
R4	0.610	0.573	0.478	1	0.335	0.485
R5	0.359	0.294	0.182	0.335	1	0.309
R6	0.512	0.528	0.505	0.485	0.309	1
Mean	***0.519***	***0.518***	***0.454***	***0.496***	***0.296***	***0.468***

Table 5.6: Cronbach's alpha α for arousal and valence in training and development datasets. All means the α calculated over the combination of train and dev sets.

	Arousal	Valence
Training	0.7643	0.8194
Development	0.7944	0.8111
All	0.7811	0.8161

achieve lower performance than arousal prediction. This is possibly owing to the fact that the annotations are evaluated by listening to the speech and watching the video simultaneously.

Additionally, the Cronbach's alpha on the training and development partition is also reported in Table 5.6. The Cronbach's alpha also suggests that the inter-rater variability was lower in valence rather than that in arousal.

5.4.2.2 Correlation between inter-rater variability and emotion states

A variety of different cluster numbers ranging from 4 to 10 is tested. GMMs with 2 mixture components are utilised to model the probabilistic distribution, where full covariance matrices are adopted to capture the correlation between arousal and valence. The symmetric KL divergence is calculated using 100,000 Monto-Carlo sampling points.

The KL divergence between one emotion cluster and all other clusters $A(i)$, between same emotion cluster $R(i)$, and between one emotion cluster and UBM $U(i)$ (Section 5.3.2) for six clusters is shown in Figure 5.7. It should be noted that a similar result has been observed for all different numbers of clusters *K*, and the result for six clusters is only presented for explanatory purposes. It can be seen that the KL divergence $A(i)$ calculated between one emotion cluster and all other emotion clusters (red bars) is significantly larger than that between one emotion cluster and UBM $U(i)$ (light blue bars), and that between same emotion states $R(i)$ (dark blue bars) for all clusters, suggesting that the inter-rater variability varies significantly among emotion clusters. U(i) is also shown larger than R(i) where R(i) is extremely small approaching to 0 since they have the similar distribution for the same emotion cluster.

In addition, the average of the KL divergence $A_f(K)$, $R_f(K)$ and $U_f(K)$ for different cluster number *K* is shown in Figure 5.8. There is no significant difference observed for different numbers of clusters. The consistent larger value of $A_f(K)$ than $R_f(K)$ and $U_f(K)$ indicates that the number of emotion clusters is not a factor affecting our findings, which further confirms our observation that inter-rater variability is emotion-dependent, which warrants deeper consideration as discussed in Section 5.3.2.

5.5 Summary

This chapter analysed the information from multiple raters in terms of the individual reaction lags and the inter-rater variability characterised by two measurements, i.e. the mean of pair-wise Pearson's CCs and Cronbach's alpha, and more importantly the correlation between the inter-rater variability

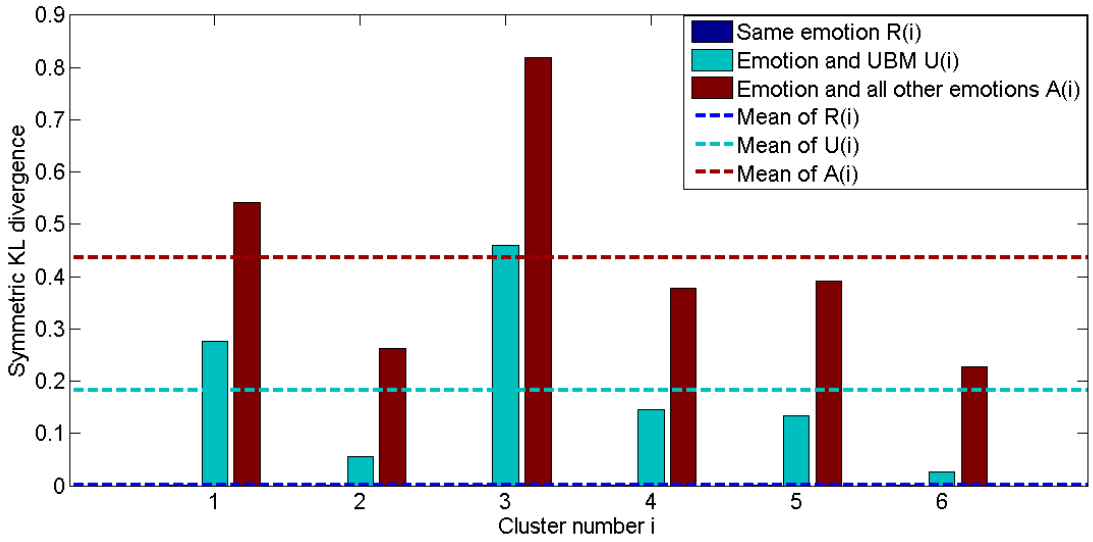


Figure 5.7: Symmetric KL divergence for emotion-dependent inter-rater distributions. Dark blue bars indicate the symmetric KL divergence calculated between the same emotion cluster $R(i)$; light blue bars represent that calculated between one emotion and UBM; red bars represent that calculated between one emotion and all other emotion clusters $A(i)$. Note that dark blue $R(i)$ is extremely small (10^{-6}) that is not observed in this Figure.

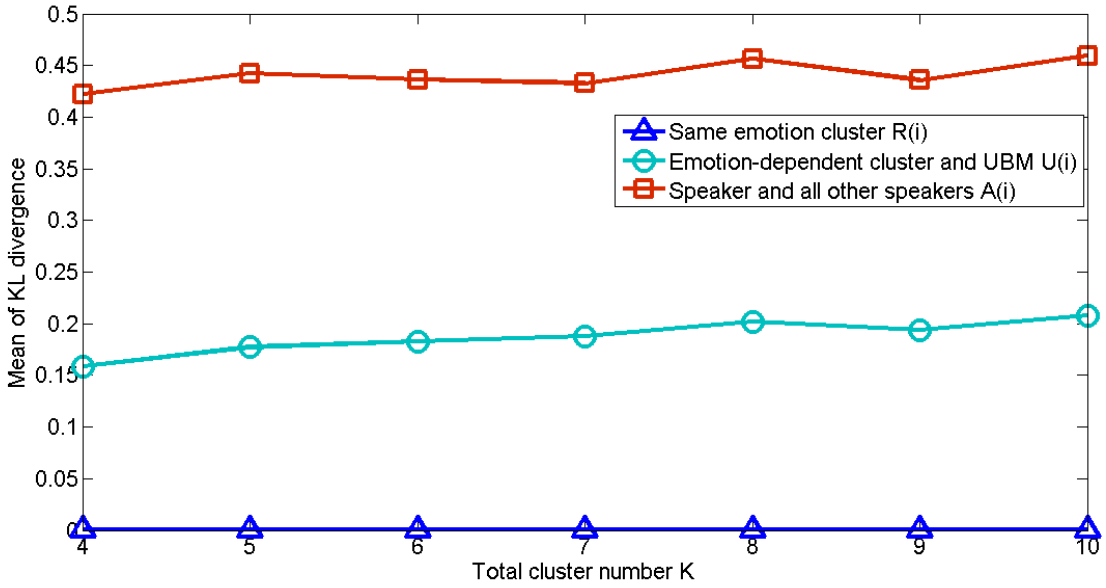


Figure 5.8: Average of symmetric KL divergence for emotion-dependent inter-rater distributions using different clusters. Dark blue bars indicate the average of symmetric KL divergence calculated between same emotion cluster $R_f(k)$; light blue bars represent that between one emotion and UBM $U_f(k)$; red bars represent that between one emotion and all other emotion clusters $A_f(k)$. $A_f(k)$ is consistently larger than $R_f(k)$ and $U_f(k)$, indicating the number of clusters is not a affecting factor.

and different emotion categories. Compensation for the mean ratings was observed to significantly improve these measures, suggesting the importance of taking reaction lag into account in continuous emotion prediction systems. However, the system using the realigned individual ratings based on the

proposed compensation technique could not outperform the baseline system without compensation, which is possibly due to intra-rater variability. Future work could focus on a small segment instead of a long utterance, a different criterion instead of the inter-rater 'correlation coefficient', or further analysis of the inconsistency of individual ratings. Additionally, the correlations found between the inter-rater variability and the emotion clusters suggested that the inter-rater variability varies significantly among different emotion clusters, which contains valuable information related to emotion uncertainty. These findings motivate future work on the incorporation of inter-rater variability into continuous emotion prediction systems, where the emotion label can be represented by a distribution using the information from multiple raters, instead of single mean ratings only.

6 UNCERTAINTY IN EMOTION PREDICTION

6.1 Introduction

As mentioned in Chapter 5, differences in perception among raters result in time-varying inter-rater variability that is shown to be correlated with emotion categories. Considering that the overall distribution of inter-rater variability is able to reflect the uncertainty in the speech frames to some extent, meaning that a high inter-rater variability indicates a high level of uncertainty about the affective attributes corresponding to that speech frame, incorporating this kind of information into a model for uncertainty prediction to support the point estimation of conventional emotion prediction systems can give us insights into the natural variability of human emotion expressions.

In current literature, inter-rater variability has been considered in continuous emotion systems, by either dealing with relative labels using preference learning [22, 87], or considering a multi-task learning framework which learns the inter-rater variability by the inter-rater standard deviation based on LSTM-NNs [151], or capturing the inter-rater variability by Convolution Neural Network [148]. One shortcoming of these methods is that they treat multiple raters variability as an additional point prediction target instead of an overall distribution that can comprehensively represent the inter-rater variability. Preference learning [22, 87] aims to predict the ranking of emotion labels while still treating the emotion intensity as a point estimation, and the LSTM-NN based system [151] predicted the mean and standard deviation of multiple raters, which still takes the predicted emotion labels (i.e. mean and standard deviation as two predicted labels) as point estimates for mean and standard deviation, though it does consider the label distribution to be Gaussian. In this chapter an emotion prediction system is proposed that targets the prediction of label distributions instead of point estimates, where the label distribution is indicated by inter-rater variability. It is expected that the prediction of the label distribution can reflect the uncertainty of the emotion content. For instance, a frame with a high inter-rater variability would correspond to a predicted label distribution with low confidence regions.

The key challenges in estimating the prediction uncertainty directly from annotated speech are: (i) finding a probabilistic model for the posterior distribution and thus providing a means to estimate the uncertainty information; and (ii) finding a way to incorporate the inter-rater variability into the model. In terms of the first challenge of a probabilistic model, the commonly used SVMs [74, 158, 159] and LSTM-NN [27, 160] are not able to handle this problem since they can only find a point estimate based on one specific structural risk minimisation. However, probabilistic models such as RVMs [29, 31, 84] and GMR [97] are capable of incorporating a probabilistic description of the target labels. Furthermore both RVMs and GMR have been shown to be effective in predicting emotions [31, 97]. Given the desire to incorporate inter-rater variability into the model, multivariate RVM [93], where multiple raters evaluations can be treated as multi-task learning, does not show good performance when arousal and valence are modelled as multi-task learning [93]. Moreover, RVMs are constrained to modelling only the label distribution as Gaussian distributions. In this chapter, GMR is adopted as the regression model. More importantly, it provides the flexibility to incorporate the inter-rater variability into the input feature space while generating the joint vectors for GMR training (Section 2.3.3), which is achieved by concatenating individual ratings with the original features on a frame basis.

6.2 Emotion uncertainty prediction

The conventional GMR will be briefly reviewed in Section 6.2.1. Then the proposed method is separated into the incorporation of inter-rater variability in Section 6.2.2 and the label distribution prediction in Section 6.2.3.

6.2.1 Conventional GMR

First I will briefly review the conventional GMR as discussed in Section 2.3.3. Let $\boldsymbol{X}_n = [\boldsymbol{x}_n^T, \Delta\boldsymbol{x}_n^T]^T$ and $\boldsymbol{Y}_n = [\boldsymbol{y}_n^T, \Delta\boldsymbol{y}_n^T]^T$ represent the features and labels consisting of the static information (low level descriptors) and dynamic information (generally delta features) at frame n. The training features and labels are then represented as $\boldsymbol{X} = \left[\boldsymbol{X}_1^{\mathrm{T}}, \boldsymbol{X}_2^{\mathrm{T}}, \cdots \boldsymbol{X}_N^{\mathrm{T}}\right]^{\mathrm{T}}$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^{\mathrm{T}}, \boldsymbol{Y}_2^{\mathrm{T}}, \cdots \boldsymbol{Y}_N^{\mathrm{T}}\right]^{\mathrm{T}}$, where N represents the

total number of frames. The GMM $\lambda^{(\boldsymbol{Z})}$ modelling the joint probability distribution of features $\boldsymbol{X}$ and labels $\boldsymbol{Y}$ is trained using all the joint features $\boldsymbol{Z}_n = [\boldsymbol{X}_n, \boldsymbol{Y}_n]$, referring to Section 3.2.3. In order to find the predictions $\widehat{\boldsymbol{Y}}_t$ of the test features $\boldsymbol{X}_t$ at frame t, the conditional probability of label $\boldsymbol{Y}_t$ as $P(\boldsymbol{Y}_t|\boldsymbol{X}_t, \lambda^{(\boldsymbol{Z})})$ is calculated for each frame as a GMM as:

$$P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(Z)}) = \sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) \tag{6.1}$$

where M represents the total mixture component of the GMM $\lambda^{(\boldsymbol{Z})}$. Here the first term on the right side of (6.1) $P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})})$ is a scalar indicating the posterior probability of $\boldsymbol{X}_n$ belonging to the m^{th} mixture component, as shown in equation (2.37). The second term on the right side of (6.1) is a Gaussian distribution as:

$$P(\boldsymbol{Y}_n|\boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}) = N(\boldsymbol{Y}_n; \boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}) \tag{6.2}$$

where the mean $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and covariance $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ of the Gaussian distribution are

$$\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})} = \boldsymbol{u}_m^{(\boldsymbol{Y})} + \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-\mathbf{1}} (\boldsymbol{X}_n - \boldsymbol{u}_m^{(\boldsymbol{X})}) \tag{6.3}$$

$$\boldsymbol{D}_m^{(\boldsymbol{Y})} = \boldsymbol{\Sigma}_m^{(\boldsymbol{YY})} - \boldsymbol{\Sigma}_m^{(\boldsymbol{YX})} \boldsymbol{\Sigma}_m^{(\boldsymbol{XX})-\mathbf{1}} \boldsymbol{\Sigma}_m^{(\boldsymbol{XY})} \tag{6.4}$$

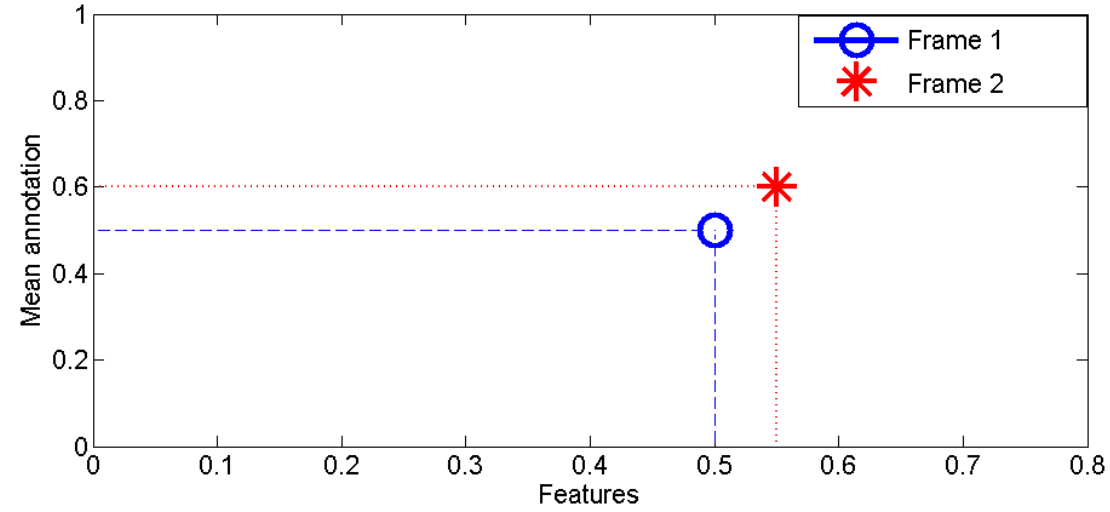
where $\boldsymbol{u}_m^{(\boldsymbol{Y})}$, $\boldsymbol{u}_m^{(\boldsymbol{X})}$, $\boldsymbol{\Sigma}_m^{(\boldsymbol{XX})}$, $\boldsymbol{\Sigma}_m^{(\boldsymbol{YY})}$, $\boldsymbol{\Sigma}_m^{(\boldsymbol{XY})}$,and $\boldsymbol{\Sigma}_m^{(\boldsymbol{YX})}$ are the GMM $\lambda^{(\boldsymbol{Z})}$modelling parameters, referring to Section 2.3.3. Finally the label prediction $\widehat{\boldsymbol{Y}}_t$ is estimated as the point which achieves the maximum probability of $P(\boldsymbol{Y}_n|\boldsymbol{X}_n, \lambda^{(Z)})$.

6.2.2 Inter-rater variability incorporation

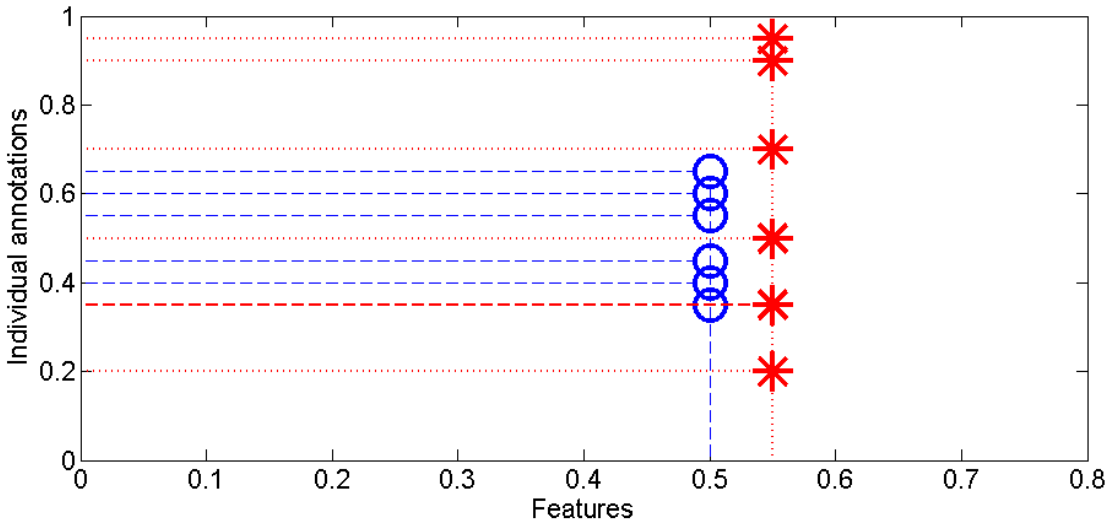
To incorporate inter-rater variability into the conventional GMR model, the r^{th} individual rating $Y_{n,r}$ at frame n is concatenated with the frame-wise features X_n to generate the joint vector $Z_{n,r}$ as:

$$\boldsymbol{Z}_{n,r} = [\boldsymbol{X}_n, \boldsymbol{Y}_{n,r}] \tag{6.5}$$

Given R raters in total, R joint vectors are generated per frame as:



(a)



(b)

Figure 6.1: An example of 2-dimensional joint vector generation with and without incorporation of inter-rater variability. Joint vectors are represented as red dots. (a) Joint vectors for two frames without incorporation of inter-rater variability in the conventional GMR system. (b) Joint vectors for two frames with incorporation of inter-rater variability in the proposed GMR system.

$$\boldsymbol{Z}_n = [\boldsymbol{Z}_{n,1}^T, \boldsymbol{Z}_{n,2}^T, \cdots, \boldsymbol{Z}_{n,R}^T]^T \tag{6.6}$$

As shown in Figure 6.1(a), the conventional GMR develops the joint vectors by concatenating the feature X_n and Y_n at frame n, which results in one joint vector for each frame n represented by the red dot. The joint vectors at these two frames were shown to be close, indicating similar emotion intensity. Assuming six annotators in total for one database, the proposed incorporation of the inter-rater variability, as in Figure 6.1(b), generates the joint vectors Z_{nr} by concatenating the features X_n with each individual ratings Y_{nr} . This results in six joint vectors for each frame sharing the same feature

$\boldsymbol{X}_n$ but different $\boldsymbol{Y}_{nr}$. For frame 1, six joint vectors displays close while they are shown apart for frame 2. Though the conventional joint vectors of these two frames in Figure 6.1(a) are similar, but our newly generated joint vectors that incorporate inter-rater variability indicate that it is of high uncertainty to obtain the emotion intensity for frame 2 when compared to frame 1. In this way, inter-rater variability is introduced in the joint vectors.

6.2.3 Predicting label distribution

During the test phase, the conditional posterior $P\left(\boldsymbol{Y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$ is estimated for each test feature vector $\boldsymbol{X}_n$ as in equation (6.1), where $\boldsymbol{Y}_n = [\boldsymbol{y}_n^T, \Delta\boldsymbol{y}_n^T]^T$. Note that $\boldsymbol{Y}_n$ contains the label and the delta label, but the aim of the multi-rater system is to find the uncertainty information related to the label y_n only, instead of $\boldsymbol{Y}_n = [\boldsymbol{y}_n^T, \Delta\boldsymbol{y}_n^T]^T$. Therefore, I focus on the prediction of $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$. In order to obtain $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$, we can marginalise $P\left([\boldsymbol{y}_n^T, \Delta\boldsymbol{y}_n^T]^T \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$ over $\Delta\boldsymbol{y}_n^T$ as:

$$\begin{aligned} P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right) &= \int_{\Delta\boldsymbol{y}_n} P\left(\boldsymbol{Y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right) d(\Delta\boldsymbol{y}_n) \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) \int_{\Delta\boldsymbol{y}_n} P\left([\boldsymbol{y}_n^{\mathrm{T}}, \Delta\boldsymbol{y}_n^{\mathrm{T}}]^{\mathrm{T}} \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right) d(\Delta\boldsymbol{y}_n) \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right) \end{aligned} \tag{6.7}$$

It can be seen that $P(m|\boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}$ is consistence as in equation (6.1). The second term on the right side of equation (6.8) $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right)$ is a Gaussian distribution. To obtain the representation of $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right)$, first $P\left(\boldsymbol{Y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right)$ (equation (6.2)) can be rewritten with separated information of $\boldsymbol{y}_n$ and $\Delta\boldsymbol{y}_n$ as:

$$\begin{aligned} P\left(\boldsymbol{Y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right) &= N\left(\boldsymbol{Y}_n; \boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}, \boldsymbol{D}_m^{(\boldsymbol{Y})}\right) \\ &= N\left([\boldsymbol{y}_n^T, \Delta\boldsymbol{y}_n^T]^T; \begin{bmatrix} \boldsymbol{u}_m^{(\boldsymbol{y}_n)} \\ \boldsymbol{u}_m^{(\Delta\boldsymbol{y}_n)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\boldsymbol{y}_n)} & \boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n\Delta\boldsymbol{y}_n)} \\ \boldsymbol{\Sigma}_m^{(\Delta\boldsymbol{y}_n\boldsymbol{y}_n)} & \boldsymbol{\Sigma}_m^{(\Delta\boldsymbol{y}_n\Delta\boldsymbol{y}_n)} \end{bmatrix}\right) \end{aligned} \tag{6.8}$$

where m is the mixture number and w_m is the weight for each mixture. $\boldsymbol{u}_m^{(\boldsymbol{y}_n)}$ and $\boldsymbol{u}_m^{(\Delta \boldsymbol{y}_n)}$represent the mean vectors of the m^{th}mixture component for original labels and delta labels respectively; the matrices $\boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n \boldsymbol{y}_n)}$and $\boldsymbol{\Sigma}_m^{(\Delta \boldsymbol{y}_n \Delta \boldsymbol{y}_n)}$ represent the covariances of the m^{th} mixture for the original and delta labels; and $\boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n \Delta \boldsymbol{y}_n)}$ and $\boldsymbol{\Sigma}_m^{(\Delta \boldsymbol{y}_n \boldsymbol{y}_n)}$ are the cross-covariance matrices of the m^{th} mixture for the original and delta labels. Then $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right)$ can be obtained as:

$$P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, m, \lambda^{(\boldsymbol{Z})}\right) = N(\boldsymbol{y}_n; \boldsymbol{u}_m^{(\boldsymbol{y}_n)}, \boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n \boldsymbol{y}_n)}) \tag{6.9}$$

The parameters of $\boldsymbol{u}_m^{(\boldsymbol{y}_n)}$ and $\boldsymbol{\Sigma}_m^{(\boldsymbol{y}_n \boldsymbol{y}_n)}$ can be gained using the $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ terms in (6.8). $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$ is found to be a GMM as in equation (6.7).

It was found that the approximated algorithm described in Section 2.3.2 shows comparable performance as the EM algorithm [81, 83]. Therefore, I adopted the approximated algorithm which simplifies the uncertainty estimation in this framework.

For each frame n, the suboptimal mixture component number $\hat{m}_n$ is firstly estimated as

$$\hat{m}_n = arg \max_{m_n} P(m | \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}) \tag{6.10}$$

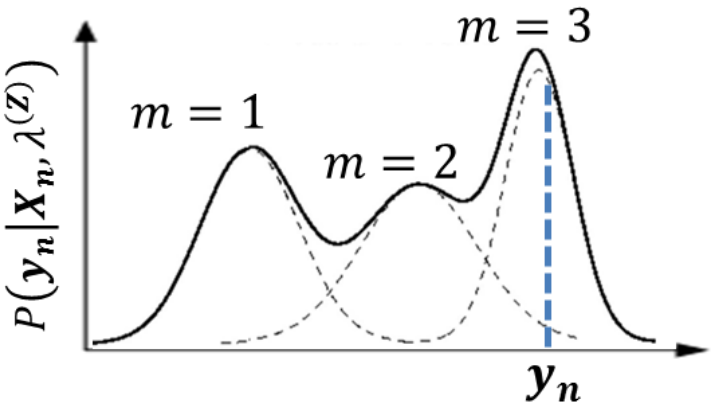
where $\hat{m}_n$ represents the mixture component to which $\boldsymbol{X}_n$ is most probably belongs. This suboptimal mixture component $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \hat{m}_n, \lambda^{(\boldsymbol{Z})}\right)$ (the $\hat{\mathrm{m}}_{\mathrm{n}}$ Gaussian mixture component) is used to approximate the overall GMM distribution $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$, as shown in Figure 6.2. The third mixture component of $P\left(\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right)$ is adopted to approach the overall distribution as in Figure 6.2(b), as $\hat{m}_n = 3$. This then allows for the estimation of emotion intensity $\hat{\boldsymbol{y}}_n$, as the expected value of $\boldsymbol{y}_n$, where

$$\hat{\boldsymbol{y}}_n = \boldsymbol{E}\left[\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \hat{m}_n, \lambda^{(\boldsymbol{Z})}\right] = \boldsymbol{u}_{\hat{m}_n}^{(\boldsymbol{y}_n)} \tag{6.11}$$

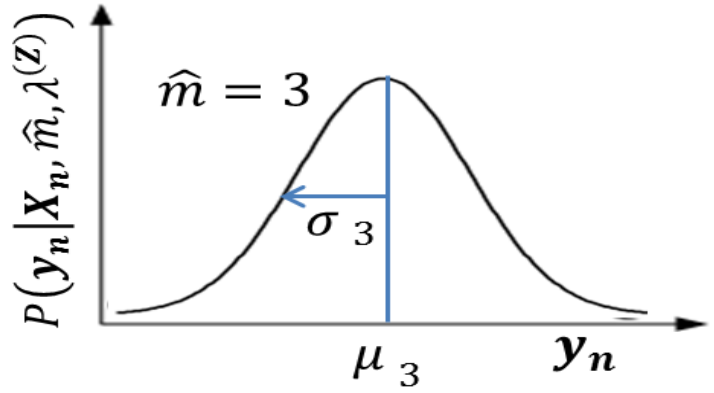
and the time-varying indicator of prediction uncertainty in $\hat{y}_n$ for each frame as its standard deviation $\hat{\boldsymbol{\sigma}}_n$:

$$\hat{\boldsymbol{\sigma}}_n = \boldsymbol{D}\left[\boldsymbol{y}_n \middle| \boldsymbol{X}_n, \hat{m}_n, \lambda^{(\boldsymbol{Z})}\right] = \boldsymbol{\Sigma}_{\hat{m}_n}^{(\boldsymbol{y}_n \boldsymbol{y}_n)} \tag{6.12}$$

where $\boldsymbol{D}[\cdot]$ in (6.12) denotes the computation of standard deviation.



(a)



(b)

Figure 6.2: An example of suboptimal mixture component $\hat{m}_n = 3$ approaching the overall distribution $P(y_n|X_n, \lambda^{(Z)})$. (a) A 3 mixture GMM representing $P(y_n|X_n, \lambda^{(Z)})$; (b) The 3rd dominant mixture is used to approach $P(y_n|X_n, \hat{m}_n, \lambda^{(Z)})$, and the mean and standard deviation parameters are estimated.

It should be noted that the uncertainty prediciton $\hat{\sigma}_n$ which is derived from the covariance matrix $D_m^{(Y)}$ in (6.4) is fixed for each mixture m across all frames and does not vary with the frame-based test features X_n. Consequently, the standard deviation $\hat{\sigma}_n$ will only take one of M distinct values as adopting the dominant mixture in each frame n. This will result in a quantised standard deviation $\hat{\sigma}_n$ for the uncertainty prediction. The quantisation can be improved by increasing the GMM mixture number, but this is constrained by the amount of available training data. The RECOLA database used in this work contains 90 minutes of speech and it was empirically found that it can only be used to train GMMs of 16 mixtures or less.

This new paradigm aims to predict the uncertainty of emotional labels, as opposed to conventional point estimation that indicates the exact emotion intensity. The proposed framework reveals the time-varying nature of the level of human emotion certainty.

6.3 Experimental Settings and Results

6.3.1 Experimental settings

The RECOLA database was used to verify the effectiveness of the proposed method (Section 4.3.2). 65 low-level descriptors (LLDs) and their first-order derivatives were extracted using OpenSMILE [130], using the same LLDs and delta features as [131]. Two second windows with 40 ms shift were used to compute the statistical features by applying five functionals: maximum, minimum, mean, standard deviation, and range. Dynamic features and labels were calculated as in [97]. PCA was used to conduct dimensionality reduction in the feature space from 650 dimensions to 40 dimensions, preserving approximately 85% of the data variance in the training dataset [17, 97]. Dynamic features $\Delta \boldsymbol{x}_n$ calculated on $\boldsymbol{x}_n$ were concatenated with the statistic feature, which results in 80 dimensional feature vector in total. The arousal and valence prediction systems are implemented independently. Thus the label vector $[\boldsymbol{y}_n^T, \Delta \boldsymbol{y}_n^T]^T$ is of 2 dimensions and further concatenated with the feature vector, generating 82 dimensional joint vectors for GMR training. The reason for using 80 feature dimensions was to preserve enough feature variability in the training dataset and to provide a sufficiently high dimensionality to train the system with a large number of parameters for GMM. Delays of 4s for arousal and 2s for valence were applied during the training phase, based on a previous study [29]. The delay thus introduced in the predicted uncertainty was compensated for by removing the corresponding frames. GMMs with different numbers of mixture components ranging from 4 to 32 were tested using HTK. The same feature set and back-end are adopted for the conventional emotion prediction system for a direct comparison.

6.3.2 Experimental results

6.3.2.1 Performance of uncertainty prediction

Under the assumption that a high inter-rater variability $\tilde{\sigma}_n$ will result in a high predicted uncertainty $\hat{\boldsymbol{\sigma}}_n$, we aim to investigate the positive correlation between the predicted standard deviation $\hat{\boldsymbol{\sigma}} = [\hat{\boldsymbol{\sigma}}_1, \hat{\boldsymbol{\sigma}}_2, \cdots \hat{\boldsymbol{\sigma}}_N]$ and the multi-rater standard deviation σ calculated from six ratings, referred as inter-rater variability $\boldsymbol{\sigma}$. Pearson's correlation coefficient (CC) was computed on the 9 development utterances individually. The final performance measure was the mean correlation coefficient averaged over the 9 evaluation utterances. The results are shown in Table 6.1. A moving averaging filter was utilised to relax the quantisation of the predicted standard deviation for each utterance, with an optimal window size determined from [100, 800] experimentally. It should be mentioned that I do not have a specific baseline for a direct comparison owing to the new paradigm which predicts distributions instead of point estimates; therefore our main aim is to reveal the essence of the uncertainty.

Table 6.1: Mean CC computed between $\hat{\boldsymbol{\sigma}}$ and $\boldsymbol{\sigma}$. No smoothing means the mean CC between the quantized $\hat{\boldsymbol{\sigma}}$ and ground truth $\boldsymbol{\sigma}$ smoothing means the mean CC between the smoothed $\hat{\boldsymbol{\sigma}}$ and groud truth $\boldsymbol{\sigma}$.

		No Smoothing		Smoothing	
		Arousal	Valence	Arousal	Valence
Mixtures	4	0.3530	0.0461	0.5173	0.0890
	8	***0.4097***	***0.0937***	***0.5684***	***0.1322***
	16	0.3998	0.0457	0.5350	0.0745
	32	0.3872	0.0476	0.4410	0.0881

The best performance of 0.4097 and 0.5684 were achieved with 8 mixture components before and after smoothing, and there was not much variation between different mixture numbers for arousal. This indicates the positive correlation between the predicted $\hat{\boldsymbol{\sigma}}$ and the inter-rater variability $\boldsymbol{\sigma}$ to some extent. However, the performance of valence is worse for all mixture numbers, agreeing with previous

studies [29] that showed valence is well predicted by a video signal. Consequently, in this chapter I mainly focus on analysing the predicted uncertainty for arousal.

In order to better understand the predicted uncertainty, the scatter plot of the predicted $\hat{\boldsymbol{\sigma}}$ with smoothing and the inter-rater variability σ is investigated over the entire test dataset as shown in Figure 6.3. It can be observed a positive correlation between the predicted $\hat{\sigma}$ in the x-axis and the inter-rater variability $\boldsymbol{\sigma}$ in the y-axis. This indicates a relatively strong correlation between the predicted $\hat{\sigma}$ and the inter-rater variability $\boldsymbol{\sigma}$, which means a high inter-rater variability will result in a high uncertainty in prediction.

In addition, one segment of the ratings from 6 raters of Speaker 2 is shown in Figure 6.4, which

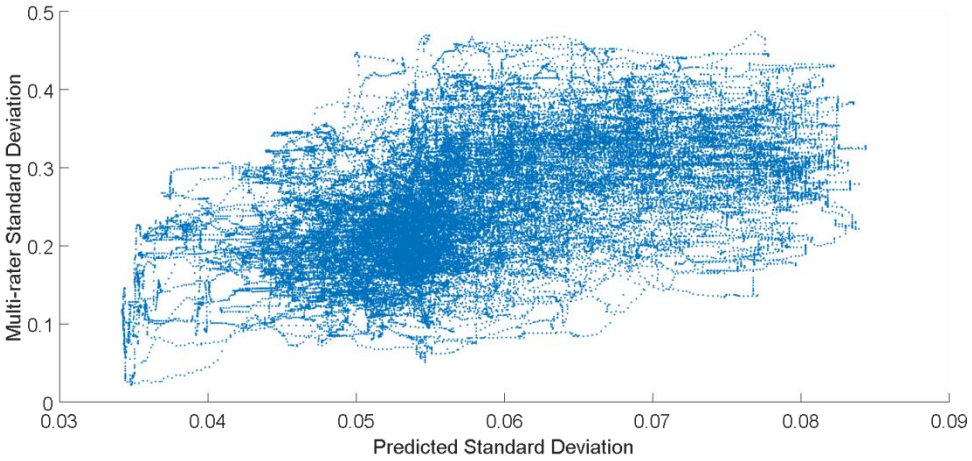


Figure 6.3: Scatter plot of the smoothed predicted standard deviation σ and the inter-rater σ̃. A positive correlation is observed.

displays the uncertainty of the emotion prediction changing over time. Only one speech segment from speaker 2 is shown in order to reduce clutter but the predicted uncertainty was observed to be generally consistent with the inter-rater variability across all speakers and speech segments. The grey error bar shows the predicted standard deviation $\pm\hat{\boldsymbol{\sigma}}_n$ from the expectation value $\hat{\mathbf{y}}_\mathrm{n}$ for each frame n as shown in equations (6.11) and (6.12). Six coloured lines indicate the individual raters' ratings. It can be seen that the speech segments with higher inter-rater variability are associated with higher variability in predicted estimates and vice versa.

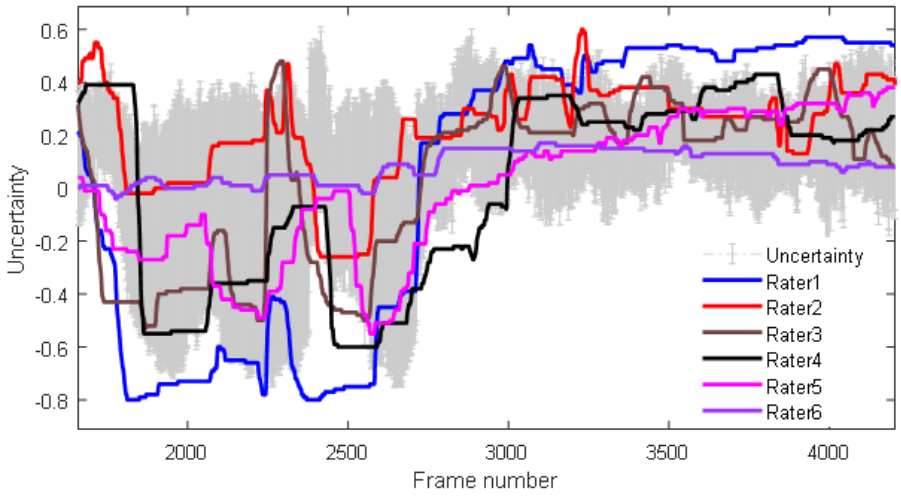


Figure 6.4: Six raters' ratings (coloured lines) and predicted uncertainty (grey error bar) from a speech segment from Speaker 2. The predicted uncertainty changes in concordance with the inter-rater variability, i.e. predicted uncertainty becomes smaller when six annotators have high agreement (frames 3000-4500).

6.3.2.2 Correlation between uncertainty and conventional emotion prediction for arousal

In order to gain an in-depth understanding of this paradigm for support of speech based emotion prediction, the performance of conventional emotion prediction systems that use the average rating as the ground truth is also analysed, by comparing the prediction accuracy on specific speech segments, which are determined based on the uncertainty prediction obtained by the proposed method. Since valence generally performs worse with speech signal only, I mainly analysed the speech based arousal prediction. Generally, low inter-rater variability is represented by low predicted uncertainty, and indicates that raters were more in agreement about the emotion expressed in those speech segments. Thus the prediction should be easier to make accurately. Given the inter-rater variability or the predicted uncertainty information, conventional emotion prediction performance for arousal is investigated by segmenting the speech frames into two regions: low variability regions where affect should be easier to predict from the speech segments since there is less uncertainty, and high variability regions, where affect should be harder to predict from the speech segments since there is greater uncertainty.

In terms of defining low and high variability regions, the predicted uncertainty should be considered instead of the inter-rater variability, since the latter is not generally accessible during the testing phase in real scenarios. I propose using percentiles based on the histogram of

predicted uncertainty $\hat{\sigma}_n$ to determine the low and high variability regions. Given the threshold ρ as a percentile, the speech segments with $\hat{\sigma}_n$ smaller than the value of the ρ^{th} percentile and higher than the value of $(100-\rho)^{th}$ percentile were clustered as low and high variability regions respectively. For instance, $\rho^{th} = 50$ indicates the split of data in half, and a lower ρ^{th} loses data from the middle range. These regions will be referred as predicted low and high uncertainty regions. Five thresholds of $\rho \in [10, 50]$ with a step increase of 10 for the segmentation of low and high uncertainty regions were investigated.

One important precondition of this analysis is that the predicted uncertainty used to define high and low uncertainty regions should be accurate. To guarantee it, similarly the low and high uncertainty regions are defined using the ground truth of six annotations. These are generally not accessible during test phase, but we aim to reveal the insights and it can be helpful to use the ground truth as a reference. The standard deviation among six annotations is computed to represent the inter-rater variability/uncertainty. Then the low and high uncertainty regions are similarly defined by applying the threshold of ρ^{th} percentile to the histogram of these standard deviations, referred as truly low and high uncertainty regions. Then the accuracy of the speech frames in the predicted low (high) uncertainty regions is calculated, with speech frames in the truly low (high) uncertainty regions serving as ground truth. However, the accuracy in predicted low uncertainty regions is found to be relatively low, as shown in Figure 6.5. Here, the black line indicating the percentage of correctly predicted uncertainty frames in low variability regions, which only achieves 20% accuracy when using $\rho = 10$ and $\rho = 20$. Thus, segmenting low and high variability regions only based on the predicted $\hat{\sigma}_n$ may result in misleading comparison of the conventional emotion systems. Therefore, the truly low and high regions were used as a reference for the predicted $\hat{\sigma}_n$ to define low and high variability regions, serving as an initial analysis of the correlation between the uncertainty and conventional point estimation.

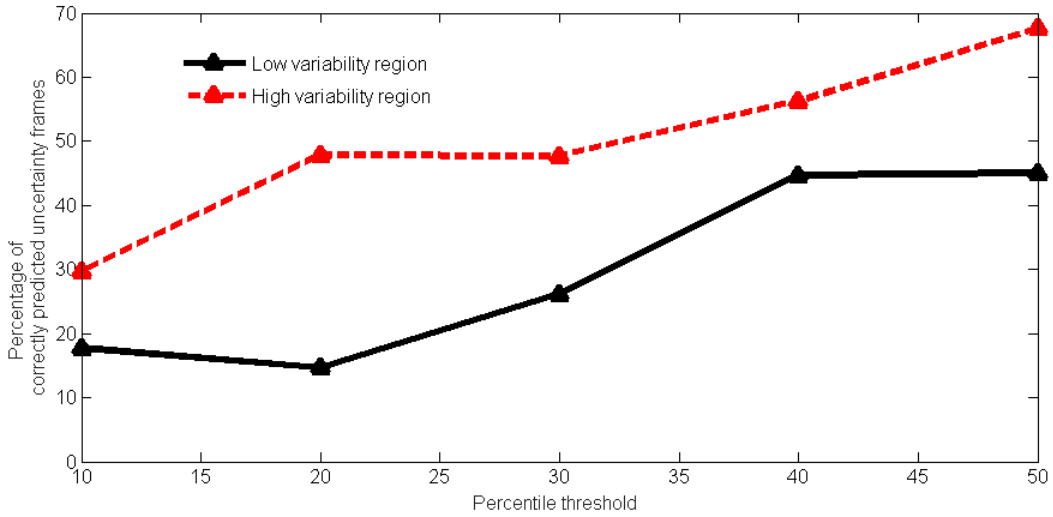


Figure 6.5: Percentage of correctly predicted uncertainty frames in low and high variability regions with 8 mixtures The low and high uncertainty regions are defined using the thresholds $(< \rho^{th})$ *and* $(> (100 - \rho^{th}))$.

To avoid mis-categorising speech segments to low and high variability regions, only the speech segments clustered to both the predicted and truly low/high uncertainty regions are selected, by finding the intersection segments that appear in these two regions. The Pearson's correlation coefficient, which is calculated between the conventional emotion predictions and the ground truth (i.e. mean ratings), is then adopted to evaluate the system performance within high and low regions separately. The CCs for each region are finally compared to reveal the impact of the uncertainty prediction on the conventional emotion prediction systems, shown in Table 6.2.

As expected, it can be seen that the performance in the lower variability regions is in general much better than in the higher variability regions when they are defined using 10th-40th

Table 6.2: CC in low and high variability regions based on the intersection of predicted $\hat{\boldsymbol{\sigma}}$ and the inter-rater variability $\boldsymbol{\sigma}$. Percentiles (ρ) 10^{th} to 50^{th} indicate the regions of the histograms of $P(\hat{\boldsymbol{\sigma}})$ and $P(\boldsymbol{\sigma})$ used to determine low and high variability regions.

Region	Percentiles ρ				
	10th	20th	30th	40th	50th
Low	***0.8013***	***0.7055***	***0.6891***	***0.6905***	0.6885
High	0.3787	0.3837	0.4829	0.6125	0.6905

percentiles. In addition, the conventional arousal prediction system that uses the mean of the labels as the ground truth is also evaluated over the entire test dataset instead of the low and high variability regions only, which achieves an optimal CC of 0.7990 using 8 mixtures. The CC of 0.8013 in the low variability region defined using 10th percentiles (Table 6.2) outperforms the CC of 0.7990 over the entire dataset, suggesting that the conventional emotion system outputs more reliable predictions in these regions with low uncertainty. It should be noted that the comparison between the conventional performance using entire test dataset and Table 6.2 is not a direct comparison, since a different number of speech frames is used to compute the CC for each uncertainty region.

6.4 Summary

This chapter proposes a novel paradigm that is able to incorporate uncertainty about the emotion labels of speech frames by explicitly accounting for inter-rater variability in the system. The results of this investigation show the effectiveness of the proposed method for uncertainty prediction. The second interesting finding is the high correlation between the uncertainty and the emotion prediction achieved by using the average value over multiple raters. The predictions are more reliable in the lower inter-rater variability regions than that in the higher inter-rater variability regions. As the first study to analyse the uncertainty related to emotion prediction, these findings provide more insights into the time-varying variability introduced in emotion prediction by multiple raters. The idea of predicting emotion uncertainty can open up opportunities for building classifiers with an "option to reject," where the classifier indicates that it is not confident enough to provide an answer. This direction may be extremely important if these methods are used in practical applications. However, further studies need to be carried out to perfect the relaxation of the constraint of the quantised uncertainty predictions, which is improved in Chapter 8.

7 MODELLING TEMPORAL DEPENDENCIES FOR EMOTION POINT ESTIMATION

7.1 Introduction

The modelling techniques in continuous emotion prediction and the prediction of uncertainty information of underlying emotion states discussed in Chapters 3 to 6 mainly focused on the statistical models such as GMR and RVM, which did not take into account temporal dependencies in emotion prediction. Several studies have reported that modelling of the temporal dependencies is beneficial for emotion prediction [24, 27, 31, 75, 160], which motivates our work on incorporating such temporal dependencies within our framework.

A variety of existing literature has focused on unidirectional and bidirectional long short term memory neural networks (LSTM and BLSTM) [24, 27, 75, 160], due to their ability to model long range temporal dependencies. While LSTMs are able to model the dependency of current prediction on past information, BLSTMs are trained using frame level features in forward and reverse order, making the system aware of both past and future events in relation to the current time step. However, in-depth analyses of BLSTMs modelling temporal dependencies with regards to affective attributes are still lacking. Nicolaou et al. [31] proposed an output-associate RVM (OA-RVM) framework that augments traditional RVM regression by incorporating the temporal dynamics of arousal and valence. It is a two stage regression modelling technique, where the outputs of the first stage regression model are captured by a temporal window and utilised as the input to the second stage modelling using RVM. Comprehensive experimental results suggest its efficiency and advantages over the statistical model of conventional RVM.

In this chapter, I aim to analyse the temporal dependencies in the feature extraction level, and in the two-stage regression framework, which expands the OA-RVM structure in a multimodal diagram. In terms of the feature sets, the most commonly adopted sets in continuous emotion prediction are the

statistical features applied on the low level descriptors (LLDs) as mentioned in Section 2.2.2. They are generally calculated over a small interval or window ranging from 2s to 8s. However, these features only capture the statistical characteristics of the LLDs in a specific window, but cannot capture time-dependent information. Thus, exploration of the dynamic features that capture long-term information is required. Two categories of dynamic features are explored for emotion prediction including the regression delta coefficients and the shifted delta coefficients [161]. In addition, the two-stage regression originating from OA-RVM is extended in a multimodal system with different OA configuration settings, aiming to investigate the effect of different OA frameworks on continuous emotion prediction systems.

The rest of this chapter is organised as follow: the feature sets that take into account the long-term dynamics are discussed in Section 7.2, the OA framework is expanded in Section 7.3. The experimental settings are discussed in Section 7.4, and the work is finally summarised in Section 7.5.

7.2 Dynamic feature extraction

7.2.1 Regression delta coefficients

In order to capture dynamic information in the feature space, regression delta coefficients are utilised. It has been shown that in the case of a single Gaussian model the incorporation of regression delta coefficients in an analogous scenario corresponds to fixed-lag Kalman smoothing [162]. The lag depends on the window length $2L-1$, over which the derivatives are approximated by

$$\Delta \boldsymbol{x}_n = \frac{\sum_{\theta=-L}^{\theta=L} \theta(\boldsymbol{x}_{n+\theta} - \boldsymbol{x}_{n-\theta})}{2\sum_{\theta=-L}^{\theta=L} \theta^2} \tag{7.1}$$

where $\boldsymbol{x}_n$ represent the feature vector at frame n, L is a parameter related to the window size, and $\Delta \boldsymbol{x}_n$ is the regression delta feature vector. In addition to incorporating temporal derivatives in feature space, it was also shown in [81] that for a GMR based emotion prediction system incorporating derivatives in the label space enables the prediction to be affected by the entire sequence of observations. The label derivatives are also calculated as:

$$\Delta \boldsymbol{y}_n = \frac{\sum_{\theta=-L}^{\theta=L} \theta(\boldsymbol{y}_{n+\theta} - \boldsymbol{y}_{n-\theta})}{2\sum_{\theta=-L}^{\theta=L} \theta^2} \tag{7.2}$$

Both the regression delta features and regression delta labels are used to train the GMR models.

The features $\boldsymbol{X}_n$ and labels $\boldsymbol{Y}_n$ at frame n that are used to train a GMR model are obtained by concatenating the original features and labels with regression delta features and labels respectively as:

$$\boldsymbol{X}_n = [\boldsymbol{x}_n, \Delta \boldsymbol{x}_n]^T \tag{7.3}$$

$$\boldsymbol{Y}_n = [\boldsymbol{y}_n, \Delta \boldsymbol{y}_n]^T$$

The relationship between the final label set $\boldsymbol{Y}_n$ and the original label $\boldsymbol{y}_n$ is a linear mapping:

$$\boldsymbol{Y}_n = \boldsymbol{W}\boldsymbol{y}_n \tag{7.4}$$

If $L = 1$ in equation (7.1), the delta coefficients are calculated between three frames. Equation (7.4) can be rewritten for a sequence of observed labels $\boldsymbol{y} = [\boldsymbol{y_1^T}, \boldsymbol{y_2^T}, \cdots \boldsymbol{y_N^T}]^{\boldsymbol{T}}$ as:

$$\boldsymbol{Y} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0.5 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ -0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & \cdots & \cdots & -0.5 & 0 & 0.5 \\ 0 & \cdots & \cdots & 0 & 0 & 1 \\ 0 & \cdots & \cdots & 0 & -0.5 & 0 \end{bmatrix} \cdot \boldsymbol{y} \tag{7.5}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, \Delta \boldsymbol{y}_1, \boldsymbol{y}_2, \Delta \boldsymbol{y}_2 \cdots \boldsymbol{y}_N, \Delta \boldsymbol{y}_N]^{\boldsymbol{T}}$. Increasing the length of the temporal window enables longer term dynamics to be incorporated in the system. The joint Gaussian mixture mode $\lambda^{(\boldsymbol{Z})}$ is developed based on the joint vector $\boldsymbol{Z}_n = [\boldsymbol{X}_n, \boldsymbol{Y}_n]$. During the test phase, the emotion prediction is estimated based on the conditional probability $P(\boldsymbol{Y}|\boldsymbol{X}, \lambda^{(\boldsymbol{Z})})$ for the entire sequence of labels using the EM algorithm as described in Section 6.2. After incorporating the delta regression coefficients of the labels, the emotion prediction is updated in each iteration as:

$$\hat{\boldsymbol{y}} = \left(\boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} \tag{7.6}$$

Compared the equation (2.42), the weight matrix $\boldsymbol{W}$ is utilised in the prediction of $\hat{\boldsymbol{y}}$, and the inverse of $\boldsymbol{W}$ enables the final prediction at each step to be affected by the entire label sequence. In equation (7.6), $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}}$ is

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} = diag[\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}}, \overline{\boldsymbol{D}_2^{(\boldsymbol{Y})^{-1}}} \cdots, \overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}}, \cdots \overline{\boldsymbol{D}_N^{(\boldsymbol{Y})^{-1}}}] \tag{7.7}$$

and

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}^{(\boldsymbol{Y})}} = [\overline{\boldsymbol{D}_1^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_1^{(\boldsymbol{Y})}}^T, \overline{\boldsymbol{D}_2^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_2^{(\boldsymbol{Y})}}^T, \cdots, \overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_n^{(\boldsymbol{Y})}}^T, \cdots \overline{\boldsymbol{D}_N^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_N^{(\boldsymbol{Y})}}^T] \tag{7.8}$$

Further the submatrices $\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}}$ and $\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_n^{(\boldsymbol{Y})}}$ are calculated as

$$\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} = \sum_{m=1}^{M} \gamma_{m,n}\, \boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}} \tag{7.9}$$

$$\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_n^{(\boldsymbol{Y})}}^T = \sum_{m=1}^{M} \gamma_{m,n}\, \boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}}\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})} \tag{7.10}$$

The scalar $\gamma_{m,n}$ in equation (7.9) and (7.10) is the posterior probability of mixture m as:

$$\gamma_{m,n} = P\left(m\middle|\boldsymbol{X}_n, \boldsymbol{Y}_n, \lambda^{[\boldsymbol{Z}]}\right) = \frac{w_m N(\boldsymbol{Z}_n; \boldsymbol{u}_m, \boldsymbol{\Sigma}_m)}{\sum_{k=1}^{M} w_k N(\boldsymbol{Z}_n; \boldsymbol{u}_k, \boldsymbol{\Sigma}_k)} \tag{7.11}$$

where w_m, $\boldsymbol{u}_m$, and $\boldsymbol{\Sigma}_m$ are the parameters of the GMM model $\lambda^{[\boldsymbol{Z}]}$, referring to equation (4.65). The $\boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}$ and $\boldsymbol{D}_m^{(\boldsymbol{Y})}$ are the mean and covariance matrix of the m^{th} mixture component at frame n, similar as in equation (4.69) and (4.70). It can be seen that the inverse of the weight matrix $\boldsymbol{W}$ in equation (7.6) is used in the prediction of $\hat{\boldsymbol{y}}$, which enables the final prediction at each step to be affected by the entire label sequence. In this way, the temporal dependencies have been incorporated in the system by using delta regression coefficients.

7.2.2 Shifted delta coefficients

The shifted delta cepstra (SDC) feature vector, which captures the dynamic information across a time segment has been shown to improve language identification systems in [163]. The SDC feature vector is created by stacking delta cepstra/features computed across multiple frames. Compared to the

standard delta feature vector, SDC feature vector is able to capture long-term temporal dependencies in the feature space, which motivates our investigation into the use of SDC feature vectors in continuous emotion prediction systems.

The SDC feature vector is controlled by a set of four parameters, N, d, P and k, as shown in Figure 7.1. N represents the number of cepstral coefficients computed at each frame, d represents the time advance or delay for the delta computation, k is the number of blocks whose delta coefficients are concatenated to form the final feature vector, and P is the time shift between consecutive blocks. The feature vector $\Delta c(t+iP)$ at the $(i+1)^{th}$ block at frame t is

$$\Delta c(t+iP) = c(t+iP+d) - c(t+iP-d) \tag{7.12}$$

where $0 \leq i \leq k-1$. The final SDC feature vector $\Delta c_f(t)$ at frame t is generated by concatenating all the $\Delta c(t+iP)$

$$\Delta c_f(t) = [c(t), c(t+P), \cdots, c(t+(k-1)P)] \tag{7.13}$$

The SDC feature vectors $\Delta c_f(t)$ are then utilised to replace the conventional statistical features for the regression modelling.

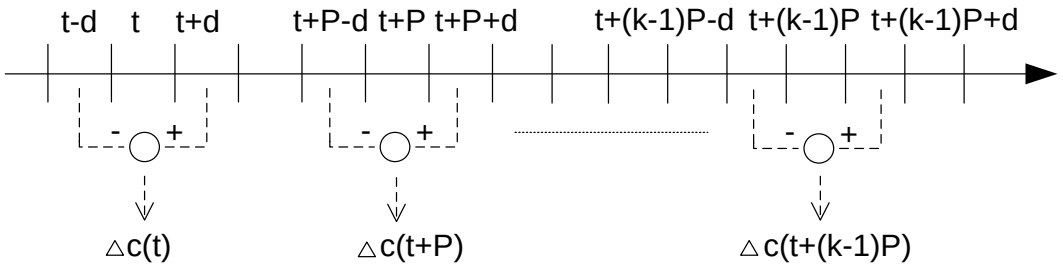


Figure 7.1: Computation of the SDC feature vector at frame t for parameters $N-d-P-k$ [161]

7.3 Output-Associate framework

Output-associative (OA) techniques are gaining popularity in continuous emotion prediction [31, 75, 147]. These techniques take into account the contextual and temporal dependencies that exist within and between predicted arousal and valence values when performing fusion. In this section, the OA

structure originated from OA-RVM [31] has been extended to three configurations in a multimodal framework: OA fusion, OA regression, and OA regression incorporating uncertainty, as shown in Figure 7.2. OA fusion utilised the temporal dependencies in the arousal and valence predictions only, OA regression additionally considers the input feature space, and OA regression with uncertainty predictions further incorporated the long-term dynamics of the predictions of uncertainty information about emotion states. These three configurations will be discussed in Sections 7.3.1 to 7.3.3 in details, and will be easily adapt to a multimodal setting in Section 7.3.4.

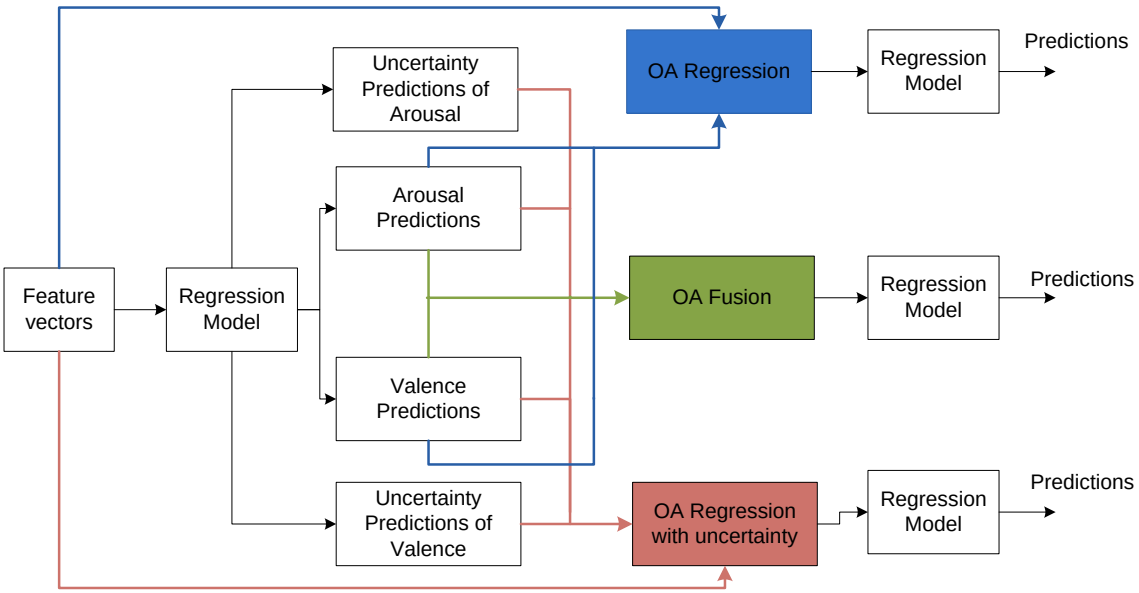


Figure 7.2: Block diagram of OA fusion, OA regression and OA regression with uncertainty. OA fusion utilised the temporal dependencies in the arousal and valence predictions only, OA regression additionally considers the input feature space, and OA regression with uncertainty predictions further incorporated the long-term dynamics of the predictions of uncertainty information about emotion states.

7.3.1 Output-associate fusion

OA-fusion is an extension of decision-level fusion and is achieved by learning the fusion weights of an OA-matrix. The OA-matrix is formed by output-associative vectors from a set of initial predictions, taken from each dimension as shown in Figure 7.3.

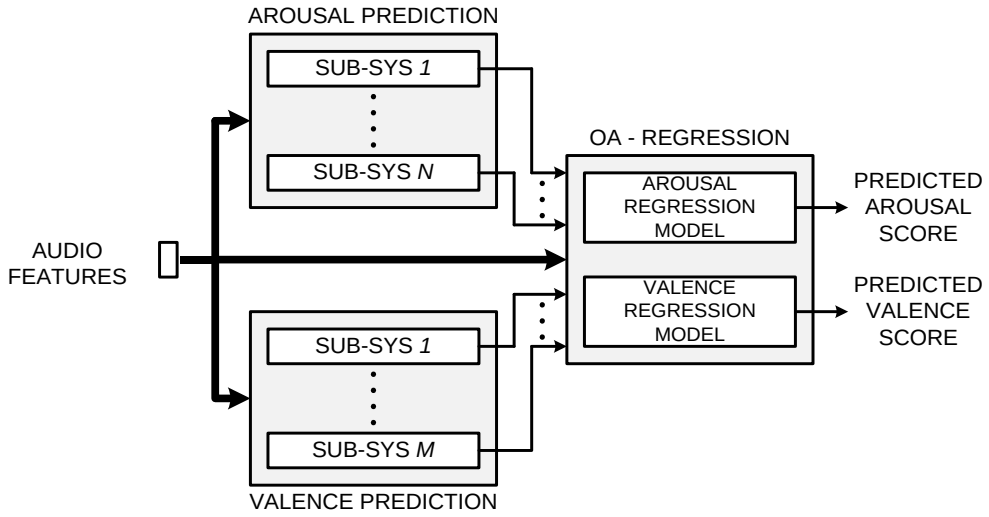


Figure 7.3: Block diagram showing the output-associative regression strategy for the task of speech-based continuous emotion prediction.

As shown in Figure 7.3, generally multiple regression modelling techniques or features are tested in speech-based continuous emotion prediction systems, resulting in N subsystems for arousal prediction and M subsystems valence prediction. An N or M equal to 1 indicates that only one main system is used for arousal or valence predictions. In this thesis, I did not treat arousal and valence differently, thus N and M were always equal. Let $\widetilde{\boldsymbol{y}}_{am}$ and $\widetilde{\boldsymbol{y}}_{vm}$ represent the arousal and valence prediction of m^{th} subsystem respectively. The associated OA-matrix $\boldsymbol{Y}_{OA}^{m}$ can be formed in terms of arousal and valence predictions as below:

$$\boldsymbol{Y}_{\boldsymbol{OA}}^{\boldsymbol{m}} = [\widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}] \tag{7.14}$$

where $\widetilde{\boldsymbol{y}}_{am}^{i+n}$ and $\widetilde{\boldsymbol{y}}_{vm}^{i+n}$ are the prediction values at time $i+n$ and i ranges in $[-\frac{\boldsymbol{P}}{\boldsymbol{2}}, \frac{\boldsymbol{P}}{\boldsymbol{2}}]$; and $\boldsymbol{Y}_{OA}^{m}$ represents the OA matrix for the m^{th} subsystem; and P denotes the length of the temporal window for OA matrix. By simply concatenating the predictions within a window, the long-term dynamics within time range $[-\frac{\boldsymbol{P}}{\boldsymbol{2}}, \frac{\boldsymbol{P}}{\boldsymbol{2}}]$ are introduced for each frame, and regression modelling techniques will automatically learn the relationship between current frame and the past and future information. In this way, the temporal dependencies are incorporated by the OA framework in continuous emotion prediction systems.

The OA fusion matrix is generated by concatenating the arousal and valence predictions of all subsystems within a longer window size as:

$$\boldsymbol{Y_{OA}} = [\boldsymbol{Y_{OA}^1}, \boldsymbol{Y_{OA}^2}, \cdots, \boldsymbol{Y_{OA}^m}, \cdots, \boldsymbol{Y_{OA}^M}] \tag{7.15}$$

The $\boldsymbol{Y}_{OA}$ is then utilised as input to the second stage regression models. With increasing length P of the temporal window, the dimension of $\boldsymbol{Y}_{OA}$ increases dramatically. In order to minimise the linearity between feature dimensions referred as multi-collinearity effects, either a RVM or regularised linear regression (RLR) is used to learn the fusion weights since RVM is able to conduct feature selection and RLR helps prevent model from over fitting the high dimensional training data.

To explain this mathematically, first let $\widetilde{\boldsymbol{Y}}_a$ and $\widetilde{\boldsymbol{Y}}_v$ represent the OA matrix of M subsystems for arousal and valence respectively:

$$\widetilde{\boldsymbol{Y}}_a = [\underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{a1}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{a1}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{a1}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ 1}, \cdots, \underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{am}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ m}, \cdots, \underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{aM}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{aM}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{aM}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ M}] \tag{7.16}$$

$$\widetilde{\boldsymbol{Y}}_v = [\underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{v1}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{v1}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{v1}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ 1}, \cdots, \underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{vm}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ m}, \cdots, \underbrace{\widetilde{\boldsymbol{y}}_{\boldsymbol{vM}}^{-\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}, \cdots, \widetilde{\boldsymbol{y}}_{\boldsymbol{vM}}^{\boldsymbol{i}+\boldsymbol{n}}, \cdots \widetilde{\boldsymbol{y}}_{\boldsymbol{vM}}^{\frac{\boldsymbol{P}}{\boldsymbol{2}}+\boldsymbol{n}}}_{subsystem\ M}] \tag{7.17}$$

where the concatenation $[\widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v]$ for frame n are referred as OA vector. Based on either RVM or RLR as the second stage regression techniques, OA fusion aims to learn a set of weights $[\boldsymbol{\varphi}_a, \boldsymbol{\psi}_a]$ and $[\boldsymbol{\varphi}_v, \boldsymbol{\psi}_v]$ for arousal and valence predictions, such that

$$\breve{\boldsymbol{y}}_{\boldsymbol{a}}^{\boldsymbol{n}} = (\boldsymbol{\varphi_a})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi_a})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + \boldsymbol{\epsilon} \tag{7.18}$$

$$\breve{\boldsymbol{y}}_{\boldsymbol{a}}^{\boldsymbol{n}} = (\boldsymbol{\varphi_a})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi_v})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + \boldsymbol{\epsilon} \tag{7.19}$$

where $\breve{\boldsymbol{y}}_a^n$ and $\breve{\boldsymbol{y}}_v^n$ represent the arousal and valence prediction at frame n, and $\boldsymbol{\epsilon}$ represents the learned noise. During the training phase, the OA vector $[\widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v]$ for each frame n is treated as the new feature vector, and training data referred as OA matrix is generated by concatenating the OA vectors of all frames. Regression techniques are applied to the OA matric to learn the fusion weights.

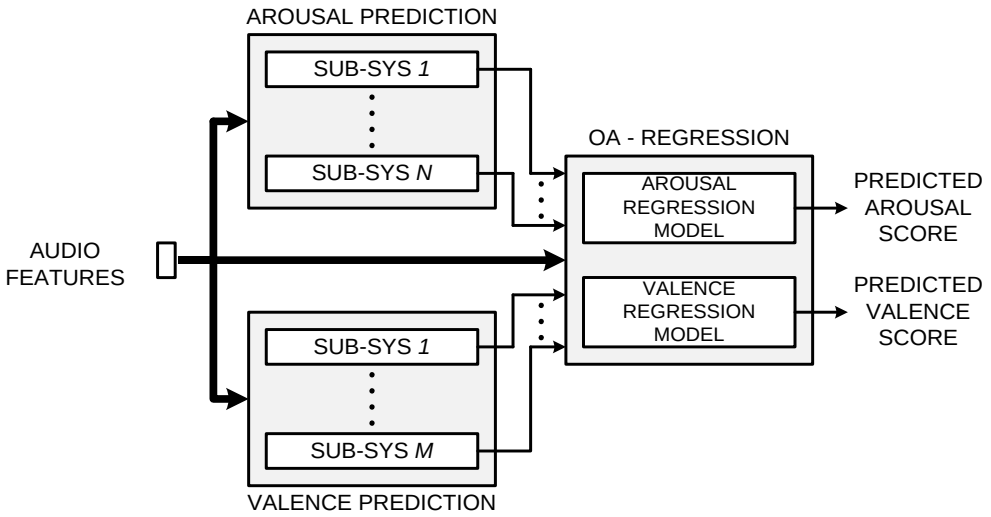


Figure 7.4: Block diagram showing the output-associative regression strategy for the task of speech-based continuous emotion prediction.

7.3.2 Output-associate regression

The other fusion strategy is a combined feature-level, decision-level and OA fusion scheme, herein referred to as output-associative regression (OA regression). This system is an extension of the OA fusion, in which the OA matrix is concatenated with the input feature space to learn the fusion weights as shown in Figure 7.4 (shown in bold arrow). Fusion of dimensions using this system will be performed using the OA-RVM [31].

The OA-RVM technique extends the contextual and temporal mapping performed in OA fusion to also incorporate the relationship between the input feature space $\boldsymbol{x}$ when updating the prediction values, such that the final prediction values $\breve{\boldsymbol{y}}_a^n$ ad $\breve{\boldsymbol{y}}_v^n$ (similar to equations (7.18) and (7.19) for OA fusion) are:

$$\breve{\boldsymbol{y}}_{\boldsymbol{a}}^{\boldsymbol{n}} = (\boldsymbol{\omega}_{\boldsymbol{a}})^{\boldsymbol{T}}\boldsymbol{\phi}(\boldsymbol{x}) + (\boldsymbol{\varphi}_{\boldsymbol{a}})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi}_{\boldsymbol{a}})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + \boldsymbol{\epsilon} \tag{7.20}$$

$$\breve{\boldsymbol{y}}_{\boldsymbol{v}}^{\boldsymbol{n}} = (\boldsymbol{\omega}_{\boldsymbol{v}})^{\boldsymbol{T}}\boldsymbol{\phi}(\boldsymbol{x}) + (\boldsymbol{\varphi}_{\boldsymbol{v}})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi}_{\boldsymbol{v}})^{\boldsymbol{T}}(\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + \boldsymbol{\epsilon} \tag{7.21}$$

where $\widetilde{\boldsymbol{Y}}_a$ and $\widetilde{\boldsymbol{Y}}_v$ are the temporal independently-learnt set of arousal and valence prediction values as in equation (7.16) and (7.17), and these are continuous on the range $\left[n - \frac{P}{2}, n + \frac{P}{2}\right]$. $\boldsymbol{\phi}(\boldsymbol{x})$ represents

the input features after applying kernel $\boldsymbol{\phi}(\cdot)$, and $[\boldsymbol{\omega}_a, \boldsymbol{\omega}_v]$ are the learnt weights for input features for arousal and valence.

OA-RVM therefore uses the past, current and future prediction context associated with the input feature frames, as well as the input features, to update a prediction result. Prediction using the non-causal relationship has been shown to be superior to RVM and SVR when performing continuous emotion prediction [164]. The work presented within this section aims to reinforce the usefulness of the OA-RVM framework and furthermore explore this paradigm in terms of a multimodal fusion technique in Section 7.3.4.

During the training phase, the OA vector $[\widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v]$ for each frame n is concatenated with the original feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ to generate the feature vector $\boldsymbol{x}'_n = [\boldsymbol{\phi}(\boldsymbol{x}), \widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v]$, and the training data referred as OA matric are generated by concatenating features $\boldsymbol{x}'_n$ of all frames. Regression techniques are applied to the OA matrix to learn the fusion weights. During the test phase, similarly the OA matrix is generated and the trained weights are used for final emotion predictions.

7.3.3 Output-associate regression incorporating uncertainty prediction

The prediction about uncertainty of the underlying emotion states can be further incorporated as additional information in OA regression. The emotion predictions about uncertainty of the underlying emotion states are adopted as an additional feature set, and OA regression will be further extended to incorporate the relationship between the predictions about uncertainty of the underlying emotion states when updating the prediction values, as in equations (7.18) and (7.19), and (7.20) and (7.21). These prediction values are

$$\breve{\boldsymbol{y}}_{\boldsymbol{a}}^{\boldsymbol{n}} = (\boldsymbol{\omega}_{\boldsymbol{a}})^T \boldsymbol{\phi}(\boldsymbol{x}) + (\boldsymbol{\varphi}_{\boldsymbol{a}})^T (\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi}_{\boldsymbol{a}})^T (\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + (\boldsymbol{\chi}_{\boldsymbol{a}})^T (\widetilde{\boldsymbol{U}}_a) + (\boldsymbol{\varsigma}_{\boldsymbol{a}})^T (\widetilde{\boldsymbol{U}}_v) + \boldsymbol{\epsilon} \quad (7.22)$$

$$\breve{\boldsymbol{y}}_{\boldsymbol{v}}^{\boldsymbol{n}} = (\boldsymbol{\omega}_{\boldsymbol{v}})^{\boldsymbol{T}} \boldsymbol{\phi}(\boldsymbol{x}) + (\boldsymbol{\varphi}_{\boldsymbol{v}})^{\boldsymbol{T}} (\widetilde{\boldsymbol{Y}}_{\boldsymbol{a}}) + (\boldsymbol{\psi}_{\boldsymbol{v}})^{\boldsymbol{T}} (\widetilde{\boldsymbol{Y}}_{\boldsymbol{v}}) + (\boldsymbol{\chi}_{\boldsymbol{v}})^T (\widetilde{\boldsymbol{U}}_a) + (\boldsymbol{\varsigma}_{\boldsymbol{v}})^T (\widetilde{\boldsymbol{U}}_v) + \boldsymbol{\epsilon} \quad (7.23)$$

where $\widetilde{\boldsymbol{U}}_a$ and $\widetilde{\boldsymbol{U}}_v$ are the temporal independently-learnt set of arousal and valence predictions about uncertainty of the underlying emotion states values (similarly developed as in equations (7.16) and (7.17)), continuous on the range $\left[n - \frac{P}{2}, n + \frac{P}{2}\right]$; the concatenation $[\widetilde{\boldsymbol{U}}_a, \widetilde{\boldsymbol{U}}_v]$ for frame n are referred as

OA uncertainty vector; and $[\chi_a, \varsigma_a, \chi_v, \varsigma_v]$ represent the weights learned from the predictions about uncertainty of the underlying emotion states of arousal and valence.

During the training phase, the OA vector $[\widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v]$ at frame n, the OA uncertainty vector $[\widetilde{\boldsymbol{U}}_a, \widetilde{\boldsymbol{U}}_v]$ at frame n are concatenated with the original feature vector $\boldsymbol{\phi}(\boldsymbol{x})$ to generate the feature vector $\boldsymbol{x}'_n = [\boldsymbol{\phi}(\boldsymbol{x}), \widetilde{\boldsymbol{Y}}_a, \widetilde{\boldsymbol{Y}}_v \widetilde{\boldsymbol{U}}_a, \widetilde{\boldsymbol{U}}_v]$, and OA matric are generated by concatenating features $\boldsymbol{x}'_n$ of all frames. Regression techniques are applied to the OA matrix to learn the fusion weights. During the test phase, similarly the OA matrix is generated and the trained weights are used for final emotion predictions.

7.3.4 Multimodal output-associate fusion and regression

The OA frameworks described in Sections 7.3.1 to 7.3.3 can be extended to fusion of multiple modalities in a straightforward manner. A complete multimodal OA-matrix $\boldsymbol{Y}_{all}$ can be formed by combining the OA matrices from all modalities:

$$\boldsymbol{Y_{all}} = [\boldsymbol{Y_{OA}}(\mathbf{1}), \boldsymbol{Y_{OA}}(\mathbf{2}), \cdots, \boldsymbol{Y_{OA}}(\boldsymbol{M})] \tag{7.24}$$

where $\boldsymbol{Y}^i_{OA}$ represents the OA matrix of the i^{th} modality and M represents the total number of

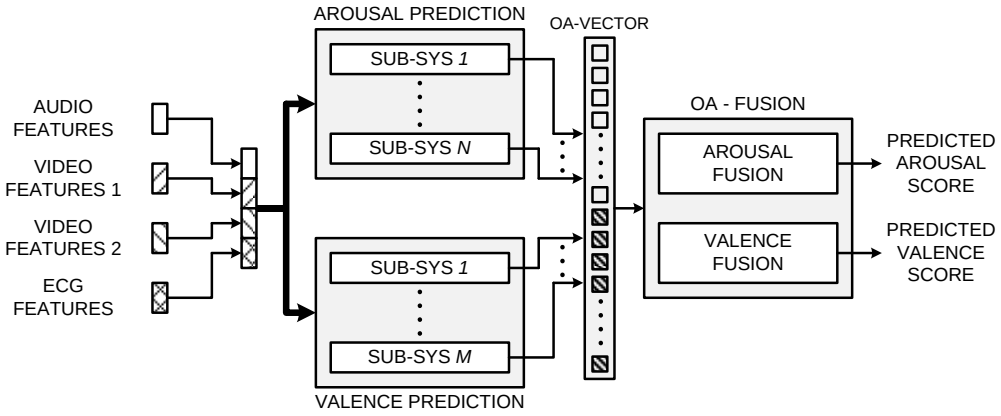


Figure 7.5: Block diagram showing the output-associative fusion strategy used to combine information from different modalities for the task of continuous emotion prediction.

modalities. The three fusion schemes outlined in Sections 7.3.1 to 7.3.3 can be extended to the multimodal case: Figure 7.5 shows the multimodal OA fusion, Figure 7.6 demonstrates the multimodal OA regression, and Figure 7.7 displays multimodal OA regression with uncertainty

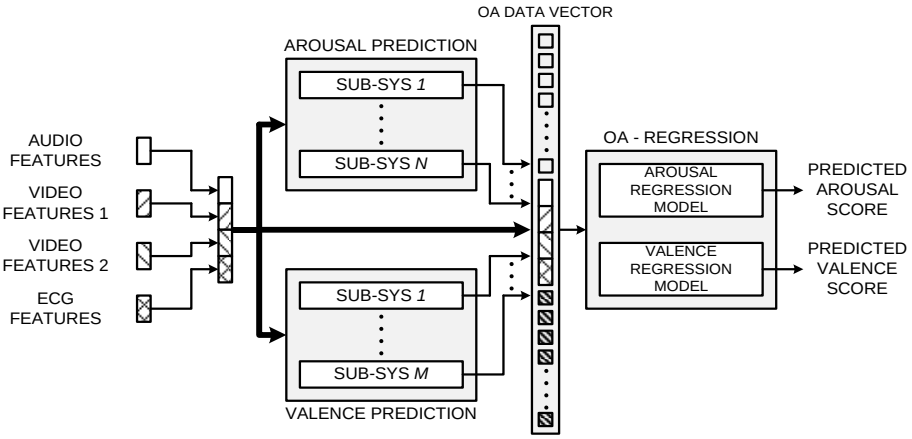


Figure 7.6: Block diagram showing the output-associative regression strategy used to combine information from different modalities for the task of continuous emotion prediction.

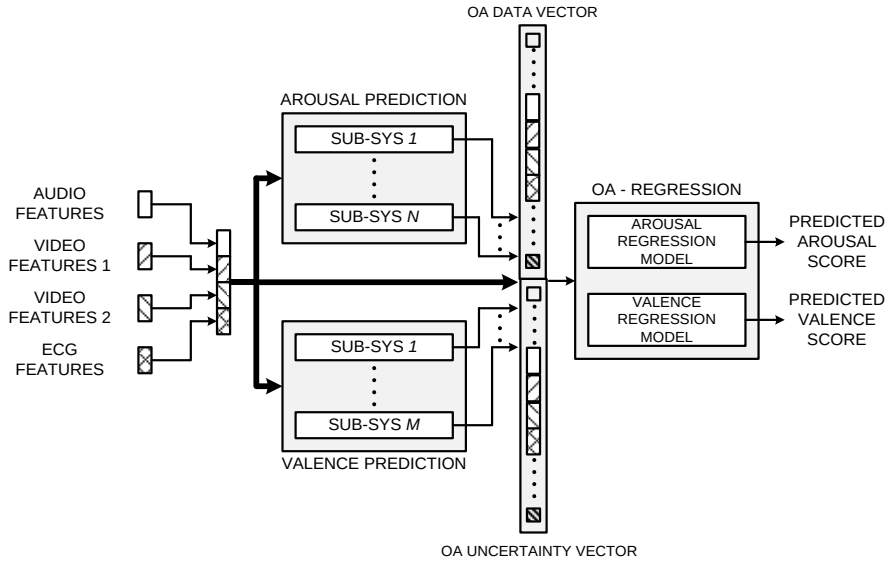


Figure 7.7: Block diagram showing the output-associative regression with uncertainty strategy used to combine information from different modalities for the task of continuous emotion prediction.

incorporation. Fusion between modalities is simply a case of training a regressor with the OA-matrix. Again, to minimise multi-collinearity effects, either a RVM or RLR was used to learn the fusion weights.

7.4 Experimental settings and results

7.4.1 Dynamic feature extraction

7.4.1.1 Experimental settings

The USC CreativeIT database was used to verify the impact of incorporating the dynamic features, and the experiments were carried out in a leave-one-session-out framework. Compensation for the reaction lags as discussed in Chapter 5 were first carried out with 4s and 2s for arousal and valence prediction respectively. The same 650 dimensional features were first extracted as in Section 6.3.1. These features were directly reduced to 80 dimensions using PCA for the baseline system without any dynamic feature extraction.

Firstly, PCA was used to reduce the feature dimension to 40. Then the regression delta coefficients were calculated for each of the 40 dimensions and concatenated with the original 40 dimensional features, resulting in a total of 80 dimension features. This aims for a fair comparison using 80 dimensional features for all systems. GMR modelling techniques are utilised as the back-end, where the three dimensional labels of arousal, valence and dominance ratings were jointly modelled. The systems with and without regression delta coefficients are compared. These three systems are: one without any regression delta coefficients, one only including regression delta coefficients of features, and another utilising both the regression delta coefficients of features and labels to verify the combined impact of features and labels. The experiments are carried out in a multi-dimension prediction framework, where arousal, valence and dominance are simultaneously predicted.

In terms of the SDC features, different combinations of $N-d-P-k$ are investigated as in Section 7.2.2. To keep the comparison consistent, PCA was still adopted to reduce the SDC feature dimensions to 80. However, the SDC features projected from an extreme high dimension could not maintain the original information comprehensively and could not be modelled properly by GMR, thus the investigation on the SDC features with GMM backend was limited. In total, GMM with 4, 8 and 16 mixture components was tested with different feature sets.

Table 7.1: Correlation coefficients of system performance with and without regression delta coefficients for a 4-mixture GMM.

	Arousal	Valence	Dominance
No regression delta coefficients	0.4765	0.1639	0.1652
Regression deltas of features only	0.4822	0.1472	0.1582
Regression deltas of features and labels	***0.5643***	***0.2148***	***0.2226***

Table 7.2: Correlation coefficients of system performance with SDC features, using different SDC parameters.

	CC		
Parameters $N-d-P-k$	**Arousal**	**Valence**	**Dominance**
65-3-2-2	0.5094	0.2061	0.1772
65-4-2-2	0.4578	0.2001	0.1316
65-6-2-2	0.3140	0.0892	0.0715
65-3-1-2	0.4787	0.2456	0.1888
65-2-1-2	***0.5103***	***0.2907***	***0.2139***
GMR baseline without regression deltas	0.4765	0.1639	0.1652
GMR baseline with regression deltas	***0.5643***	***0.2148***	***0.2226***

7.4.1.2 Experimental results

The comparison of systems with and without regression delta coefficients is shown in Table 7.1. Different mixture components were tested and a system with 4 mixture components was found to achieve the best performance, thus the comparison was conducted based on the 4-mixture GMM. It can be seen that only incorporating the regression delta coefficients for features improves the results for arousal slightly, but not for valence and dominance. However, further incorporating the delta label coefficients dramatically improves the system performance for all three affective attributes, suggesting the significance of the dynamic label information. Therefore, the GMR based continuous emotion prediction systems all incorporate the regression delta coefficients throughout this thesis, including the GMR based experiments in Chapters 4, 6, 7, and 8.

The results of the system performance with SDC features are shown in Table 7.2. Compared to the GMR baseline without regression deltas, the SDC features showed superior performance with some

experimental settings, highlighting the effectiveness of incorporating more temporal information. However, it did not outperform the system with regression delta coefficients for arousal and dominance predictions. These results suggest that that valence prediction may benefit more from the incorporation of longer-term temporal dependencies.

7.4.1 Output-associate framework

7.4.1.1 Experimental settings

The RECOLA [76] and SEWA [65] databases are used to verify the extension of the OA framework to a multimodal context. Three modalities in the RECOLA database are used in our experiments: audio, video and physiology. The SEWA database does not provide physiological signals and instead audio,video and text modalities are used in the experiments run on this database. Note that the extension of OA framework has been mainly proposed for the Audio/Video Emotion Challenge (AVEC) 2015 and 2017, thus the databases utilised are different. AVEC is a competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual and physiological depression and emotion analysis, with all participants competing under strictly the same conditions. The RECOLA databased was used in AVEC 2015 and the SEWA database was used in AVEC 2017. In AVEC 2015, multimodal OA-fusion and OA regression are carried out in the RECOLA database to verify the effectiveness of OA framework for temporal dynamics modelling; and In AVEC 2017, OA regression with and without predictions about uncertainty of the underlying emotion states are investigated in the SEWA database.

In the RECOLA database, audio, video and physiology signals are provided in the RECOLA database. All the experiments are conducted using the training partition and evaluated using the validation partition. Four feature sets from three modalities have been adopted: the 102-dimensional *Extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) (adopted in AVEC 2015 [76]), two types of video descriptors (84-dimensional set of facial based appearance features and 316-dimensional set of facial based geometric features); and the 54-dimensional set of electro-cardiogram (ECG) features. Apart from the provided feature sets, audio features were also extracted using

VoiceSauce and openSMILE. The reader is referred to [76] for a complete description of the challenge feature sets. SVR and RVM were initially used as the first stage regression methods. Before training, 4s and 2s delay compensation for arousal and valence was introduced as in Chapter 5. All training features were then normalised to [0, 1] and the normalisation coefficients were subsequently used to normalise the test data. A binomial filter was used for post processing.

For SVR, one out of each 20 frames of training data was selected for training for reasons of computational efficiency, with negligible performance drop. A linear kernel was used and *C* was set to 0.005 and 0.05 for arousal and valence respectively, based on optimising the performance with delay and smoothing in the development set in the range of $[10^{-4},1]$. The number of RVM training iterations was set to 30 for arousal and 40 for valence, based on the best performance on the development data. When performing OA fusion and OA regression, the number of RVM training iterations was tailored for each system, ranging between 10-100.

In the SEWA database, the training and evaluation partitions are the same as the training and development partitions from AVEC 2017 [65]. OA regression with and without predictions about uncertainty of the underlying emotion states is verified using audio, video and text modalities, and the system overview is shown in Figure 7.8. The audio features include a set of 88-dimensional eGeMAPS features and a set of BoW representation applied on the LLDs. Video features included a set of 121-dimensional features describing facial position and expression of subjects, and a set BoW representations was applied on the video features. For text-based features, a set of BoW representations was again applied on transcripts to form 521-dimensional frame-level features. Readers can refer to [65] for the details of these provided features. Accompanying the provided features, another novel feature set were applied for emotion prediction: phonetically-aware acoustic features [165]. GMR, Gaussian process (GP) and RVM are the three probabilistic models that output the additional emotion predictions about uncertainty of the underlying emotion states as discussed in Section 3.3 (readers can refer to [166] for GP), but RVM was experimentally found to perform worse for emotion uncertainty prediction. Therefore, the predictions about uncertainty of the underlying

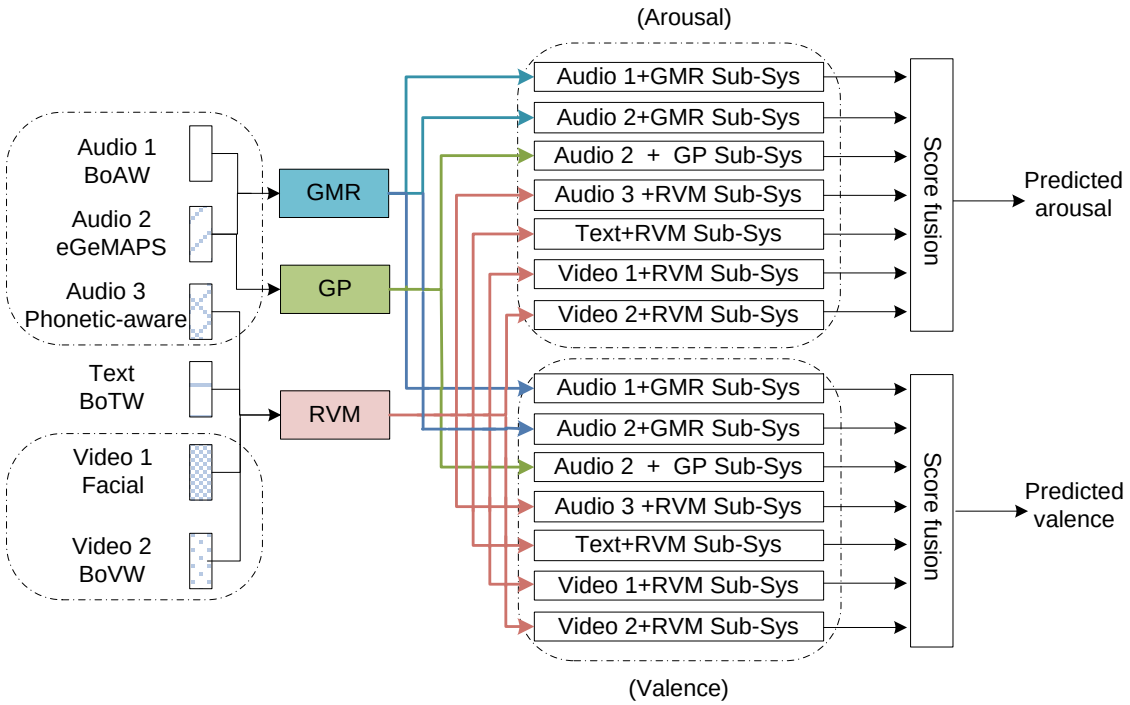


Figure 7.8: Overview of arousal and valence prediction systems on the SEWA database. Three modalities including three audio, one text and two video feature sets are investigated, and three back-ends consisting of GMR, GP and RVM are applied to different feature sets which are chosen experimentally. In total, seven subsystems are developed for both arousal and valence.

emotion states obtained by GMR and GP are used in the OA fusion framework. The performance measure adopted is the concordance correlation coefficient (CCC).

7.4.1.2 Experimental results

Output-associative (OA) fusion techniques seek to utilise the temporal dynamics among first stage predictions and the correlations between arousal and valence values [167]. As discussed in Section 7.4.2.1, four subsystems are developed using either SVR or RVM based on the four features individually. The results are shown in Table 7.3.

All OA fusion systems were tested using either RLR or RVM, represented as OA-RLR and OA-RVM, outperformed the decision-level fusion (using either RLR or RVM without OA structure) for both arousal and valence predictions. These results indicate that consistent significant improvements can be obtained by incorporating the temporal dependencies using OA fusion.

Table 7.3: CCC for OA fusion using either a RVM or RLR to learn the fusion weights. SVR and RVM are utilised as the 1st stage regression techniques; Fusion techniques include decision-level fusion and OA fusion; decision-level fusion using both RLR and RVM are compared to OA fusion using RLR (OA-RLR) and RVM (OA-RVM).

1st stage regression	**Fusion**	**CCC**	
		Arousal	**Valence**
SVR	**RLR**	0.684	0.524
	RVM	0.667	0.436
	OA-RLR	***0.736***	***0.615***
	OA-RVM	0.718	0.509
RVM	**RLR**	0.274	0.495
	RVM	0.648	0.458
	OA-RLR	0.447	***0.578***
	OA-RVM	***0.710***	0.535

Furthermore, the extension of OA-fusion to OA regression has also been tested. These systems combined feature-level fusion and OA fusion in order to further improve performance. Owing to the high dimensional feature set of the OA matrix, RVM instead of RLR is adopted due to its ability to perform inherent feature selection within the modelling techniques. Thus all OA regression systems used an OA-RVM framework [31], which was extended to fuse a set of predictions learnt from each modality.

Output-associative (OA) regression results with different system configurations are reported in Table 7.4, where C1 describes a system employing OA-RVM to the multimodal arousal and valence predictions obtained using four SVR for each modality at the first stage; and C2 describes a system employing OA-RVM to the multimodal arousal and valence predictions using four RVM for each modality at the first stage; and C3 describes a system employing OA-RVM to the concatenated multimodal features. A further increase in system performance was observed when compared to OA fusion methods. The superior results for OA regression appear very consistent across the different system configurations compared to OA fusion. These results confirm the usefulness of RVM based

OA regression for performing continuous emotion prediction. Taken together, it is concluded that incorporating temporal dependencies by either OA fusion or OA regression improves the system performance, with OA regression a little superior.

In the SEWA database, firstly the performance of features from different modalities is investigated, including audio, video, and text using three types of regression models. The selection of backend regression models was empirical. Among all seven subsystems shown in Table 7.5, system V1 performed the best, achieving 0.518 for arousal and 0.583 for valence in terms of CCC. Note that this result outperformed the multimodal baseline system of AVEC 2017 by a considerable margin [65], and far exceeded that of the video baseline, which uses the same video feature set with LSTM back-ends. Interestingly, system T1 also achieved very good performance, suggesting the suitability of RVMs for emotion prediction. The two acoustic systems of A1 and A2 showed similar performance for arousal, while A1 was outperformed by A2 for valence. This may indicate that the BoW method applied to LLDs is able to capture a more informative representation of emotion information compared to functionals. Although the BoW representation in general provided reasonably good

Table 7.4: Comparison of CCC's using different OA regression systems using a OA-RVM set-up to learn the fusion weights. OA-RVM is used as the fusion technique.

System number	1st stage regression	Arousal	Valence
C1	**SVR (4 subsystems)**	***0.766***	***0.655***
C2	**RVM (4 subsystems)**	0.742	0.608
C3	**RVM (1 system with concatenated features)**	0.743	0.600

performance for both arousal and valence prediction, the representation has a very high dimension, which carries a greater risk of over-fitting.

Table 7.5: Subsystem performances with different features and back-ends in terms of CCC.A indicates audio modalities; T indicates text modalities; V indicates video modalities. In total, four audio, one text and two video systems are presented.

System Number	Features	Back-end	Arousal	Valence
A1	**eGeMAPS**	GMR	0.454	0.446
A2	**BoAW**		***0.451***	***0.515***
A3	**eGeMAPS**	GPR	0.315	0.368
A4	**PA**	RVM	0.400	0.362
T1	**BoTW**		0.441	0.499
V1	**Norm Facial**		***0.518***	***0.583***
V2	**BoVW**		0.397	0.422

Owing to the better system performance of OA regression over OA fusion, OA regression was then further applied to subsets of the seven subsystems to take into account the temporal dynamics of emotion predictions, including up to four audio, two video and one text subsystem; the chosen subsets will be described later. Probabilistic predictions were incorporated within the OA regression as presented in Section 7.2.2.4. Several combinations of the seven subsystems were investigated and the optimal combinations were determined primarily according to the fusion performance of the probabilistic predictions. A comparison of selected subsystem combinations, with and without predictions about uncertainty of the underlying emotion states is reported in Table 7.6. Note that the optimal combination of multiple subsystems is initially determined according to the evaluation performance. However, a more reasonable method could be to employ statistical analyses of how complementary the subsystems are., which can be referred to [168]. The results presented in Table 7.6 are for multimodal fusion of 1) system C1 comprising two audio-only subsystems, 2) system C2 of four audio and one text subsystem, and 3) system C3 containing all audio, text and video subsystems. A comparison among these three combined systems (C1, C2 and C3) suggests that including additional modalities improves CCC performance for both arousal and valence, achieving 0.620 and 0.682 respectively when fusing all subsystems. However, subsystems within the same modality may carry similar information, which could be somewhat redundant during fusion, especially for the very high dimensional BoAW and BoVW features. For this reason, a system is evaluated, where one

subsystem from each modality was selected for fusion, i.e. C4 consisting of A4, T1 and V1. Further, system A1 was included in C5 owing to its good performance shown in Table 7.5. The performance of C4 and C5 improves the CCC of arousal from 0.620 to 0.672 compared to C3, suggesting that dropping the multiples of the same modality did improve system performance; however, this did not aid valence prediction where the performance was actually degraded from 0.682 to 0.605.

Of particular interest is that consistent improvements were observed by incorporating prediction uncertainty across different system configurations for predicting arousal and valence, with relative improvement for arousal between 0.8% and 2.3%, and for valence of maximum 1.4%. There is no uncertainty comparison for the C4 system (**A4+T1+V1**), since there are no probabilistic predictions. Overall, the OA regression results in Table 7.6 considerably outperformed the multimodal baseline reported in AVEC 2017 [65] for both arousal and valence on the development set.

In conclusion, incorporating temporal dynamics of uncertainty of predicted emotion labels in OA regression further improves system performance for both arousal and valence predictions, with little superior performance for arousal.

Table 7.6: Fusion performance of systems from Table 7.5, without and with uncertainty for arousal and valence in terms of CCC.

		CCC			
System number	**Combined Systems (Fused Systems)**	**Arousal**		**Valence**	
		Without uncertainty	With uncertainty	Without uncertainty	With uncertainty
C1	**A1+A2**	**0.490**	**0.494**	**0.500**	**0.507**
C2	**A1+A2+A3+A4+T1**	**0.551**	**0.560**	**0.597**	**0.597**
C3	**A1+A2+A3+A4+T1+V1 +V2**	**0.609**	**0.620**	**0.676**	***0.682***
C4	**A4+T1+V1**	**0.657**	**N/A**	**0.602**	**N/A**
C5	**A1+A4+T1+V1**	**0.657**	***0.672***	**0.605**	**0.605**

7.5 Summary

This chapter aims to incorporate the temporal dependencies into conventional emotion prediction systems, with the main focus on dynamic feature extraction and the output-associate framework. Two dynamic feature sets of regression delta coefficients and shifted delta coefficients are verified based on the GMR backend. The regression delta coefficients were shown to increase the arousal and valence prediction systems dramatically, suggesting their usefulness in capturing changing emotion and the relative dependencies of several adjacent frames. However, the shifted delta coefficients could not outperform the typical statistical features, possibly owing to the high dimensionality of the SDC features not being properly modelled by the GMR.

The output-associate (OA) framework was proposed and extended to OA fusion and OA regression, which have been verified in multimodal settings. It was shown that incorporating temporal dynamics in both OA fusion and OA regression all outperform the typical methods in multiple experimental configurations of different backends, with OA regression having slightly superior results to OA fusion. Furthermore, incorporating the emotion predictions about emotion uncertainty in the OA regression yields better results than that without uncertainty incorporation. All these results together suggest that incorporating temporal dependencies of emotion predictions in OA framework benefits continuous emotion prediction systems, and additionally modelling the long-term dependencies of predictions about emotion uncertainty further improves system performance.

In summary, incorporating temporal dependencies benefits emotion prediction, either in the feature level or the two state regression modelling techniques such as the OA framework. Further, the predictions about uncertainty of the underlying emotion states obtained from the probabilistic model were shown to be useful in emotion prediction when combined in the OA framework, which motivates an in-depth analysis on the temporal dependencies or the evolving nature of the emotion predictions about uncertainty of the underlying emotion states , discussed in Chapter 8.

8 MODELLING TEMPORAL DEPENDENCIES FOR EMOTION UNCERTAINTY PREDICTION

8.1 Introduction

As discussed in Chapter 6, predicting emotion attributes as a distribution using inter-rater variability instead of hard label has shown promising result in the uncertainty information of the emotion labels. However, the GMR regression technique as a statistical model cannot take into account temporal dynamics. The superior performance achieved by incorporating temporal dependencies into conventional emotion prediction systems in Chapter 7 motivated further investigations on the temporal modelling for the label distribution of uncertainty information.

The previous work in Chapter 6 [169] developed the GMR system that incorporated information from multiple raters to predict emotion uncertainty, under the assumption that multi-ratings reflect the uncertainty of speech frames, which showed potential in predicting the emotion uncertainty. However, these methods all assumed that label distribution obtained from multiple raters is a single Gaussian, which may not always be true in reality. Though the work in Chapter 6 estimated the label distribution as a GMM, it was still carried out by taking the dominant Gaussian mixture component of GMM. In addition, incorporation of the long-term dynamics either by LSTM [27, 79] or the extension work on OA structure [31, 93] cannot be directly used to explore the temporal dependencies of uncertainty in emotions, quantified as a distribution. Thus, exploring the temporal uncertainty about emotional states aims to reveal the evolving process of label distributions.

Thus far, only a limited number of papers have considered the prediction of uncertainty in emotions, and even fewer studies have considered their temporal dynamics. Since human emotion is a slowly varying process where current emotion state evolves from the past emotion states, the regression model developed should be able to model the temporal dynamic of emotions. LSTM as one of the effective techniques to model long-term dependencies could not predict label distributions directly. Thus, incorporating filter techniques that considers the temporal dependencies within the GMR

system in Chapter 6 has been proposed. Kalman filters are one of the most widely adopted techniques in time series analysis [170]. They have been explored as a multi-modal or multi-subsystem fusion technique for emotion prediction in recent years, since they are ideally suited for continuous state tracking. Good performance for predicting arousal and valence was observed [171-173]. However, this was still carried out for hard labels of emotion attributes. The work presented in this chapter explores the use of Kalman filters to model the temporal dynamics of a distribution over emotional states that captures the uncertainty in predicted emotions, which are applied to the emotion label distributions instead of hard labels of emotion attributes. Forward and backward Kalman filters are adopted to take into account both past and future information. In addition, the proposed method assumes that the emotion label distribution is a GMM instead of single Gaussian distribution, generalising the assumptions made about the distribution.

Another challenge that arises from this new framework is the question of evaluation metrics between the predicted GMM label distribution and the ground truth GMM distributions. Two measurements are proposed to evaluate the new system performance, in terms of the KL divergence between the predicted and ground truth distributions, and the correlation between the local variability of the predicted and ground truth distributions.

The rest of this chapter is organised as follow: the incorporation of temporal dynamics by Kalman filter is discussed in Section 8.2, and the experimental setting and results are presented in Section 8.3, and finally they are summarised in Section 8.4.

8.2 Incorporation of temporal information for probabilistic estimation

As mentioned in Section 6.2, the GMR system incorporates multi-rater variability in the feature concatenation level, and a GMR is developed to capture the label variability. In order to obtain the uncertainty prediction for test speech, the conditional distribution $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$ of label $\boldsymbol{y}_t$ for each frame t is estimated as a GMM, where λ represents the joint model and $\boldsymbol{x}_t$ represents the feature vector at frame t. An approximation with the dominant mixture component of the GMM is adopted as discussed in Section 6.2.3, shown in Figures 8.1(a) and 8.1(b). Figure 8.1(a) displays the ratings from

6 raters of one speech segment. Figure 8.1(b) shows the prediction $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ approximated as a Gaussian distribution for each frame. This allows for a time-varying indicator of uncertainty prediction as the standard deviation of each frame-wise Gaussian distribution. It is expected that a small standard deviation reflects a low inter-rater variability.

Emotion uncertainty is generally captured by a distribution, thus incorporating the temporal dependencies of emotion uncertainty is focused on the evolving process of label distributions $P(\boldsymbol{y}_t)$. Kalman filters are used to estimate the hidden states $P(\boldsymbol{y}_t)$ based on the previous states $P(\boldsymbol{y}_{1:t-1})$ and current observations, where the predicted conditional distribution $\mathrm{P}(\mathrm{y_t}|\mathrm{x_t},\lambda)$ is treated as a current noisy observation of $P(\boldsymbol{y}_t)$. This framework also relaxes the assumption of label distribution being a Gaussian distribution to allowing it to be a GMM. i.e., Instead of approximating $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$ by a Gaussian distribution, the proposed dynamic multi-rater GMR treats $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$ as GMMs. Readers are referred to Chapter 6 for the details of estimation on the conditional probability $\mathrm{P}(\mathrm{y_t}|\mathrm{x_t},\lambda)$. The vector representation $\mathrm{v_t}$ of $\mathrm{P}(\mathrm{y_t}|\mathrm{x_t},\lambda)$ can then be generated by concatenating their GMM parameter weights $\bar{\mathrm{w}}_{\mathrm{mt}}$, means $\bar{\mathrm{u}}_{\mathrm{mt}}$ and vectorised covariances $\bar{\Sigma}_{\mathrm{mt}}$ of each mixture component m, and $\mathrm{s_t}$ of $\mathrm{P}(\mathrm{y_t})$ can be similarly generated by concatenating $[w_{mt}, \boldsymbol{u}_{mt}, \boldsymbol{\Sigma}_{mt}]$ as:

$$\boldsymbol{v}_t = [\bar{w}_{1t}, \cdots \bar{w}_{M_1 t}, \bar{\boldsymbol{u}}_{1t}^T, \cdots \bar{\boldsymbol{u}}_{M_1 t}^T, Vec(\bar{\boldsymbol{\Sigma}}_{1t})^T, \cdots Vec(\bar{\boldsymbol{\Sigma}}_{M_1 t})^T]^T \tag{8.1}$$

$$\boldsymbol{s}_t = [w_{1t}, \cdots w_{M_2 t}, \boldsymbol{u}_{1t}^T, \cdots \boldsymbol{u}_{M_2 t}^T, Vec(\boldsymbol{\Sigma}_{1t})^T, \cdots Vec(\boldsymbol{\Sigma}_{M_2 t})^T]^T \tag{8.2}$$

where M_1 and M_2 represents the number of mixture components for $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ and $P(\boldsymbol{y}_t)$. Prediction of the hidden states $\boldsymbol{s}_t$ can be formulated as a Kalman filter:

$$P(\boldsymbol{s}_t|\boldsymbol{s}_{t-1}) = N(\boldsymbol{s}_t; \boldsymbol{F}\boldsymbol{s}_{t-1}, \boldsymbol{Q}) \tag{8.3}$$

$$P(\boldsymbol{v}_t|\boldsymbol{s}_t) = N(\boldsymbol{v}_t; \boldsymbol{H}\boldsymbol{s}_{t-1}, \boldsymbol{R}) \tag{8.4}$$

where matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}$ and $\boldsymbol{R}$ are the process matrix, observation matrix, process noise covariance and observation noise covariance, which can be estimated during the training phase. The label distribution $\boldsymbol{s}_t$ can be updated sequentially based on equations (8.3) and (8.4). An illustration of the proposed dynamic multi-rater GMR is shown in Figure 8.1(c). The Kalman filter ensures that the

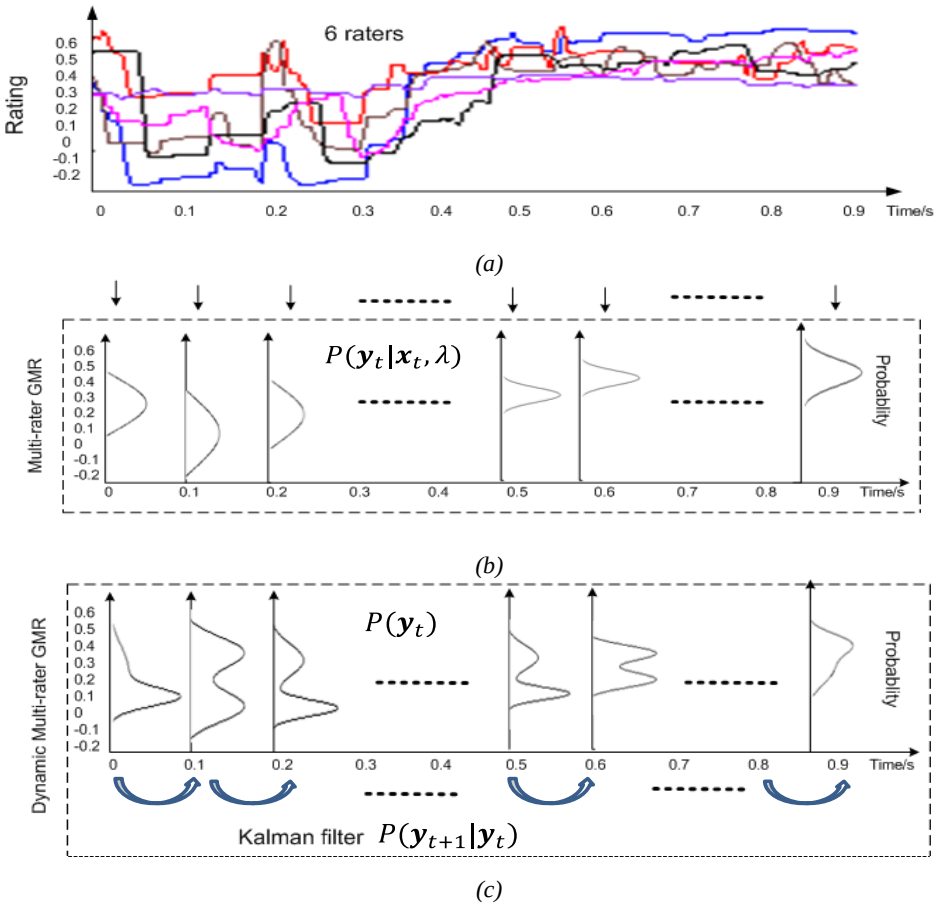


Figure 8.1: Comparison of multi-rater GMR and dynamic multi-rater GMR; (a) six ratings for one speech segment; (b) multi-rater GMR based prediction $P(y_t|x_t,\lambda)$; (c) proposed dynamic multi-rater GMR based prediction $P(y_t)$ using Kalman filter.

hidden state $\boldsymbol{s}_t$ is dependent on previous states, which reduces the negative effect of sudden misleading frames. It should be noted that the Kalman filter is utilised to predict the label distribution $\boldsymbol{s}_t$ instead of hard labels. The uncertainty about predictions can then be quantified based on the label distribution $\boldsymbol{s}_t$, namely $P(\boldsymbol{y}_t)$.

8.2.1 Training phase

As in Section 6.2.2, a joint GMM $\lambda = P(\boldsymbol{x}, \boldsymbol{y})$ is developed and the prediction $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$ is estimated using validation partition. Recall that $P(\boldsymbol{y}_t|\boldsymbol{x}_t,\lambda)$, represented by $\boldsymbol{v}_t$, is regarded as the noisy observation of the hidden states $\boldsymbol{s}_t$. Vectors $\boldsymbol{s}_t$ and $\boldsymbol{v}_t$ are required to train the Kalman matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}$ and $\boldsymbol{R}$ from equations (8.3) and (8.4). Ideally, $\boldsymbol{s}_t$ can be trained directly using the labels from multiple raters at each frame t. However, there are generally a limited number of raters in existing

databases (i.e. 3 or 6), thus it is not reliable to directly train $\boldsymbol{s}_t$ as a GMM. Maximum-a-posterior (MAP) adaptation is used to obtain $\boldsymbol{s}_t$ for each frame based on a UBM trained using all labels in the training partition. $\boldsymbol{v}_t$ can be obtained by predicting $P(\boldsymbol{y}_t|\boldsymbol{x}_t, \lambda)$ for each frame.

Given $\boldsymbol{v}_t$ and $\boldsymbol{s}_t$, the matrices $\boldsymbol{F}, \boldsymbol{H}, \boldsymbol{Q}$ and $\boldsymbol{R}$ of the Kalman filter can be estimated as in [174]. As suggested by [173], introducing an internal delay d during estimation of the process matrix $\boldsymbol{F}$, benefits emotion prediction systems since $\boldsymbol{F}$ cannot be an identity matrix. This is owing to the fact that emotion is a slowly changing process where two adjacent frames are extremely similar. Let $\boldsymbol{A} = (\boldsymbol{s}_{1:t-1-d})^T$ and $\boldsymbol{B} = (\boldsymbol{s}_{d+1:t})^T$. $\boldsymbol{F}$ and $\boldsymbol{Q}$ can be estimated as:

$$\boldsymbol{F} = (\boldsymbol{A}^T\boldsymbol{A} + \lambda \boldsymbol{I})^{-1}\boldsymbol{A}^T\boldsymbol{B} \tag{8.5}$$

$$\boldsymbol{Q} = cov(\boldsymbol{B} - \boldsymbol{A}\boldsymbol{F}) \tag{8.6}$$

where the regularisation parameter λ can be determined experimentally. Similarly, let $\boldsymbol{C} = (\boldsymbol{s}_{1:t})^T$ and $\boldsymbol{D} = (\boldsymbol{v}_{1:t})^T$, $\boldsymbol{H}$ and $\boldsymbol{R}$ can be estimated as:

$$\boldsymbol{H} = (\boldsymbol{C}^T\boldsymbol{C} + \lambda \boldsymbol{I})^{-1}\boldsymbol{C}^T\boldsymbol{D} \tag{8.7}$$

$$\boldsymbol{R} = cov(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{H}) \tag{8.8}$$

8.2.2 Test phase

During the test phase, the predicted label distribution $\boldsymbol{v}_t$ is estimated using the GMR based system outlined in Chapter 6. Initial values of the hidden states $\boldsymbol{s}_0$ and the covariance $\boldsymbol{Q}_0$ are given and the Kalman filter is applied to sequentially predict the hidden states $\boldsymbol{s}_t$, given $\boldsymbol{v}_t$ and the Kalman matrices. The algorithm used to estimate hidden states $\boldsymbol{s}_t$ can be found in [174]. Finally, the predicted distribution $P(\hat{\boldsymbol{y}}_t)$ can be reconstructed by decomposing $\boldsymbol{s}_t$ into GMM parameters and emotion uncertainty can be obtained.

Additionally, the KL divergence was estimated between the predicted GMM distributions and the fitted GMM distributions, which are viewed as the ground truth on a frame basis. This is compared to the KL divergence that estimated between the predicted Gaussian distributions and the fitted Gaussian distributions as in [14].

8.2.3 Forward and backward Kalman filters

Since the Kalman filter only considers the temporal dependencies on past information, two Kalman filters, one trained in the forward direction (KF1), and another in the backward direction (KF2) are proposed to consider the temporal dependencies of both past and future information.

During the test phase, the label distribution $\boldsymbol{s}_t^{KF1}$ and $\boldsymbol{s}_t^{KF2}$ were estimated using KF1 and KF2 respectively. A linear combination of $\boldsymbol{s}_t^{KF1}$ and $\boldsymbol{s}_t^{KF2}$ is used as the final estimation $\boldsymbol{s}_t$,

$$\hat{\boldsymbol{s}}_t = \alpha \boldsymbol{s}_t^{KF1} + (1-\alpha)\boldsymbol{s}_t^{KF2} \tag{8.9}$$

where the linear coefficient α was determined experimentally as

$$\alpha = \underset{\alpha}{argmin} \sum_{t=1}^{N} \|\boldsymbol{s}_t - \hat{\boldsymbol{s}}_t\|_2 \tag{8.10}$$

8.2.4 Uncertainty prediction

It is supposed that a broad GMM indicates a high uncertainty prediction corresponding to high disagreement among multiple raters, while a narrow GMM represents a low uncertainty prediction corresponding to low disagreement. The broadness of a GMM is quantified in terms of the acoustic volume, which can be estimated as the local variability of the GMMs. This local variability of the GMM can also be utilised as the uncertainty prediction. The acoustic volume AVL_t is adopted [126].

Figure 8.2 illustrated $P(\hat{\boldsymbol{y}}_1)$ and $P(\hat{\boldsymbol{y}}_2)$ for time t_1 and t_2. Given a threshold θ, the AVL_t of $P(\hat{\boldsymbol{y}}_t)$ is the red area under θ (Figure 8.2) as:

$$AVL_t = \int f(\boldsymbol{y}) d\boldsymbol{y}, \;\; f(\boldsymbol{y}) = \begin{cases} 1, P(\boldsymbol{y}_t) > \theta \\ 0, P(\boldsymbol{y}_t) \leq \theta \end{cases} \tag{8.11}$$

Details of AVL_t can be found in equation (3.3) in Section 3.2.2, where a set of thresholds have been applied for a PAV profiles, but only one threshold is experimentally selected to the AVL_t calculation. As shown in Figure 8.2, AVL_1 for a broad GMM is larger than AVL_2 for a narrow GMM. These are expected to correspond to two frames with high and low inter-rater variability respectively. The inter-rater variability is treated as the ground truth of emotion uncertainty which is similarly estimated as the acoustic volume of $P(\boldsymbol{y}_t)$ obtained using only multi-ratings.

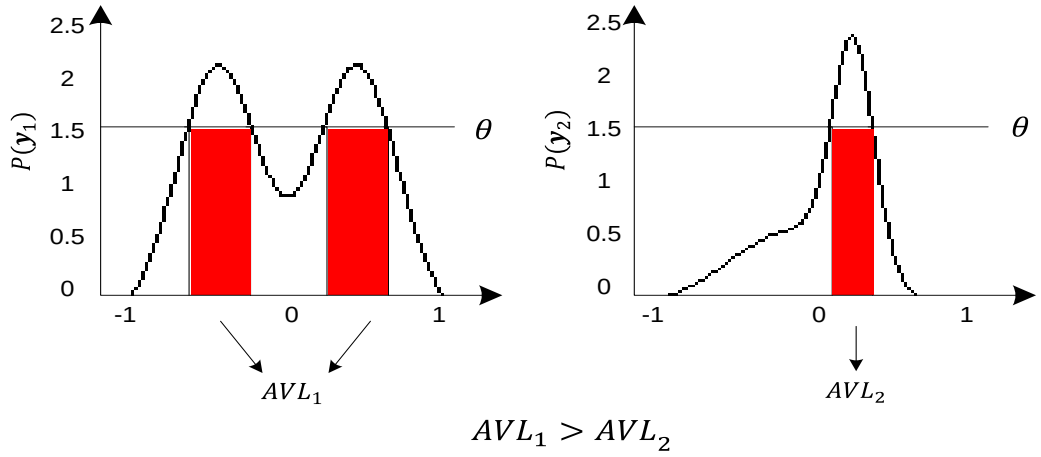


Figure 8.2: Acoustic volume (AVL) of two distributions $P(y_1)$ and $P(y_2)$. Red area under threshold θ is the AVL for GMM.

Additionally to estimate whether the predicted distribution matches the true distribution, Kullback-Leibler (KL) divergence is estimated numerically as in equation (3.1), and the mean and standard deviation of the KL divergence over entire validation partition are reported.

8.3 Experimental settings and results

8.3.1 Experimental settings

The RECOLA database was used to verify the effectiveness of the proposed methods (Section 4.4.1). 65 low-level descriptors (LLDs) and their first-order derivatives are extracted using Opensmile [130]. Five functionals are used to calculate the statistic features [169]. Dynamic features and labels are calculated as in [81]. PCA is used to conduct dimensionality reduction in the feature space from 650 to 40 dimensions. Delays of 4 s for arousal and 2 s for valence are applied as discussed in Chapter 5. GMMs with 2, 4 or 8 full covariance mixture components are tested to model $P(\boldsymbol{y}_t)$, and GMMs with 8 mixture components are used for the joint distribution λ since they showed good performance in Chapter 6.

During the training phase, internal delays of 1, 3, 5, 7 and 9 seconds have been tested for the Kalman filter. The fusion coefficients α for these filters are tested in the range of [0,1] with a step increase of 0.1. A regularisation term for the filters is optimized in the range $[10^{-10}, 10^{5}]$. During the

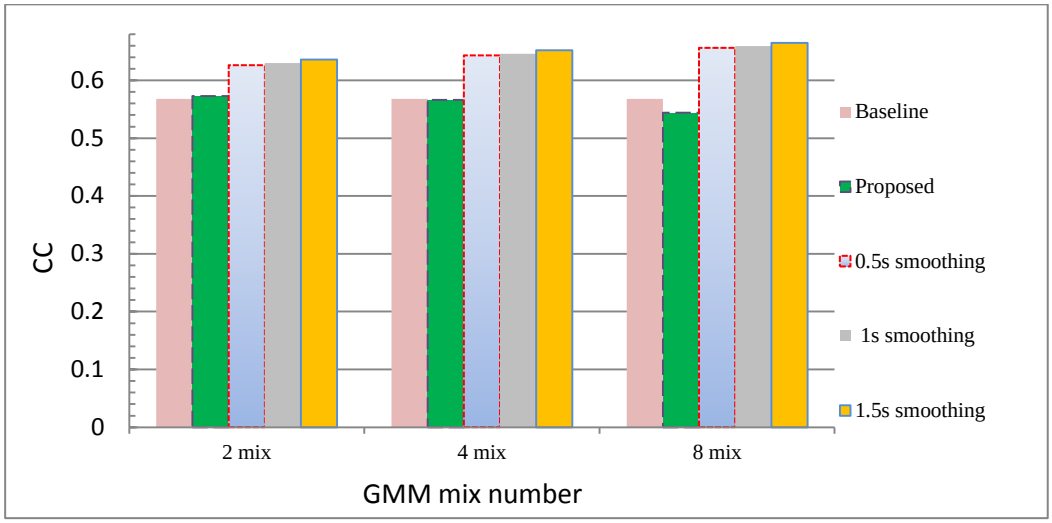
test phase, acoustic volume is estimated by sampling 100,000 points based on Monte-Carlo approach. The threshold θ, used to estimate acoustic volume of the $P(\boldsymbol{y}_t)$ over the entire test partition, is optimized in the range of [1,99] percentiles with a step increase of two.

All the experiments are trained and validated using the 9 speakers in the training partitions, and evaluated using the development dataset. The evaluation metrics for the uncertainty prediction and emotion prediction are Pearson's correlation coefficient (CC) and the concordance correlation coefficients (CCC) respectively. KL divergence is estimated using 100,000 Monte Carlo sampling points as in Section 3.3.

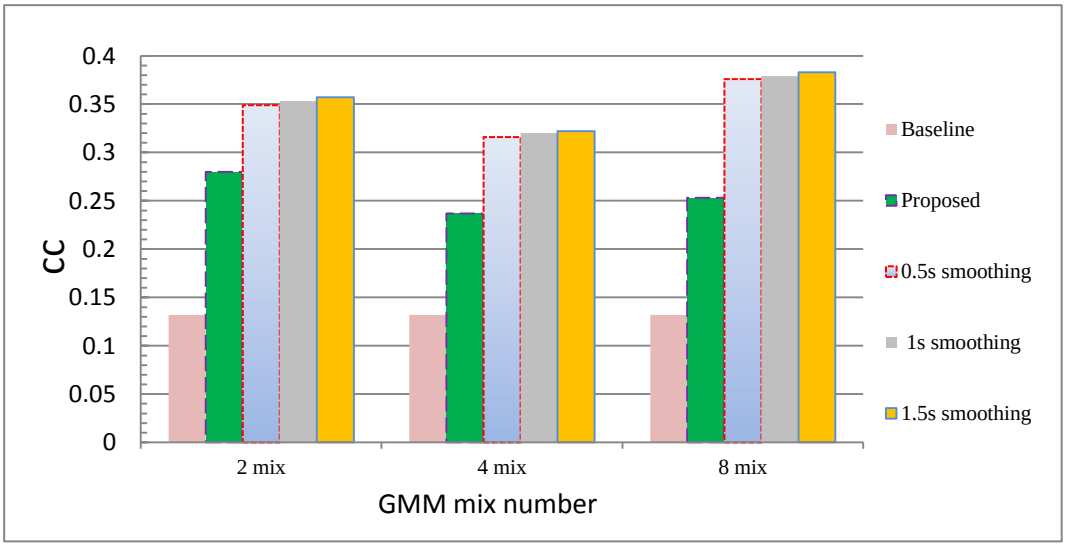
8.3.2 Experimental results

Given the assumption that high inter-rater variability produces a high uncertainty prediction, I aim to investigate is a positive correlation exists between the predicted uncertainty $\widehat{AVL_t}$ computed from Kalman prediction $P(\hat{\boldsymbol{y}}_t)$, and the multi-rater uncertainty AVL_t computed from the distribution $P(\boldsymbol{y}_t)$ obtained using test labels only. A moving average filter was used to smooth the prediction $\widehat{AVL_t}$. It was observed that the 'ground truth' AVL_t is noisy, which is likely due to MAP adaptation using 6 ratings only. Thus I additionally apply a mean filter with 0.5s, 1s and 1.5s windows to smooth, but not to over smooth the 'ground truth' AVL_t.

It can be observed in Figure 8.3 that the proposed method on its own outperforms the baseline with raw (unsmoothed) AVL_t, for both arousal and valence, suggesting that incorporating temporal dependencies does benefit uncertainty prediction, especially for valence. With the increased smoothing, the system performance was further improved. No significant performance differences were observed when using different numbers of mixture components to model $P(\boldsymbol{y}_t)$ for arousal, while the model with eight mixtures outperforms all other configurations for valence, suggesting that predicting valence uncertainty from speech is a more complex problem. Surprisingly, the internal delays of the Kalman filter were not shown to be an influential factor, possibly due to that Kalman filters are applied to a complex representation of the labels in terms of the model parameters instead of the point estimations of the emotion attributes.



(a)



(b)

Figure 8.3: Uncertainty prediction performance in terms of CC with x axis indicating the mixture components of (y_t) : (a) arousal, and (b) valence. The baseline (salmon) system is the one described in Chapter 6 and the performance is presented in Table 6.1. the proposed system (green) indicates evaluation on raw AVL_t, and 'smoothing' systems indicate evaluation on smoothed AVL_t.

Additionally, the KL divergence between the ground truth (modelled as a GMM) and predicted label distributions $P(\hat{\boldsymbol{y}}_t)$, for the proposed systems was compared to that of the baseline system in Chapter 6 [169]. The results given in Table 8.1, indicate that the proposed system leads to a more reliable and smoothed distribution prediction.

The system performance using a single feedforward is also compared to the bidirectional Kalman filters under the optimal system configurations. Bidirectional Kalman filters showed a slightly better performance of 0.665 over 0.662, and 0.383 over 0.381 for arousal and valence respectively, when

Table 8.1: Comparison of mean and standard deviation (SD) of the KL divergence between the predicted distribution and the ground truth distribution.

	Arousal		**Valence**	
	Proposed	**Baseline**	**Proposed**	**Baseline**
Mean	0.1439	1.6872	0.2085	1.8628
SD	0.1818	7.2714	0.2044	1.1236

compared to the feedforward Kalman filter alone. The optimal fusion coefficient α was found to be 0.5, suggesting equal levels of influence for both directional filters.

In order to investigate the effectiveness of the proposed framework for emotion attribute prediction, the point estimations of arousal were obtained from $P(\hat{\boldsymbol{y}}_t)$ by the expectation–maximization algorithm [83]. Valence prediction was not analyzed since it is hard to predict form speech as in Chapter 6. The performance for arousal prediction achieves 0.70 and 0.43 in terms of CC and CCC respectively, which is calculated between predicted $\hat{\boldsymbol{y}}_t$ and the mean ratings. Though it could not outperform the state-of-the-art arousal prediction system with a CCC of 0.796 [92], it still shows potential in predicting emotion attributes without directly using mean ratings.

8.4 Summary

This chapter proposes a dynamic multi-rater GMR which takes into account the temporal dependencies of the emotion uncertainty prediction. The main contributions of this chapter are: (1) incorporation of both feedforward and backward Kalman filters into multi-rater GMR to account for the temporal dependencies of label distributions; (2) estimating label distribution as a GMM instead of single Gaussian assumption; (3) adoption of two new measurements to estimate uncertainty prediction from GMM in terms of the KL divergence between the predicated and the ground truth GMM, and the correlation between the acoustic volumes of the predicted and ground truth GMMs. The results indicate a 17% relative improvement for arousal uncertainty prediction and more than 100% relative improvement for valence uncertainty prediction. Moreover, it shows that predicting

emotion label distributions are more informative than predicting a single mean ‘ground truth’, and this chapter develops methods that can do it while incorporating temporal dynamics and proposes measures to quantify performance under this new paradigm. As the pioneer study considering the temporal dependencies of emotion uncertainty, the work presented in this chapter provides insights into the time-dependent variability introduced by multi-raters.

9 CONCLUSIONS AND FUTURE WORK

9.1 Conclusion

This thesis has described a series of investigations into continuous emotion prediction systems, specifically the impact of variability in the expression and perception of emotions by humans that manifest as speaker variability and inter-rater variability. In addition, the temporal modelling of emotion predictions was also investigated. The aims of these investigations can be summarised as: 1) determining the effect of speaker variability in continuous emotion prediction systems and further compensating the speaker variability based on those analyses; 2) examining the effect of inter-rater variability, and proposing a new framework to obtain uncertainty in predicted emotion that are to some extent indicated by the inter-rater variability and 3) developing techniques to capture the temporal dependencies of both hard labels of emotions and label distributions quantifying emotions.

9.1.1 Effect of speaker variability on the feature space

Chapter 3 explores speaker variability within a probabilistic framework and explores the hypotheses that speaker variability can be characterised in terms of both the differences in how speakers express their gamut of emotional states, as well as the differences in how the same emotional state is expressed by different speakers. These two aspects are captured in the feature and model spaces respectively. A GMM based probabilistic framework was adopted to quantify the speaker variability, since it involves a generative model of the joint distribution over the features and labels, which lends itself well to quantitative analysis of the effect of speaker variability on the feature space. Furthermore, two measures to compare the speaker-dependent distributions in terms of inter- and intra-speaker variability were proposed: the *Kullback-Leibler* divergence and *probabilistic acoustic volume*.

A key insight into the effect of the speaker variability revealed in Chapter 3 is that speaker variability showed similar effect in both feature and model spaces. This results in distinct speaker-dependent distributions in the (joint) feature space, and the speaker-dependent distributions are

'narrower' than the speaker-independent distribution modelled over (joint) features from all speakers, indicated by the smaller value of the slope of the PAV profiles. Thus, pooling data from multiple speakers to train a single model without compensation of speaker variability will lead to a model vulnerable to speaker variability and is likely to lead to less accurate predictions.

9.1.2 Novel approaches for compensating speaker variability

Chapter 4 developed a range of novel approaches to compensate speaker variability in continuous emotion prediction systems: *factor analysis* based normalisation techniques, *partial least squares dimension reduction* based normalisation techniques (PLSDR), and *feature mapping* based normalisation techniques. Among the three compensation techniques, factor analysis based normalisation method was shown to reduce the discrimination between speaker-dependent distributions, and PLSDR and feature mapping based normalisation techniques were shown to compensate for speaker variability in both feature and model spaces simultaneously. Furthermore, the PLSDR based method was shown to be slightly superior to feature mapping. More specifically, the proposed PLSDR based speaker normalisation had the greatest effect on the features from speakers who were the furthest from the other speakers, bringing them in line with the other speakers, while feature mapping based speaker normalisation reduced the differences between feature distributions across all speakers. In terms of the intra-speaker variability, both methods reduced the difference in 'widths' of speaker-dependent distributions.

The results presented in Chapter 4 validate the effectiveness of the proposed compensation techniques. Compared with the baseline system without compensation for speaker variability and the state-of-the-art compensation techniques implemented under the same experimental configurations, all three methods outperformed the state-of-the-art systems on three databases. PLSDR based normalisation techniques in particular achieved relative improvements of 11.7%, 33% and 12% over baseline for arousal, valence and dominance in the USC CreativeIT database, 106%, 11.2% and 3% for arousal, valence and dominance in the SEMAINE database, and 4.6% and 33% over the baseline for arousal and valence in the RECOLA database.

9.1.3 Analysis of inter-rater variability

Chapter 5 analysed emotion annotation variability, with a focus on the inter-rater variability, and the correlation between the inter-rater variability and emotion categories/clusters. First, the reaction lag of both the mean ratings and the individual ratings was investigated, aiming to realign the emotion ratings for a subsequent analysis of the inter-rater variability. It was observed that the optimal delay for the mean rating in the RECOLA database is approximately 4s and 2s for arousal and valence respectively. A similar delay for arousal and valence was also observed in the SEWA database, altogether suggesting a more rapidly changing nature of the arousal intensity. The system performance was improved significantly with the delay compensation for mean ratings. However, the proposed compensation technique for individual raters could not demonstrate improvement in terms of the system performance, which is probably owing to intra-rater variability. Furthermore, inter-rater variability was evaluated in terms of the mean of pair-wise Pearson's correlation coefficients and Cronbach's alpha over the entire RECOLA database, serving as support for the framework for emotion uncertainty predictions presented in Chapters 6 and 8. The most important finding in Chapter 5 is the correlation between the inter-rater variability and the emotion categories/clusters. A significant difference was observed among the inter-rater variability for different emotion clusters, motivating reconsideration of the nature of inter-rater variability in continuous emotion prediction systems.

9.1.4 Novel framework for prediction of uncertainty in emotion labels

Chapter 6 proposed a novel framework for emotion uncertainty prediction based on a probabilistic Gaussian mixture regression (GMR) model. This was implemented by incorporating individual ratings into a conventional GMR system, and then estimating the uncertainty from a GMM, which were approximated by the dominant Gaussian mixture component. The Pearson's correlation coefficient between the predicted uncertainty of emotion labels and the inter-rater variability (computed as the standard deviation among six raters) was adopted as the evaluation metric to quantify system performance. High correlations of 0.56 and 0.13 were observed between the predicted

uncertainty and inter-rater variability for arousal and valence respectively, suggesting the proposed framework is a promising approach, especially for arousal uncertainty prediction. Additional analysis of the correlation between the uncertainty prediction and conventional emotion systems were also explored. The conventional emotion prediction system was found to perform better in the regions where inter-rater variability is low, and worse in the regions where inter-rater variability is high. This also provides a path for using the uncertainty information to improve conventional emotion prediction systems.

The proposed framework for emotion uncertainty prediction has shown to be promising for predicting the uncertainty of emotion labels using inter-rater variability, and changes the perspective from which to view continuous emotion prediction systems. Instead of predicting emotion in terms of hard labels, it is potentially of more interest to predict human emotion as a distribution that implicitly represents uncertainty about the underlying emotion intensity.

9.1.5 Temporal modelling of hard emotion labels

Chapter 7 analysed temporal modelling techniques used for hard emotion label prediction, with a focus on the dynamic feature extraction and the output-associate (OA) frameworks. Regression delta coefficients and shifted delta coefficients (SDCs) were evaluated in conjunction with a GMR back-end with a GMR back-end. The regression delta coefficients calculated on both the feature and label spaces improved the system performance significantly. Shifted delta coefficients outperformed feature vectors without dynamic information, but not regression delta coefficients. This is possibly due to a limited number of speech blocks being used to obtain the SDC features, since a larger number of speech blocks will lead to higher dimensional feature vectors, which cannot be properly modelled using a GMR back-end. All the results reported in this chapter consistently suggest that modelling long-term dynamics in the feature space benefits continuous emotion prediction.

Additionally, the OA framework originated from [31] for continuous emotion prediction was further extended to OA-fusion and OA-regression in a multimodal framework, and were tested with relevance vector machines (RVM) and regularised linear regression (RLR) back-ends. Experiments

results showed significant improvement could be achieved by using the OA-framework for incorporating long term information.

9.1.6 Temporal modelling of emotion label distributions

Chapter 8 extended the emotion uncertainty prediction system introduced in Chapter 6 (a static model) to a dynamic GMR system that takes into account the evolving nature of human emotion. In order to consider the temporal dependencies of the emotion uncertainty and to relax the Gaussian assumption made in the inter-rater variability in Chapter 6, Chapter 8 incorporated Kalman filters into the framework. Feedforward and backward Kalman filters were estimated to model the current emotion state based on past and future information. Two measures, acoustic volume and KL divergence between the predicted and the ground truth distributions, were proposed to compare the proposed dynamic GMR system to the static multi-rater GMR system, and the dynamic system was shown to be superior, especially for valence uncertainty prediction, indicating that valence may benefit more from modelling temporal dependencies. We also found that uncertainty prediction systems for valence performed better with more complex model parameters, suggesting that it is a harder problem compared to arousal uncertainty predictions.

9.2 Future work

Future work efforts could focus on generalising and verifying the proposed compensation techniques for speaker variability to other tasks, where the task-unrelated factors are required to be eliminated. Moreover, the framework for emotion uncertainty prediction can be employed in any research field that suffers from defining the ground truth as hard labels. One of the major factors that potentially limits the validation of the reported investigations is the static regression modelling techniques used in this thesis, such as GMR and RVM. Both of these techniques ignore the temporal dynamics of time series signals, which was shown to be a significant factor for improving continuous emotion prediction systems in Chapters 7 and 8. Therefore, with the rapid developments in deep learning, multiple deep neural networks can be investigated for emotion prediction or related affect modelling

fields. Another key consideration in any future research effort is improving these systems for use in real applications. The developed technologies should ultimately be employed in the real world, either in interactive human-computer interface design, a call centre or clinical diagnosis tools.

There are several specific avenues for extending the work presented in this thesis. Firstly, appropriate evaluation techniques for emotion uncertainty prediction should be investigated, as a matter of urgency. The evaluation metrics used in this thesis are the correlation coefficient of the uncertainty prediction and the inter-rater variability, and the KL divergence between the predicted distribution and actual fitted distribution. However, an optimal evaluation metric that can estimate both the distribution similarity and the accuracy of uncertainty prediction should still be developed based on current work. Furthermore, a definitive method to compare evaluations of the proposed distribution estimation and that of the conventional point estimation is yet to be settled on.

Secondly, though many of the insights into the emotion uncertainty prediction were gained and a Kalman filter was incorporated to take into account the temporal dependencies of emotion's evolving nature based on the GMR framework, it still assumes that emotion uncertainty as a linearly evolving process, though this may not be true in reality. Deep learning especially LSTM and RNN is well established and has been shown to be potentially useful for modelling temporal dependencies in continuous emotion prediction. However, the analysis on modelling label distributions with deep learning structure has just started, and it requires more in-depth insights.

The emotion uncertainty prediction system presented in this thesis adopted the emotion ratings without any delay compensation in individual ratings, which is known to be a confounding factor in emotion uncertainty prediction systems. This will introduce noise into the ratings for each frame. It is expected that compensating the individual reaction lag and realigning the individual ratings will improve the emotion distribution estimation.

Another issue with the use of multiple ratings is that humans tend to perceive emotion changes better than the absolute emotion intensity. This leads to a different type of absolute ratings that measure the trends in emotion change for similar emotion states within duration. For instance, a high

agreement among raters is generally observed with a common increasing trend of arousal intensity, but with different absolute values for each rater. Thus, exploring the underlying emotion states instead of the absolute values may be more appropriate. Any future work in this vein should focus on investigating the hidden emotion intensity based on emotion change for each rater individually. In this way, though the absolute values of individual emotion intensity are different, the hidden emotion state is still estimated as similar for each individual rater. Systems developed using hidden emotion uncertainty can benefit from this representation, including conventional emotion prediction systems and uncertainty prediction systems.

A key challenge in emotion prediction is the reliability of the affect labels perceived by humans, which influences the robustness of the regression model developed based on it. Supervised learning, often adopted in continuous emotion prediction, takes in the label information when training the model, and is highly dependent on the quality of the provided labels. Thus, reinforcement learning ought to be considered, since they use information in the feature space only, and train the neural network with the feedback of the output. Reinforcement learning in particular takes into account the long-term consequences, which suits emotion's evolving nature.

Finally, the proposed compensation techniques for speaker variability presented in this thesis can be tested in other paralinguistic tasks. The compensation techniques described in Chapter 4 can be well generalised to other tasks where the factors unrelated to the task are shown to be negative confounding factors in the framework. These could include gender variability in depression detection, speaker variability in sentiment analysis, etc., since most classification or regression modelling techniques themselves cannot handle such specific variability. The factor analysis and the feature mapping based normalisation techniques can be directly applied in other applications since they are conducted in the feature space, while the PLSDR based normalisation technique can be applied in those fields in which the ground truth is also a time series.

APPENDIX – EM FOR GMR

The auxiliary function of equation (2.41) is written as:

$$
\begin{aligned}
Q(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) &= \sum_{m=1}^{M} P(m|\boldsymbol{X}, \boldsymbol{Y}, \lambda^{(\boldsymbol{Z})}) log\, P\left(\widehat{\boldsymbol{Y}}, m \middle| \boldsymbol{X}, \lambda^{(\boldsymbol{Z})}\right) \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} P\left(m \middle| \boldsymbol{X}_n, \boldsymbol{Y}_n, \lambda^{(\boldsymbol{Z})}\right) log P\left(\widehat{\boldsymbol{Y}}_n, m \middle| \boldsymbol{X}_n, \lambda^{(\boldsymbol{Z})}\right) \\
&= \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{m,n} \left(-\frac{1}{2} \widehat{\boldsymbol{Y}}_n{}^T \boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}} \widehat{\boldsymbol{Y}}_n + \widehat{\boldsymbol{Y}}_n{}^T \boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_{m,n}^{(\boldsymbol{Y})}\right) + K \\
&= \sum_{n=1}^{N} -\frac{1}{2} \widehat{\boldsymbol{Y}}_n{}^T \overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}} \widehat{\boldsymbol{Y}}_n + \widehat{\boldsymbol{Y}}_n{}^T \overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_n^{(\boldsymbol{Y})}} + K \\
&= -\frac{1}{2} \widehat{\boldsymbol{Y}}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{Y}}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} + K \\
&= -\frac{1}{2} \widehat{\boldsymbol{y}}^T \boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W} \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{y}}^T \boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}} + K
\end{aligned}
$$

where $\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}}}$, $\overline{\boldsymbol{D}_n^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_n^{(\boldsymbol{Y})}}$, $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}}$, $\overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}}$, and $\gamma_{m,n}$ are shown in equations (2.42) – (2.47), K is independent of $\widehat{\boldsymbol{y}}$. Note the regression delta coefficients are incorporated in this derivation as in equation (7.6). In terms of the normal labels without regression delta coefficients, $\boldsymbol{W}$ is an identity matrix. In order to find $\widehat{\boldsymbol{y}}$, the first derivative of the auxiliary function is set to zero with respect to $\widehat{\boldsymbol{y}}$:

$$
\frac{\partial Q(\widehat{\boldsymbol{Y}}, \boldsymbol{Y})}{\partial \widehat{\boldsymbol{y}}} = -\boldsymbol{W}^T \overline{\boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W} \widehat{\boldsymbol{y}} + \boldsymbol{W}^T \overline{\boldsymbol{D}_m^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}_m^{(\boldsymbol{Y})}} = 0
$$

and the estimation of $\widehat{\boldsymbol{y}}$ is computed as

$$
\widehat{\boldsymbol{y}} = \left(\boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^T \overline{\boldsymbol{D}^{(\boldsymbol{Y})^{-1}} \boldsymbol{E}^{(\boldsymbol{Y})}}
$$

Finally the label $\widehat{\boldsymbol{y}}$ is estimated as the point that maximises the auxiliary function.

REFERENCE

[1] O. Spindler and T. Fadrus, "Grimace project documentation," *Vienna University of Technology,* 2009.

[2] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing,* vol. 6, no. 2, pp. 97-108, 2015.

[3] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE,* vol. 101, no. 5, pp. 1203-1233, 2013.

[4] R. Jhangiani, H. Tarry, and C. Stangor, "Principles of Social Psychology-1st International Edition," ed: Campus Manitoba, 2015.

[5] R. Pekrun, "The impact of emotions on learning and achievement: Towards a theory of cognitive/motivational mediators," *Applied Psychology,* vol. 41, no. 4, pp. 359-376, 1992.

[6] P. Ekman, "An argument for basic emotions," *Cognition & emotion,* vol. 6, no. 3-4, pp. 169-200, 1992.

[7] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 827-834: IEEE.

[8] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology,* vol. 14, no. 4, pp. 261-292, 1996.

[9] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology,* vol. 17, no. 3, pp. 715-734, 2005.

[10] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science,* vol. 18, no. 12, pp. 1050-1057, 2007.

[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, no. 3, pp. 572-587, 2011.

[12] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework withknowledge-inspired vocal features," *IEEE transactions on affective computing,* vol. 5, no. 2, pp. 201-213, 2014.

[13] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification," in *INTERSPEECH*, 2008, pp. 617-620.

[14] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-Analysis and normalisation," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7522-7526: IEEE.

[15] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication,* vol. 57, pp. 1-12, 2014.

[16] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE transactions on Affective computing,* vol. 4, no. 4, pp. 386-397, 2013.

[17] H. Khaki and E. Erzin, "Continuous Emotion Tracking Using Total Variability Space," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] M. Valstar *et al.*, "AVEC 2016-Depression, Mood, and Emotion Recognition Workshop and Challenge," *arXiv preprint arXiv:1605.01600,* 2016.

[19] Z. Huang *et al.*, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 19-26: ACM.

[20] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "iVectors for Continuous Emotion Recognition," *Training,* vol. 45, p. 50, 2014.

[21] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natlae, "Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models."

[22] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Proc. Interspeech*, 2016.

[23] E. Mower *et al.*, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1-8: IEEE.

[24] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters,* vol. 66, pp. 22-30, 2015.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS),* vol. 2, no. 1, p. 6, 2012.

[26] M. Wöllmer *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech*, 2008, vol. 2008, pp. 597-600.

[27] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing,* vol. 31, no. 2, pp. 153-163, 2013.

[28] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, 2015, pp. 698-704: IEEE.

[29] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 41-48: ACM.

[30] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, no. 3, pp. 439-446, 2016.

[31] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing,* vol. 30, no. 3, pp. 186-196, 2012.

[32] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition,* vol. 17, no. 2, pp. 484-495, 2008.

[33] D. C. Rubin and J. M. Talarico, "A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words," *Memory,* vol. 17, no. 8, pp. 802-808, 2009.

[34] A. Ben-Zeev, "The nature of emotions," *Philosophical Studies,* vol. 52, no. 3, pp. 393-409, 1987.

[35] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

[36] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[37] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior research methods,* vol. 45, no. 4, pp. 1191-1207, 2013.

[38] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, 2009, pp. 1-4: IEEE.

[39] I. Fonagy, "Emotions, voice and music," *Research Aspects on Singing, Royal Swedish Academy of Music* no. No. 33, pp. 51–79, 1981.

[40] J. R. Davitz, "Personality, perceptual, and cognitive correlates of emotional sensitivity " presented at the The Communication of Emotional Meaning, 1964.

[41] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker identification using mel frequency cepstral coefficients," *variations,* vol. 1, no. 4, 2004.

[42] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America,* vol. 57, no. S1, pp. S35-S35, 1975.

[43] R. Cowie *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine,* vol. 18, no. 1, pp. 32-80, 2001.

[44] A. Batliner *et al.*, "Whodunnit–searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language,* vol. 25, no. 1, pp. 4-28, 2011.

[45] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review,* vol. 43, no. 2, pp. 155-177, 2015.

[46] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on audio, speech, and language processing,* vol. 17, no. 4, pp. 582-596, 2009.

[47] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, no. 10, pp. 1175-1191, 2001.

[48] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, p. 4 pp.: IEEE.

[49] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 1500-1503: IEEE.

[50] H. Hu, M.-X. Xu, and W. Wu, "Fusion of global statistical and segmental spectral features for speech emotion recognition," in *INTERSPEECH*, 2007, pp. 2269-2272.

[51] D.-N. Jiang and L.-H. Cai, "Speech emotion classification with the combination of statistic features and temporal features," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, 2004, vol. 3, pp. 1967-1970: IEEE.

[52] M. Wöllmer *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[53] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[54] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[55] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[56] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[57] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 22, no. 1, pp. 4-37, 2000.

[58] R. J. Schalkoff, *Pattern recognition*. Wiley Online Library, 1992.

[59] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems,* vol. 2, no. 1-3, pp. 37-52, 1987.

[60] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biological psychology,* vol. 87, no. 1, pp. 93-98, 2011.
[61] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing,* vol. 37, no. 3, pp. 328-339, 1989.
[62] L. R. Rabiner and B.-H. Juang, "Fundamentals of speech recognition," 1993.
[63] Z. Huang and J. Epps, "A PLLR and multi-stage Staircase Regression framework for speech-based emotion prediction," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 5145-5149: IEEE.
[64] M. Schmitt and B. Schuller, "openXBOW—Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research,* vol. 18, no. 96, pp. 1-5, 2017.
[65] F. Ringeval *et al.*, "AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 3-9: ACM.
[66] V. S. a. E. A. K.W. Gamage, "Modeling variable length phoneme sequences - a step towards linguistic information for speech emotion recognition in wider world," presented at the Affective Computing and Intelligent Interaction (ACII), 2017.
[67] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5200-5204: IEEE.
[68] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia,* vol. 16, no. 8, pp. 2203-2213, 2014.
[69] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 2741-2745: IEEE.
[70] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, 2013, pp. 511-516: IEEE.
[71] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters,* vol. 21, no. 9, pp. 1068-1072, 2014.
[72] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4153-4156: IEEE.
[73] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing,* vol. 14, no. 3, pp. 199-222, 2004.
[74] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV-1085-IV-1088: IEEE.
[75] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing,* vol. 2, no. 2, pp. 92-105, 2011.
[76] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1335-1336: ACM.
[77] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks,* vol. 5, no. 2, pp. 157-166, 1994.
[78] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," ed: A field guide to dynamical recurrent neural networks. IEEE Press, 2001.

[79] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing,* vol. 4, no. 5, pp. 867-881, 2010.

[80] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, 2015, pp. 65-72: ACM.

[81] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing,* vol. 31, no. 2, pp. 137-152, 2013.

[82] H. G. Sung, "Gaussian mixture regression and classification," Rice University, 2004.

[83] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[84] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of machine learning research,* vol. 1, no. Jun, pp. 211-244, 2001.

[85] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Relevance Vector Machine for Depression Prediction," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[86] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," in *Robust Speech Recognition and Understanding*: InTech, 2007.

[87] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5205-5209: IEEE.

[88] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions," in *Int. Conference on Affective Computing and Intelligent Interaction*, 2017.

[89] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing,* 2017.

[90] P. Sedgwick, "Pearson's correlation coefficient," *Bmj,* vol. 345, no. 7, 2012.

[91] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE signal processing magazine,* vol. 26, no. 1, pp. 98-117, 2009.

[92] M. Valstar *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3-10: ACM.

[93] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, no. 3, pp. 461-468, 2016.

[94] M. E. Tipping, "Bayesian inference: An introduction to principles and practice in machine learning," *Lecture notes in computer science,* vol. 3176, pp. 41-62, 2004.

[95] D. J. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural computation,* vol. 4, no. 3, pp. 448-472, 1992.

[96] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm," *IEEE transactions on medical imaging,* vol. 20, no. 1, pp. 45-57, 2001.

[97] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 2288-2291: IEEE.

[98] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter," in *Acoustics, Speech, and*

Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, 2005, vol. 1, pp. I/9-I12 Vol. 1: IEEE.

[99] J. H. Hansen and S. E. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[100] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Interspeech*, 2005, vol. 5, pp. 1517-1520.

[101] S. T. Jovicic, Z. Kasic, M. Dordevic, and M. Rajkovic, "Serbian emotional speech database: design, processing and evaluation," in *9th Conference Speech and Computer*, 2004.

[102] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005, pp. 474-477: IEEE.

[103] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 114-125: Springer.

[104] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus," in *Proc. of a Satellite Workshop of LREC*, 2008, pp. 28-31.

[105] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, 2008, pp. 865-868: IEEE.

[106] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation,* vol. 42, no. 4, p. 335, 2008.

[107] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality,* p. 55, 2010.

[108] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, 2013, pp. 1-8: IEEE.

[109] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[110] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[111] S. E. Eskimez, M. Sturge-Apple, Z. Duan, and W. B. Heinzelman, "WISE: Web-based Interactive Speech Emotion Classification," in *SAAIP@ IJCAI*, 2016, pp. 2-7.

[112] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, and S. Narayanan, "Automatic recognition of emotion evoked by general sound events," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 341-344: IEEE.

[113] T. Dang, Stasak, B., Huang, Z., Jayawardena, S., Atcheson, M., Hayat, M., Le, P., Sethu, V., Goecke, R.and Epps, J., "Investigating Word affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," presented at the In Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain View, California, USA, 2017.

[114] F. Eyben *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing,* vol. 7, no. 2, pp. 190-202, 2016.

[115] E. A. Butler, T. L. Lee, and J. J. Gross, "Emotion regulation and culture: are the social consequences of emotion suppression culture-specific?," *Emotion,* vol. 7, no. 1, p. 30, 2007.

[116] R. A. Emmons, "Emotion and religion," *Handbook of the psychology of religion and spirituality,* pp. 235-252, 2005.
[117] M. R. Taghavi, A. R. Moradi, H. T. Neshat-Doost, W. Yule, and T. Dalgleish, "Interpretation of ambiguous emotional information in clinically anxious children and adolescents," *Cognition & Emotion,* vol. 14, no. 6, pp. 809-822, 2000.
[118] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition," *arXiv preprint arXiv:1708.07050,* 2017.
[119] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006.
[120] J. J. Gross, L. L. Carstensen, M. Pasupathi, J. Tsai, C. Götestam Skorpen, and A. Y. Hsu, "Emotion and aging: experience, expression, and control," *Psychology and aging,* vol. 12, no. 4, p. 590, 1997.
[121] R. Parkins, "Gender and emotional expressiveness: An analysis of prosodic features in emotional expression," *Pragmatics and Intercultural Communication,* vol. 5, no. 1, pp. 46-54, 2012.
[122] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.
[123] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, 2011, pp. 523-528: IEEE.
[124] D. Bone, C.-C. Lee, and S. S. Narayanan, "A Robust Unsupervised Arousal Rating Framework using Prosody with Cross-Corpora Evaluation," in *INTERSPEECH*, 2012, pp. 1175-1178.
[125] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *INTERSPEECH*, 2014, pp. 1238-1242.
[126] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic acoustic volume analysis for speech affected by depression," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
[127] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 2, pp. II-53-6 vol. 2: IEEE.
[128] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006, vol. 1, pp. I-I: IEEE.
[129] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, no. 4, pp. 788-798, 2011.
[130] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462: ACM.
[131] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," 2013.
[132] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE transactions on speech and audio processing,* vol. 2, no. 2, pp. 291-298, 1994.
[133] M. Brookes, "VOICEBOX: A speech processing toolbox for MATLAB. 2006," ed.
[134] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of chemometrics,* vol. 17, no. 3, pp. 166-173, 2003.
[135] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," *Lecture notes in computer science,* vol. 3940, p. 34, 2006.

[136] B. J. a. Y. Goegebeur. (2007). *Module 8: Partial least squares regression II*. Available: http://statmaster.sdu.dk/courses/ST02/module08/
[137] R. Manne, "Analysis of two partial-least-squares algorithms for multivariate calibration," *Chemometrics and Intelligent Laboratory Systems,* vol. 2, no. 1-3, pp. 187-197, 1987.
[138] G.-Z. Li and X.-Q. Zeng, "Feature selection for partial least square based dimension reduction," *Foundations of Computational Intelligence Volume 5,* pp. 3-37, 2009.
[139] Bent Jørgensen and Yuri Goegebeur. (2007). *Module 8: Partial least squares regression II*. Available: http://statmaster.sdu.dk/courses/ST02/module08/
[140] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, 2003, vol. 2, pp. II-53: IEEE.
[141] N. Ding, V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in emotion recognition-an adaptation based approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 5101-5104: IEEE.
[142] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing,* vol. 3, no. 1, pp. 5-17, 2012.
[143] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification," *Speech communication,* vol. 50, no. 4, pp. 312-322, 2008.
[144] C. Busso, A. Metallinou, and S. S. Narayanan, "Iterative feature normalization for emotional speech detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011, pp. 5692-5695: IEEE.
[145] T. Dang, V. Sethu, and E. Ambikairajah, "Factor Analysis Based Speaker Normalisation for Continuous Emotion Prediction," in *INTERSPEECH*, 2016, pp. 913-917.
[146] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 449-456: ACM.
[147] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 501-508: ACM.
[148] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*, 2012, pp. 485-492: ACM.
[149] B. Zhang, G. Essl, and E. Mower Provost, "Predicting the distribution of emotion perception: capturing inter-rater variability," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 51-59: ACM.
[150] I. Siegert, R. Böck, and A. Wendemuth, "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements," *Journal on Multimodal User Interfaces,* vol. 8, no. 1, pp. 17-28, 2014.
[151] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," in *Proceedings of the 2017 ACM on Multimedia Conference*, 2017, pp. 890-897: ACM.
[152] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics,* vol. 34, no. 4, pp. 555-596, 2008.
[153] K. L. Gwet, "Computing inter‐rater reliability and its variance in the presence of high agreement," *British Journal of Mathematical and Statistical Psychology,* vol. 61, no. 1, pp. 29-48, 2008.
[154] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 19-26: ACM.

[155] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Audio/Visual Emotion Challenge, Proceedings of the 4th International Workshop on* 2014, pp. 19-26: ACM.
[156] J. M. Bland and D. G. Altman, "Statistics notes: Cronbach's alpha," *Bmj,* vol. 314, no. 7080, p. 572, 1997.
[157] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics),* vol. 28, no. 1, pp. 100-108, 1979.
[158] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home,* vol. 6, no. 2, pp. 101-108, 2012.
[159] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.
[160] M. Wöllmer *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, 2008, pp. 597-600.
[161] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Seventh International Conference on Spoken Language Processing*, 2002.
[162] Y. Minami, E. McDermott, A. Nakamura, and S. Katagiri, "A theoretical analysis of speech recognition based on feature trajectory models," in *Eighth International Conference on Spoken Language Processing*, 2004.
[163] M. A. Kohler and M. Kennedy, "Language identification using shifted delta cepstra," in *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, 2002, vol. 3, pp. III-69: IEEE.
[164] F. Wang, W. Verhelst, and H. Sahli, "Relevance vector machine based speech emotion recognition," in *Affective computing and intelligent interaction*: Springer, 2011, pp. 111-120.
[165] T. Dang *et al.*, "Investigating Word Affect Features and Fusion of Probabilistic Predictions Incorporating Uncertainty in AVEC 2017," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 27-35: ACM.
[166] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE transactions on pattern analysis and machine intelligence,* vol. 30, no. 2, pp. 283-298, 2008.
[167] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing,* vol. 31, no. 2, pp. 120-136, 2013.
[168] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine learning,* vol. 51, no. 2, pp. 181-207, 2003.
[169] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An Investigation of Emotion Prediction Uncertainty Using Gaussian Mixture Regression," *Proc. Interspeech 2017,* pp. 1248-1252, 2017.
[170] M. S. Grewal, "Kalman filtering," in *International Encyclopedia of Statistical Science*: Springer, 2011, pp. 705-708.
[171] K. Somandepalli, R. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. S. Narayanan, "Online affect tracking with multimodal kalman filters," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 59-66: ACM.
[172] K. Brady *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97-104: ACM.

[173] Z. Huang and J. Epps, "An Investigation of Emotion Dynamics and Kalman Filtering for Speech-based Emotion Prediction," *Proc. Interspeech 2017,* pp. 3301-3305, 2017.

[174] M. Oveneke, I. Gonzalez, V. Enescu, D. Jiang, and H. Sahli, "Leveraging the Bayesian Filtering Paradigm for Vision-Based Facial Affective State Estimation," *IEEE Transactions on Affective Computing,* 2017.