Continuous Emotion Prediction from Speech: Modelling Ambiguity in Emotion

Deboshree Bose

A thesis in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Electrical Engineering and Telecommunications

Faculty of Engineering

The University of New South Wales

September 2022
Abstract

There is growing interest in emotion research to model perceived emotion labeled as intensities along the affect dimensions such as arousal and valence. These labels are typically obtained from multiple annotators who would have their individualistic perceptions of emotional speech. Consequently, emotion prediction models that incorporate variation in individual perceptions as ambiguity in the emotional state would be more realistic. This thesis develops the modeling framework necessary to achieve continuous prediction of ambiguous emotional states from speech.

Besides, emotion labels, feature space distribution, and encoding are an integral part of the prediction system. The first part of this thesis examines the limitations of current low-level feature distributions and their minimalistic statistical descriptions. Specifically, front-end paralinguistic acoustic features are reflective of speech production mechanisms. However, discriminatively learned features have frequently outperformed acoustic features in emotion prediction tasks, but provide no insights into the physical significance of these features. One of the contributions of this thesis is the development of a framework that can modify the acoustic feature representation based on emotion label information.

Another investigation in this thesis indicates that emotion perception is language-dependent and in turn, helped develop a framework for cross-language emotion prediction. Furthermore, this investigation supported the hypothesis that emotion perception is highly individualistic and is better modeled as a distribution rather than a point estimate to encode information about the ambiguity in the perceived emotion. Following this observation, the thesis proposes measures to quantify the appropriateness of distribution types in modeling ambiguity in dimensional emotion labels which are then employed to compare well-known bounded parametric distributions. These analyses led to the conclusion that the beta distribution was the most appropriate parametric model of ambiguity in emotion labels.

Finally, the thesis focuses on developing a deep learning framework for continuous emotion prediction as a temporal series of beta distributions, examining various parameterizations of the beta distributions as well as loss functions. Furthermore, distribution over the parameter spaces is examined and priors from kernel density estimation are employed to shape the posteriors over the parameter space which significantly improved valence ambiguity predictions.

The proposed frameworks and methods have been extensively evaluated on multiple state-of-the-art databases and the results demonstrate both the viability of predicting ambiguous emotion states and the validity of the proposed systems.

Declaration relating to the disposition of project thesis/dissertation

I hereby grant the University of New South Wales or its agents a non-exclusive license to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or hereafter known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyrighted material used in this thesis, written permission for use has been obtained, or the copyrighted material is removed from the final public version of the thesis.

Signature Deboshree Bose Witness Date 19 September, 2022

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 published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project’s design and conception or in style, presentation and linguistic expression is acknowledged.

Deboshree Bose
19 September, 2022
Copyright Statement

I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or hereafter known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

Deboshree Bose
19 September, 2022

Authenticity Statement

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.

Deboshree Bose
19 September, 2022
ORIGINALITY STATEMENT

☑️ I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

COPYRIGHT STATEMENT

☑️ I hereby grant the University of New South Wales or its agents a non-exclusive licence to archive and to make available (including to members of the public) my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known. I acknowledge that I retain all intellectual property rights which subsist in my thesis or dissertation, such as copyright and patent rights, subject to applicable law. I also retain the right to use all or part of my thesis or dissertation in future works (such as articles or books).

For any substantial portions of copyright material used in this thesis, written permission for use has been obtained, or the copyright material is removed from the final public version of the thesis.

AUTHENTICITY STATEMENT

☑️ I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis.
UNSW is supportive of candidates publishing their research results during their candidature as detailed in the UNSW Thesis Examination Procedure.

Publications can be used in your thesis in lieu of a Chapter provided:

- You contributed **greater than 50%** of the content in the publication and are the "primary author", i.e. you were responsible primarily for the planning, execution and preparation of the work for publication.
- You have approval to include the publication in their thesis in lieu of a Chapter from your Supervisor and Postgraduate Coordinator.
- The publication is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in the thesis.

☑ Some of the work described in my thesis has been published and it has been documented in the relevant Chapters with acknowledgement.

A short statement on where this work appears in the thesis and how this work is acknowledged within chapter/s:
The following 3 papers have been explicitly listed in the thesis and also the chapters they specifically contribute to have been mentioned.

A Novel Bag-of-Optimised-Clusters Front-End for Speech-based Continuous Emotion Prediction was published by Deboshree Bose, T. Dang, V. Sethu, E. Ambikairajah, S. Fernando in the 8th ACII, Cambridge, UK. The work was done in the first year of my Ph.D. and the experiments, analyses and findings are included in chapter 3 with citations.

Parametric Distributions to Model Numerical Emotion Labels was authored by Deboshree Bose, V. Sethu, E. Ambikairajah in Interspeech 30 August 2021. The work was done in the third year of my Ph.D. and the experiments, analyses and findings are included in chapter 5 with citations.

Continuous Emotion Ambiguity Prediction: Modeling with Beta Distribution was authored by Deboshree Bose, V. Sethu, E. Ambikairajah and is submitted for publication in IEEE Transactions on Affective Computing. The paper has been acknowledged with an explicit mention under journal papers and that it contributed to the material in chapter 6.

CANDIDATE’S DECLARATION

☑️ I declare that I have complied with the Thesis Examination Procedure.
Abstract

There is growing interest in emotion research to model perceived emotion labelled as intensities along the affect dimensions such as arousal and valence. These labels are typically obtained from multiple annotators who would have their individualistic perceptions of emotional speech. Consequently, emotion prediction models that incorporate variation in individual perceptions as ambiguity in the emotional state would be more realistic. This thesis develops the modelling framework necessary to achieve continuous prediction of ambiguous emotional states from speech.

Besides, emotion labels, feature space distribution and encoding are an integral part of the prediction system. The first part of this thesis examines the limitations of current low-level feature distributions and their minimalistic statistical descriptions. Specifically, front-end paralinguistic acoustic features are reflective of speech production mechanisms. However, discriminatively learnt features have frequently outperformed acoustic features in emotion prediction tasks, but provide no insights into the physical significance of these features. One of the contributions of this thesis is the development of a framework that can modify the acoustic feature representation based on emotion label information.

Another investigation in this thesis indicates that emotion perception is language-dependent and in turn, helped develop a framework for cross-language emotion prediction. Furthermore, this investigation supported the hypothesis that emotion perception is highly individualistic and is better modelled as a distribution rather than a point estimate to encode information about the ambiguity in the perceived emotion. Following this observation, the
thesis proposes measures to quantify the appropriateness of distribution types in modelling ambiguity in dimensional emotion labels which are then employed to compare well-known bounded parametric distributions. These analyses led to the conclusion that the beta distribution was the most appropriate parametric model of ambiguity in emotion labels.

Finally, the thesis focuses on developing a deep learning framework for continuous emotion prediction as a temporal series of beta distributions, examining various parameterizations of the beta distributions as well as loss functions. Furthermore, distribution over the parameter spaces is examined and priors from kernel density estimation are employed to shape the posteriors over the parameter space which significantly improved valence ambiguity predictions.

The proposed frameworks and methods have been extensively evaluated on multiple state-of-the-art databases and the results demonstrate both the viability of predicting ambiguous emotion states and the validity of the proposed systems.
Acknowledgement

My deepest gratitude goes to my supervisor Dr. Vidhyasaharan Sethu for his extraordinary patience, encouragement and direction and to my joint-supervisor Professor Eliathamby Ambikairajah for his excellent suggestions and timely support always. I cannot thank them enough for helping me navigate hard circumstances while inspiring me to outdo myself during my PhD. Many thanks for my ultra intelligent and humane review panel - Professor Andrew Dempster who’s quick humour made progress reviews funnily awaited and pain-free, Professor Julien Epps and Dr. Beena Ahmed for their indispensable feedback and smart suggestions.

Thank you my dearest colleagues turned friends of our UNSW Signal Processing Research group! Kaavya Sriskandaraja, Ting Dang, Zhaocheng Huang- thanks for guiding a newbie to emotion prediction and Australia; my lunch and coffee chat gangs - Sadari Jayawardena, Iresha Pasquel, Tharshini Gunendradasan, Hang Li and Namalka Liyanage, Tharmakulasingam Sirojan, Jingyao Wu, Zheng Nan, Miao Jing, Hanyu Meng, Antoni Dimitriadis for wholesome fun and activities and Dr Phu Ngoc Le, Gajan Suthokumar, Mia Atcheson, Kalani Gamage, Sarith Fernando, Saad Irtza, Jianbo Ma for tons of moral support.

I thank our School of Electrical Engineering and Telecommunications UNSW for supporting me throughout my studies. Thank you Rohan Dhuri and Mukul Bose for your invaluable feedback on my presentations regardless of the odd hours in your continents and your unshakeable beliefs in my abilities. It means more than I can put into words.
To my beloved mother Tanushree Bose and wonderfully supportive father Devabrata Bose with a heart of gold.
Publications and Presentations

Journal Papers


The experiments, analyses, and findings of this paper are reported in chapter 6.

0.1 Conference Papers Presentations


The experiments, analyses, and findings of this paper are reported in chapter 5.


The experiments, analyses, and findings of this paper are reported in chapter 3.
Contents

Abstract iii

Acknowledgement v

Publications and Presentations vii

0.1 Conference Papers Presentations vii

Contents viii

List of Figures xv

List of Tables xxii

1 Introduction 1

1.1 Speech based continuous emotion prediction - An Overview 1

1.2 Thesis Objectives 6

1.3 Thesis Organization 8
2 Speech based Emotion Prediction - A Review

2.1 Emotion : its expression, perception and representation

2.1.1 Emotion Expression and Perception

2.1.2 Emotion Representation

2.2 Emotion Ambiguity: Motivation and Introduction

2.2.1 Point-Estimation of Emotion

2.2.2 Ambiguity-Aware Estimation of Emotion

2.3 Speech-based emotion prediction – system overview

2.3.1 Pre-processing

2.3.2 Feature Extraction

2.3.3 Feature Extraction Tools

2.3.4 High-Level Feature Representation

2.3.5 Supra-Frame Level Features

2.3.6 Voice Quality Features

2.3.7 Linguistic Features

2.3.8 Non-linguistic Acoustic Events

2.3.9 Feature Selection
2.4 Regression Modelling Techniques ............................................ 35

2.4.1 Relevance Vector Machines .............................................. 36

2.4.2 Gaussian Mixture Regression ............................................ 36

2.4.3 Gaussian Process Regression ............................................ 37

2.4.4 Support Vector Regression .............................................. 38

2.4.5 Autoregression Exogenous Back-end .................................. 41

2.4.6 Long Short Term Memory - Recurrent Neural Network ............ 45

2.5 Evaluation Metrics ............................................................. 47

2.5.1 Point-Prediction Metrics ................................................. 48

2.5.2 Distribution Comparison Metrics ...................................... 51

2.6 Emotion Databases ............................................................. 52

2.6.1 The RECOLA database .................................................... 56

2.6.2 Limitations of current databases ...................................... 58

2.7 Summary ................................................................. 60

3 Bag-of-Optimized Codebooks - Emotion based Clustering of Features 63

3.1 Introduction ................................................................. 63

3.2 Capturing the joint-LLD Distribution ................................... 67

3.3 Proposed Bag-of-Optimised-Clusters Framework ....................... 67
3.3.1 Framework Description ............................................. 68
3.3.2 Error Backpropagation ............................................. 72
3.4 Experiment Settings ..................................................... 72
3.4.1 Database ............................................................ 72
3.4.2 Annotation Delay Compensation ................................. 73
3.4.3 Performance Metric .................................................. 73
3.4.4 Continuous Emotion Recognition System Configuration ...... 73
3.5 Experimental Results and Discussion .............................. 74
3.5.1 Results .............................................................. 75
3.5.2 Codebook Analyses .................................................. 76
3.5.3 Model Fitting ........................................................ 79
3.6 Summary ............................................................... 80

4 Cross Culture Emotion Prediction using the Linear ARX back-end 81
4.1 Introduction and Motivation ........................................... 81
4.2 System Overview ....................................................... 83
4.2.1 ARX Model ........................................................ 86
4.2.2 Label normalization ................................................. 86
4.3 Experimental Settings .................................................. 88
4.4 Baseline System .................................................. 88

4.5 Development Results ......................................... 89
  4.5.1 Feature and Decision-Level Fusion ..................... 89
  4.5.2 Language-Dependent Models ............................ 90

4.6 AVEC Challenge Test Set Result ............................ 91

4.7 Summary .......................................................... 92

5 Modelling Ambiguity with Bounded Parametric Distributions 93
  5.1 Motivation and Introduction .............................. 93

  5.2 Modeling Ambiguity with Gaussians ....................... 94

  5.3 Comparing Distributions .................................. 96
    5.3.1 Comparing Maximum Likelihood Estimates .......... 97
    5.3.2 Expected Log-Likelihood ............................ 100

  5.4 Experimental Settings .................................... 102
    5.4.1 Prior Choices .......................................... 103

  5.5 Results and Discussion .................................. 104

  5.6 Summary ..................................................... 106

6 Predicting Ambiguity Modelled with Beta Distributions 107
  6.1 Motivation and Introduction ............................ 107

xii
7.1 Conclusion ............................................................................................................. 145

7.1.1 Effect of Discriminative Learning of Acoustic Feature Encoding .................. 146

7.1.2 Investigation into Cross-Language Affect with a Linear Back-end ............... 147

7.1.3 Impact of Modelling Ambiguity with Parametric Distributions ................. 147

7.1.4 Predicting Ambiguity Modeled with Beta Distributions ............................... 148

7.2 Future Work ......................................................................................................... 149

References ..................................................................................................................... 152
List of Figures

1.1 Minimalistic emotion class locations in the arousal-valence dimensional plane in study [1] ........................................... 2

1.2 Plutchik’s Wheel of Emotions. [2] ........................................... 4

1.3 Illustration of different emotion class locations and boundaries in the arousal-valence dimensional plane. [3] ........................................... 5

2.1 Illustration of placement of emotion categories on the plane of emotion dimensions. ........................................... 14

2.2 Arousal ratings recorded over time in seconds from 3 annotators in the RECOLA dataset. ........................................... 17

2.3 Overview of a traditional continuous emotion prediction system ........................................... 20

2.4 Illustration of how speech waveform is framed depending on the chosen LLD types, then LLD vectors are constructed over overlapping speech segments and statistical features are computed over a larger window of a few seconds. 26
2.5 The figure illustrates the scenario of a BoAW single assignment with the top plot showing three cluster centers or codewords in green on a two-dimensional feature space and an input feature vector in red. The bottom plot is an illustration of how the BoAW histogram assignment is done based on the top plot.

2.6 Speech feature vectors are represented by green markers on a two-dimensional plane. A regression line represented by the dashed line and a tube - the solid black lines, is fitted around the regression line. The feature vectors on the tube represent the support vectors. The vectors present outside the tube contribute to errors based on their distance from the regression lines and are represented by slack errors $\xi$.

2.7 Overview of the ARX backend. There is a set of parallel FIR filters implemented, one for filtering each feature dimension and finally cascaded with an all-pole filter.

2.8 Concept of a recurrent neural network is illustrated with its and the unfolded structure in time during forward computation. $X_t$ represents an $N$ dimensional input feature vector at time $t$, $H_{mt}$ hidden unit at time $t$ for $m = 1, 2, ..., M$ and $Y_t$ 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.

2.9 The limitations of the MSE and $\rho$ measures are illustrated in the above two plots respectively. The plot on top displays the instance when the predictions follow the trend of the mean labels but the MSE is high. Whereas the plot below shows the instance where the $\rho$ had decreased but the MSE has improved.
2.10 Four normal distributions are displayed. The value of the KL divergence between each of the red, yellow and purple distributions and the blue distribution are equal.

3.1 A vector of pre-selected LLDs is extracted from each short time frame of the speech waveform. Functionals are calculated for LLDs over several time frames within an emotion window. [4]

3.2 Left: $C_0$ magnitude for 10 ms frames within a randomly selected 3s segment. Right: Histogram of above plotted $C_0$ values. A suprasegmental feature representing the $C_0$ values plotted in the above segment, i.e. the mean value is indicated in black.

3.3 A schematic diagram of the NN showing the first phase, the BoOC creation by accumulating the histogram of the LLDs of all frames, and the prediction phase using the BoOC features [4]

3.4 Illustration of BoOC assignment. Top: Plot illustrates a hypothetical 2-dimensional speech feature space where the locations of the input feature vector is shown by the red mark and the cluster centers in green. Bottom: Plot illustrates the audio word assignment of the feature vector to the codewords for the scenario demonstrated above. [4]

3.5 The plot illustrates the location of the input feature vectors in yellow on an 88-dimensional feature space reduced to a 2-dimensional visualization space with t-distributed stochastic neighbor embedding (t-SNE). The position of the codewords derived from k-means++ is shown in red. The varying positions of the optimized clusters over 20 optimization epochs are shown in blue. [4]
3.6 CCC over 30 epochs of the test set when different data partitions were used to train and test the model. 79

4.1 Proposed system for emotion prediction (System 1 and System 2). 84

4.2 Proposed system for emotion prediction (System 3). 85

4.3 Histograms of Arousal (top), Valence (middle), and Liking(bottom) mean annotations for development (left) and training (right) sets. The blue histograms denote the annotations for German speech and the orange for Hungarian speech. 87

5.1 Examples frame from RECOLA showing six arousal and valence ratings as well as maximum likelihood Gaussian distribution fit to the ratings. In both cases, the best fitting distributions do not agree with the fact that the ratings are also bounded. 95

5.2 Histograms constructed from all instances or frames of emotion ratings when at least by 1 percent of the total area under the Gaussian Maximum Likelihood distributions lie outside the domain $[0, 1]$. 97

5.3 A typical U-shaped Beta Distribution is obtained when both shape parameters are each less than one, indicating the unlikely case that mostly opposite extreme emotions are perceived amongst humans. 99

5.4 Maximum Likelihood Trapezoidal Distribution Fit over Emotion Ratings. 100

5.5 Histograms of log-likelihood ratios of Beta distributions to others based on ML parameter estimates for arousal ratings. 102

5.6 Histograms of log-likelihood ratios of Beta distributions to others based on ML parameter estimates for valence ratings. 103
5.7 Histograms of expected log-likelihood ratios of Beta distributions to others over parameter posteriors for arousal ratings. . . . . . . . . . . . . . . . . . 104

5.8 Histograms of expected log-likelihood ratios of Beta distributions to others over parameter posteriors for valence ratings. . . . . . . . . . . . . . . . . . 105

6.1 An overview of the key aspects of the ambiguity prediction system that need addressing and their position within the flow of the emotion prediction system.109

6.2 MLE of shaper parameter $a$ for current frame ratings (6 ratings) in blue vs. four neighboring frames with the current frame (54 ratings) in red. . . . . 113

6.3 Top: Maximum likelihood distributions over four frames of emotion ratings. Bottom: Gradients of the log-likelihood functions where $a = b$ with respect to shape parameter $a$ over four frames of emotion ratings. . . . . . . . . . . 115

6.4 Illustration of (left-most column) frame-wise MLE of shape parameters in red over the shape parameter space; KDE of the prior distribution over MLE in shape $(a, b)$ space (center-column); KDE of the prior distribution mode-concentration MLE in the $(w, k)$ space. The top row of plots are for arousal ratings and the bottom is for valence. . . . . . . . . . . 116

6.5 This is the block diagram of the continuous emotion prediction system that is proposed for predicting emotion ambiguity as beta distributions. The input to the system is the front-end speech features and the time continuous emotion labels from multiple raters and the output prediction are beta distribution parameters. . . . . . . . . . . . . . . . . . 120
6.6 Heat-map of log-likelihoods over the beta parameter space for widely distributed valence ratings on top plots and clumped ratings in the bottom plots. The parameterization on the left column of plots is shape parameters and the right column is mode-concentration.

6.7 Top: The grey lines are the arousal ratings. The green shade is over the area over the arousal support that falls within the top 40% of the cumulative density of the predicted arousal ambiguity distribution. Bottom: The variance of the MAP estimates on the development set are plotted in blue and the predicted variance is plotted in red.

6.8 Top: The grey lines are the valence ratings. The green shade is over the area over the valence support that falls within the top 40% of the cumulative density of the predicted valence ambiguity distribution. Bottom: The variance of the MAP estimates on the development set are plotted in blue and the predicted variance is plotted in red.

6.9 A histogram of the maximum likelihoods for all frames of the valence development partition.

6.10 Bar plots of Mean Squared Error (MSE) of prediction mode (left column) and standard deviation (right column) for frames grouped into deciles on the standard deviation of the MAP estimates of arousal (top row) and valence ratings (bottom row). The x-axis denoted the decile and y-axis the MSE. The red crosses depict the MSE of the frames in the indicated decile. The bar plot shows the MSE computed over all frames accumulated up to the indicated decile.
6.11 Illustrating that the mean value of an utterance’s ambiguity’s spread increases it may not result in the framework being able to learn the standard deviation of the ambiguity as demonstrated by the individual $\rho$ and $\text{CCC}$ values per utterance.

6.12 The figure displays the histograms constructed from the standard deviations of all frames of utterances 4 and 9 shown in blue and orange respectively.

6.13 The Relative Error per frame of arousal rating is exhibited with a $\ast$ that is color-coded according to the decile it belongs to. The x–axis is the standard deviation of the MAP estimate and the y–axis is the RE percentage. The RRMSE is calculated collectively over all frames within a decile and is shown with black $\ast$.

6.14 Frame-wise standard deviation of arousal MAP estimates shaded differently for each decile.

6.15 The Relative Error per frame of valence ratings is exhibited with a $\ast$ that is color-coded according to the decile it belongs to. The x–axis is the standard deviation of the MAP estimate and the y–axis is the RE percentage. The RRMSE is calculated collectively over all frames within a decile and is shown with black $\ast$.

6.16 Frame-wise standard deviation of valence MAP estimates shaded differently for each decile.

6.17 Histograms of the frame-wise skewness calculated over the training partition for arousal MAP estimates on the top and valence MAP estimates at the bottom.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Key differences between the RECOLA and SEWA databases.</td>
<td>59</td>
</tr>
<tr>
<td>2.2</td>
<td>Key differences between the RECOLA and SEWA databases.</td>
<td>60</td>
</tr>
<tr>
<td>3.1</td>
<td>CCC reported for arousal and valence predictions on RECOLA development partition [5]</td>
<td>74</td>
</tr>
<tr>
<td>3.2</td>
<td>CCC reported for arousal and valence predictions on RECOLA validation partition in [6]</td>
<td>75</td>
</tr>
<tr>
<td>3.3</td>
<td>CCC obtained with proposed BoOC +LSTM framework for different codebook sizes on RECOLA development partition [5]</td>
<td>77</td>
</tr>
<tr>
<td>4.1</td>
<td>Single modality - audio or video, performance with the ARX back-end</td>
<td>89</td>
</tr>
<tr>
<td>4.2</td>
<td>Feature and decision-level fusion systems with and without label normalisation in terms of CCC</td>
<td>89</td>
</tr>
<tr>
<td>4.3</td>
<td>Comparison between language-independent (LI) and language-dependent (LD) models</td>
<td>90</td>
</tr>
<tr>
<td>4.4</td>
<td>Test performance for CES challenge.</td>
<td>91</td>
</tr>
</tbody>
</table>
5.1 Aggregate measures over all frames .................................................. 98

6.1 Concordance Correlation Coefficient on the training set of arousal MLE and predicted (unsmoothed) standard deviations for different parameterization choices. ................................................................. 128

6.2 Concordance Correlation Coefficient on the development set predicted mode and variance for different loss functions based on MLE GT labels. ......................................................... 128

6.3 Concordance Correlation Coefficient (CCC) and Pearson’s Correlation Coefficient (\(\rho\)) measure between predicted standard deviation (unsmoothed) and the standard deviation of ambiguity model. .................................................. 134

6.4 Concordance Correlation Coefficient (CCC) and Pearson’s Correlation Coefficient (\(\rho\)) measures between predicted and standard deviation and the standard deviation of the ambiguity model after post-processing. .......................... 134
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARX</td>
<td>AutoRegressive-eXogenous</td>
</tr>
<tr>
<td>AVEC</td>
<td>Audio/Visual Emotion Challenge</td>
</tr>
<tr>
<td>BLSTM</td>
<td>Bidirectional Long Short-Term Memory</td>
</tr>
<tr>
<td>BoAW</td>
<td>Bag-of-Audio Words</td>
</tr>
<tr>
<td>BoOC</td>
<td>Bag-of-Optimized Clusters</td>
</tr>
<tr>
<td>CC</td>
<td>Correlation Coefficient</td>
</tr>
<tr>
<td>CCC</td>
<td>Concordance Correlation Coefficient</td>
</tr>
<tr>
<td>CEPS</td>
<td>Continuous Emotion Prediction System</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Network</td>
</tr>
<tr>
<td>COMPARE</td>
<td>COMputational PARa-linguistics Challenge set</td>
</tr>
<tr>
<td>DNN</td>
<td>Deep Neural Network</td>
</tr>
<tr>
<td>ECG</td>
<td>ElectroCardioGram</td>
</tr>
<tr>
<td>EDA</td>
<td>Electro-Dermal Activity</td>
</tr>
<tr>
<td>EEG</td>
<td>ElectroEncephaloGram</td>
</tr>
<tr>
<td>eGEMAPS</td>
<td>Extended Geneva Minimalistic Acoustic Parameter Set</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximisation</td>
</tr>
<tr>
<td>EWE</td>
<td>Evaluator Weighted Estimator</td>
</tr>
<tr>
<td>FAU</td>
<td>Facial Action Units</td>
</tr>
<tr>
<td>GEMAPS</td>
<td>Geneva Minimalistic Acoustic Parameter Set</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>GMR</td>
<td>Gaussian Mixture Regression</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian Process</td>
</tr>
<tr>
<td>GPR</td>
<td>Gaussian Process Regression</td>
</tr>
<tr>
<td>HCI</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>KDE</td>
<td>Kernel Density Estimate</td>
</tr>
<tr>
<td>LLD</td>
<td>Low-Level Descriptor</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short-Term Memory</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MAP</td>
<td>Maximum A Posterior</td>
</tr>
<tr>
<td>MFCC</td>
<td>Mel Frequency Cepstral Coefficients</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum Likelihood Estimate</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>PAD</td>
<td>Pleasure, Arousal and Dominance</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>RECOLA</td>
<td>REmove COLlaborative and Affective interactions</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>RLR</td>
<td>Regularized Linear Regression</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Network</td>
</tr>
<tr>
<td>RRMSE</td>
<td>Relative Root Mean Square Error</td>
</tr>
<tr>
<td>RVM</td>
<td>Relevance Vector Machines</td>
</tr>
<tr>
<td>SAL</td>
<td>Sensitive Artificial Listener</td>
</tr>
<tr>
<td>SDC</td>
<td>Shifted Delta Coefficients</td>
</tr>
<tr>
<td>SEMAINE</td>
<td>Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-Distributed Stochastic Neighbour Embedding</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Speech based continuous emotion prediction - An Overview

Emotion is the social display of the affective state of a person [7] and affects expressions of speech, face, body language, etc. Since speech forms an integral part of human interactions, is easy to collect non-invasively [8] with the omnipresence of speech recording devices such as telephones and mobile phones for a long time, and humans are particularly good at perceiving nuances of emotion from voice [9] speech emotion prediction research has become integral for the development of futuristic human-computer- interaction-based applications [10]. Speech emotion prediction can be leveraged in healthcare, education, marketing, customer service, gaming, and a plethora of applications [10] improving human-computer interfaces. This has led to extensive speech-based emotion research over the last decades. Within speech, linguistic information is obtained from what is being said, i.e., words spoken, and paralinguistic information refers to how the words are being spoken. In some studies, it has been reported that the felt emotion of the speaker is revealed less deceptively in the paralinguistic aspects of speech [10] since it may be harder to disguise true emotion than with linguistic information. Traditional systems, aim to learn human emotion from paralinguistic features and compare it to human emotion ratings [11]. Recently, end-to-end frameworks that automatically learn speech feature representations to
So how are emotions represented in affective computing? While a universal agreement on maximizing performance on prediction evaluation metrics but are not theoretically backed to provide information on the underlying mechanisms in the speech that are affected by emotion.

Figure 1.1: Minimalistic emotion class locations in the arousal-valence dimensional plane in study [1]
1.1. SPEECH BASED CONTINUOUS EMOTION PREDICTION - AN OVERVIEW

how emotions should be represented in computational systems is not yet existing [12] [1], there are two widely adopted approaches - namely, categorical and dimensional representations. Categorical representations group emotions into classes or categories such as Angry, Happy, etc. [13]. The minimalistic collection of emotion categories consists of six broadly segregated emotion states, anger, fear, disgust, joy, sadness and surprise, employed by most research as illustrated in Figure 1.1. Plutchik’s [2] defined eight primary color-coded emotion categories as illustrated in Figure 2 with increasing intensity of emotion towards the wheel’s center. Other research, such as [14] consists of 154 distinct emotion categories including separate categories for subtle emotion expressions e.g. annoyance, aggressiveness, and complex emotions such as embarrassment. More subtle or even complex emotional states are hard to categorize. One reason is that emotion perception is highly individualistic and fine nuances distinguishing subtle emotion categories are obscure with no universally agreed upon boundaries. This may be viewed as an issue of insufficient information being available to ascertain the location and boundaries of emotion categories. Moreover, emotion perceptions change over time depending on preceding affective states which compounds the problem. For example, the distinction between self-consciousness and embarrassment is illustrated by the differing location of the class of 'neutral emotion" according to the level of excitation and pleasantness in two studies [1] and [3] illustrated in Figure 1.1 and Figure 1.3. Moreover, some emotions may be considered as a subset of another emotion category, e.g. happiness may be rated as joyful.

The dimensional approach allows quantifying how differently the emotions were perceived, provides the flexibility of choosing the number of categories, and also provides the choice of selecting emotion categories based on dimensional rating [15]. Dimensional emotion representations are emotion intensities graded by continuous real values along orthogonal affect dimensions such as arousal which perceived excitation of the speech, and valence which is how pleasant the speech was perceived to be [15]. The most accepted dimensional representation in speech emotion prediction systems consists of two basic emotion dimensions, arousal, and valence. In recent years more interests have been directed at the problem of continuously predicting emotion over time as it changes and this has been the focus of some of the more recent emotion challenges [16] [17] [5]. This work focuses on speech
Figure 1.2: Plutchik’s Wheel of Emotions. [2]
1.1. SPEECH BASED CONTINUOUS EMOTION PREDICTION - AN OVERVIEW

Figure 1.3: Illustration of different emotion class locations and boundaries in the arousal-valence dimensional plane. [3]

Continuous emotion prediction tasks require that the intensity of the perceived emotion is tracked at very short time intervals in the order of milliseconds to achieve the continuous time effect, emotion is a relatively slowly varying quantity in the order of several seconds [13]. A typical emotion database would contain multi-modal (speech, video, EEG, etc.) recordings of emotionally colored expressions or interactions. For each such utterance, multiple annotators are requested to grade their momentary emotion perception with a real number bounded within a fixed interval, throughout the recording. Ratings are
elicited from multiple annotators for each recording since annotations are expected to indi-
individualistically vary because of bio-physical factors and the unique history of experiences
of each annotator. When emotion is clearly expressed, usually, the agreement between
ratings is high versus low agreement for subtle expression where rater interpretations may
vary more. In short, emotion ambiguity arises due to the subtlety of the expressed emotion
and the individualism in its perception and varies over the duration of the speech.

Conventional deep learning and state-of-the-art machine learning frameworks \[4,18,19\]
employ the mean annotation over raters as the ‘ground truth emotion labels’ to learn the
relationship between emotion and the speech features \[20\] with the intention of attenu-
ating individualistic variations and the research focusing on ambiguity-aware systems is
rather limited. There exists a plethora of research in affective computing that employs
deep learning algorithms owing to the increasing availability of computational power and
quality speech recordings with state-of-the-art deep learning algorithms achieving com-
petitive performances in the task of continuous emotion prediction. However, the affective
computing systems holding the mean rating to be the only quantity of interest, attenu-
ating individualistic variation in opinions of raters ignore valuable information regarding
how ambiguously emotion is expressed and perceived. This inter-rater variability is most
likely to be considered a reasonable reflection of the ambiguity in the expressed emotion
since more subtly expressed emotions are less obvious and harder to identify by individ-
uals. Ambiguity may carry useful information, adding a layer of complexity to current
affective computing systems that shows potential in making the predicted emotion more
human-like.

1.2 Thesis Objectives

Given the limitations discussed in section 1.1, the three principle objectives of this thesis
are presented in this section. Firstly, using discriminatively learned feature representa-
tions learned on the training set has been shown to outperform standard acoustic feature
representations when tested on the most popular emotion databases \[6\]. However, unlike
1.2. THESIS OBJECTIVES

standard acoustic features, discriminatively learned feature representations are not backed up by the theoretical mechanisms that produce speech [11]. Thus, it was of interest to:

- Investigate feature distributions for standard acoustic feature sets
- Develop a representation or choose one of the existing representations to adequately represent the acoustic feature sets
- Develop a framework to incorporate label information with acoustic features

Following this, an alternate feature representation was proposed, based on para-linguistic features partitioned in the feature space based on discriminative training with emotion annotation information.

Following the investigation of feature distributions, were investigations on emotion label distributions. The speech utterances to examine label distributions were from different languages to investigate the language-wise label distributions and to develop systems across the chosen languages. The various partitions of the language-wise speech data consistently showed similarity in label distributions within a chosen language, that differed between languages, indicating that the individualistic variations in each language were uniquely distributed. Although, the state-of-the-art systems on cross-language effect were achieving competitive performances, they predominantly employed the mean annotation of all available ratings in the given language as the ground truth rating.

As explained in section [1.1] most current systems employ mean over available ratings as ground truth emotion labels as the emotion that need to be learned and that ignores individualistic variations in emotion perception losing the richness of information on the ambiguity of the expressed emotion and its perception. One of the main focuses of my thesis is finding an appropriate model for emotion ambiguity, addressing the challenges associated with learning and predicting the chosen ambiguity models, and finally proposing a system to predict ambiguity. The current limitations of existing ambiguity models are explored and measures to compare them have been designed. The in-depth analyses and proposed techniques will be further explained in detail in the following chapters.
1.3 Thesis Organization

The remainder of the thesis is organized as follows:

Chapter 2 describes the different types of emotion representations and provides an overview of the different aspects of the traditional continuous emotion prediction systems and current techniques for evaluating and comparing their performances in terms of the predicted emotion intensity in terms of three affective attributes; namely, arousal, valence, and dominance. The most popular and commonly employed emotion databases are introduced, focusing on the RECOLA dataset in [20] and the SEWA dataset [21] that was used to generate all of the experimental results presented in this thesis. Besides the descriptions of the subsystems the section also highlights the challenges and limitations currently associated with them.

Chapter 3 explores the most commonly employed high-level feature representations in terms of being able to capture the probability distributions of the low-level features and proposing suitable representations. Additionally, it proposes a novel framework that is able to incorporate label information in the feature representation that is made from a selection of standard acoustic features and investigates if the new representation is able to incorporate beneficial information for the task of emotion prediction.

Chapter 4 investigates cross-language emotion with linear back-end - autoregressive exogenous models (ARX) which were investigated for robustness in generalizing across languages while still being able to capture key emotion information from speech and facial expressions.

Chapter 5 analyses the limitations of the most commonly employed parametric distribution in modeling ambiguity. Secondly, it proposes a methodology to compare families of distributions for their appropriateness in modeling emotion ambiguity. Finally, an exhaus-
tive list of bounded distributions and their subsets are compared and the most suitable model for emotion ambiguity is proposed.

Chapter 6 explores the different aspects of the beta distribution most relevant to ambiguity characteristics largely relevant for emotion prediction systems in terms of their usefulness and comparability with current ambiguity models. Additionally, a novel probabilistic framework is proposed to predict emotion dimensional intensities, and an analysis is presented on why the predicted emotion ambiguity may have been able to outperform current emotion ambiguity models.

Chapter 7 concludes the thesis with a summary of the research contributions and presents potential future research directions to follow up beyond this thesis.

1.4 Major Contributions

The research presented in this thesis explains the original contributions to the assessment of continuous emotion intensity predictions in terms of arousal, valence, and liking based on paralinguistic speech cues. The major contributions can be summarised as follows:

- A novel framework was developed that learned a new feature representation based on label information for the prediction of point estimates of emotion. The proposed feature representation is called the Bag-Of-Optimised clusters which are histograms created based on the bag-of-words concept with the novelty that the codebook is learned discriminatively. Results suggest that the discriminatively learned cluster centers are indeed helping improve the mean emotion prediction and the experiments visualized on a reduced dimensional space showed that there is significant movement of cluster centers when they are discriminatively trained. The details of this contribution are reported in chapter 3.
• Emotion label distributions were examined across languages and frameworks proposed and compared with different feature fusion strategies with language-dependent and language-independent emotion models. Cross-culture affect was investigated with a linear regression backend that can incorporate the temporal evolution of affect. Moreover, the cross language affective labels that were jointly trained with label normalization across languages yielded higher prediction performances on languages from similar cultures (German and Hungarian) versus the other cultures (Chinese). Also, the language-dependent training models were superior in predicting the emotion in unseen test languages (Chinese). The details of this contribution are reported in chapter 4.

• Measures were proposed for comparing families of distributions to assess the distribution family’s suitability in modeling emotion ambiguity. The proposed measures are based on the entropy estimates on the parameter spaces of the families of distribution or the sharpness of the distributions in the parameter space. The measures were employed to compare an exhaustive list of bounded distributions and the normal distribution families. Results suggest that the beta distribution is the most suitable distribution for modeling ambiguity. The details of this contribution are reported in chapter 5.

• A framework based on time-varying emotion ambiguity modeled with beta distributions for both arousal and valence is proposed. Results indicate that with the alternate representation of beta distribution parameters, namely mode-concentration instead of mean-variance, it is possible to enhance the loss function. Moreover, the importance of incorporating prior beliefs is demonstrated while modelling ambiguity for valence. Experimental validation of the proposed framework is found to be able to track the level of ambiguity in the labels over time and predict the emotional state accurately within regions of high agreement. The details of this contribution are reported in chapter 6.
Chapter 2

Speech based Emotion Prediction
- A Review

Chapter 2 introduces the concept of emotion in this study’s context, its expression, perception, and representation, and then describes continuous emotion prediction systems (CEPS), the widely employed emotion databases, and evaluation metrics to assess the performance of prediction systems. The CEPS description encompasses discussions on different cues present in speech indicative of emotion including acoustic cues and verbal and non-verbal vocalization cues and their representations as it is essential to understand the underlying aspects we are trying to capture by the standard features used for emotion recognition and most popular back-ends used for emotion prediction.

2.1 Emotion: its expression, perception and representation

Emotion may be described as a strong feeling deriving from one’s circumstances, mood, or relationships with other people. Amongst the limited well established theories in emotion research, Darwin’s theories \cite{22} are one of the earliest, suggesting that there are universal expressions of basic emotions amongst animals and that it is extremely hard to conceal
the true emotion felt based on behavioral observations encouraging speech-based emotion research. However, plenty of consequent literature on shows differences in agreement on the labeling of emotion and aspects such as boundaries and number of emotion classes, and duration of emotion since emotion is such a subjective concept. The subjective definitions of emotion boundaries make research in artificial emotion hard to compare with each other. An important characteristic of learning emotion is the time duration over which emotion may be considered stationary and determine the rate of its evolution. suggests that emotion lasts for a few seconds and may extend for a few minutes but not more. Moods on the other hand last over a more prolonged duration of several hours and might provoke the frequency and type of emotions experienced. For example, a person in a melancholic mood expresses emotions such as sadness and nostalgia more often within the duration that the mood lasts. Although it may be a good idea to identify the mood of a person according to, the mood is not expressed but just a prolonged mental state that provokes short bursts of expression such as emotion and therefore is hard to detect. The following section discusses the sources introducing subjectivity in emotion definitions, that is, its expression and perception.

2.1.1 Emotion Expression and Perception

In this study, the term "expressed emotion" refers to the conscious mental reaction (such as anger or fear) based on one’s circumstances, mood, or relationships with others that is expressed in outward behavioral cues such as speech and facial expression. Emotion perception on the other hand is the listener’s interpretation of the expressed emotion and is influenced by individualistic variations based on their bio-physical factors and personal circumstances. A majority of the speech emotion databases within affective computing contain annotations of perceived emotion from multiple observers of each speech utterance. So what are the different representations of emotion for which ratings are collected? This will be discussed in the following section 2.1.2.
2.1.2 Emotion Representation

As described in section 1.1, due to the absence of a universally established theory of emotions, there are differences in opinion with representing emotions. However there are three widely accepted representation schemes namely, categorical, dimensional and appraisal-based.

2.1.2.1 Categorical Representation of Emotion

Of these, the categorical emotion representation is the most intuitive wherein, emotions are grouped by class labels such as sad, happy etc. The most basic set of emotions consist of six categories namely - anger, disgust, fear, happiness, sadness and surprise [27]. and the set of emotion categories encompassing more subtle, complex [23] and culture-specific emotions may exceed 165 emotion categories. One of the issues with categorizing emotions is that the class labels, number of emotion classes, and their boundaries are not universally agreed upon and the chosen class labels for study may not be nuanced enough or account for all the emotions perceived from the speech [28]. Additionally, there may be emotions that are subsets of multiple emotion categories and few emotions such as embarrassment, are not covered in a simplistic categorical framework.

2.1.2.2 Dimensional Representation of Emotion

Alternatively, emotions may be more flexibly represented by the dimensional approach [29] wherein, instantaneous emotion is a point plotted on an n-dimensional vector of emotion primitives [30]. The widely adopted emotion primitives are arousal which indicates how active or energetic the behavioral cue is, valence which is an indicator of pleasantness and dominance which is an indication of how controlling the behavior is. One of the theories supporting such a representation is that the emotional state is a result of interconnected neurophysiological activity rather than separate neural systems which were thought to be responsible for different emotion categories [31]. Several dimensional models of emotion
Figure 2.1: Illustration of placement of emotion categories on the plane of emotion dimensions.
2.1.2 Emotion Representation

have been proposed and one of the most widely accepted models in emotion prediction research is the circumplex model which assumes that any emotion may be represented by the arousal and valence \[32\] \[33\] \[21\] \[4\]. The relationship between the dimensional and categorical emotion representation is elucidated by Figure 2.1 which illustrates that the levels of arousal and valence may be combined to reflect the emotional category that is being perceived. In addition to the circumplex model is the dimension of dominance besides arousal and valence, resulting in the pleasure-arousal dominance (PAD) emotion model that is able to represent a wider range of emotions \[34\]. The dominance quantity is how controlling or submissive the emotional cue is. For example, anger and fear both have high arousal and low valence values however, anger is a dominant emotion whereas fear is a submissive one \[35\]. The PAD model \[34\] is shown in Figure 2.1. Previous literature has reported that arousal and dominance appear to be highly correlated \[36\] \[37\] which is why most continuous emotion research is in the arousal-valence space. However, these existing dimensional emotion models may not be sufficiently capturing emotions \[28\] since a two-dimensional space may be too simplistic, and also there is currently little agreement on the number and choice of emotion primitives beyond the arousal-valence space \[28\].

2.1.2.3 Appraisal-based Representation of Emotion

The third representation type is based on the appraisal theory which assumes emotions to be generated through continuous, recursive subjective evaluation of both the person’s internal state and the external environment \[38\]. The OCC model \[39\] is a computationally tractable model based on appraisal theory and has mostly been used in affect synthesis \[27\] which distinguishes 22 emotion types differentiated by their respective psychologically significant situations. However, comparatively limited literature has investigated the appraisal theory in affective computing research to other emotion representations.

Emotion recognition \[3\] \[6\] \[40\], may refer to a classification task that is to identify the category of emotion learned from multi-modal behavior signals \[41\] or a regression task
called emotion prediction which generally refers to the task of predicting the intensity of the perceived emotion dimension (i.e., arousal and valence) from the behavioral cues \[4,6\]. The intensity of emotion is measured by a number within a scale. A dimensional emotion prediction system \[4\] typically employs regression models to capture the relationship between the multi-modal behavior signals and the dimensionless numerical valued emotion intensities.

I focus on emotion prediction systems based on dimensional emotion representations, specifically the two-dimensional circumplex model \[31\], throughout the thesis.

2.2 Emotion Ambiguity: Motivation and Introduction

As briefly explained in chapter 1, emotion databases \[42,43\] typically contain ratings based on individual perceptions from several raters. The reason for collecting emotion ratings from multiple raters is that there may be individualistic variations in the perception of the emotion of the same speech utterance which may be a result of the raters’ bio-physical factors and socioeconomic circumstances \[15\]. It is reasonable to expect that the emotion is more ambiguous when there is a high variation in perceptions. Ambiguity in emotion may arise due to:

- clarity of the expressed emotion in the behavioral cue of the speaker.
- the interpretation of the emotion influenced by the listeners’ personal bio-physical factors and history.

Figure 2.2 demonstrates how the perception differences between individuals vary with time. Figure 2.2 shows 800 seconds of arousal annotations from 3 annotators illustrated by the blue, red, and yellow lines. The green window between 100-200 seconds highlights a region reflecting low ambiguity in perceived arousal versus the purple window between 600-700 seconds where the ambiguity is high. Although all speakers tend to perceive arousal positively there are changing individualistic variations over time.
2.2.1 Point-Estimation of Emotion

Traditional CEPS research predominantly assumes the ground truth labels to be the mean emotion rating, that is the average rating over all raters to attenuate individual variations. Techniques to assess inter-rater agreement such as intra-class correlation coefficient (ICC(3,1)) \[44\], and Cronbach’s α \[45\] are applied to measure the degree of concordance between two (or more) series of ratings. Ratings with large deviations from the mean ratings or most other ratings are considered to be unreliable and discarded. The inherent assumption here is that the inter-rater perceptual variations are treated to be constant over the duration of the utterances whereas in reality as illustrated in Figure 2.2 which conveys useful information on emotion ambiguity. An obvious observation in such a model is that the mean ratings may be the same for certain sections of speech but the inter-rater agreement may be vastly different.

2.2.2 Ambiguity-Aware Estimation of Emotion

Ambiguity information may help identify universal emotion cues from more culture-specific emotion cues. Previous studies such as \[19\] have stated that ambiguity in perceived emo-
tion plays a crucial role in determining the uncertainty in CEPS. In order to model emotion ambiguity, there needs to be a sufficient number of emotion ratings available from different raters for a given snippet of expressed emotional speech [32]. The literature on ambiguity-aware systems is fairly limited but is gaining increasing interest recently [46]. Finally finding an appropriate model for emotion ambiguity and building a framework treating emotion such as a distribution warrants deeper consideration. It is worth mentioning that in ambiguity-aware research, techniques to penalize relatively larger individualistic variations mentioned in section 2.2.1 are not employed and the normalization techniques are not used.

2.2.2.1 Non-Parametric Distribution Models

Broadly speaking, both parametric [32] [33] [19] and non-parametric distribution models [47] [48] can be used to represent distributions over emotion ratings. Non-parametric models are less restrictive in terms of prior assumptions about the shape of the distribution but also tend to be less efficient computationally. Amongst the modelling of ambiguity with non-parametric models ambiguity, has been modelled using Gaussian Process-Particle Filters which have shown compelling performances [49] and also most commonly been modeled as a Gaussian process [47,50] which found that in addition to treating time-varying arousal and valence labels as Gaussian processes, explicitly modelling the temporal dynamics leads to better models than [40].

2.2.2.2 Parametric Distribution Models

Parametric models may be more appropriate when the amount of data available to estimate the distribution is limited [51]. Additionally, the parameters of most parametric models tend to each describe definite distribution characteristics [52]. Firstly, parametric probability distributions by the virtue of their generality are usually more powerful than their non-parametric counterparts as their formulations can be extended to unseen data [51]. Then, parametric probability distributions are defined by a set number of parameters each
responsible for determining definitive characteristics of the distributions and hence they are compact and generalized representations of variations and uncertainty in data [52]. Moreover, parametric distributions and their statistical parameters have existing closed-form expressions and are computationally efficient and less time-consuming to process and therefore the use of parametric models are more commonly used in emotion recognition systems to model ambiguity. Amongst the limited research in ambiguity prediction, the parametric methods of modelling ambiguity of arousal and valence have most commonly the Gaussian distribution. Most research has assumed the ambiguity distribution is Gaussian [19] employing multi-task learning to train a BLSTM-based system to predict the mean and standard deviation of annotator ratings, or treated the final prediction to be the dominant Gaussian component of the Gaussian Mixture Model moreover assuming that the prediction at each instance to be independent of all other instances [53]. In [53], Gaussian mixture models were used to represent arousal and valence distributions and did not model temporal continuity at first and then Kalman filters were added to model temporal variations in the distribution parameters as a linear dynamical system [54]. Most commonly, Gaussian distributions are employed for their computational convenience [55–57], but other alternatives include Gaussian mixture models [53,54]. However, despite the computational convenience, assuming that numerical emotion ratings are distributed normally may not be accurate since the numerical scales themselves are bounded (for e.g., valence ranging from −1 to 1) [57] and a symmetrical distribution might not be the best fit for the ratings. There needs to be more research on quantitatively determining the appropriateness of the assumption of Gaussianity that is often made are defined and employed compared to other types of distributions. Most recently [32] found that representing a set of emotion ratings as a Gaussian implicitly implies that there is a non-zero probability that emotion ratings are available beyond the allowable emotion limits due to an infinite Gaussian span which in practice, might be an acceptable approximation as long as the variance of the distribution is small which may not be the case always. Moreover, the Gaussianity assumption is less likely to be suitable when the numerical labels are close to the edge of the interval of allowable values. It was found from a quantitative comparison of an exhaustive list of parametric distributions, that beta-distributions were most suitable
in modelling ambiguity of arousal/valence ratings [32].

The choice to employ beta distribution within a continuous emotion prediction system, however, comes with a number of challenges that needs addressing. Unlike the better-understood formulation for Gaussian distributions and corresponding choice of loss functions for training neural-network to predict them since [58] [59], the choice of beta distribution’s parameterization paired with loss function choices and output activation and their gradient definitions are not understood and call for further research.

2.3 Speech-based emotion prediction – system overview

Just as a human hears speech and the brain is able to identify the emotionally relevant parts of the uttered speech and gauge the intensity of the emotion, a speech-based emotion prediction system aims to identify the aspects of uttered speech that are affected by emotion and then learn the relationship between the uttered speech and the perceived arousal and valence levels by humans. The traditional methodology is a supervised learning-based approach and its overview is illustrated in Figure 2.3.
2.3.1 Pre-processing

As Figure 2.3 displays, the traditional CEPS system is comprised of two phases, first - the training phase, and second, the prediction (test) phase. The inputs for both these phases are raw speech with the difference that and emotion labels are available for the speech inputs in the training phase. The first task in both these phases is to extract the emotionally relevant aspects of speech which include pre-processing steps and feature extraction described in sections 2.3.1 and 2.3.2 respectively. Pre-processing may refer to one or more of the following processes such as pre-emphasis, de-noising, de-reverberation, down-mixing, re-sampling, source- separation, truncating the speech signal, delay compensation and cleaning, etc. some of which will be described more in detail in section 2.3.1. The next step in the training phase focuses on learning the relationship between the extracted speech features and their corresponding emotion labels with regression modeling techniques which are described in more detail in section 2.4. Once the regression model is trained, it is utilized in the prediction (test) phase to predict emotion labels for speech data whose emotion labels have not been made available to the system.

2.3.1 Pre-processing

The pre-processing step is a sequence of pre-processing phases to prepare the raw speech waveforms for further processing (feature extraction) to improve model accuracy and efficiency. Some of the pre-processing phases may be performed on the raw digital speech waveform to enhance the speech signal i.e. de-noising (reducing background noise) the signal or de-reverberation(reducing the impact of varying room impulse responses), extracting parts of interest, and clean the audio signal to remove the background noises, silent portion and other irrelevant information from speech signal such as separating sources. Popular source separation algorithms include independent component analysis [60] in the case of multiple microphones/arrays, and non-negative matrix factorization [61] in the case of single microphones. Automatic gain control may be performed to mitigate the impact of varying recording levels and microphone gains. The most commonly employed pre-processing step before the feature extraction from the state-of-the-art emotion databases is pre-emphasis which is achieved by passing the signal through a first-order FIR filter and
delay compensation. This is succeeded by framing the speech, a method of partitioning the speech signal over short time frames for the next phase of feature extraction as explained in 2.3.2.

2.3.1.1 Pre-emphasis filter

The pre-emphasis filter is applied to amplify the high frequencies in the speech signal primarily to balance the frequency spectrum as generally higher frequency components have smaller magnitudes and consequently prevent numerical problems during the Fourier Transform computation. The generic equation of the pre-emphasis filter is given by (2.1).

\[ \tilde{s}[m] = s[n] - \alpha . s[n - 1] \]  

(2.1)

Where \( \tilde{s}[m] \) is the \( m^{th} \) sample value of the speech signal. Typically the value of \( \alpha \) or the pre-emphasis filter coefficient is chosen to be 0.95 or 0.97. In this thesis the value of 0.97 has been chosen to be consistent with 62, 53, 54.

2.3.1.2 Delay Compensation

Annotation procedures involve a series of wherein the annotators have to view the stimuli, perceive the underlying emotional behavior by making judgments and finally move the cursor to annotate the value and record the annotated level 63. It is reasonable to expect that these processes introduce a delay between the viewing of expressed speech and the collected annotations. This is termed as evaluators’ reaction lag 64. 64 compensated the evaluators’ reaction lag using an approach based on the maximum mutual information criterion 65 and observed the average reaction lag to be over 2s and noted improvements in predictions with shorted reaction lag times.
2.3.2 Feature Extraction

Of the two core steps of speech-based continuous emotion prediction systems (CEPS), the first is the parameterization of the emotion-specific information present in uttered speech or feature extraction. Raw speech signals contain a large amount of data, some of which may be superfluous for the task of emotion prediction that not only makes speech emotion prediction computationally expensive and slow to process but also makes it hard to discern the desired emotion information directly from raw speech. Features are succinct parametric representations of raw speech waveforms that capture the emotionally relevant aspects of emotionally rich raw speech. The selection of feature types may:

- factor in the mechanisms involved in producing speech and the effect that emotion may have on the mechanisms that produce speech.
- learns a representation discriminatively from speech based on emotion labels.

The aspects of speech impacted by emotion are related to sub-glottal pressure, trans-glottal airflow, and vocal fold vibration [66], [67], [68], [69], [70], [71], [72], [11]. A number of studies have proposed to use neural networks as the feature extractor, such as CNNs and variational autoencoders [75]. These features extracted from these networks are not backed by the theory of speech but have demonstrated competitive performance in speech emotion recognition on certain datasets. The following sections will describe the different types of speech features based on their time spans.

2.3.2.1 Short-term Speech Features

As explained before speech is a result of several complex non-stationary processes. However, the features may represent information resulting from one or two of those processes [76] and change at slower rates called frame rates than the sampling rate of the speech signal [77]. Within these short frame duration of a few milliseconds (25ms -60
ms \[11\]), the processes are assumed to be quasi-stationary \[76\]. Since each feature is process dependent the time interval where the signal is assumed to be quasi-stationary for that feature’s extraction depends on the process responsible for it. For windowing of speech into short frames, mostly Rectangular and Hamming windows are used \[78\]. The window span selection depends on the parameter that is being extracted. This is illustrated in Figure 2.4 where the speech signal denoted in blue is split into overlapping short-time frames spaced apart at every 10ms. The set of features obtained with the short-time analysis of speech signals is called low-level descriptors (LLD).

### 2.3.2.2 Frequency Related Features

One of the most important features in this category is pitch \[79\] which although defined as the fundamental frequency perceived by the listener is often taken as the fundamental frequency of the speaker in the context of feature extraction. Typically a single pitch value is extracted from frames about 60 ms long. Some of the other important frequency-related features are jitter and different formants.

### 2.3.2.3 Energy Related Features

Loudness is another very important acoustic for emotion recognition \[80\] which is the sound pressure level perceived by the listener depending both on the intensity of the speech and the distance between the speaker and the listener. Shimmer and Harmonics-to-Noise Ratio are some of the other important energy-related features \[11\].

### 2.3.2.4 Spectral and Cepstral Features

Spectral features generally contain information on the power spectrum such as alpha ratio, Hammarberg Index, spectral slope and harmonic differences. These LLDs represent some
of the spectral energy distribution over the range of speech frequencies. The most popular spectral features for an emotion task historically are the linear predictive coefficients (LPCs) that model the character of the human vocal tract \[81\] and recently the most popular spectral features are the Mel-frequency cepstral coefficients (MFCCs) along with their derivatives \[82\], that are spectral energy measures augmented with the frequency scale of the human auditory system \[83\]. The cepstrum is obtained from the logarithm and then the inverse Fourier transform of the spectrum which is useful for investigating periodic structures in frequency spectra \[84\]. Cepstral features provide a detailed description of short-term spectrum \[85\] and their dimensions are less correlated with each other compared to spectral feature dimensions.

The optimal length of the frame should be short enough to ensure quasi-stationarity of the signal with respect to the LLD of interest, and long enough so that the frame has enough data to compute the LLD.

### 2.3.3 Feature Extraction Tools

Amongst many speech feature extraction tools employed in speech research, some of the toolkits are the Hidden Markov Model Toolkit (HTK) \[86\], the PRAAT Software \[87\], the Speech Filing System3 (SFS) \[88\], a MATLAB toolbox by Raul Fernandez \[89\], the Tracter framework \[90\], and the open-source tools such as the SNACK package and openSMILE \[91\]. Of these the most commonly employed toolboxes for feature extraction in current emotion prediction systems are the HTK, PRAAT and most recently the openSMILE toolbox. It is worth mentioning that there exist many extraction algorithms for each of the speech features. For instance, pitch estimation may be performed based on short-time autocorrelation functions (ACF) \[92\], zero-crossing rate with autocorrelation \[93\], weighted autocorrelation \[94\], direct time domain fundamental frequency estimation \[95\], autocorrelation of the cepstrum \[96\], Average Magnitude Difference Function (AMDF) detectors \[97\], Comb Transformation \[98\], sub-harmonic summation (SHS) \[99\], etc.
Figure 2.4: Illustration of how speech waveform is framed depending on the chosen LLD types, then LLD vectors are constructed over overlapping speech segments and statistical features are computed over a longer window of a few seconds.
2.3.4 High-Level Feature Representation

Different research methods use different feature extraction algorithms with their own assumptions, approximations, advantages and limitations. Moreover, although there is an approximate estimate for the appropriate time intervals to be considered for each LLD, the exact choice of the interval lengths and windowing functions to extract the speech segment of interest varies which unsurprisingly leads to a difference in predictions in experiments of this study. There needs to be more research done for a standard choice of windowing functions and their duration for the LLD extractions or at least the choice of windowing needs to be specified for replication of results.

2.3.4 High-Level Feature Representation

The short-time LLDs extracted over tens of milliseconds are incapable of capturing the slowly varying long-term temporal evolution of emotion over several seconds. A popular approach is to compute statistical aggregates over LLDs present in longer emotion-relevant time windows \[100, 101, 102, 103\]. The temporal patterns may not be captured if the higher-dimensional functional representation is not carefully designed. The other approach is to employ sophisticated machine learning back-ends with the ability to model the temporal patterns of LLD feature sequences.

2.3.4.1 Functional Representation

With the assumption that the emotional effect remains the same over a window of a few seconds based on arguments presented in \[13, 22, 24, 25\], the statistical descriptions of the LLDs are calculated with the assumption that they will better reflect the emotional effect in the long window and early emotion prediction literature employing statistical features demonstrated good performance \[104, 105, 106\] and the use of statistical features popularly continues \[21\]. Numerous statistical descriptors or functionals are calculated for each LLD type over the longer window of several seconds, e.g., means (arithmetic mean, absolute mean etc.), measures of spread such as ranges, percentiles and standard deviation, extrema (minima and maxima), range, etc., higher moments (skewness, kurtosis, etc.),
peaks (number, distances, etc.), regression (coefficients, error, etc.), spectrum (discrete cosine transformation coefficients, etc.) and tempo (durations, positions, etc.). The next section will explain what feature types are commonly chosen for speech emotion prediction and indicate reasons why they are relevant.

Figure 2.4 illustrates how statistical functionals are computed from the extracted LLDs. A frame size of 10 ms is selected for this illustration. Pitch estimation per frame for example is computed over 60 ms (shown by the blue dashed lines) of the speech signal shifted by the frame size of 10 ms and the duration of the signal considered for MFCC is 25 ms shown by the red dotted lines. Thus are values of pitch, MFCC and other LLDs at every 10 ms. In Figure 2.4 the emotion is considered to be almost the same over a 3-second window and hence statistical functionals are calculated over 300 frames of LLDs indicated by the green arrows.

Other approaches to describe the probability distribution of an LLD statistically such as i-vectors may be used as suprasegmental features [107] [108] but it is more common to employ a small set of statistical measures to form a suprasegmental feature vector, especially as a univariate statistical model with functionals such as [11] [109] like mean, variance, percentiles, etc. to represent the distribution of each of the LLDs.

### 2.3.4.2 GMM Supervectors

Alternatively, the entire probability distribution over the LLDs in an emotion window may be represented by Gaussian mixture models (GMMs). Generally, MFCCs, and their delta and delta-delta features are modelled with GMM since they provide detailed descriptions of short-term spectra and also their dimensions are decorrelated and provided competitive performances in [33]. Since GMM supervectors are powerful representations of feature distributions, variability compensation techniques with GMM-based systems for speaker and language identification tasks are developed and discriminative training techniques can be used with GMM systems, exploring GMM-based feature representation in emotion recognition is worth investigating further.
One problem with GMMs is their speed and complexity due to parametrizations in terms of means and covariance matrices’ computations with expectation maximization per window. Moreover, there are a high number of hyper-parameters that may require experimental tuning. There is an alternate representation that captures feature distributional information by partitioning the feature space into clusters that is far less computationally intensive than is described in section 2.3.4.3.

2.3.4.3 Bag-of-Audio Words

The bag-of-words (BoW) concept was first introduced in natural language processing where each document was represented by a histogram of the words in the document thereby efficiently representing a large number of low-level data e.g. words, images, video pixels, etc. that were used to successfully recognise high-level information e.g. document type, recognizing human action, etc. The bag-of-audio-words (BoAW) is an alternative representation of the speech feature distributions. The BoAW features have resulted in competitive performances in acoustic event detection, multimedia event detection, snore sound classification, speech-based emotion recognition, and many emotion prediction systems such as where the BoAW representation has been enhanced with context to provide even better results than with the BoAW representation. The openXBOW toolkit is generally used for the BoAW extraction. One important advantage of using the BoAW representation is that it can produce fixed-length feature vectors even when the length of the speech recordings varies. Other techniques proposed to solve this problem are GMM supervectors, i-vectors, x-vectors, etc.

From the frame-level acoustic feature vectors i.e. LLDs (e.g. MFCCs) that are extracted over the entire duration of the emotion windows, the BoAW representation is obtained as follows. The frame-level LLD vectors over the whole training data are computed and the feature vectors are clustered based on their location in the feature space with an unsu-
Supervised algorithm like k-means \cite{129}, k-means++ \cite{130} or random sampling as employed in \cite{120,131}. Each LLD vector within an emotion window is then assigned to the closest cluster center and one hot vector is created. All one hot vectors in an emotion window are added to form a histogram to form a compact bag of audio words \cite{121}. Since the length of the BoAW feature depends upon the number of clusters chosen, the BoAW representation may be useful in reducing the dimension of the front-end features and may achieve a robust representation of features that reduce the risk of overfitting when the size of the training set is small. Figure 2.5 illustrates BoAW single assignment where the top plot shows three codewords K1, K2 and K3 in green or codewords on a two-dimensional feature space and a feature vector illustrated with a red cross. Since the feature vector is closest to the cluster center or codeword K3, one hot vector has one assigned to K3 as demonstrated in the bottom plot of Figure 2.5. In the case of multiple assignments, e.g. double assignment, the vector elements corresponding to the closest two codewords are set to 1. Several LLDs have been employed in the context of BoAW extraction, and the most commonly adopted ones are the ones described in the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) \cite{11}. However, recent work has shown that using MFCC LLDs could be better \cite{131}.

The basic BoAW representation is heavily sensitive to the choices for front-end features, codebook sizes for generating the codewords, histogram normalization techniques and the number of training samples. Having too few clusters may be insufficient in capturing sufficient details of the utterance features, while too many clusters may lead to overfitting for a small training data size. When the training data size is large the feature clustering needs enormous execution times. To improve the clustering speed random sampling may be used as it has similar accuracy scores versus using traditional clustering for codebook creation \cite{118} or initialize codebook creation with k-means++ as in \cite{131} wherein, the first codeword is chosen randomly, then each additional center is chosen as the input vector that is the farthest away from the already chosen cluster centers \cite{130}. Regardless of the choice of clustering algorithms employed to obtain the BoAW representation, the cluster positions are sensitive to the randomness of the initialization process \cite{124}. Simple random sampling is a highly stochastic procedure, k-means is sensitive to the cluster center initialization \cite{132}.
and the same may apply for k-means++ initialization for smaller codebook sizes and this adversely affects paralinguistic classification \cite{124}. With this train of thought it was hypothesised in this study that perhaps features that are clustered only using training features may be less optimal compared to clustering the features discriminatively wherein the emotion label information is incorporated with the training features for the task of emotion prediction \cite{4}. The hypothesis and results are further explained in chapter \ref{chapter:emotion_prediction}. 

Figure 2.5: The figure illustrates the scenario of a BoAW single assignment with the top plot showing three cluster centers or codewords in green on a two-dimensional feature space and an input feature vector in red. The bottom plot is an illustration of how the BoAW histogram assignment is done based on the top plot.
2.3.5 Supra-Frame Level Features

Supra-Frame Level Features are estimated over a longer duration than a frame but shorter than an utterance and are commonly employed in emotion recognition systems. These may be classified as follows:

2.3.6 Voice Quality Features

Prosodic features are suprasegmental [133] since they often extend over syllables, words, or phrases and are not limited to single sounds. Therefore, these features are typically extracted over longer speech frames e.g. pitch or interchangeably used term fundamental frequency (f0) of the vibration of the vocal folds are indicative of emotional state [134] [135] [136] [137] [138] [70]. Additionally, other resonant frequencies of the vocal tract i.e. formant frequencies are commonly employed prosodic features [134]. Emotion researchers refer to shimmer and jitter as prosodic features unlike other speech research referring to them as voice quality features as prosody refers to the rhythm, stress and intonation of the speech whereas shimmer and jitter are the qualities of the voice such as harshness, breathiness and nasality.

2.3.7 Linguistic Features

The lexical content of speech focuses on what is being said and may be a means of directly conveying emotions. Such features are extracted at the phoneme or word level by identifying the phoneme or word sequences first. However, since phoneme or word recognisers show poor performance [139] [140], linguistic feature extraction is rendered unreliable for the task of continuous emotion prediction. With the increasing availability of transcripts with databases there is scope for improvement in linguistic feature extractions. Phoneme level Phone Log-Likelihood Ratios (PLLR) features have promising performance in arousal and valence predictions [141]. A bag-of-text-words (BoTW) feature representation based on transcripts has also been proposed in AVEC 2017 [142] and generated with
2.3.8 Non-linguistic Acoustic Events

OpenXBOW \[125\] using 521 single words off which histograms are created. Significant improvements in arousal, valence and likability predictions were observed using the BoTW features over the standard acoustic feature set of Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) \[142\].

2.3.8 Non-linguistic Acoustic Events

Non-linguistic acoustic events refer to descriptions of occurrence and position of events such as disfluencies, breathing, laughter etc. within an utterance \[143\] and are not used as often as linguistic features in emotion recognition. Incorporating nonverbal vocal gestures such as laughter or filler have shown to boost the performance on using the acoustic and linguistic features \[144\], however, their use in emotion recognition needs further research.

2.3.9 Feature Selection

Most speech emotion prediction research focuses on identifying emotion from aspects of speech depending on how the speech is being spoken rather than what is exactly being spoken. Therefore, for the task of speech emotion prediction, speech is most commonly parameterized with acoustic features that are independent of linguistic information. Although there needs to be much more research by psychologists and emotion recognition researchers to ascertain the specific acoustic features affecting impacting particular emotion characteristics, there has been some research indicating certain acoustic features being correlated with the expression of stress \[145\] \[146\] and other emotions \[147\] \[148\] \[149\] \[11\] found their proposed feature set was more accurate for binary arousal than valence classification and \[135\] found that spectral and fundamental frequency (F0) features discriminate valence levels more accurately. Some research \[150\] \[134\] \[151\] indicate that spectral and prosodic acoustic features are able to reflect specific emotional states. For example, \[151\] found spectral features to have high energy for high frequencies in happy speech and early research \[152\] \[153\] found low spectral energy at high frequencies in sad speech.
CHAPTER 2. SPEECH BASED EMOTION PREDICTION - A REVIEW

2.3.9.1 Popular Acoustic Feature Sets

Within emotion recognition research, a universal agreement on which statistical functionals are most relevant in preserving emotional information over LLDs in a given emotion window is still lacking. Some research aims to solve this problem by suggesting a set of pre-selected acoustic parameters and functionals based on their theoretical relevance and experimental performance. Some of the most popular used feature sets are the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [11], which is a minimalistic set of 88-dimensional feature vectors, Interspeech computational paralinguistics challenge features set (COMPARE) [154], Interspeech 2010 paralinguistic challenge feature set (IS10) [155] and Interspeech 2009 emotion challenge feature set (IS09) [155]. Of these, the eGEMAPS feature set is the most recent minimalistic feature set with selected acoustic features based on their theoretical significance, potential to reflect physiological changes in voice production, performance in previous research, and ease of extraction [11] whereas the other feature sets have thousands of dimensions. Dimensionality reduction techniques are often applied to the extensive feature vectors to reduce feature dimensions retaining as much emotionally relevant information as possible. Some approaches to reducing feature dimensions include transforming the feature space.

2.3.9.2 Feature Reduction

The primary objective of feature selection systems is to find the feature subset that achieves the best possible prediction accuracy. The drawback is the requirement of initially building a regression model for a very high dimensional feature space which is computationally expensive and time-consuming. Another method of selecting suitable features for classification tasks is by applying a feature selection criterion which is maximizing the mutual information between the feature type and the emotion class [156].

Feature transformation refers to transforming a high dimensional feature space to a lower
2.4 Regression Modelling Techniques

dimensional space by applying an appropriate linear or nonlinear mapping function to the original feature space preserving as much of the emotion-relevant information in the feature set as possible. Some feature selection methods that were employed for emotion tasks are the Forward Feature Selection (FFS) and Backward Feature Selection (BFS) [157], Sequential Floating Forward Selection (SFFS) [158], wrapper approach with forward selection [159], Principal Component Analysis (PCA) [156] or independent component analysis (ICA) [160] [161] and Linear Discriminate Analysis (LDA) [162], Sequential Forward Selection (SFS) and Fast Correlation-Based Filter (FCBF) [163]. Of these the PCA is the still the most popular technique to perform the feature mapping onto lower dimension spaces [164] [165] [166] [167] [168] [169] [170] [171] [172] [173]. There also exist a number of machine learning techniques such as random forest and decision trees, missing valued ratios, etc. aimed at this task. PCA essentially orthogonally transforms the high dimensional feature vectors to a set of basis vectors such that they are arranged in order of their impact on the variance in the original feature set. The basis vectors with less impact on the variance in the original feature set are discarded resulting in reduced dimensionality. The drawback of this method is the loss of interpretability of the new basis vectors i.e. the resulting principle components cannot be related to speech production systems.

2.4 Regression Modelling Techniques

The regression predictive modeling task here is to find an approximate mapping function that maps speech features to output the continuous-valued emotion dimension annotations of arousal and valence. The regression model may be based on any of the following principles: (a) define a hard margin in the feature space by a set of parameters obtained by optimizing an objective function, e.g. support vector regression, neural networks, autoregressive exogenous (ARX) models; (b) model the joint probability distribution function over the speech feature and emotion space, e.g. Gaussian mixture regression models; or (c) represents a probabilistic mapping function from the features to the labels, e.g. RVM. All approaches show comparable results. LSTMs [75] [174] [175] or Gated Recurrent Units.
(GRUs) remain the most commonly adopted back-ends, as they can capture the temporal dynamics of emotion \cite{176}, occasionally followed by an advanced attention mechanism \cite{6} and SVR described in section \cite{2.4.4.1} \cite{26} owing to its robustness. Apart from the neural network-based back-ends, a significant number of regression modeling techniques such as RVMs \cite{141} and GMR \cite{53} have also shown good performance and been widely adopted in emotion prediction tasks. The general idea behind each of the aforementioned regression models shall be discussed in the following sub-sections of this section.

### 2.4.1 Relevance Vector Machines

The relevance vector machine (RVM) \cite{177} is a probabilistic model based on Bayesian inference to obtain sparse regression probabilistic solutions. The aim of training an RVM system is to determine the set of model weights for each of the feature dimensions and a bias term during the training process and then use the trained weights and biases during the test phase to make the probability predictions. A Bayesian formulation of a linear model is introduced by assigning an appropriate Gaussian prior to the weight of each dimension and a Gaussian assumption to the bias term. One of the key advantages of RVM is of implicit feature selection. Since the weights are usually sparse after training, the model ends up being characterized by only a subset of features thereby achieving feature selection automatically during the training phase. Additionally, the probabilistic output of the RVM is able to provide an estimate of the prediction uncertainty as a Gaussian distribution, which is useful in a plethora of applications such as in clinical applications \cite{178}.

### 2.4.2 Gaussian Mixture Regression

GMR \cite{179} \cite{180} \cite{181} is based on finding the best fitting Gaussian mixture model (GMM) over the joint probability distribution over the speech- features and the affective dimensions during the training phase. Then during the test phase, the conditional density functions and the regression functions are derived from the GMM model developed during
the training phase. This method is primarily a non-linear regression modelling technique over a multi-variate space offering the flexibility of prediction over multiple chosen dimensions. One continuous emotion prediction-based study found that the GMR yields superior performance compared to neural networks [179] but very limited research based on such comparisons has been done so far.

2.4.3 Gaussian Process Regression

Gaussian process regression (GPR) is a non-parametric, Bayesian inference supervised learning technique to estimate and predict posterior distributions over test outputs after constructing a model over the training input and output points. In [40] a detailed covariance structure has been proposed to be able to apply GPR to continuous emotion prediction that is able to model both temporal relationships between points nearby in time and non-temporal relationships between points at any temporal distance, using separate covariance functions that are additive. The work was continued in [50] where GPR was combined with LSTM networks to predict the emotional content of speech as a Gaussian Process (GP). This approach achieved slightly superior accuracy for arousal point predictions compared to only using LSTM and also showed that the probabilistic predictions of the LSTM-GP system were able to more accurately represent the underlying distribution of the target annotations than the scalar predictions of the LSTM network by measuring the log-likelihood of the individual annotations under the predicted distribution over an emotional dimension.

Besides the regression modelling techniques listed above, other regression approaches have been developed for the task of continuous emotion prediction such as distance-based fuzzy k-nearest neighbours and rule-based fuzzy-logic estimators [182]. Upon investigation by [183] these backends were found to be less effective than SVR systems.
2.4.4 Support Vector Regression

The theory of support vector regression (SVR) is derived from the concept of support vector machine (SVM) classification \[184\]. Support vector regression (SVR) \[185\] is one of the most popular and extensively used regression techniques employed by speech-based CEPS \[186\] \[33\] \[187\] and also have been widely used as default baseline backends besides LSTMs. The main advantage of SVR is that it generalises well in many applications \[188\] \[189\] \[190\]. Support vector regression is also a sparse approach which is described more in detail in section 2.4.4.1 is adept at handling high-dimensional feature spaces, making them suitable for emotion prediction front-ends with high-dimensional feature representations of utterances. Moreover, the SVR training is formulated as a convex problem \[185\] which may lead to globally optimal solutions with the training algorithm.

2.4.4.1 SVR Problem Formulation

The goal of the SVR is to find a linear mapping function between the speech features and corresponding emotion annotation. The difference is that while a linear regression model aims to minimize the error between the actual and predicted emotion labels based on the line of best fit, SVR aims to fit the best line within a threshold of values called the \textit{epsilon} ($\epsilon$)-insensitive tube. The problem is formulated as follows.

If there are $x$ speech feature vectors over $N$ training time frames then let the feature vector at the $n^{th}$ time frame be represented by $x_n$ and the corresponding emotion label be $y_n$, where $y \in [-1, 1]$. Figure 2.6 illustrates the speech vectors on a two-dimensional plane with green markers. In linear regression, the features would be fit by a line but the SVR aims to fit the points within a tube that is illustrated with black solid lines in Figure 2.6. In higher dimensions, the two tube edges are two hyperplanes. The orientation of the hyperplane is defined by its normal vector $w$ and its location is defined by bias $b$. Support vectors are those green markers that fall on the tube’s surface. SVR has a tuneable parameter $\epsilon$ which determines the width of the tube around the estimated hyperplane. A
slack variable $\xi$ is introduced to allow for training vectors to be present on the outside of the tube such that, $\xi_n \geq 0$ thereby incorporating a soft margin.

Speech vectors, shown with green markers, that fall within the tube are considered correct predictions and are not penalized by the algorithm. Support vectors are those green markers that fall on the tube’s surface. For cases where the linear function with a margin of $\epsilon$ is not able to accommodate all the training data, slack variable $\xi$ may be introduced. The slack variable $\xi$ measures the distance of the speech vectors to the borders of the tube and is introduced to cope with the infeasible constraints of the optimisation problem in SVR which applies to all feature vectors lying outside the tube. The error of the slack variable is penalized and the penalty may be controlled by tuning the complexity or regularization parameter $C$. The complexity parameter $C$ is a tuneable parameter. A higher $C$ defines the importance of separating all instances of training data i.e. a narrower margin, while a lower $C$ indicates the importance of allowing a softer margin.

For the task of linear regression, a regression function $f(x)$ is as flat as possible given all the $N$ speech training vectors. This is illustrated in Figure 2.6.

$$f(x) = w^T x + b$$ \hspace{1cm} (2.2)

where, $w, b$ are the regression parameters. The emotion label for the $n^{th}$ frame is $y_n \in \mathbb{R}_{[-1,1]}$. An $\epsilon$-SVR aims to find a function as flat as possible with a deviation of $\epsilon$ from the expected original emotion dimensional label for all the possible training data speech vectors. The regression parameters $w$ and $b$ can be estimated by minimizing the margin width as given by (2.3).

$$J = \frac{1}{2}||w^T w||^2 + C \sum_{n=1}^{N} |\xi_n|$$ \hspace{1cm} (2.3)

(2.3) is subject to the constraints given in (2.4) and (2.5) as in [185].

$$y_n - w^T x_n - b \leq \xi$$ \hspace{1cm} (2.4)
CHAPTER 2. SPEECH BASED EMOTION PREDICTION - A REVIEW

Figure 2.6: Speech feature vectors are represented by green markers on a two-dimensional plane. A regression line represented by the dashed line and a tube - the solid black lines, is fitted around the regression line. The feature vectors on the tube represent the support vectors. The vectors present outside the tube contribute to errors based on their distance from the regression lines and are represented by slack errors $\xi$.

$$-y_n + w^T x_n + b \leq \xi$$ (2.5)

The $\epsilon - tube$ is the space within the black solid lines. It is modeled using the speech feature vectors present only on the hyperplanes. Since only the subset of speech features present on the edges of the $\epsilon - tube$ are employed to characterize the regression mode, SVR is said to be a sparse approach. In cases where the tube of separation is non-linear, kernel functions are adopted to transform the original feature space to a higher dimensional feature space to make the data linearly separable. This may be achieved using a non-linear transformation function such as $\Phi : \mathbb{R}^P \rightarrow \mathbb{R}^Q; (P << Q)$. The explicit transformation of the feature space is unnecessary as it is sufficient to apply a kernel function $K^\Phi((x, x') = \langle \Phi(x), \Phi(x') \rangle$ on the data which is more computationally efficient.

Different types of kernel functions may be employed such as Gaussian, polynomial, radial basis kernels [191] on the basis of experimental performance. Similarly, the complexity $C$ and soft margin $\epsilon$ parameters have to be set and again there is no direct approach to obtain the optimal values other than experimentally. The parameter selection significantly affects the performance of SVR [192]. Prior knowledge of the data distribution and the
complexity of the task may be helpful in selecting the parameter values [192].

The main advantages of SVR are the good generalization that fits well in many applications since the width of the $\epsilon$-tube is modifiable to yield the most fitted region around a regression line, and additionally, non-linear mapping problems can easily be addressed by employing appropriate kernel functions. Moreover, the number of parameters that require to be tuned are limited by the number of support vectors or feature vectors present on the edges of the $\epsilon$-tube, and thus the SVR modeling is an efficient approach in terms of computation. The limitations with SVR are that the error permitted and the kernel functions selected need to be carefully chosen as SVR heavily relies on the choice of these parameters. Although the SVR model is more likely to achieve the globally optimal solution, it is not context-sensitive [189].

2.4.5 Autoregression Exogenus Back-end

The Autoregression exogenous (ARX) [193] model models the arousal (or valence) value at the current time step as a linear combination of the past arousal (or valence) values, past input speech features, and the current input speech features. Thus the ARX model may be viewed as having an eXogenous part which is a multivariate linear regression model consisting of a set of parallel FIR filters for the speech features cascaded with an all-pole filter as the AR model to work with the past arousal (or valence) values. A more mathematically detailed description will be provided later in this section.

This ARX model was validated on the RECOLA, SEWA and USC CreativeIT datasets described in section 2.6 and the performance was comparable to state-of-the-art LSTM systems. Since the theory of linear systems is very well-established, this sort of model can be interpreted. Moreover, unlike complex non-linear models such as LSTM-RNN and GMR, these models need fewer model parameters to be estimated, do not have significantly larger model training times with small increases in training data size as in SVR and are able to factor in temporal dependencies within emotion labels unlike most RVM and SVR based systems. Despite the advantages of using linear models, the models were perhaps
considered too simplistic to be able to adequately model the complex relationship between speech features and emotions due to the limited previous research on using linear models in speech emotion recognition such as multivariate linear regression, which focuses on the current emotion, linear models that take into account temporal dynamics such as ARX models have only been invested in the context of categorical emotion classification systems and only recently has it been demonstrated that a linear model is capable of capturing most of the relationship between speech features and emotion labels in the continuous arousal-valence space and consistent results were found in this thesis.

As explained before the ARX model represents the current emotion value as a function of the past sequence of emotion values and the input speech features. Let us consider the current time stamp as $t$, the current emotion label as $y(t)$ and the input feature vector at time $t$ to be $x_0(t)$. Then the current emotion label may be represented as a function of the sequence of the past emotion labels and delayed speech features as in

$$y(t) = \sum_{i=1}^{n_a} a_i y(t - i) + \sum_{j=n_d}^{n_b} b_j^T x_0(t - j)$$  \hspace{1cm} (2.6)

Where, $b_j$ is the weight vector corresponding to the $j^{th}$ feature frame with scalar weights for each dimension of the speech feature vector, $n_a$ represents the order of AutoRegressive (AR) sub-model, $n_d$ is the delay between input features and predicted output which is typically referred to as annotators’ reaction lag. In typical emotion prediction systems, the reaction lag is compensated either by manually shifting the feature frames backward by adding "lag" frames at the start of the feature vectors or by shifting the labels forward by truncating the "lag" before training the model. However, the ARX model is directly able to incorporate the reaction lag. The order of the eXogenous sub-model is denoted by the series of time frames considered of the speech features which is $n_b - n_d + 1$. The model may be implemented as a set of parallel FIR filters, with one filter for each dimension of the speech vector as in Figure 2.7.
2.4.5 Autoregression Exogenous Backend

Figure 2.7: Overview of the ARX backend. There is a set of parallel FIR filters implemented, one for filtering each feature dimension and finally cascaded with an all-pole filter.
2.4.5.1 ARX Training

The parameters of a linear prediction model may be estimated with least-squares estimation or with regularised least-squares estimation. Furthermore L1 regularised least-squares estimation may be adopted in order to encourage sparsity \[197\] in the parameters which makes the model easier to interpret \[193\]. The estimate of the parameters of the ARX model, \( \hat{\Theta} \) is given in equation \[2.7\].

\[
\hat{\Theta} = \arg \min_{\Theta} \frac{1}{N} \| y - X \hat{\Theta} \|_2^2 + \lambda \| \Theta \|_1 \quad (2.7)
\]

where, \( \lambda \) is a regularization parameter controlling the strength of shrinkage of less influential parameters, i.e. a larger \( \lambda \) encourages greater sparsity, \( N \) is the number of training frames available, and \( y \) is the emotion label vector i.e. a vector made out of all arousal (or valence) labels at different time frames and can be written as \( y = [y_1, y_2...y_N]^T \); \( \Theta \) is the vector of ARX model parameters and is given by equation \[2.8\].

\[
\Theta = [a_1, a_2...a_{n_a}, b_{n_d}^T, b_{n_d+1}^T...b_{n_b}^T] \quad (2.8)
\]

\( X \) is the matrix of the training data. The \( t^{th} \) row of \( X \) is given by \( x_t \) as in equation \[2.9\].

\[
x_t = [y_t, x_0^T(t-n_d), x_0^T(t-n_d+1)...x_0^T(t-n_b)] \quad (2.9)
\]

Where, \( y_t \) is a sub-vector of \( y \) consisting of \( n_a \) elements prior to the current time frame \( t \). Then \( y_t \) is written as equation \[2.10\].

\[
y_t = [y(t-1), y(t-2)...y(t-n_a)]. \quad (2.10)
\]

The recursive Alternating Direction Method of Multipliers (ADMM) \[198\] may be chosen as in \[193\] to estimate the best-fitting parameter arrays, owing to its computational
2.4.6 Long Short Term Memory - Recurrent Neural Network

Recurrent neural networks (RNNs) and more specifically LSTM-RNNs, have attracted a growing interest in recent years to be employed as a back-end in CEPS, due to their capability of automatically capturing long-term temporal dependencies that become relevant in the context of emotion due to its evolutionary characteristic. These are derived from recurrent neural networks (RNNs) having feedback loops that have the advantage of capturing sequential information wherein the output depends upon the previous computations. This is achieved by the ‘memory cells’ within the RNN which store information calculated until the current time step. A typical RNN is shown in Figure 2.8. The formulation of the problem solved with RNNs is explained in the following paragraph.

Let the number of speech features at time \( t \) be \( N \), and the speech features at time \( t \) be denoted by \( X_t = [X_{1t}, X_{2t}, \ldots, X_{Nt}]^T \) and the corresponding prediction be denoted by \( y_t \). If \( W_{in} \) and \( W_{out} \) denote the weight matrices connecting two conservative layers and \( W_r \) the weight matrix that connects the previous time step to the current time step, the \( W_r \) is able to memorize the past information. This is illustrated in the structure in Figure 2.8. However, it was found by several studies such as [199] [200] that RNNs can only remember short-term temporal dependencies.

Therefore as the gap between the relevant past information and the current frame grows, RNNs become unable to learn to connect the information and thus another special type of RNN was proposed called the LSTM-RNN that is able to capture long-term information and dependencies. LSTM-RNNs [201] [105] [202] [203] are capable of modelling long-term
Figure 2.8: Concept of a recurrent neural network is illustrated with its and the unfolded structure in time during forward computation. $X_t$ represents an $N$ dimensional input feature vector at time $t$, $H_{mt}$ hidden unit at time $t$ for $m = 1, 2, ..., M$ and $Y_t$ 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.
dependencies due to the presence of gates in the network that are able to control how much of information to let through based on relevance. Moreover, while LSTMs and BLSTM-RNNs are good at capturing the correlation between past and future contexts [189], the BLSTM-RNN model may easily lead to a solution at any local minima and has a risk of over-fitting [204].

Some potential shortcomings of non-linear back-ends are summarized below:

- they are complex nonlinear models that involve learning a large number of model parameters, such as LSTM-RNN and GMR.
- suffers from a significant increase in computation time with a relatively small increase in training data size such as $\epsilon - SVR$ or in [205].
- they ignore the temporal dependencies within emotion labels such as with most RVM and SVR-based systems [33,62,183].

Additionally, the lack of interpretability of these non-linear models also limits the development of emotion prediction systems.

2.5 Evaluation Metrics

As explained earlier in section 2.2.1 the vast majority of CEPS research considers the mean rating as the only quantity of interest thereby ignoring individualistic differences between emotion ratings provided by multiple raters. The reliability of the emotion ratings is assessed from the (ICC(3,1)) [44] and Cronbach’s $\alpha$ [45] over all the ratings or the root-mean-square error (RMSE), Pearson’s correlation coefficient $\rho$ and the Concordance Correlation Coefficient (CCC) [206] calculated pair-wise and then averaged over all pairs [20]. In [20] the inter-rate reliability of the ratings in [43] was found, to be high. Furthermore, prior to the ground truth calculations a normalisation technique based on the Evaluator Weighted Estimator [207], was employed [201]. This technique significantly
improved the inter-rater reliability for both arousal and valence however with losses in individualistic variations of emotion ratings. Since most existing CEPS focus on predicting the ground truth emotion labels or focus on the trend of ground truth labels, the most widely adopted evaluation metrics are suitable to evaluate single-overalledictions or time series of point estimates.

2.5.1 Point-Prediction Metrics

Let the emotion labels over $N$ time frames be represented by vector $y$ and the estimated label be represented as $\hat{y}$.

2.5.1.1 Mean Square Error

Mean squared error (MSE) measures \[208\] the average of the squares of the error terms, i.e. the difference between the ground truth and the predictions as shown in (2.11). The smaller the value of MSE is, the better the prediction performance is.

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \quad (2.11)$$

2.5.1.2 Pearson’s Correlation Coefficient

Pearson’s correlation coefficient $\rho_{cc}$ \[209\] indicates the strength and direction of the linear relationship between the emotion ground truth labels and the emotion prediction labels. It is calculated as in equation 2.12

$$\rho_{cc} = \frac{\text{cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \quad (2.12)$$

where $\text{cov}()$ is the covariance function between the two variable vectors, and $\sigma_y$ and $\sigma_{\hat{y}}$ are the standard deviations of $y$ and $\hat{y}$. A value of $\rho_{cc}$ close to one indicates a strong correlation between the ground truth labels and their respective predictions indicating a
better performance of the emotion prediction system in terms of being able to capture
the trend of the ground truth variable. The limitations of the MSE and \( \rho \) measures are
illustrated in the plots of Figure 2.9. The plot on top displays an instance when the
predictions follow the trend of the mean labels and therefore have a very high \( \rho \) but the
MSE is high. Whereas the plot below shows the instance where the \( \rho \) had decreased but
the MSE has improved. It is hard to compare the predictions displayed in the top and
bottom plots since one measure shows an improvement while the other deteriorates. It is
thus necessary to use a measure that combines the two measures.

2.5.1.3 Concordance Correlation Coefficient

The \( \rho_{cc} \) and the MSE have been the two most commonly used metrics to evaluate the
performance of CEPS. However, \( \rho_{cc} \) and MSE measures used alone may only be able to
partially indicate the emotion prediction system performance. Some of the limitations of
these measures are illustrated in Figure and Figure. For example, a high \( \rho_{cc} \)

\[
CCC = \frac{2\rho_{cc}\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}
\]  

(2.13)

In equation 2.13, \( \mu_y \) and \( \mu_{\hat{y}} \) denote the mean of the ground truth labels and the predic-
tions respectively. Since CCC is able to incorporate the characteristics of MSE and \( \rho_{cc} \), it
has become the most commonly adopted measure to evaluate overall CEPS’ performances
since AVEC 2016.

Although \( \rho_{cc} \), MSE and CCC have been adopted to evaluate CEPS for years, there is still
not adequate support on the optimal evaluation metric. More importantly, as discussed in
section 2.2.2 I have investigated emotion prediction as a prediction of distributions that
model inter-rater ambiguity instead of assuming there to be single-valued true emotion
represented as a point estimate for every time stamp based on the mean rating as the
ground truth. For such scenarios, there need to be evaluation metrics for distribution
predictions that should be able to compare sequences of distributions over time, wherein
Figure 2.9: The limitations of the MSE and $\rho$ measures are illustrated in the above two plots respectively. The plot on top displays the instance when the predictions follow the trend of the mean labels but the MSE is high. Whereas the plot below shows the instance where the $\rho$ had decreased but the MSE has improved.

the three aforementioned evaluation metrics may not be appropriate since they are only applicable for the time sequence of point estimations.
2.5.2 Distribution Comparison Metrics

As explained before in section 2.2, emotion ambiguity may effectively be modeled by distributions that call for evaluation metrics to compare the predicted ambiguity distributions with the ground truth distributions. There exist many methods to measure the difference between distributions but the most widely employed distance measure is the KL Divergence.

2.5.2.1 Kullback-Leibler Divergence

Given two distributions $P$ and $Q$, the Kullback-Leibler Divergence ($D_{KL}(P||Q)$) is a non-negative score that quantifies by how much distribution $P$ differs from $Q$. For distributions $P$ and $Q$ with $p$ and $q$ denoting the probability distribution functions respectively of a continuous random variable $x$, the $D_{KL}(P||Q)$ is described in (2.14). It is worth noting the scenario where there are data points available from an unknown distribution and we may assume the distribution that these data points are drawn from. In this case, minimizing the KL divergence is equivalent to maximizing the log-likelihood of each data point under the target distribution.

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \left( \frac{q(x)}{p(x)} \right) dx \quad (2.14)$$

Although metrics such as the KL divergence are able to measure the difference between distributions in terms of distance, the measure itself is little insightful of why the distributions are different namely what aspects of the distribution have changed. Figure 2.10 illustrates the scenario where the KL divergence measure is unable to provide information on what aspects of the distributions have changed. The KL divergence measure of the yellow and red distributions are centered equidistantly at opposite sides of the blue distribution and therefore have the same KL divergence from the blue distribution. The purple distribution with a larger variance than the blue distribution also has the same KL divergence from it as the blue-yellow and the blue-red pair. There are infinite normal
distributions possible with the same KL divergence measure from the blue distribution centered at different points and with different variances but they are not able to reflect information on the mean and variance of the distribution. This makes KL divergence a less suitable metric to be measured for emotion prediction.

2.6 Emotion Databases

The earliest emotion databases [210] contain recordings of acted emotion wherein several actors were given scripts to read off and enact the emotion states. The annotations of such early databases were primarily focused on category-based emotion representation. [210] is a Dutch Emotional database and the emotion categories are anger, contempt, disgust, fear, interest, joy, sadness, shame, surprise and neutrality. There were four speaker- gen-

Figure 2.10: Four normal distributions are displayed. The value of the KL divergence between each of the red, yellow and purple distributions and the blue distribution are equal.
2.6. EMOTION DATABASES

The earliest database with the dimensional emotion representation was the Modified LDC CallFriend corpus prepared by [218] where discrete numerical values from 1 to 5 were assigned each for arousal, valence and engagement from the telephone conversations recorded between friends. About 1,011 utterances from 4 female and 877 utterances from 4 male speakers were collected. Some of the most popular databases before 2015 were the Vera am Mittag, German Audio-Visual Spontaneous Speech Database (VAM) database and the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [219]. The VAM
database \cite{220} contains 12 hours of recordings of a German TV talk-show. Discrete valued annotations ranging from -1 to +1, were collected for three emotion dimensions: arousal, valence and dominance. IEMOCAP \cite{219} is an audio-visual database collected in dyadic interactions between professional actors. About 12 hours of recordings were collected from 10 subjects in 5 dyadic sessions, and it was one of the first databases with dyadic interaction. The annotations for each emotion dimension would be a natural number until 5 and the emotion dimensions considered were: arousal, valence and dominance.

The most commonly employed databases in CEPS since 2015 are the USC CreativeIT database \cite{221}, the RECOLA database \cite{43}, the SEMAINE database \cite{222} and the SEWA database \cite{42}. The continuous annotations of the emotion dimensions for utterances are collected from several annotator who are shown the video of the speech or made to listen to the audio recordings of the utterance and then move a cursor within a 1D or 2D emotional space to rate the momentary emotion intensity perceived from the speech over the duration of the utterance. The ground truth of the speech sample is obtained as the weighted or unweighted average among several evaluators \cite{223} \cite{43} \cite{224}.

The USC CreativeIT multimodal database \cite{221} provides a novel bridge between the study of theatrical improvisation and human expressive behaviour in dyadic interaction \cite{221}. The theatrical improvisation technique of Active Analysis is able to induce natural affective states and spontaneous goal-driven dyadic interactions. The USC CreativeIT database consists of two different theatrical techniques; the two-sentence exercise and the paraphrase exercise. In the two-sentence exercise, each actor is given a verb to describe their emotion and actions and then say one predefined sentence with the emotion. In the paraphrase exercise actors are provided with a script and instructed to act it out based on their own words and interpretation. The database contains multimodal behaviour signals including speech and video, with 8 sessions of 90 sentences recorded by 16 speakers in English. The affective dimensions of arousal, valence and dominance are annotated for each session both frame-level and utterance-level. All continuous annotations at frame-level are performed by watching the session videos and using Feeltrace software \cite{225} that enables the user to continuously move a mouse around a computer screen and rate the emotion intensity.
on a nearly continuous scale ranging from -1 to 1 over the duration of the session or recordings. The discrete annotations were labels provided at the utterance-level, where the labels ranged from 1 to 5. Each recording have evaluation only by 2 to 4 evaluators. It was determined that the annotations were highly correlated in pairs regardless of the type of model. The final continuous attribute values are obtained by averaging all individual annotations. Although this database is popular it has the limitation of having speech between actors and not realistic in the wild scenarios besides having only 2-4 annotations per recording.

The SEMAINE database \[222\] is collected based on the Sensitive Artificial Listener (SAL) paradigm. In the SAL paradigm, the artificial listener is the role-playing operator interacting with a human subject. This sort of spontaneous interaction is thought to elicit more natural emotional responses. The database consists of three basic scenarios wherein the artificial listener’s interaction is defined: i) solid SAL, where human operators play the roles of the SAL characters; ii) semi-automatic SAL, where the system speaks to the 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 to the human. The subject or participant interacts with four operators each having a different ‘personality’. From the 141 conversation sessions recorded in total, speech from only 20 speakers recorded over 26 sessions using the solid SAL scenario could have been used to evaluate models for spontaneous and more natural emotion from dyadic speech. Five affective dimensions that are arousal, valence, dominance, expectation and intensity and were rated separately each on one-dimensional emotion support ranging from 0 to 10. Final annotations were the average of the individual ratings. 2-8 raters were available for arousal, valence and dominance and the annotations were obtained using the FEELTRACE tool \[225\].

The RECOLA and SEWA databases are described in more detail in Section \[2.6.1\] and \[2.6.1.1\] respectively; and the limitations of current databases is presented in Section \[2.6.2\]. The RECOLA and SEWA databases are used for the continuous emotion prediction tasks
in this thesis. Both databases contain spontaneous conversations where emotions are elicited implicitly by the data collection design (details mentioned below in Table 2.1) and are also not constrained by linguistic content, such that both acoustic and linguistic features can be analysed as emotional cues. Moreover, the RECOLA database was used for the AVEC 2015, AVEC 2016 20 and AVEC 2018 220 challenges and SEWA database were used for AVEC 2017 142, AVEC 2018 226 and AVEC 2019 21 challenges. Therefore using these datasets provide the opportunity to compare the evaluations on the proposed systems against state-of-the-art systems.

2.6.1 The RECOLA database

The RECOLA database consists of multimodal dyadic 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 electrocardiogram (ECG) and electro-dermal activity (EDA) data. The two participants were placed in separate rooms and engaged in a discussion remotely over a simple task paradigm. The annotation tool developed for and employed to annotate the RECOLA database is ANNEMO which is similar to WISE 227. This tool offers one-dimensional continuous affective annotation instead of joint two-dimensional annotations to reduce the cognitive load on the annotators of having to rate along two emotion dimensions at the same time. The two dimensions for which annotations were provided were arousal and valence. The dimensions were annotated separately every 40 milliseconds with the emotion support ranging from 1 to +1 and quantised to a step size of 0.01. For the gold standard, the mean rating was assumed to be the true rating where post-processing was performed to reduce the "unwanted" variability of each annotation, and finally, the ground truth of an utterance was estimated by taking the mean of the annotations provided by all six gender-balanced annotators based on the evaluator weighted estimator (EWE) approach 228. The RECOLA database is used in AVEC 2015 and AVEC 2016 challenges. Speech data from 27 speakers were equally divided into training, development and test partitions. The affect labels of the designated test partition were not released for the challenge purpose.
2.6.1 The RECOLA database

This partition of RECOLA database is most commonly used in the existing literature. Further information about RECOLA can be found in [43].

2.6.1.1 The SEWA database

The SEWA database is an audio-visual cross-cultural emotion database which has been collected ‘in-the-wild’ i.e. based on interactions recorded on webcams and microphones from computers in the participants’ homes or offices. The interactions were in German, Hungarian and Chinese. The participants were divided into pairs based on if they were friends or relatives, shown an advertisement and then assigned the task of discussing the commercial they had viewed at the same time but remotely without any scripts or constraints in regards to emotion provisions. Therefore the dyadic interactions were spontaneous and quite natural behaviour signals. These dynamic conversations were limited to 3 minutes. The database is annotated with respect to arousal, valence and likability in continuous time i.e. there is a rating available for every 0.01 second. Each speech utterance (clip) is annotated by 6 annotators in German, 5 annotators in Hungarian, and 6 annotators in Chinese. Post-processing is applied to each individual annotation to normalize them to the same range and to remove any bias. Then the ground truth is calculated as a weighted average over all the normalised annotations based on the evaluator weighted estimator (EWE) approach [228].

The SEWA database was released in 2017 and since then it was used for the AVEC 2017, 2018 and 2019 challenges. Initially, only the German data was released for the AVEC 2017 challenge, the Hungarian data was released and Chinese data was released for the AVEC 2019 challenge. In total, the utterance of the participants aged from 18 to 60 was divided into three partitions: training, development and test, which contain 34, 14 and 16 utterances respectively in German and 34, 14 and 18 utterances in Hungarian respectively. There are 70 Chinese utterances available in the test set. It should be noted that the labels of the test partition were not provided since AVEC 2019 challenge organizers held
them to validate the system performance of participants’ submissions. An improvement of the SEWA database over other databases is that transcripts have been provided which have been manually transcribed from the video chats. Timestamps have been provided to indicate which subject is speaking. This has motivated more investigations on the linguistic analysis and experimental results have shown superior performance [19] [33]. A summary of the details of the data portions of RECOLA and SEWA databases provided with AVEC 2016 and AVEC 2019 challenges are given in Table 2.1.

2.6.2 Limitations of current databases

Almost all current and even the most state-of-the-art and popular databases may have some limitations. It is of interest to be aware of some of the possible limitations as they may affect the performance of emotion prediction systems. A few of these limitations commonly occurring in databases are summarised as follows:

- Even in databases where some context is provided to subjects to have to watch film clips, or imagine specific emotion-prone situations to influence their mood one might induce more realistic emotion, but the "extent of naturalness" cannot be determined.
- Prompted speech is mostly monologic, not interactive and even then interactive emotions are elicited based on some specific activities.
- Since emotion for databases are not simulated naturally, techniques such as the sensitive artificial listener (SAL) paradigm and the theatrical improvisation for active analysis are used to elicit emotion as naturally as possible. However, the techniques have not been compared to determine the most appropriate technique for eliciting the most natural expressions of emotion.
- Transcripts were not provided in most early databases [229], making the linguistic content analysis difficult [134].
- Many databases are recorded in a controlled environment, where the recording quality may not exactly match a realistic scenario.
### 2.6.2 Limitations of current databases

<table>
<thead>
<tr>
<th>Frame</th>
<th>RECOLA</th>
<th>SEWA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recorded Language</td>
<td>French</td>
<td>German</td>
</tr>
<tr>
<td>Speakers’ Native Language</td>
<td>33 French, 8 Italian, 4 German and 1 Portuguese speakers.</td>
<td>408 speakers: British, German, Hungarian, Greek, Serbian, and Chinese.</td>
</tr>
<tr>
<td>Recording Condition and Experimental Setting</td>
<td>Dyadic interactions between two participants in separate rooms who are remotely connected via audio-video devices to complete a survival task according to instructions given.</td>
<td>Dyadic interactions recorded through webcams and microphones of a computer-mediated face-to-face dyadic interaction session wherein the subjects were assigned the task of watching the same selected three advertisements and discussing the last advertisement watched with their partner through the webcam.</td>
</tr>
<tr>
<td>Emotion elicitation</td>
<td>Emotion changes are expected to occur in the process of completing the survival task (mentioned above) by two speakers as a team (could be positive, negative, or neutral). (Pre-selected video clips were used to induce mood before starting the sessions, selected based on self-reported prior moods, and to have a balanced number of participants with positive and negative moods for a debatable level of success.)</td>
<td>The advertisements were chosen to elicit mental states including amusement, empathy, liking and boredom. After a person watched the advertisement, they then discuss it with another participant through the video-chat device. The discussion is intended to elicit further reactions and opinions about the advertisement and the advertised product, such as whether the product is to be purchased, whether it is to be recommended to others, what the best parts of the ad are, whether it is appropriate, how it can be enhanced, etc.</td>
</tr>
<tr>
<td>Recording modalities</td>
<td>Multimodal (audio, video, and physiological data)</td>
<td>Audio, video through webcam</td>
</tr>
<tr>
<td>Utterance length</td>
<td>The first five minutes of the task is used by each speaker.</td>
<td>Limited to three minutes per speaker.</td>
</tr>
<tr>
<td>Data Released in Challenges</td>
<td>AVEC 2016 challenge released Training, Development, and Testing partitions with 9 utterances of non-overlapping speakers in each.</td>
<td>AVEC 2019 challenge released Training, Development, and Testing partitions with 34, 14 and 16 non-overlapping speakers in German respectively</td>
</tr>
</tbody>
</table>

Table 2.1: Key differences between the RECOLA and SEWA databases.
### 2.7 Summary

This chapter has briefly explained the overview of speech-based CEPS and introduced the current developments in feature extraction, emotion databases, regression modelling.
techniques and evaluation metrics commonly employed in this field. Continuous emotion prediction from speech has received growing interest in the last few decades and has mainly be focused on predicting emotion dimensions as point estimates, improving the accuracy of predictions of such systems by developing representative acoustic and linguistic features, or have focused on developing more advanced regression modelling techniques for prediction. A large fraction of speech emotion literature since 2016 have adopted the eGeMAPS acoustic feature set [11] as a standard minimalistic set of voice parameters to enable comparison of results from proposed systems across studies, however raw speech signal representations discriminatively learnt with neural networks were found to have better performances [6]. Increasing research has focused on linguistic features owing to the increased availability of speech database transcripts [231]. Moreover, nonverbal vocal gestures such as ‘laughter’, ’pause’ or ‘filler’ provide complementary information in predicting emotion [144].

Amongst regression modelling techniques for emotion prediction, recurrent neural networks are most widely adopted owing to their ability to model the short-term and long-term temporal dependencies in expressed emotion. Regression techniques such as GMR, RVM, and SVR are also commonly employed in emotion prediction systems.

Many challenges exist in CEPS, some of which are listed as follows:

- Discriminatively learning speech representations have been shown to outperform standard acoustic feature sets on emotion databases. However, discriminatively learning feature representations do not give any insights into the influence of emotion on the speech production aspects unlike selections of paralinguistic features.

- The LSTM back-end is the most successful and widely adopted of the deep learning structures in CEPS indicating the long-term and short-term dependencies in the nature of emotional expression. However, there is still limited knowledge of the emotional aspects captured in literature by the LSTM and the underlying mechanism with which the LSTM captures emotion-related information.
In psychology research, cross-cultural differences in emotion expression have consistently been found especially between individualist cultures versus collectivist cultures in arousal \cite{232} based on their goals to influence other people or adjust and conform to the community respectively. Moreover, certain eastern cultures conceptualized happiness as experiencing low arousal positive emotions rather than high arousal positive emotions in western cultures \cite{233}.

Most continuous emotion prediction literature to date is focused on predicting point estimates of emotion. However, modelling perceived emotion as distribution would be more useful in real-world scenarios as prediction certainty has been shown to rely on the difficulty level in ascertaining the expressed emotion.

There is limited research on modelling emotion as distributions over the human population rather than predicting a point estimate as the perceived emotion.

A key limitation of most current systems is the assumption of the ground truth emotion labels to be the average of all available annotations, which motivates our continued research on the analyses of inter-rater ambiguity in this thesis. One challenge I focus on in this thesis is modelling the inter-rater ambiguity as it has consistently shown in part literature to be correlated to the prediction uncertainty of the system. The current limitations of existing ambiguity models are explored and measures to compare them have been designed. Moreover, discriminatively learnt speech representations outperform standard acoustic feature sets on emotion databases but, are not backed up by the theoretical mechanisms that produce speech. This calls for further investigation where in this thesis, an alternate representation of features was proposed that although based on a set of para-linguistic feature set selection, were discriminatively trained based on emotion label information. The in-depth analyses and proposed techniques will be further explained in detail in the following chapters.
Chapter 3

Bag-of-Optimized Codebooks -
Emotion based Clustering of
Features

3.1 Introduction

Most traditional speech-based emotion prediction systems typically consist of a front-end
to extract emotionally relevant aspects of speech as explained in section 2.3.2 as speech
features and a back-end that models the relationship between the extracted speech features
and the target emotion labels [234].

Of the two broad types of features extracted, one is acoustic features that are backed by
theoretical understanding of speech production [66, 67, 68, 69, 70, 71, 72, 11] and the other may be discriminatively learnt features from CNNs and [6, 73, 74] vari-
tional autoencoders [75] which although not not backed by speech processing theory, have
achieved the best performances with respect to the widely accepted evaluation metrics
described in section 2.5.1.3. These features may capture many affective characteristics,
but the complexity of the network makes it impossible to discern the role of individual
acoustic features in emotion and interpret the underlying mechanisms by which they may represent aspects of speech and relate them with emotion. Therefore, it was of interest to find an acoustic feature based representation that could be discriminatively modified based on label information to see if this enhanced the learning of the emotion labels on unseen speech based on the new feature representation.

Of the most widely extracted acoustic feature representations are the LLDs described in section 2.3.2.1 from which high-level statistical functionals are usually derived because emotion is a slowly evolving quantity (illustrated in Figure 2.4). The rationality behind using statistical functionals to describe the distribution of the LLDs is that the functionals are able to capture the distribution of the LLDs sufficiently and are a succinct representation of speech features in each emotion window. For example, Figure 3.1 illustrates how LLDs are extracted per 10 ms over speech windows of 3s meaning each emotion window of 3s has 300 frames and each frame has a vector of LLDs.

Constraining the suprasegmental features to simple functionals such as mean and standard
deviations implicitly assumes that, the distribution of the LLDs is fairly simple, such as Gaussian distributions. However, from investigation of the speech features extracted from the RECOLA dataset [43], it was found that the distributions of various LLDs in a time segment may not be best represented by a simple Gaussian distribution. To illustrate some of the distribution shapes, histograms are plotted over 5 different emotion windows of the mel frequency cepstral coefficients (MFCC), $C_0$ in Figure 3.2. $C_0$ is chosen for illustration because there is a general consensus amongst emotion prediction research that of the paralinguistic feature set the MFCCs and the pitch strongly reflect the magnitude of the perceived emotion dimension [235]. It is clear that this distribution is neither unimodal nor symmetric. It may be that for the emotion task at hand, the statistical functionals are unable to comprehensibly capture the statistical information of the LLDs.

Due to the above finding the BoAW representation of LLD distributions was explored. This is based on partitioning the LLD feature space based on clustering algorithms and representing the cumulative cluster memberships over a set of frames [236] which roughly captures the joint distribution over the features (essentially the zeroth order statistics) and is further explained in section 3.2. The bag-of-audio-words (BoAW) [131] feature representation which takes this approach has been successfully employed as a front-end for emotion recognition systems [115] [237].

Given that there are advantages in terms of interpretability when using selected acoustic features versus most competitive performances when features are discriminatively learnt, it was of interest to examine if the network would be able to learn a superior set of features that are derived from traditionally selected feature representation but incorporate label information as well. To achieve this goal, this work proposes a novel front-end that is based on discriminatively learning feature clusters to generate a codebook optimised for emotion prediction, which is then used to generate a novel feature set called the Bag-of-Optimised-Clusters (BoOC).

Just as a range of front-ends have been investigated for use in emotion recognition systems, several back-ends have also been investigated. Among these, the most popular back-ends are support vector regression (SVR), gaussian mixture regression and more recently
Figure 3.2: Left: $C_0$ magnitude for 10 ms frames within a randomly selected 3s segment. Right: Histogram of above plotted $C_0$ values. A suprasegmental feature representing the $C_0$ values plotted in the above segment, i.e. the mean value is indicated in black.
LSTM recurrent neural networks (LSTM-RNN). Of these, LSTM-RNNs explicitly capture temporal dependencies between the acoustic observations that evolve over time. However, LSTM can easily overfit when data size is small \[238\] and care must be taken when training them.

Finally, in most existing continuous emotion prediction systems the front-end and the back-end are generally separate entities and are typically not jointly designed. The novel framework proposed in this thesis entails the front-end capturing statistical information about LLDs more effectively with the label information and an LSTM-RNN back-end both of which are jointly trained. The Bag-of-Optimised-Clusters front-end is tested on the RECOLA database and results show that it outperforms the well-established BoAW features.

### 3.2 Capturing the joint-LLD Distribution

As discussed in chapter 2 section 2.3.4.3 in the BoAW method, the feature space is partitioned into a chosen number \((C_s)\) of \(C_s\)-dimensional spaces with a unique codeword representing a space. Each LLD frame vector is located in one of the spaces and is represented by the codeword representing the space which is akin to the quantization of the feature vector. A bag is a histogram created based on the number of times each codeword is input within a given emotion window. So in a sense the bag is representing the joint probability distribution function of the LLDs in each emotion frame. Therefore, the absolute location of codewords in the feature space and their relative positions influence the pdf of the front-end speech features contained within each emotion window.

### 3.3 Proposed Bag-of-Optimised-Clusters Framework

An overview of the proposed system \[4\] is depicted in Figure 3.3 which although can be functionally partitioned as a front-end and a back-end, is structurally a single deep
learning system. The initial four layers form the front-end which output the finally learned speech features and the final two layers which are the LSTM layers, constitute the back-end which predicts emotion. The input to the system is the sequence of LLDs from consecutive frames of speech and the output produced are the arousal or valence predictions for every given emotion window. The front-end layers are designed to estimate BoAW-type feature representations where codewords are network parameters. The output of the front-end is the BoOC features which are then passed on to the back-end layers. LSTM is adopted as the backend as it was proven to be effective in emotion prediction systems [16] owing to its capability to model the temporal dynamics of emotion. Furthermore, the neural network-based back-end also offers the advantages of jointly training of the front-ends and back-ends which as discussed in the previous section has given the best performances. Though the BoW representation may lose temporal information in the feature space to some extent due to the histograms created not being temporally arranged, the window size chosen as 4 seconds with a frame shift of only 40 milli-seconds can still guarantee a highly correlated feature representation which preserves the similarity between adjacent frames, therefore, the temporal properties.

This proposed framework brings a number of key potential advantages over conventional emotion prediction systems. Namely, by making the codewords as network parameters of the training phase and by utilizing a unified network comprising both the front-end and the back-end, the codebook and the LSTM layers are both jointly estimated to complement each other. Furthermore, codewords optimized for the task of emotion prediction can be learned directly from the training data via backpropagation. The proposed framework here is referred to as the bag-of-optimized clusters (BoOC) framework.

### 3.3.1 Framework Description

In Figure [3.3](#), the first layer \( a \) depicts the input layer which consists of the elements of the LLD vectors, where each LLD vector is \( N \)-dimensional. The input feature vectors are fed into the network one frame at a time. The second layer \( b \) consists of radial basis neurons. Each neuron \( b_j \) is an entry in the codebook and is called a codeword and the total number
3.3.1 Framework Description

Figure 3.3: A schematic diagram of the NN showing the first phase, the BoOC creation by accumulating the histogram of the LLDs of all frames, and the prediction phase using the BoOC features [4].
of codewords in the codebook shown are \( M \). These radial basis neurons compute the Euclidean distances between the input vector and their corresponding codeword in layer \( c \). These distances are then translated into probabilities via a suitable \textit{softmin} layer \cite{236}. The probabilities over a set of consecutive input feature vectors are then accumulated by the next layer to form the BoOC representation in layer \( d \), which is fed to subsequent LSTM back-end layers. A detailed description of each layer follows:

Let \( N \) denote the dimensionality of the input vectors, \( N_W \) denote the number of consecutive frames/input vectors over which we will accumulate probabilities to form the BoOC vector, then the distances measured by the radial basis neurons are given by

\[
b_j = \| a - w_j \|^o
\] (3.1)

where, \( a = [a_1, a_2...a_N]^T \) denotes input vector and \( w_j \in \mathbb{R}^{N_W} \) denotes the \( j^{th} \) codeword, with \( 1 \leq j \leq C_s \) where \( C_s \) denotes the size of the codebook (which in Figure 3.3 is \( M \)) and \( o \) denotes the order of the euclidean distance.

Following the estimation of the distances to the codewords, the set of \( C_s \) distance, \( \{b_j; j = 1,...,C_s\} \) is translated to a \( C_s \) dimensional vector, \( c = [c_1,c_2...c_{C_s}]^T \), of probabilities via \textit{softmin} functions \cite{236} as follows:

\[
c_j = \frac{e^{x(p(-b_j/T_{\text{decay}}))}}{\sum_{k=1}^{C_s} e^{x(p(-b_k/T_{\text{decay}}))}}
\] (3.2)

where, \( T_{\text{decay}} \) is a tuneable parameter that controls radius around each codeword beyond which probability of membership to the ‘cluster’ corresponding to that codeword diminishes rapidly. Appropriate choice of \( T_{\text{decay}} \) avoids the denominator in 3.2 becoming zero thereby causing numerical problems. From 3.2 it can be inferred that a decrease in distance to the \( j^{th} \) codeword translates to a higher probability at the \( j^{th} \) node of the third layer. This idea is depicted in Figure 3.4 using a 2 dimensional input vector space and 3 codewords. A detailed explanation of the rationale behind using the Softmin layer and
Figure 3.4: Illustration of BoOC assignment. Top: Plot illustrates a hypothetical 2-dimensional speech feature space where the locations of the input feature vector is shown by the red mark and the cluster centers in green. Bottom: Plot illustrates the audio word assignment of the feature vector to the codewords for the scenario demonstrated above. [4]

euclidean distances can be found in the following section 3.5.2.

The following layer accumulates probability vectors $c$ of $N$ consecutive frames to form the final BoOC vector, $d = [d_1, d_2...d_{C_s}]^T$. Here $N$ is chosen to correspond to the time segment of around a few seconds to effectively capture the slowly varying dynamics of emotion [223]. Specifically,

$$d_i = \log \left( \sum_{n=1}^{N} c_i^n + \epsilon \right)$$

(3.3)
where, we use the overloaded notation $c^n_i$ to denote the $i^{th}$ element of the probability vector, $c$ estimated from the $n^{th}$ frame. Adding a positive quantity $\epsilon$, usually 1, to the accumulated probabilities prior to taking the logarithm ensures that the problem with the logarithm of 0 is never encountered in the network. This logarithmic compression of the histogram follows [17].

Finally, two LSTM layers serve as the back-end, modeling the relationship between the BoOC feature vectors produced by the previous layers and the desired emotion labels.

### 3.3.2 Error Backpropagation

During the training phase of the model, estimation of the codebook and back-end weights occur simultaneously. The mean squared error between the gold standard and the prediction labels is chosen as the loss function and the error is backpropagated to adjust the weights of the back-end as well as the codeword vectors in the same passes, parameterizing each radial basis neuron. Specifically, the three layers for which the weights are iteratively updated are the codewords in the front-end layer $b$ and the weights of the two LSTM layers $e$ and $f$ in the back-end.

### 3.4 Experiment Settings

#### 3.4.1 Database

The proposed model was tested on the database used in the AVEC 2016 challenge [5] from the RECOLA corpus [43] to enable comparisons with other studies [5, 131, 234, 239]. The 9 audio recordings from in the training partition [5] were used as the training set and 9 development partition recordings [5] were used as the test set. Arousal and valence were annotated every 40 ms.
3.4.2 Annotation Delay Compensation

To align the lagging gold standard with the features, the gold standard scores, in the beginning, were dropped and the last score was replicated at the end to compensate for the length. The predicted labels were thus shifted forward in time containing high-frequency noise. To smooth and realign the predictions with the gold standard, a binomial filter as described in [62] was employed.

3.4.3 Performance Metric

CCC as described in section 2.5.1.3 was adopted to measure the performance of the proposed BoOC framework as the framework predicts point estimates for emotion per time frame and as the measure is currently the most widely used [234] [240] metric in emotion prediction research.

3.4.4 Continuous Emotion Recognition System Configuration

While any set of LLDs can be employed as the input vector to the proposed framework, MFCCs were chosen in this study to evaluate the viability of this framework since they are described by previous studies to carry information about the emotional aspects of speech and have since been used for BoAW representations [131] for emotion recognition and prediction research. The openSMILE toolkit was used to extract the MFCC features to ensure that the features are reproducible [91]. The first 20 MFCC and their 20 delta coefficients were extracted per frame to serve as the input feature vectors for the proposed framework with a spacing of 10ms between consecutive frames. In several latest studies based on [17] [16], BoAW features were calculated over 4s segments and these features were used with LSTM back-end. For this reason, a segment of length 4s was chosen to obtain BoOC features for arousal-based models in this study. In the preliminary investigation of this study, BoOC extracted over 6s segments performed better for valence as opposed to 4s for all choices of $C_s$. Hence, all reported valence results in this chapter correspond to
features being extracted over 6s segments. The framework was developed in Python with Keras. The size of the codebook \( (C_s) \) varied from 20 to 500. \( T_{\text{decay}} \) was varied from 0.5 to 10. The order of the Euclidean distance \( o \) used in 3.2 was assigned values 0.5, 1, 2, and 4 wherein, 1 resulted in the highest CCC and was thus used for further experiments. For different \( T_{\text{decay}} \) keeping other parameters fixed, there was no significant difference in the highest CCC within the range of [0.690.71], hence, \( T_{\text{decay}} \) was chosen to be 2. The ADAM optimizer was used to minimize the loss and a learning rate of 0.01 was adopted. Each of the 2 LSTM layers comprises 64 neurons and this back-end is used to obtain test results for both BoAW and BoOC. The codebooks were initialized with random feature vectors from the training set. Additionally, initialization of the codebook using cluster centers of the feature vectors from the training set obtained via kmeans++ was also explored. Finally, the optimal delay compensation for annotators’ reaction lags was taken to be 2.8s for arousal and 3.6s for valence [5] while obtaining the optimized cluster center with the described LSTM back-end.

3.5 Experimental Results and Discussion

Table 3.1: CCC reported for arousal and valence predictions on RECOLA development partition [5]

<table>
<thead>
<tr>
<th>Model</th>
<th>( C_s )</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoAW+SVR [131]</td>
<td>100</td>
<td>0.673</td>
<td>0.471</td>
</tr>
<tr>
<td>BoOC+SVR</td>
<td>100</td>
<td>0.743</td>
<td>0.406</td>
</tr>
<tr>
<td>BoAW [131]+LSTM</td>
<td>100</td>
<td>0.685</td>
<td>0.152</td>
</tr>
<tr>
<td>BoOC+LSTM</td>
<td>100</td>
<td>0.757</td>
<td>0.417</td>
</tr>
<tr>
<td>BLSTM-RNN [239]</td>
<td>–</td>
<td>0.800</td>
<td>0.398</td>
</tr>
<tr>
<td>BoOC+LSTM</td>
<td>300</td>
<td>0.758</td>
<td>0.477</td>
</tr>
</tbody>
</table>
3.5.1 Results

The CCC values for arousal and valence predictions obtained using the novel BoOC frontend and novel framework of BoOC+LSTM are shown in Table 3.1. \( C_s \) denotes the number of codewords employed or codebook size. Since in several latest studies based on [17] [16], BoAW features were calculated with a \( C_s \) of 100, the results reported in Table 3.1 are based on 100 as chosen \( C_s \). A CCC of 0.743 for arousal and 0.406 for valence with the proposed BoOC frontend and SVR backend was obtained. The performance of the proposed framework, i.e. BoOC+LSTM is a CCC of 0.757 for arousal and 0.417 for valence. A relatively higher CCC for arousal and valence has been obtained with the BoAW+SVR system in [131] as shown in Table 3.2.

Table 3.2: CCC reported for arousal and valence predictions on RECOLA validation partition in [6].

<table>
<thead>
<tr>
<th>Model</th>
<th>( N_a )</th>
<th>( C_s )</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoAW+SVR [131]</td>
<td>200</td>
<td>5000</td>
<td>0.793</td>
<td>0.543</td>
</tr>
<tr>
<td>CNN+LSTM [6]</td>
<td>–</td>
<td>–</td>
<td>0.741</td>
<td>0.325</td>
</tr>
</tbody>
</table>

However, the reported CCCs in [131] is the highest, with a large codebook size of 5000 and with multiple assignment of 200 which is possible because of the large codebook size. However, the results presented in this study are not directly comparable to these results in Table 3.2 as the partitions used for modeling the BoAW+SVR system are different. When modeled on the same data, the BoOC is clearly outperforming BoAW irrespective of the backend employed that is, whether the SVR or the LSTM backends are being used. The results reported by an end-to-end CNN+LSTM system [6], a BLSTM-RNN system [239] and a BoAW-SVR system [131] were chosen as baselines since: (a) front-end of the BoAW+SVR is the closest in concept to the proposed front-end but lacks the joint discriminative training of the codewords; (b) the CNN-LSTM is one of the best performing end-to-end systems where the front-end and back-end are jointly trained; and (c) the BLSTM-RNN system employs a dynamic back-end with a typical front-end. These
comparisons together can help ascertain whether the proposed framework can serve as a suitable deep learning approach for continuous emotion prediction. Although again a direct comparison is not possible due to the different partitions used in \[6\], the proposed framework of BoOC+LSTM and the proposed front-end BoOC+SVR might be able to outperform the CNN+LSTM \[6\] system as shown in Table \[3.2\]. Another observation is that choosing a higher number of cluster centers until a point, such as a \(C_s\) of 300, improves the model performance for both arousal and valence predictions of the development set as shown in Table \[3.1\]. The proposed framework outperforms the BLSTM-RNN \[239\] system in Table \[3.1\] when predicting valence. It may be, that the arousal predictions can improve to outperform the arousal prediction of the BLSTM-RNN \[239\] if the number of cluster centers is optimized while modeling the proposed framework for arousal.

### 3.5.2 Codebook Analyses

The traditional BoAW is created by using hard single-assignment where each word is a \(C_s\)-dimensional binary vector initialized to zero. Each element of this vector corresponds to a unique codeword. The element of this vector that corresponds to the index of the codeword closest to the input feature vector, is set to 1 as illustrated in Figure \[2.5\]. The binary vectors for input frames in a segment are added up to create a ‘bag’ or a histogram. The problem with hard assignment in a deep learning system is that gradients cannot be back-propagated past this layer which in turn would prevent the codewords from being updated.

To circumvent this problem, soft assignment with the Softmin Layer was used and a probability distribution from the likelihood of the feature vectors belonging to each of the codewords is created as previously illustrated in Figure \[3.4\]. The numerator of \(3.2\) exponentially reduces with the increasing distance between the feature vector and the codeword. The denominator is the sum of these functions over all codewords. This results in a normalized value at each node of \(c\), so that the value of all nodes of \(c\) in Figure \[3.3\] adds up to 1 for any given segment. This concept is illustrated in Figure \[3.4\] where the
3.5.2 Codebook Analyses

top plot shows that the feature vector lies closest to cluster center K2 and farthest from K1 in a 2 dimensional feature space. The bottom plot illustrates the ‘audio word’ created for the scenario of the top plot. The magnitude of the audio word is highest at K2 which is the closest cluster center. Finally, it can be seen that the magnitudes of all elements of the optimised cluster word sum up to 1.

Table 3.3: CCC obtained with proposed BoOC +LSTM framework for different codebook sizes on RECOLA development partition [5].

<table>
<thead>
<tr>
<th>$C_s$</th>
<th>20</th>
<th>35</th>
<th>50</th>
<th>70</th>
<th>100</th>
<th>300</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.757</td>
<td>0.678</td>
<td>0.666</td>
<td>0.731</td>
<td>0.757</td>
<td>0.685</td>
<td>0.730</td>
</tr>
<tr>
<td>V</td>
<td>0.322</td>
<td>0.361</td>
<td>0.383</td>
<td>0.435</td>
<td>0.406</td>
<td>0.387</td>
<td>0.309</td>
</tr>
</tbody>
</table>

Table 3.3 shows the effect of the codebook size $C_s$ on the CCC obtained using the proposed BoOC+LSTM framework on the development partition [5]. The highest CCC over 20 epochs for each $C_s$ choice is reported. A smaller $C_s$ means a smaller feature size and although is more robust feature set can lead to over-generalization of the feature space. However, in table 3.3 we see that choice of $C_s$ is robust in the range of 20 to 500 with respect to CCC. The best CCC over 20 epochs for arousal is when $C_s$ is 100 and for valence when $C_s$ is 70.

To visually investigate if the back-propagated error gradients were successfully updating the codewords, the codewords were plotted alongside a scatter plot of the feature space as shown in Figure 3.5. The yellow points represent the 40-dimensional input feature vectors projected down to 2 dimensions using t-Distributed Stochastic Neighbor Embedding (tSNE). The red points are the 35 cluster centers of the input feature vectors generated using k-means++. The blue points are the 35 codewords that are updated over 20 epochs. The changing clusters can be discerned from the movement of the optimized codewords or cluster centers in the feature space as shown in Figure 3.5. It is observed in 3.5 that there is a significant migration of the optimized codewords from the original codewords especially evident in the right of the first dimension equaling 20. Finally, it should be noted that when the codewords were initialized as centers obtained via k-means++ they
CHAPTER 3. BAG-OF-OPTIMIZED CODEBOOKS - EMOTION BASED CLUSTERING OF FEATURES

Figure 3.5: The plot illustrates the location of the input feature vectors in yellow on an 88-dimensional feature space reduced to a 2-dimensional visualization space with t-distributed stochastic neighbor embedding (t-SNE). The position of the codewords derived from k-means++ is shown in red. The varying positions of the optimized clusters over 20 optimization epochs are shown in blue. [4]
3.5.3 Model Fitting

Figure 3.6: CCC over 30 epochs of the test set when different data partitions were used to train and test the model.

do not migrate much over the training epochs since they are reasonably good center to start off with. However, when the codewords are initialized as random feature vectors, they migrate to a much greater degree to reach much more optimal points, thereby, clearly demonstrating that the supervised training of the codewords has an appreciable effect on them. A similar effect on the cluster centers is observed when valence labels are used to train the model instead of arousal labels.

3.5.3 Model Fitting

To ensure that the 64-neuron double-layered LSTM was not likely to be overfitting the data, the optimized model parameters were further verified by swapping the training and
development datasets, i.e. model was trained on the development partition and tested on the training partition. The performance comparison between these two cases is shown in Figure 3.6. Since the performance using swapped dataset even achieves a slightly better CCC, it suggests no overfitting.

3.6 Summary

The Bag-of-Optimised-Clusters framework proposed in [4] included in this thesis has been shown to be an effective deep learning framework within which a front-end captures statistical information about low-level descriptors that are backed by solid theory in speech production and a back-end that models the dynamics of emotion can be jointly trained. The combination of the ability to capture distributional information of theoretically backed features with discriminative learning based on label information makes this a compelling approach for continuous emotion prediction. The framework allows for the space of low-level descriptors to be partitioned such that the statistical information captured by the front-end in the form of histograms of cluster memberships, is optimally utilized by the back-end to predict emotions. Comparisons with other systems that individually employ similar bag-of-words type features, LSTM back-ends, and end-to-end systems demonstrate that the proposed framework achieves similar state-of-the-art results while having all three of these characteristics.

Avenues for future investigation that look promising include supervised learning of the distance metric in addition to the codebook entries instead of simply employing Euclidean (or other) distance metrics and investigating the location and translation of codewords affected by discriminative training on other emotion datasets. Additionally, other approaches to representing statistical information of low-level descriptors within similar deep learning approaches are also worth exploring.
Chapter 4

Cross Culture Emotion Prediction using the Linear ARX back-end

4.1 Introduction and Motivation

Although there is a plethora of research in emotion prediction centred on high-level feature representations and neural network-based back-ends, the focus on building systems bridging cultural gaps in continuous emotion prediction is limited. Cross-culture emotion prediction has been a major challenge in continuous emotion prediction due to cultural differences in emotion expression and perception. For instance, in the western culture high arousal emotions are valued and promoted more than low arousal emotions, while by contrast, in Eastern culture, low arousal emotions are valued more than high arousal emotions [232]. Additionally, in terms of emotion perceptions, American subjects have been shown to rate the same expressions of happiness, sadness, and surprise more intensely compared to Japanese subjects [241]. Unsurprisingly, a limitation on current emotion prediction systems is that the models are non-transferable across different languages or cultures, owing to the great variabilities in emotion expressions and perceptions among different cultures. Approaches in the Audio-Visual Emotion Recognition Challenge (AVEC 2018) for
continuous emotion systems (CES), have mainly focused on detecting laughter segments as emotion salient regions \cite{242}, or LSTM-RNN-based data-driven approaches \cite{243,244}. In AVEC 2019 CES, Chinese utterances were introduced in addition to German and Hungarian utterances introduced in AVEC 2018, leading to a greater cultural variety based on these three languages.

The limitations in cross-cultural emotion prediction call for a robust model that is able to generalize well across cultures. A simple linear model that captures the fundamental but important information might be expected to generalize well across different cultures. Non-linear models may involve learning large numbers of model parameters, may suffer from significant increase in computation times with a relatively small increase on training data size such as $\epsilon - SVR$ or \cite{205}, may ignore the temporal dependencies within emotion labels \cite{33,62,183} and finally are hard to interpret which limits the development of emotion prediction systems. Though it may not be adequate to capture the complex relationship between speech and emotions, the linear models have been proven to be effective in previous studies, such as multivariate linear regression \cite{41} the autoregressive exogenous (ARX) models \cite{193} invested in the context of categorical emotion classification systems \cite{196}. The use of linear models in the back-end would help with interpretability but current wisdom suggests that they may not be adequate to model the complex relationship between speech and emotions. This is supported by the observation that previous studies on linear models in emotion recognition such as multivariate linear regression \cite{245} only focus on the emotion content at the current time step; and linear models that consider temporal dynamics such as ARX models \cite{195} have only been invested in the context of categorical emotion classification systems \cite{193}. In this chapter, it is demonstrated that generalized linear models such as ARX models can serve as suitable back-ends for the continuous prediction of emotion labels in the arousal-valence space across cultures and additionally lend themselves to a variety of analyses that aid with the interpretability of the model. Further, it also aims to compensate for cultural differences in emotion perception via label normalization, which transforms different label distributions across languages into a normal distribution to mitigate the label variability/mismatch in the model training process.
4.2. SYSTEM OVERVIEW

It was unclear if the cross-culture differences would be reflected in the label spaces and therefore we examine the label distributions in the different language spaces. Secondly, perhaps the perception of emotion was varying according to the culture that the people belonged to for similar expressions of speech features. So it was of interest to normalize the label distributions across languages and compare them with language-dependent cross-culture emotion systems. Furthermore, since video modalities might add complementary emotion information, especially in the valence space, feature and decision level modality fusions were employed.

4.2 System Overview

There are three proposed fusion strategies based on ARX models for examining the cross-culture effect:

- Feature-level fusion of audio and video modalities using dyadic information and label normalization
- Decision-level fusion of audio and video modalities
- Language-dependent decision-level fusion of audio and video modalities.

All three systems use BoAW-e audio features and FAUs video features. The AXR model is trained using the entire training dataset which includes both the German and Hungarian training files, kept consistent across all three systems. It should be noted that the audio features BoAW-e used in all three proposed systems were manually set to zero at the time frames where the interlocutor speaks or no one speaks, as it is assumed to supply no relevant information to the speaker’s emotion state at those given time. This was completed by chunking the speech files first using the turn information provided. Let $S_A(t)$ represent the acoustic features at a given time frame $t$ for the speaker. Then the final acoustic features $A(t)$ used for all three systems are:
As shown in Figure 4.1, the first system, referred as AFA system, introduces the dyadic information by using the interlocutor’s audio information. The speaker’s audio features BoAW-e, video features FAUs and the interlocutor’s BoAW-e features are concatenated to form a new feature vector prior to training, referred as the feature level fusion. Let $S_F(t)$ represent the facial features at a given time frame $t$ for the speaker, and $I_A(t)$ represents the facial features at time $t$ for the interlocutor. Then the concatenated features $S(t)$ are represented as:

$$S(t) = [A(t), S_F(t), I_A(t)] \quad (4.2)$$

The ARX model is trained with the normalized labels to eliminate the difference in the label space. The second system referred as A+F carried out the decision level fusion,
Figure 4.2: Proposed system for emotion prediction (System 3).

where the ARX models are developed for BoAW-e and FAUs separately. The decision-level fusion at the second stage is completed by regularized linear regression (RLR). The label normalization is also applied.

The third system named as AF shown in Figure 4.2 explores the language-dependent models, where the German model and Hungarian model were developed separately. The concatenated features of BoAW-e and FAUs are modified to

\[ S(t) = [A(t), S_F(t)] \]  

are fed into ARX to train a language-dependent model. This language-dependent model is then used to obtain the predictions for both German and Hungarian training partitions. The predictions from both models are concatenated as the input for the RLR predictor to conduct the second-level fusion. This model aims to explore language-specific information.

The reason why label normalization is not helping in this system is most probably owing to the independent training process for German and Hungarian models, which do not jointly train the model using two languages.
CHAPTER 4. CROSS CULTURE EMOTION PREDICTION USING THE LINEAR ARX BACK-END

4.2.1 ARX Model

The ARX model treats the emotion attributes such as arousal $y(t)$ time step $t$ as a combination of several past arousal values and the past input feature vectors as described in section 2.4.5. As a consequence, the most commonly employed methods for delay compensation accounting for the evaluator’s reaction lag mentioned in 2.3.1.2 are implicitly accomplished by the auto-regressive parameter choices.

4.2.2 Label normalization

As mentioned in section 4.1, the emotions perceived in different cultures would vary, where the differences in emotion perception are also a confounding factor in cross-culture emotion prediction systems. It was unclear if the cross-culture differences would be reflected in the label spaces. In order to examine this, the histogram inferring the distribution of the German and Hungarian labels was examined as displayed by Figure 4.3. Due to the differences observed between these two distributions, a normalization scheme was applied to both of the label spaces as:

$$y_{\text{norm}}(t) = \frac{y(t) - \mu_y(t)}{\sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y(t) - \mu_y(t))^2}}$$  \hspace{1cm} (4.4)

where $\mu_y(t) = \frac{1}{N} \sum_{i=1}^{N} y(n)$, is the mean of $y(t)$ over all time frames. It is worth noting here that language-based normalization considering the labels spread over each dimension was applied rather than over all the dimensions. This is conducted during the training process for each language. During the test phase, the predictions are de-normalized by using the mean and standard deviation computed from the training data partitions.
Figure 4.3: Histograms of Arousal (top), Valence (middle), and Liking (bottom) mean annotations for development (left) and training (right) sets. The blue histograms denote the annotations for German speech and the orange for Hungarian speech.
4.3 Experimental Settings

Three subsystems developed using audio and visual modalities were investigated. As shown in the baseline paper, BoAW with a codebook size of 100 was calculated over the eGeMAPS and was seen to perform best with respect to cross-cultural affect prediction. Similarly, 17-dimensional facial action units (FAUs) achieve the best results for the video modality. Hence, BoAW-e and FAUs are two adopted feature sets in the three subsystems owing to their superior performance. The data in the train partition, i.e. 34 German and 34 Hungarian, were used to train the models. The models are validated using the 14 German and 14 Hungarian speakers’ data provided in the development partition. For the submission of the test results, the data in both training and development partitions were used to train the ARX model and the test labels were generated for the 16 German, 18 Hungarian, and 70 Chinese test data that the organizers provided. ARX model was used as the back-end for each monomodal and multimodal system throughout the paper. The ARX model involves the estimation of three sets of parameters, namely, the order of the autoregressive filter, the order of the FIR filter determining the number of past feature-frames that influence the current emotion attributes emotion label, and the delay between input feature space and label space which is referred to as reaction lags. The AR filter order was varied in the range of \([0 \ 20]\), the FIR filter order was varied in the range of \([2 \ 20]\), and the delay was varied in the range of \([0 \ 20]\), all with a step of 2. The performance was evaluated using CCC.

4.4 Baseline System

The baseline system for this study was a 2-layer LSTM-RNN \((64 / 32 \text{ units})\) as a time-dependent regressor of the three targets (learned together) for each representation of the audiovisual signals as described in [21] with FAUs and BoAW-e front-ends.
4.5 Development Results

4.5.1 Feature and Decision-Level Fusion

First, the single modality system performance using ARX model was shown in Table 4.1. Compared to the baseline system using LSTM, ARX achieves better performance for audio modality, and relatively similar performance for video modality for arousal, valence and likability.

Table 4.1: Single modality - audio or video, performance with the ARX back-end

<table>
<thead>
<tr>
<th>Language</th>
<th>Front-end</th>
<th>Back-end</th>
<th>Arousal</th>
<th>Valence</th>
<th>Liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>BoAW-e</td>
<td>LSTM</td>
<td>0.434</td>
<td>0.455</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARX</td>
<td>0.523</td>
<td>0.442</td>
<td>0.317</td>
</tr>
<tr>
<td></td>
<td>FAUs</td>
<td>LSTM</td>
<td>0.606</td>
<td>0.639</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARX</td>
<td>0.576</td>
<td>0.601</td>
<td>0.060</td>
</tr>
<tr>
<td>Hungarian</td>
<td>BoAW-e</td>
<td>LSTM</td>
<td>0.291</td>
<td>0.135</td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARX</td>
<td>0.388</td>
<td>0.113</td>
<td>0.226</td>
</tr>
<tr>
<td></td>
<td>FAUs</td>
<td>LSTM</td>
<td>0.425</td>
<td>0.463</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ARX</td>
<td>0.305</td>
<td>0.368</td>
<td>0.110</td>
</tr>
</tbody>
</table>

Table 4.2: Feature and decision-level fusion systems with and without label normalisation in terms of CCC

<table>
<thead>
<tr>
<th>Language</th>
<th>Fusion</th>
<th>Arousal</th>
<th>Valence</th>
<th>Liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>AFA</td>
<td>0.584</td>
<td>0.603</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>AFA(norm)</td>
<td>0.598</td>
<td>0.608</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td>A+F</td>
<td>0.627</td>
<td>0.629</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>A+F(norm)</td>
<td>0.632</td>
<td>0.633</td>
<td>0.221</td>
</tr>
<tr>
<td>Hungarian</td>
<td>AFA</td>
<td>0.413</td>
<td>0.481</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td>AFA(norm)</td>
<td>0.431</td>
<td>0.464</td>
<td>0.260</td>
</tr>
<tr>
<td></td>
<td>A+F</td>
<td>0.423</td>
<td>0.455</td>
<td>0.180</td>
</tr>
<tr>
<td></td>
<td>A+F(norm)</td>
<td>0.444</td>
<td>0.442</td>
<td>0.157</td>
</tr>
</tbody>
</table>

The first and second fusion systems named AFA and A+F employ feature-level fusion which adopts the dyadic information as well as label normalization. The histograms of
CHAPTER 4. CROSS CULTURE EMOTION PREDICTION USING THE LINEAR ARX BACK-END

Table 4.3: Comparison between language-independent (LI) and language-dependent (LD) models

<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Arousal</th>
<th>Valence</th>
<th>Liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>LI</td>
<td>0.604</td>
<td>0.610</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>LD</td>
<td>0.631</td>
<td>0.619</td>
<td>-0.120</td>
</tr>
<tr>
<td>Hungarian</td>
<td>LI</td>
<td>0.417</td>
<td>0.478</td>
<td>0.228</td>
</tr>
<tr>
<td></td>
<td>LD</td>
<td>0.435</td>
<td>0.462</td>
<td>0.213</td>
</tr>
</tbody>
</table>

German and Hungarian arousal training labels were shown in Figure 4.3. The German labels are more condensed near 0 values indicating a neutral activated state while the Hungarian labels are more spread out suggesting a richer emotion expression range. A similar distribution was observed in the development partitions.

The difference in these two label spaces verifies the assumption of perception differences among cultures, which may lead to less accurate predictions in a cross-culture prediction system. Therefore, the normalization techniques were applied to transform these two label distributions into a normal distribution. The first and second systems with and without label normalization are shown in Table 4.2. It can be seen here that normalizing the labels before training improves the performance for all three attributes in the German dataset, and for arousal in the Hungarian dataset. The performance was slightly decreased for valence and likability for valence and likability.

4.5.2 Language-Dependent Models

The performance of language-dependent models AF are compared to the language-independent models, which is trained using the combined German and Hungarian training data, shown in Table 4.3. Relative improvements of 4.5% and 4.3% are observed for arousal in German and Hungarian respectively, and a 1.5% improvement for valence in German culture, which suggests that capturing the language-specific information benefits the emotion prediction tasks.
4.6 AVEC Challenge Test Set Result

Table 4.4: Test performance for CES challenge.

<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Arousal</th>
<th>Valence</th>
<th>Liking</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>Baseline</td>
<td>0.517</td>
<td>0.622</td>
<td>-0.019</td>
</tr>
<tr>
<td></td>
<td>AFA</td>
<td>0.421</td>
<td>0.606</td>
<td>0.127</td>
</tr>
<tr>
<td></td>
<td>A+F</td>
<td>0.602</td>
<td>0.686</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>0.604</td>
<td>0.678</td>
<td>0.009</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Baseline</td>
<td>0.525</td>
<td>0.397</td>
<td>-0.220</td>
</tr>
<tr>
<td></td>
<td>AFA</td>
<td>0.435</td>
<td>0.438</td>
<td>0.091</td>
</tr>
<tr>
<td></td>
<td>A+F</td>
<td>0.548</td>
<td>0.475</td>
<td>0.047</td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>0.544</td>
<td>0.439</td>
<td>0.043</td>
</tr>
<tr>
<td>Chinese</td>
<td>Baseline</td>
<td>0.238</td>
<td>0.423</td>
<td>-0.012</td>
</tr>
<tr>
<td></td>
<td>AFA</td>
<td>0.238</td>
<td>0.155</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>A+F</td>
<td>0.251</td>
<td>0.193</td>
<td>-0.018</td>
</tr>
<tr>
<td></td>
<td>AF</td>
<td>0.330</td>
<td>0.390</td>
<td>0.040</td>
</tr>
</tbody>
</table>

Three emotion prediction systems for arousal, valence, and likability were submitted as CES sub-challenge entries, shown in Table 4.4. The best systems for German culture yielded the CCC of 0.604, 0.686, and 0.127 for arousal, valence, and likability respectively, with relative improvements of 16.8%, 10.3%, and 7.7 times better performance for likability. Similarly, the best CCC for Hungarian culture was shown relative improvements of 4.4% and 19.7% for arousal and valence, and 8.6 times better performance for likability. This suggests that the proposed models can perform well in both German and Hungarian cultures. In terms of the Chinese culture which is not included in the training and validation sets, the CCC was 0.330, 0.390, and 0.040 for arousal, valence, and likability respectively, which was achieved by the AF model, indicating that training language-specific models benefit the transferable capability among the unseen languages. The performance in valence prediction in Chinese culture could not outperform the baseline system, which may indicate that valence could be a more complex task when transferring among different cultures which cannot be well captured by a linear model.
4.7 Summary

This chapter presents investigations into continuous emotion prediction across cultures and languages. For cross-culture emotion prediction systems with a linear ARX back-end was proposed, which is a linear regressor that incorporates the temporal dynamics of emotion labels. The performance was proven to outperform or be similar to the complex LSTM-based systems. Further, it was seen that label normalization across cultures was proven to be effective while joint training the emotion models with data from different cultures and yielded better prediction performances on the test set of those languages for which the labels were normalized (German and Hungarian). However, for the performance in unseen languages (Chinese) it was seen that the language-dependent training models which capture the specific language information were superior in predicting the emotion in unseen test languages (Chinese).
Chapter 5

Modelling Ambiguity with Bounded Parametric Distributions

5.1 Motivation and Introduction

As seen in chapter 4 for the same language, the individualistic ratings had similar distributions which may indicate that the distribution in emotion ratings is not an error that needs to be attenuated but rather individualistic variations that were influenced by the rater’s cultural background. This leads to the idea that the individualistic variations in the perception of emotions are influenced differently not only by individual observers’ cultures but also depends on their personal history of experiences, mood, personality, culture, involvement, context, the environment and physiological factors. Everyday emotional expressions are often subtle which can result in increased ambiguity of their perception. In typical emotion prediction systems, often emotion labels from multiple annotators are combined to mean ratings based on the assumption that the ‘mean’ rating is a better representation of the perceived emotion which is evenly felt across all raters thereby ignoring the inherent ambiguity (individual variations) in emotion as noise and these reductionist assumptions may be unrealistic for practical emotion
Emotion representations need to be capable of reflecting the diversity of human annotation, due to the inherently subjective nature of affective experiences, both while expressing and perceiving emotions. Recently, systems that employ distributions over the numerical scales to represent emotional states have been proposed. In this study, the common and widespread assumption that this distribution is Gaussian may not be suitable since the underlying numerical scales are bounded has been examined. Following this, a range of well-known distributions defined on bounded domains to ascertain which of them would be the most suitable alternative has been compared. Finally, statistical measures are proposed to enable quantifiable comparisons and the results are reported.

5.2 Modeling Ambiguity with Gaussians

As previously mentioned, the arousal and valence scales along which emotions are rated numerically are all bounded, while the support of a Gaussian distribution is over all real numbers. Consequently, representing the distribution of a set of emotion ratings as a Gaussian implicitly implies that there is a non-zero probability that emotion rating can be outside the allowable range. In practice, this might be an acceptable approximation as long as the variance of the distribution is small, but it is also worth noting that the Gaussianity assumption is less likely to be suitable when the numerical labels are close to the edge of the interval of allowable values. This is illustrated in Figure 5.1 which shows an example each for arousal (blue) and valence (red) where the Gaussian fit to the six ratings also implies there is a non-trivial probability of the rating falling outside the interval \([0, 1]\), which does not reflect reality (ratings are bounded to be within \([0, 1]\)).

To quantify the extent of this potential issue, the RECOLA dataset is analysed which includes a set of continuous time-varying arousal and valence ratings obtained from six independent raters, sampled every 40\(\text{ms}\), as emotion labels. In the RECOLA dataset both arousal and valence ratings are constrained to the interval \(x \in [-1, 1]\). For convenience (distributions with bounded support, refer section 5.3 are usually described by...
5.2. MODELING AMBIGUITY WITH GAUSSIANS

assuming the interval of support is $[0, 1])$, and without any loss of generality, a linear transform ($y = 0.5x + 0.5$) is applied to map the ratings to the interval $y \in [0, 1]$. Following this, for every frame (40ms) the maximum likelihood Gaussian fit over the six arousal (and valence) ratings is estimated by (5.1).

$$
\theta_{ML} = \arg \max_{\theta = [\mu, \sigma]} \prod_k \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y_k - \mu)^2}{2\sigma^2}}
$$

where, $y_k$ denotes the $k^{th}$ arousal/valence rating, $k = 1, ..., 6$, and $\theta$ denotes the mean ($\mu$) and standard deviation ($\sigma$) of the Gaussian.

Next, the probability implied by this Gaussian distribution that the rating falls outside

Figure 5.1: Examples frame from RECOLA showing six arousal and valence ratings as well as maximum likelihood Gaussian distribution fit to the ratings. In both cases, the best fitting distributions do not agree with the fact that the ratings are also bounded.
[0, 1] is computed as $P_{y \in [0, 1]}$ in (5.2).

$$P_{y \in [0, 1]} = 1 - \int_{0}^{1} \frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{(y-\hat{\mu})^2}{2\hat{\sigma}^2}} dy \quad (5.2)$$

where, $\hat{\mu}$ and $\hat{\sigma}$ are the maximum likelihood mean and standard deviation estimates for the Gaussian, that is, $\theta_{ML} = [\hat{\mu}, \hat{\sigma}]$.

An analysis on the RECOLA dataset revealed that for almost 10% of data, the maximum likelihood (ML) Gaussian fit to arousal ratings indicates a greater than 1% chance that arousal ratings will fall outside the allowed interval ([0, 1]). Similarly, the ML Gaussian fits to valence ratings for more than 1.5% of the data indicates a greater than 1% chance that valence rating will fall outside [0, 1]. Figure 5.2 shows a histogram of the $P_{y \in [0, 1]}$ for all the instances where $P_{y \in [0, 1]} > 0.01$. Finally, it is worth noting that this measure does not indicate goodness of fit within [0, 1], which is explored in section 5.3.

### 5.3 Comparing Distributions

The analyses in section 5.2 suggest that distributions with bounded support may be more appropriate than a Gaussian to model a distribution over numerical emotion ratings on bounded scales. In this section, an extensive list of commonly used parametric distributions with bounded support (refer Table 5.1) is examined, having reasonable shapes and degrees of freedom that would allow for the ambiguity in the perceived emotion ratings to be captured. These distributions are compared to each other and to the Gaussian based on how well they fit the ratings from each annotator which is quantified in terms of log-likelihoods. These comparisons are carried out using the best fitting distribution parameters, obtained as maximum likelihood estimates (MLE), as well expected value over the entire parameter posteriors. Note that most of the distributions listed in Table 5.1 have two parameters except the Trapezoidal distribution which has four parameters and the Logit Metalog family of distributions which admits a variable number of parameters with a greater degree of freedom in its shape with an increase in the number of parameters.
5.3.1 Comparing Maximum Likelihood Estimates

Intuitively the most straightforward approach to comparing these distributions is to determine the MLE of the distribution parameters at each point and compute the overall log-likelihood across the entire dataset. The maximum log-likelihood for each distribution is estimated, $L_{ML}$, as:

$$L_{ML} = \frac{1}{N} \sum_n \sum_k \ln p(y_{n,k}|\psi_{n,ML})$$

(5.3)

Figure 5.2: Histograms constructed from all instances or frames of emotion ratings when at least by 1 percent of the total area under the Gaussian Maximum Likelihood distributions lie outside the domain $[0, 1]$.

The Logit Metalog with 2 and 3 parameters is evaluated [32].
Table 5.1: Aggregate measures over all frames

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$L_A$</th>
<th>$L_V$</th>
<th>$E_A$</th>
<th>$E_V$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>6.0329</td>
<td>8.1728</td>
<td>5.0340</td>
<td>7.1299</td>
</tr>
<tr>
<td>Beta</td>
<td>6.1208</td>
<td>8.2510</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Truncated Gaussian</td>
<td>6.0548</td>
<td>8.1778</td>
<td>5.0336</td>
<td>7.0984</td>
</tr>
<tr>
<td>Logit Normal</td>
<td>6.0774</td>
<td>8.2345</td>
<td>4.6924</td>
<td>6.6195</td>
</tr>
<tr>
<td>Kumaraswamy</td>
<td>6.1160</td>
<td>7.9443</td>
<td>5.1044</td>
<td>7.0481</td>
</tr>
<tr>
<td>Raised Cosine</td>
<td>6.0658</td>
<td>8.1464</td>
<td>4.2919</td>
<td>6.4628</td>
</tr>
<tr>
<td>Trapezoidal</td>
<td>7.6986</td>
<td>10.1690</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logit Metalog 2 param.</td>
<td>5.9902</td>
<td>8.1277</td>
<td>4.7860</td>
<td>6.7603</td>
</tr>
<tr>
<td>Logit Metalog 3 param.</td>
<td>5.9931</td>
<td>8.1180</td>
<td>4.7497</td>
<td>6.6831</td>
</tr>
</tbody>
</table>

where, $y_{n,k}$ denotes the $k^{th}$ annotator’s rating at time frame $n$; $N$ is the total number of time points in the database; and $\psi_{n,ML}$ is MLE estimate of the distribution parameters at $n$:

$$\psi_{n,ML} = \arg \max_{\psi} \prod_k p(y_{n,k} | \psi)$$ (5.4)

The average maximum log-likelihoods estimated for all the distributions of the RECOLA dataset for arousal, $L_A$, and valence, $L_V$, are listed in the first two columns in Table 5.1.

Additionally, it is ensured that the overall log-likelihood score over the entire database is not being skewed by a few extreme data points; also, plots displaying the histogram of the frame-wise log-likelihood ratios between two distributions are presented for comparison. The histogram was expected to reveal a preponderance of positive values when the first distribution fits better than the second for a preponderance of data points.

Moreover, the beta distribution shape parameter MLE was subject to the constraint of always being greater than 1 since both shape parameter less than 1 causes the beta distribution to be U-shaped and any one shape parameter less than one causes the beta likelihood to tend to infinity at one of the edges. This is illustrated in Figure 5.3 where the shape parameters less than one result in a u-shape beta distribution alluding to the unlikely case of humans most likely perceiving extreme opposites of emotion for a given
5.3.1 Comparing Maximum Likelihood Estimates

Figure 5.3: A typical U-shaped Beta Distribution is obtained when both shape parameters are each less than one, indicating the unlikely case that mostly opposite extreme emotions are perceived amongst humans.
CHAPTER 5. MODELLING AMBIGUITY WITH BOUNDED PARAMETRIC DISTRIBUTIONS

Figure 5.4: Maximum Likelihood Trapezoidal Distribution Fit over Emotion Ratings.

5.3.2 Expected Log-Likelihood

A challenge with comparing distributions based on the MLE of the distribution parameters, as outlined in section 5.3.1, is that it may be skewed by overfitting. This is illustrated in Figure 5.4, which shows the trapezoidal maximum likelihood fit over a set of emotion ratings on the left plot. A very slight noise is added to one of the ratings as demonstrated by the red star in the plot on the right. The likelihood of the new set of emotion ratings under the previous maximum likelihood distribution estimate becomes 0 or the log-likelihood becomes infinity thus demonstrating that the trapezoidal distribution is prone to overfitting a limited number (6) of emotion ratings. This is of particular significance in this problem since the number of annotators, and consequently ratings, is usually very small (the RECOLA dataset only has 6 ratings) and the risk of over-fitting cannot be overlooked. That is, there may be certain parameters $\psi$ where the log-likelihood of the ratings given those parameters may be uncharacteristically high whereas other similar parameter values would be associated with a low log-likelihood. Furthermore, the chance of over-fitting increases with distributions with more parameters, such as the Trapezoidal distribution.
To overcome this limitation, the expected log-likelihood (ELL) was estimated as follows:

\[
E_p(\psi | y_n) [ln(p(y_n|\psi))] = \int p(\psi | y_n) \ln(p(y_n|\psi)) d\psi
\]  

(5.5)

where, \(E_p(\psi | y_n) \cdot\) denotes the expected value with respect to the posterior probability over the distribution parameters, \(p(\psi | y_n)\).

Further, it is seen that by applying Bayes’ theorem and rearranging the terms in (5.5), the following is obtained:

\[
E_p(\psi | y_n) [ln(p(y_n|\psi))] = \int p(\psi | y_n) \ln(p(\psi | y_n)) d\psi
- \int p(\psi | y_n) \ln(p(\psi)) d\psi
+ \int p(\psi | y_n) \ln(p(y_n)) d\psi.
\]  

(5.6)

Since, for any \(\zeta\), \(\int \zeta p(\zeta | y_n) d\zeta = 1\), the third term in (5.6) simplifies to \(ln(p(y_n))\).

\[
E_p(\psi | y_n) [ln(p(y_n|\psi))] = -H[p(\psi | y_n)] + ln(p(y_n))
- \int p(\psi | y_n) \ln(p(\psi)) d\psi
\]  

(5.7)

Integrating the third term by parts (5.8) is obtained,

\[
\int p(\psi | y_n) \ln(p(\psi)) d\psi = ln(p(\psi)) - \int \frac{p'(\psi)}{p(\psi)} d\psi
\]  

(5.8)

Substituting the result of (5.8) in (5.7),

\[
E_p(\psi | y_n) [ln(p(y_n|\psi))] = -H[p(\psi | y_n)] + ln(p(y_n)) + C
\]  

(5.9)
where, \( C \) is a constant of integration.

From (5.9), it is noted that the proposed measure \( E_{p(\psi|y_n)}[p(\psi|y_n)] \) depends upon the entropy \( H \) of the posterior distribution. A distribution with lower \( H \), is sharper and since ELL depends on \(-H\), a higher ELL indicates a better fit over a range of parameter values.

### 5.4 Experimental Settings

Experimental simulations were run over the arousal (and valence) ratings to compare different distribution families to determine their appropriateness in modeling ambiguity according to the measures explained in sections 5.3.1 and 5.3.2.
5.4.1 Prior Choices

Distribution parameters could be classified as location parameters describing the distribution’s position in its region of support, and scale and/or shape parameter(s) describing the distribution’s spread and shape. Location parameters were assigned uniform priors since the region of support is bounded and there is no reason to prefer one position over another.

Inverse-gamma priors were chosen for scale parameters as in hierarchical models for every Gaussian-like distribution because Inverse-Gamma prior has a desirable closed form expression due to its conditional conjugacy versus Half-Cauchy densities and the uniform distribution which was implicitly assumed for the MLE estimates. Gamma priors were chosen for shape parameters as in [249].
In the experiments described in the following sections, all relevant quantities of interest for all distributions were computed numerically over a uniform grid over the parameter space. Additionally, the leading frames of each utterance, where the ratings were all zero-valued, were discarded.

5.5 Results and Discussion

The average log-likelihood based on maximum likelihood fits for the various distributions, \( L_A \) and \( L_V \), and the expected log-likelihood over the parameter posteriors, \( E_A \) and \( E_V \), are listed in Table 5.1. In terms of the log-likelihood of the MLE fits, the Trapezoidal distribution best describes the ratings, followed by the Beta distribution represented as, \( \beta \). However, it is worth remembering that the Trapezoidal distribution has four parame-
5.5. RESULTS AND DISCUSSION

Figure 5.8: Histograms of expected log-likelihood ratios of Beta distributions to others over parameter posteriors for valence ratings.

The frame-wise log-likelihood ratios (based on MLE parameter estimates) were also analyzed by comparing the Beta distribution to all other distributions, for both arousal and valence, and plot the histograms in Figures 5.5 and 5.6. Similarly, the histograms of frame-wise expected log-likelihood ratios comparing the Beta distributions to the others are shown in Figures 5.7 and 5.8. All four plots across the majority of the data points or frames in the RECOLA dataset reveal that the Beta distribution is consistently a better choice than all the others.
5.6 Summary

Affective computing systems that model the distribution of numerical emotion ratings often assume a Gaussian distribution which as demonstrated is often incorrect. The most well-known families of bounded distributions were investigated as alternatives and the Beta distribution was found to be the most suitable to model the ambiguity of numerical emotion labels. In particular, the Beta distribution consistently proved to be a better fit for the emotion ratings, across time, in the RECOLA dataset. Mathematical reasoning was provided for the choice of measures selected to compare distribution families. Furthermore, while the RECOLA dataset was chosen for this study since it provides the highest number of label annotations amongst all the commonly employed datasets providing continuous affect annotations, the characteristics of continuous affect labels do not change with context. Therefore, the tools developed, analyses and findings reported in this chapter regarding comparing distribution families may be expected to be relevant to similar scenarios. This work demonstrates that the assumption of a symmetric ambiguity distribution model is sub-optimal to the beta distribution model since the beta family outperforms symmetric bounded and unbounded distributions. A consequence of such a finding may lead to investigating alternate measures of central tendencies other than the mean. As mathematical frameworks that consider ambiguity in affective computing systems develop, it is reasonable to expect that the choice of distribution to model affect labels would play an increasingly crucial role.
Chapter 6

Predicting Ambiguity Modelled with Beta Distributions

6.1 Motivation and Introduction

As described in chapter [5] a typical emotion database [43] contains multi-modal (speech, video, EEG, etc.) recordings of emotionally colored interactions graded with multiple real numbered emotion annotations within bounded support, over the duration of the recording. Most current CEPS such as [4] [18] [19] assume the average ratings as “true emotion” targets, being intentionally indifferent to the individualistic perceptual differences amongst raters which leads to a loss of valuable information on the diversity of opinions or and the clarity of the expressed emotion in speech [46]. The differences in ratings could be a reasonable indication of the ambiguity in the expressed emotion, since more subtly expressed emotions are less obvious and harder to identify by individuals of different backgrounds [32]. Ambiguity may be modeled as a distribution over affective ratings [53] and the investigation in chapter [5] demonstrated that amongst parametric distributions, the beta distribution is the most suitable distribution to model ambiguity [32]. Currently, however, since there is no framework to predict beta distributions for emotion
ambiguity, the main focus of this chapter is to develop a) a framework to predict ambiguity as beta distributions and present discussions on the quantities of interest from the predicted distribution, b) suitable measures to quantify the accuracy of the predictions have been revisited and c) present and improve the results by investigating performance and additionally processing the emotion data.

6.2 Proposed Framework - System Overview

As mentioned, there are a few challenges associated with developing a neural network to predict beta distributions as outputs for continuous emotion prediction (illustrated as colored blocks in Figure 6.1). Namely, (a) the choice of parametrization of the beta distribution to employ; (b) the choice of the loss function to minimize to train the neural network; and (c) the estimation of target labels, for neural network training, from affect ratings obtained from multiple annotators. In this section, we describe the novel approach we developed to address these challenges and develop a continuous emotion prediction able to predict emotion labels represented as beta distributions over the arousal and valence dimensions.

6.2.1 Beta Distribution Parameterization

The most well-known and commonly used formulation of the probability density function (PDF) of a beta distribution \cite{251} is in terms of two shape parameters, \(a\) and \(b\), as in (6.1).

\[
p_{\beta}(y) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1 - y)^{b-1} B(a, b) \tag{6.1}
\]

where, \(\Gamma\) is the Gamma function \cite{251}.

However, alternate parameterizations with other pairs of parameters, such as mean and variance or mode and concentration, are also possible. While it is straightforward to
Figure 6.1: An overview of the key aspects of the ambiguity prediction system that need addressing and their position within the flow of the emotion prediction system.
convert one set of parameters to another, they differ in terms of interpretability, ease of formulation of the loss function, choice of output activation function, and estimation of gradients. Specifically, we investigate three specific parameterizations with the beta distributions represented in terms of: (a) shape parameters \((a\) and \(b)\); (b) mean \((\mu)\) and variance \((\sigma^2)\); and (c) mode \((w)\) and concentration \((k)\). The relationships of these parameterizations to each other are given by the following equations:

\[
\mu = \frac{a}{a+b} \tag{6.2}
\]

\[
\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)} \tag{6.3}
\]

\[
w = \frac{a - 1}{a + b - 2} \tag{6.4}
\]

\[
k = a + b \tag{6.5}
\]

Shape parameters are most commonly employed for the simplest analytical expression of the beta distribution. However, the shape parameters themselves do not lend themselves to direct interpretation of the distribution. The mean-variance parameterization is of interest in this application since the variance (or standard deviation) of the distribution over arousal and valence has previously been employed as a measure of ambiguity [33]. The mode-concentration formulation is similarly of interest since the mode represents the most likely arousal and valence rating. The mean-variance and mode-concentration parameterizations also separate central tendency (mean and mode) and spread (variance and concentration). In order to compare the three choices, we implement and evaluate three versions of the emotion prediction systems. In terms of the system architecture, they only differ in terms of the activation function of the output layer since the different parameters differ in terms of the range of possible values they can take on. Noting that arousal and
6.2.2 Loss Functions

Valence ratings are bounded to a finite interval, [0, 1], mean, mode, and variance will also be bounded, and consequently, we employ a Sigmoid function as the activation function in the output neurons corresponding to these parameters. The shape parameters and concentration are all constrained to be positive but otherwise unbounded. Hence we employ ReLU activation, with a translation factor $T$ as given by (6.6), in the output neurons corresponding to these parameters.

\[
y = \begin{cases} 
  x & x > 0 \\
  T & x \leq 0 
\end{cases}
\]  

(6.6)

Where $T$ is set to 1 for predicting $a$ and $b$ (shape parameters). This constrains the neural network output to only bell-shaped beta distributions as explained in [32]. When predicting concentration ($k$), $T$ is chosen as 2, corresponding to the conditions $a > 1$ and $b > 1$.

6.2.2 Loss Functions

Since the output of the system is a probability distribution and the ground truth is a collection of ratings that the distribution is meant to represent, two approaches to defining a loss function for training the system can be envisioned: (a) quantifying the goodness of fit of the prediction distribution to the set of ground truth ratings provided by the human annotators; and (b) estimating the difference between the predicted distribution and the distribution that ‘best fits’ the ground truth ratings.

An obvious candidate for the first approach is a Negative Log-Likelihood (NLL) loss, whereby minimizing the loss function is equivalent to maximizing the log probability of the ground truth ratings given the prediction distribution:

\[
L(y|\beta) = -\frac{1}{N} \sum_n \sum_m \ln p(y_{n,m}|\beta_n)
\]  

(6.7)
where, $y_{n,m}$ denotes the $m^{th}$ annotator’s rating at time frame $n$; $N$ is the total number of time points in the data set, and $\beta_n$ represents the beta distribution predicted as the output at time $n$.

An obvious candidate also exists for the second approach. Namely, to estimate the maximum likelihood estimate (MLE) of beta distribution parameters from the ground truth ratings and minimize the KL divergence between the MLE and the predicted distribution. This is, in fact, equivalent to minimizing the Negative Log-Likelihood loss. Apart from this, it is also possible to minimize the mean square error (MSE) or mean absolute error (MAE) between the MLE of the distribution parameters and the neural network outputs. MSE or MAE losses are interpretable when the distribution parameterization is in terms of central tendency and spread. The MSE is selected as the loss function of the central tendency parameters because its easily interpretable, and differentiable globally and the support of the central tendency is [0,1] for a bell-shaped distribution. The MAE was the chosen loss for spread parameters due to the large concentrations and small standard deviation values for a fraction of frames where squaring the frame-wise error before summing (as in MSE or RMSE) for mode-concentration representations is likely to diminish learning of broader distribution estimates. This is explored in the experiments reported in section 6.4.1. Finally, it should be noted that maximum a posteriori (MAP) estimates may be employed instead of maximum likelihood estimates. A comparison of both possibilities is presented in section 6.4.1.

On the face of it, the NLL loss function ($-L_{a,b}$) appears to be entirely suitable. However, the sensitivity of log-likelihood to changes in the beta distribution parameters may vary quite significantly based on the actual value the parameters since given a set of $M$ ratings, $y = [y_1, y_2...y_M]$, the log-likelihood $L_{a,b}$ is given by

$$L_{a,b} = M \ln \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} + (a - 1) \sum_m \ln y_m + (b - 1) \sum_m \ln(1 - y_m)$$ 

(6.8)
6.2.2 Loss Functions

Figure 6.2: MLE of shaper parameter $a$ for current frame ratings (6 ratings) in blue vs. four neighboring frames with the current frame (54 ratings) in red.

which consequently translates to a non-linear sensitivity of the $L_{a,b}$ not only on changes of any one of the shape parameters as illustrated by (6.9) where the function is differentiated with respect to $a$, but also depends non-linearly on the magnitude of complex combination of the absolute values of both $a$ and $b$ and the spread and position of the ratings within the emotion support.

$$\frac{\delta L_{a,b}}{\delta a} = M[\psi(a + b) - \psi(a)] + \sum_m \ln y_m \tag{6.9}$$

where $\psi$ represents the Digamma function $\psi$. 

The top plot in Figure 6.3 shows the maximum likelihood distribution over four sets of ratings. The bottom plot shows the gradient of the log-likelihood function with respect to increasingly sharp distributions with $a = b$. It is observed that regardless of the spread and location of the ratings on the emotion support, the gradient is large for small values of shape parameters. Figure 6.3 illustrates that large changes in $a$ and $b$ may only correspond to small changes in the log-likelihood when $a$ and $b$ are already large. On the other hand, when $a$ and $b$ are small, the log-likelihood is much more sensitive to changes in $a$ and $b$. 

113
Consequently, when training a neural network on data where both large and small values of parameters may be expected, we may end up with the neural network unable to learn parameterization mapping appropriately for parts of the data where emotion ambiguity is low since errors in regions of large ambiguity lead to greater penalties with this loss function. The MSE and MAE loss functions do not suffer from this limitation but they ignore the fact that the output is a probability distribution. Combining NLL loss with MSE and MAE loss functions helps address the shortcomings of each of the individual loss functions.

6.2.2.1 Auxiliary Loss

As mentioned above, the auxiliary losses are the normalized mean square error of the central tendency parameters or the normalized absolute error of the spread parameters. The ground truth parameter $c_{GT}$ estimates employed here are either maximum likelihood (MLE) or maximum a posteriori estimates (MAP) over the frame-wise annotations. If the predicted parameters $c$ may be the mean $\hat{\mu}$ or the mode $\hat{w}$, and GT may refer to MLE or MAP parameters, the loss $L_c$ is given by (6.10).

$$L_c = \left( \frac{\hat{c} - c_{GT}}{c} \right)^2$$ \hspace{1cm} (6.10)

Similarly when $s$ may denote the spread parameters $\hat{\sigma}^2$ or $\hat{k}$, the mean absolute loss is calculated as in (6.11),

$$L_s = \left| \frac{\hat{s} - s^2}{s^2} \right|.$$ \hspace{1cm} (6.11)

6.2.3 Target Label Representation

As previously mentioned, the ground truth labels used as training sets in emotion corpora are the affect ratings obtained from multiple annotators. To train a neural network to
predict distributions over the affect ratings, an essential intermediary step is to infer target distributions from the ground truth ratings by obtaining the ‘best fitting’ distribution over the ground truth ratings in each frame. An intuitive approach to finding this ‘best’ fit is to find the maximum likelihood estimate (MLE) of beta distribution parameters (for any of the chosen parameterizations) given the ground truth ratings. i.e., the beta distribution that the ratings are most likely to belong to \[253\].
CHAPTER 6. PREDICTING AMBIGUITY MODELLED WITH BETA DISTRIBUTIONS

Figure 6.4: Illustration of (left-most column) frame-wise MLE of shape parameters in red over the shape parameter space. KDE of the prior distribution over MLE in shape parameter space (center-column); KDE of the prior distribution mode-concentration MLE in the \((w,k)\) space. The top row of plots are for arousal ratings and the bottom is for valence.
6.2.3 Target Label Representation

\[ \beta_n^{ML} = \arg \max_{\beta} \prod_n p(y_{n,m}|\beta) \]

\[ = \arg \max_{\beta_n=[a,b]} \prod_n \frac{y_{n,m}^{a_n-1}(1-y_{n,m})^{b_n-1}}{B(a_n,b_n)} \]  

(6.12)

where, \( \beta \) is a generic representation of the beta distribution’s parameter space. For example in the shape parameter space, \( \beta = a,b \) where \( a, b \in [1, \infty) \). \( \beta_n^{ML} \) denotes the maximum likelihood estimate of beta distribution in the \( n^{th} \) frame; and \( y_{n,m} \) denotes the rating from the \( m^{th} \) annotator in the \( n^{th} \) frame and \( y_{n,m} \) denotes the \( m^{th} \) annotator’s rating at time frame \( n \); \( B(,) \) is the beta function \[254\].

However, the MLE may have a tendency to ‘overfit’ in this case since the number of ratings available is quite limited (at most 5 or 6 in most emotion corpora). To somewhat alleviate this limitation, and taking into account that emotion is not expected to vary much across consecutive frames (frames are in the order of tens of milliseconds and it is reasonable to assume emotion does not change over tens or hundreds of milliseconds), ratings from \( F \) neighboring frames may be incorporated into the \( n^{th} \) frame as in (6.13) before computing the maximum likelihood estimation. Figure [6.5] illustrates the effect on the MLE of the shape parameter \( a \), which although unsurprisingly demonstrates a smoothing effect for extremely sharp distributions also demonstrates almost no effect on other frames, indicating that the sharp distributions may be unlikely occurrences.

\[ \tilde{y} = [y_{n-F}, y_{n-F+1}, y_{n-1}, y_n, y_{n+1}, \ldots, y_{n+F}] \]  

(6.13)

Alternatively, or even additionally, a reasonable prior distribution on the beta parameter space may be employed to obtain maximum a posteriori (MAP) estimates from the ground truth rating rather than MLE.

In [32], the priors were chosen according to the parameter type namely, a uniform prior for location parameters, inverse- gamma prior for spread parameters and the gamma prior
CHAPTER 6. PREDICTING AMBIGUITY MODELLED WITH BETA DISTRIBUTIONS

for shape parameters. The prior for each of the shape parameters in [32] was obtained as gamma distributions whose parameters were determined by the maximum likelihood estimates over the ratings in the training set. The joint distribution in the \( a - b \) space was obtained by multiplying the distribution by assuming that the shape parameters are independent. There are some limitations to this prior estimate as the MLE shape parameters are correlated as shown by the plot of the frame-wise MLE shape parameter plots as red points in Figure 6.4. Moreover, it is desired that prior distribution has a zero probability for all parameter values less than 1, which is not the case for the gamma priors. Kernel density estimation, a non-parametric method, may be employed to find a bivariate kernel distribution estimate (KDE) to represent PDF (prior distribution) of MLE parameters over the beta parameter space by summing component kernel smoothing functions \( K(.) \) for each MLE [255]. The smoothing function’s shape and bandwidth \( h \) depend on the MLE parameters calculated over training frames.

Kernel Density Estimation

Kernel density estimation aims to represent probability density functions (pdf) of random variables by summing component kernel smoothing functions \( K(.) \) for each data sample to produce a smooth, continuous probability curve [255]. The shape of kernel smoothing function and its bandwidth \( h \) is appropriately selected based on the data. For any real values of distribution parameters \( \beta \), the kernel density estimate \( \hat{f}_h \) is given by (6.14)

\[
\hat{f}_h(\beta) = \frac{1}{Nh} \sum_{n=1}^{N} K\left(\frac{\beta - \beta_n}{h}\right)
\]

(6.14)

where \( \beta_1, \beta_2, \ldots, \beta_n \) are random samples from an unknown distribution (the prior), \( N \) - the sample size is the number of MLE estimate pairs. Fig 6.4 illustrates the KDE priors obtained over the shape parameter space in the middle column of plots and the prior distribution obtained in the mode-concentration space in the right-most column of plots. The top plots show the priors over beta parameters for arousal and the bottom for valence. There are many possible choices for the kernel smoothing function \( K \) and in this study,
the normal distribution is chosen.

Fig 6.4 illustrates the KDE priors obtained over the shape parameter space (center column) and the mode-concentration space (right-most column) with the top row showing arousal priors and the bottom showing valence priors. The Gaussian kernel was chosen as the smoothing function \( K \) and its bandwidth was estimated as in [256]. The estimated bandwidth when too narrow may result in a rough distribution as illustrated by arousal priors in Figure 6.4. Hence priors based on multiples of the estimated bandwidths may be computed for MAP-based experimentation considering that the bandwidth is not too large leading to over-smoothing and loss of information. Moreover, since Gaussian kernels span unbounded supports, the KDE over the \( a, b \) space may be non-negative in the region below 1. Boundary correction methods [257] may be employed to contain the prior in the defined region.

6.2.4 Proposed System

A block diagram of the proposed framework is shown in Fig. 6.5, to predict beta distributions over the arousal (or valence) labels for each speech frame and its description is as follows. In the training phase, BoAW front-end features (described in section 6.3.2) are extracted from raw speech and input into the LSTM back-end. The LSTM block consists of one 128-neuron layer for valence and two 32-neuron layers for arousal prediction. The output of the LSTM block is input into two separate Linear blocks to output each beta parameter. The activation function choice at the output of the Linear block depends upon the type of parameters being predicted as described in section 6.2.3.

The total loss \( J \) is a function of the predicted parameters \( \hat{p}_{1n}, \hat{p}_{2n} \) and the ground truth (GT) parameters \( p_{1n}, p_{2n} \) and includes the NLL of the predicted parameters and additionally includes MSE and MAE losses weighted by hyperparameters \( \eta_{p_2} \) and \( \eta_{p_2} \) as given in (6.15). \( p_{1n}, p_{2n} \) denote the GT beta parameters for a frame for any of three parameterizations listed in section 6.2.1.
CHAPTER 6. PREDICTING AMBIGUITY MODELLER WITH BETA DISTRIBUTIONS

Figure 6.5: This is the block diagram of the continuous emotion prediction system that is proposed for predicting emotion ambiguity as beta distributions. The input to the system is the front-end speech features and the time continuous emotion labels from multiple raters and the output prediction are beta distribution parameters.
6.2.4 Proposed System

\[ J = -\frac{1}{N} \sum_n \left[ -L_{\hat{p}_1, \hat{p}_2} + \eta \hat{p}_1 M_1(\hat{p}_1, p_1) + \eta \hat{p}_2 M_2(\hat{p}_2, p_2) \right] \] (6.15)

where \( M_1 \) and \( M_2 \) are losses on the individual beta parameters that may be either the MSE or MAE.

The GT estimates employed in the MSE and MAE losses are either the MLE as in (6.12) or MAP estimates as obtained by (6.16).

\[ \beta_n^{MAP} = \arg \max_{\beta} \prod_m p(y_{n,m} | \beta) p(\beta) \] (6.16)

where, \( p(\beta) \) is a prior distribution over the parameter space obtained as per the description in section 6.2.3. The net loss is computed from the predicted frame-wise mode \( \hat{w}_n \) and concentration \( \hat{k}_n \) estimates within a training epoch and is given by (6.26).

6.2.4.1 Loss with Mean-Variance Parameterization

In terms of the \((\mu - \sigma^2)\) parameterization, the total loss is given by (6.17).

\[ J_{\beta(\mu, \sigma^2)} = -\frac{1}{N} \sum_n \left[ -L_{\hat{\mu}, \hat{\sigma}^2} + \eta \mu \left( \frac{\hat{\mu}_n - \mu_n}{\mu_n} \right)^2 + \eta \sigma^2 \left( \frac{\hat{\sigma}^2_n - \sigma^2_n}{\sigma^2_n} \right) \right]. \] (6.17)

The mean-variance \((\mu - \sigma^2)\) parameterization choice may be intuitive due to direct comparison with existing state-of-the-art systems evaluating their system performances in terms of mean emotion point estimates and ambiguity spread with standard deviation, but this may not be a straightforward way of parameterization of predicting beta distributions. Re-arranging (6.2) and (6.3) to obtain shape parameters \( a \) as in (6.18) and \( b \) as in (6.19) in terms of \( \mu \) and \( \sigma^2 \).
CHAPTER 6. PREDICTING AMBIGUITY MODELED WITH BETA DISTRIBUTIONS

\[ a = \left[ \frac{\mu(1 - \mu)}{\sigma^2} - 1 \right] \mu \]  \hspace{1cm} (6.18)

\[ b = \left[ \frac{\mu(1 - \mu)}{\sigma^2} - 1 \right] (1 - \mu) \]  \hspace{1cm} (6.19)

However, since we require that \( a \) should be greater than equal to 1 then from (6.18) we get the constraint (6.20).

\[ \sigma^2 \leq \frac{1 - \mu}{1 + \mu} \]  \hspace{1cm} (6.20)

Since we require that \( b \) should be greater than equal to 1 then from (6.19) we get the constraint (6.21)

\[ \sigma^2 \leq (1 - \mu)^2 \frac{\mu}{2 - \mu} \]  \hspace{1cm} (6.21)

Constraints (6.20) and (6.21) are hard to implement due to the complexity of the functions at the output activation. The simple constraints of having sigmoid activations at the output node for \( \mu \) and \( \sigma^2 \) predictions may be insufficient for training.

6.2.4.2 Loss with Shape Parameterization

Since the shape parameters should both be greater than 1, the MAE loss is chosen for both \((a - b)\) parameters to penalize higher errors more which would be reduced by the effect of squaring if MSE was the chosen loss function. The total loss \( J_{\beta(a,b)} \) is given by

\[ J_{\beta(a,b)} = -\frac{1}{N} \sum_{n} \left[ -L_{\hat{a}_n, b_n} + \eta_{a} \left| \frac{\hat{a}_n - a_n}{a_n} \right| + \eta_{b} \left| \frac{\hat{b}_n - b_n}{b_n} \right| \right] \]  \hspace{1cm} (6.22)
Where the NLL in an epoch is the average $L_{\hat{a}_n, \hat{b}_n}$ over $N$-frames as in (6.23).

$$J_{\beta(a,b)} = M \ln \frac{\Gamma(\hat{a}_n + \hat{b}_n)}{\Gamma(\hat{a}_n)\Gamma(\hat{b}_n)} + (\hat{a}_n - 1) \sum_m \ln(y_{n,m})$$

$$+ (\hat{b}_n - 1) \sum_m \ln(1 - y_{n,m}) \quad (6.23)$$

Including single-parameter losses such as MSE or MAE may not improve the mode prediction performance as it is a function of both $a$ and $b$.

### 6.2.4.3 Loss with Mode-Concentration Parameterization

Since each shape parameter is required to be greater than 1, the beta concentration parameter given by (6.5) is greater than 2. This is achieved by a ReLU activation with $T = 2$ as described in section 6.2.1. Moreover, the mode of a beta distribution is given by (6.4) and that is always in the interval $[0,1]$ as elucidated in (6.24).

$$0 < \frac{a - 1}{a + b - 2} < 1 \quad (6.24)$$

Since we have $a + b > 2$ enforced by the ReLU activation, from the left side of the inequality (6.24) we have $a > 1$ which results in the right side of the inequality so that $b > 1$.

The $w-k$ parameterization allows the inclusion of additional interpretable losses on each parameter such as the MSE and MAE formerly explained in section 6.2.1. Thus, the total loss $J_{\beta(w,k)}$ is given by (6.25),

$$J_{\beta(w,k)} = -\frac{1}{N} \sum_n \left[ -L_{\hat{w}_n, \hat{k}_n} + \eta_w \left( \frac{\hat{w}_n - w_n}{w_n} \right)^2 
+ \eta_k \left| \frac{\hat{k}_n - k_n}{k_n} \right| \right] \quad (6.25)$$
CHAPTER 6. PREDICTING AMBIGUITY MODELLLED WITH BETA DISTRIBUTIONS

\[L_{\hat{w}_n,k_n} = M \ln \left( \frac{\Gamma(\hat{k}_n)}{\Gamma(\hat{w}_n(\hat{k}_n - 2))\Gamma((1 - \hat{w}_n)(\hat{k}_n - 2) + 1)} \right) \]
\[+ \hat{w}_n(\hat{k}_n - 2) \sum_m \ln(y_{n,m}) + (1 - \hat{w}_n)(\hat{k}_n - 2) \sum_m \ln(1 - y_{n,m}) \]  

(6.26)

6.3 Experimental Results and Discussions

6.3.1 Database

The RECOLA multi-modal database [43] was used in all the experiments reported in this paper. It consists of over 9.5 hours of spontaneous dyadic conversation recordings in French of which, the expressed affective states in the first 5 minutes of each conversation are annotated by six French-speaking assistants and by the participants themselves. Of these, 18 conversations were provided as part of the Audio-Visual Emotion Recognition Challenge (AVEC 2016) [20] with 9 distinct utterances in the training and development partitions each. The test utterances were not used as the labels are not publicly available. Amongst the most widely used and publicly available datasets employed in continuous emotion prediction research, the RECOLA dataset has the largest number of individual annotations per sample (six annotations). Consequently, we chose it for all the analyses reported in this paper. All 18 utterances from the training and development partitions of the AVEC 2016 [20] were used in this study. 6 individual time continuous annotations sampled at 40ms were provided for arousal and valence for each utterance. Each rating is a numerical value in the interval \([-1, 1]\).

6.3.2 System Implementation

Bag-of-audio-words (BoAW) feature representations are extracted with 100 clusters from 20 dimensional MFCCs using OpenXbow [125]. Their first-order derivatives are computed and concatenated with the original feature vectors. Delay compensation is applied with 4
6.3.3 Beta Distribution Estimation

In a previous study that investigated the use of beta distributions over the arousal and valence ratings to describe emotion state, the distributions were characterized in terms of shape parameters $a$ and $b$, such that $a, b \in (0, \infty)$ over a fixed support $[0, 1]$ [32]. The same representation and support are retained in this work. Since both arousal and valence ratings are constrained to the interval $x \in [-1, 1]$ in the RECOLA dataset and the beta distribution family is typically described over $[0, 1]$, as in [32], without any loss of generality, a linear transform ($y = 0.4975x + 0.5$) was applied to map the ratings to the interval $y \in (0, 1)$. The value of $F$ was determined on prediction performance to be 8 for arousal and 1 for valence which means 6 ratings from the former and latter frames each are combined with the 6 ratings in the current frames. There are 102 ratings and 18 ratings available per frame (per 40ms) for arousal and valence respectively. Following this, the MLE beta distribution parameters were obtained over the eighteen arousal (and valence) ratings 6.13 in each frame as in (6.12).

The other method of label estimation discussed in section 6.2.3 is MAP estimation. The procedure followed to MAP estimate for the $n^{th}$ frame is as follows. First, the distribution
of log-likelihoods over the $n^{th}$ frame-ratings were computed over the 2-dimensional $a - b$ and $k - w$ parameter spaces. The approximate distribution was obtained for each of the two-dimensional grids of parameter combinations. The shape parameters were each varied from 1 to 1000 in steps of 1. In the $w - k$ space, the mode varied from 0.001 to 1 in steps of 0.001 and the concentration parameter varied from 1 to 2000 with a step size of 2. The MAP estimates were computed from the posterior distribution $p(\beta|y_n)$ which is proportional to the product of the grid of likelihood values over the parameter space and the prior distribution. The MAP estimates are the parameter values with the highest magnitude in the posterior grid product. The prior is a KDE as formerly explained in section 6.2.3. The reflection boundary correction method in [257] was employed to contain the prior in the region where both the shape parameters were greater than one. The kernel bandwidth $h_{\text{est}}$ for prior $p(\beta)$ estimation was calculated as in [256] for Gaussian Kernels. Since the prior visualization with the computed bandwidth indicated a rough distribution, the optimal $h$ was determined by repeating experiments with priors based on multiples of $h_{\text{est}}$. The brute-force search revealed that the prior estimated for arousal was $4h_{\text{est}}$ and for valence $h_{\text{est}}$.

6.3.4 Performance Evaluation

The prediction performance was evaluated with the commonly employed measures formerly described in section 2.5. The goodness of the model fit in each decile was evaluated using the relative root mean square error (RRMSE). The measures are described in the following sub-sections.

6.3.4.1 Correlation Coefficients

The most frequently used measures to evaluate the time series of predicted emotion quantities as formerly described in section 2.5 are the Pearson’s Correlation Coefficient ($\rho$) and Concordance Correlation Coefficient (CCC) described by (2.13) in section 2.5.1.3 which is based on the $\rho$ described in (2.12) and Mean Squared Error (MSE) described in section 2.6.1.
6.3.4 Performance Evaluation

2.5.1.1

\[ CCC = \frac{2 \rho \sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \]  \hspace{1cm} (6.27)

where the two-time series are \( x \) and \( y \), their means are \( \mu_x \) and \( \mu_y \) respectively and their standard deviations are \( \sigma_x \) and \( \sigma_y \). While the \( \rho \) tracks the similarity in trend between the two-time series, \( CCC \) incorporates the magnitude error between the two series. To evaluate the ambiguity in emotion labels there still exists no measures that are used most frequently by studies for example [53] has presented the ambiguity spread results by computing the \( \rho \) over each of the utterances between the predicted standard deviation and the standard deviation Gaussian MLE. Given the aim of continuous prediction of ambiguous emotion states, evaluation measures must account for two principles: (a) When ambiguity in the labels (inter-rater disagreement) is low, the mean rating is a representation of the emotional state, and variance of the predicted distribution must also be low and the central tendency of the predicted distribution must be close to the mean rating; and (b) when ambiguity in the labels is high, the mean rating is not as representative of the emotional state and difference between means of the predicted distribution and the ratings is more tolerable.

6.3.4.2 Relative Root Mean Square Error

The goodness of fit of the model may be evaluated using the Relative Root Mean Square Error (RRMSE). It may be calculated as follows. Let the standard deviation of the MAP estimates in the \( n^{th} \) frame of \( N_d \) frames in the \( d^{th} \) decile be denoted by \( \sigma_{d,n} \) and the predicted standard deviation of the corresponding frame be \( \hat{\sigma}_{d,n} \). To examine the prediction error for the frames in each of the deciles, the RRMSE within each decile is calculated as given in (6.28).

\[ RRMSE_{d} = \sqrt{\frac{1}{N_d} \sum_{n=1}^{N_d} \left( \frac{\hat{\sigma}_{d,n} - \sigma_{d,n}}{\sigma_{d,n}} \right)^2} \]  \hspace{1cm} (6.28)
CHAPTER 6. PREDICTING AMBIGUITY MODELLED WITH BETA DISTRIBUTIONS

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Beta</th>
<th>Gaussian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>(\mu - \sigma)</td>
<td>(a - b)</td>
</tr>
<tr>
<td>(\rho_\sigma)</td>
<td>0.002</td>
<td>0.500</td>
</tr>
<tr>
<td>(CCC_\sigma)</td>
<td>1.9e-05</td>
<td>0.207</td>
</tr>
</tbody>
</table>

Table 6.1: Concordance Correlation Coefficient on the training set of arousal MLE and predicted (unsmoothed) standard deviations for different parameterization choices.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Loss</th>
<th>((\eta_{p1}, \eta_{p2}))</th>
<th>(CCC_w)</th>
<th>(CCC_\sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arousal</td>
<td>(J_{\beta(a,b)})</td>
<td>(0,0)</td>
<td>0.603</td>
<td>0.334</td>
</tr>
<tr>
<td>Arousal</td>
<td>(J_{\beta(a,b)})</td>
<td>(1,1)</td>
<td>0.583</td>
<td>0.331</td>
</tr>
<tr>
<td>Arousal</td>
<td>(J_{\beta(w,k)})</td>
<td>(0,0)</td>
<td>0.612</td>
<td>0.414</td>
</tr>
<tr>
<td>Arousal</td>
<td>(J_{\beta(w,k)})</td>
<td>(1,1)</td>
<td>0.637</td>
<td>0.384</td>
</tr>
<tr>
<td>Arousal</td>
<td>(J_{\beta(w,k)})</td>
<td>(14,0)</td>
<td><strong>0.637</strong></td>
<td><strong>0.436</strong></td>
</tr>
<tr>
<td>Valence</td>
<td>(J_{\beta(a,b)})</td>
<td>(0,0)</td>
<td>0.320</td>
<td>7.1205e-05</td>
</tr>
<tr>
<td>Valence</td>
<td>(J_{\beta(w,k)})</td>
<td>(0,0)</td>
<td>0.311</td>
<td>0.008</td>
</tr>
<tr>
<td>Valence</td>
<td>(J_{\beta(w,k)})</td>
<td>(100,1)</td>
<td><strong>0.356</strong></td>
<td><strong>0.053</strong></td>
</tr>
<tr>
<td>Valence</td>
<td>(J_{\beta(w,k)})</td>
<td>(14,10)</td>
<td><strong>0.321</strong></td>
<td><strong>0.044</strong></td>
</tr>
</tbody>
</table>

Table 6.2: Concordance Correlation Coefficient on the development set predicted mode and variance for different loss functions based on MLE GT labels.

As a general rule, the model accuracy is considered to be excellent when RRMSE is less than 10%, good when RRMSE is between 10% and 20%, fair when RRMSE is between 20% and 30% and poor when RRMSE is greater than 30% \[259\] \[260\].

6.4 Results and Discussions

6.4.1 Result: Beta Distribution Parameterization

The simulation experiments with the proposed frameworks based on the parameterizations described in section 6.2.1 were performed by employing the log-likelihood loss function to determine if the networks were able to learn the beta distribution standard deviation over the arousal ratings. The results in terms of the correlation coefficients between the
standard deviation after training and the standard deviation of the ratings are listed in Table 6.1 where the first three columns are results of the beta loss function and the last column is based on Gaussian loss. The results demonstrate that the network is not training with $\mu - \sigma$ based beta parameterizations as discussed in section 6.2.4.1, unlike the Gaussian case. The networks are trained when the shape parameters and mode-concentration parameters are based on log-likelihood losses.

6.4.2 Result: Loss Functions

Since the results in section 6.4.1 demonstrated that the shape and mode-concentration parameters resulted in network training with the log-likelihood loss, these parameterizations were used in further simulations to determine the choice of loss function. Experiments with the $a - b$ parameterization were simulated separately based on the NLL loss function and the total loss function given by (6.25) based on MLE estimates as ground truth labels.

The result of the standard deviation prediction performance based on the NLL are given in TABLE 6.2 for the $a - b$ parameterization in rows 1 and 4 versus the $w - k$ parameterization in rows 2 and 5. The simulations employing total loss given by (6.25) showed comparable performance in standard deviation predictions but demonstrated a deteriorated performance on the mode prediction. This may be since the mode (6.4) depends on the combination of both $a, b$ and the mean square and mean absolute losses are operating on each shape parameter independently as explained in section 6.2.4.2.

The $w - k$ parameterization results based on MLE as ground truth labels and weighting scalars $\eta_w$ and $\eta_k$ determined by a grid search. For both arousal and valence, the results demonstrate that the loss function is sensitive to the choice of $\eta_w$ and $\eta_k$ due to the enhanced system performance with the incorporation of weighted auxiliary losses. The $CCC$ of valence spread although improved remains low.
CHAPTER 6. PREDICTING AMBIGUITY MODELLED WITH BETA DISTRIBUTIONS

Figure 6.6: Heat-map of log-likelihoods over the beta parameter space for widely distributed valence ratings on top plots and clumped ratings in the bottom plots. The parameterization on the left column of plots is shape parameters and the right column is mode-concentration.
6.4.3 Result: Target Label Representation

Figure 6.7: Top: The grey lines are the arousal ratings. The green shade is over the area over the arousal support that falls within the top 40% of the cumulative density of the predicted arousal ambiguity distribution. Bottom: The variance of the MAP estimates on the development set are plotted in blue and the predicted variance is plotted in red.

6.4.3 Result: Target Label Representation

TABLE 6.2 demonstrates a low CCC for valence spread prediction which may be due to a combination of factors such as the issue with gradients associated with high log-likelihood losses in section 6.2.2 and possible over-fitting of ratings by MLE. Therefore, the experiments with mode-concentration parameterization based on the loss given by (6.26) were simulated with MLE and MAP estimates as ground truth respectively. The results are displayed in TABLE 6.3 for raw predictions without post-processing and with post-processing in TABLE 6.4. For comparison with [53, 261, 262], the utterance-wise CCC and $\rho$ are averaged over 9 development utterances. TABLE 6.3 demonstrates that based MAP predictions outperform not only MLE predictions, but also the proposed
system predictions outperform the best GMM system [53] regardless of the ground truth parameterization for both arousal and valence ratings. It is worth noting that the standard deviation over the 9 utterances was in the order of 2 decimal places however, the tables report results up a precision of three decimal places for comparison with other research based on ambiguity [33, 263].

The results of MAP ground truth estimates for arousal are displayed in Figure 6.7. The top plot of Figure 6.7 demonstrates the span of the top 40% of the predicted distributions in green over the arousal ratings in grey. The span of the distribution has been calculated by finding the cumulative density function for each of the predicted distributions, then finding the desired percentiles on either side of the mean of the distribution, and
6.4.3 Result: Target Label Representation

![Maximum Likelihood valence Frames](image)

**Figure 6.9:** A histogram of the maximum likelihoods for all frames of the valence development partition.

then finding the corresponding projections on the arousal support which is the region to be shaded. There are 10 layers of shades starting from the $96^{th}$ percentile to the $60^{th}$ percentile. The bottom plot demonstrates the variance of the predicted distributions for arousal in red versus the variance of the MAP distributions in blue. Figure 6.8 shows a similar visualization for valence predictions.

In order to compare the ambiguity performance with previous studies [53], the utterance-wise $CCC$ and the $\rho$ over each of the 9 utterances are calculated, and then the average $CCC$ and $\rho$ of the standard deviation predictions without post-processing is presented in TABLE 6.3 and the measures for predictions with post-processing is presented in TABLE 6.4. The TABLE 6.3 shows the Beta MAP based representation of ambiguity outperforming ambiguity predictions on Arousal and Valence compared to both the GMR 8 and MLE ground truth estimations. TABLE 6.4 shows that even after smoothing the Beta MAP estimates display a competitive performance with the the GMR 8 outperforming the proposed representation only in with regards to the Pearson’s Correlation Coefficient for Arousal which could be because the in [33], the loss function employed for ambiguity
CHAPTER 6. PREDICTING AMBIGUITY MODELLED WITH BETA DISTRIBUTIONS

<table>
<thead>
<tr>
<th>System</th>
<th>Arousal</th>
<th>Valence</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho$</td>
<td>$CCC_\sigma$</td>
<td>$\rho$</td>
<td>$CCC_\sigma$</td>
</tr>
<tr>
<td>GMR 8 [33]</td>
<td>0.410</td>
<td>–</td>
<td>0.094</td>
<td>–</td>
</tr>
<tr>
<td>$\beta_{MLE}$</td>
<td>0.425</td>
<td>0.337</td>
<td>0.111</td>
<td>0.063</td>
</tr>
<tr>
<td>$\beta_{MAP}$</td>
<td><strong>0.440</strong></td>
<td><strong>0.393</strong></td>
<td><strong>0.161</strong></td>
<td><strong>0.123</strong></td>
</tr>
</tbody>
</table>

Table 6.3: Concordance Correlation Coefficient ($CCC$) and Pearson’s Correlation Coefficient ($\rho$) measure between predicted standard deviation (unsmoothed) and the standard deviation of ambiguity model.

<table>
<thead>
<tr>
<th>System</th>
<th>Arousal</th>
<th>Valence</th>
<th>Arousal</th>
<th>Valence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho$</td>
<td>$CCC_\sigma$</td>
<td>$\rho$</td>
<td>$CCC_\sigma$</td>
</tr>
<tr>
<td>BLSTM [264]</td>
<td>–</td>
<td>0.103</td>
<td>–</td>
<td>0.075</td>
</tr>
<tr>
<td>GMR 8 [33]</td>
<td><strong>0.568</strong></td>
<td>–</td>
<td>0.132</td>
<td>–</td>
</tr>
<tr>
<td>t-LU [261]</td>
<td>–</td>
<td>0.375</td>
<td>–</td>
<td>0.0481</td>
</tr>
<tr>
<td>SMC [262]</td>
<td>0.456</td>
<td>0.403</td>
<td>0.201</td>
<td>0.195</td>
</tr>
<tr>
<td>$\beta_{MLE}$</td>
<td>0.486</td>
<td>0.435</td>
<td>0.111</td>
<td>0.181</td>
</tr>
<tr>
<td>$\beta_{MAP}$</td>
<td>0.490</td>
<td><strong>0.488</strong></td>
<td><strong>0.318</strong></td>
<td><strong>0.314</strong></td>
</tr>
</tbody>
</table>

Table 6.4: Concordance Correlation Coefficient ($CCC$) and Pearson’s Correlation Coefficient ($\rho$) measures between predicted and standard deviation and the standard deviation of the ambiguity model after post-processing.

was based on a correlation measure.

The significant improvement in valence ambiguity predictions may be due to the 1). the enhancement due to the incorporation of auxiliary losses 2). MAP estimates over-fitting the ratings less than the MLE. Figure 6.9 shows the histogram of maximum likelihoods over valence training frames, demonstrating a large fraction of frames with clumped ratings introducing the limitations described in section 6.2.2. Figure 6.6 illustrates the log-likelihood distribution on the beta parameter shape parameter (left column) and mode-concentration (right column) space for broadly spread valence ratings on top and narrowly spread at the bottom. The log-likelihood distribution in the top plots is sharper for the region where the gradients are sharper demonstrated in Figure 6.3. Since there are issues with the log-likelihood loss function in this region, incorporating auxiliary losses may improve the
Figure 6.10: Bar plots of Mean Squared Error (MSE) of prediction mode (left column) and standard deviation (right column) for frames grouped into deciles on the standard deviation of the MAP estimates of arousal (top row) and valence ratings (bottom row). The x-axis denoted the decile and y-axis the MSE. The red crosses depict the MSE of the frames in the indicated decile. The bar plot shows the MSE computed over all frames accumulated up to the indicated decile.
learning as demonstrated in Table 6.2. Most importantly, the MAP-based simulation results are outperforming the MLE ground truth due to the MLE overfitting the valence ratings.

6.4.4 Decile-Based Analyses

The arousal (and valence) frames of rating were partitioned into 10 equal groups (deciles) based on their ambiguity level, i.e. standard deviations of their MAP estimates. The left column of Figure 6.10 demonstrates the error in mode-predictions for each decile indicated on the x–axis on top for arousal and bottom for valence. The bar plot indicates the average accumulated error over all frames until the indicated decile v.s. the red crosses which denote the average error in that decile. The plots demonstrate lower mode-prediction errors in lower deciles which is desirable as the point-predictions have more significance when the system can be sure of the perceived emotion. Similar plots for standard deviation predictions are shown in the right column of plots. The standard deviation errors are much higher in the lower deciles which may be a consequence of very high concentration values for the sharpest distributions. It may be worth remembering that although percentiles were chosen for grouping so as to have roughly equal-sized groups, comparisons across datasets may be limited because the range of standard deviation may be different.

Analysis of the valence ratings showed that the ranges between the maximum log-likelihood (ML) between frames differed in the order of a few tens as demonstrated in Figure 6.9 which is a very large difference in likelihoods and distribution sharpness. The large discrepancy may create problems in learning the maximum likelihood estimates for the clumped vs the broadly spread ratings as the log-likelihood loss is averaged over all frames. Another possible drawback is that because the log-loss is averaged over all frames, the log-loss for clumped frames may slightly dominate the small loss values over several broad-distribution frames. The possible limitations of the log-likelihood loss lead to the further examination of auxiliary losses.
6.4.5 Utterance-Based Analyses

Since the correlation coefficients to evaluate the performance was calculated per utterance and then averaged, the dependence of the mean standard deviation MAP variance estimate per utterance was compared with the prediction performance as is plotted in Figure 6.11 for arousal ratings. We see for utterance 9 that the average standard deviation is high compared to utterance 4 but the prediction performance in terms of correlation coefficients is much higher for utterance 4 than for utterance 9. This was unexpected because as discussed in section 6.2.2 the gradients for higher standard deviation values are expected to converge closer to the MAP estimate. To illustrate the reason for this behavior a histogram of the frame-wise standard deviation of the MAP estimates is plotted Figure 6.12 for the frames of utterances 4 and 9 shown in blue and orange respectively. Most of the frames of utterance 9 have a standard deviation of greater than 0.135 which are broad distributions, however, a significant portion of utterance 9 has extremely sharp MAP estimate distribution with very low standard deviations unlike utterance 4. However, each utterance is only about 7501 frames long. To confirm if the trends observed here are statistically relevant, the analysis needs to be repeated on larger datasets in the future and statistically significant measurements need to be obtained.

To find the performance of the predictions in each decile the relative error (RE) of the \( n^{th} \) frame in the \( d^{th} \) is found using 6.29 and the results are displayed in Figure 6.13. In Figure 6.13 and Figure 6.15 the Relative error per frame is exhibited with a * that is colour coded according to the decile it belongs to. The x-axis is the standard deviation of the MAP estimate and the y-axis is the relative error percentage. Figure 6.13 displays the results for arousal and Figure 6.15 displays the results for valence. The RRMSE is calculated collectively over all frames within a decile and is shown in black.

\[
RE_{d,n} = \left| \frac{\hat{\sigma}_{d,n} - \sigma_{d,n}}{\sigma_{d,n}} \right| 
\]

(6.29)

It is seen in Figure 6.13 that the RRMSE of deciles 20 to 100 is between 19.2% and 29.65% which indicates that the model has acceptable performance for these deciles. The
Figure 6.11: Illustrating that the mean value of an utterance’s ambiguity’s spread increases it may not result in the framework being able to learn the standard deviation of the ambiguity as demonstrated by the individual $\rho$ and $CCC$ values per utterance.

Error in the first decile is about 195% which indicates that the model performance on the predictions is extremely poor [260]. To investigate the result further Figure 6.14 is plotted which shows the MAP variances in each decile that the neural network is aiming to learn. For a bell-like distribution the standard deviation of the first and last deciles’ support are naturally large. In particular, we notice that the first decile shown in blue below 0.1 standard deviations encompasses points at the start of every utterance where the standard deviation becomes extremely low.

It is seen in Figure 6.15 that the RRMSE of deciles 20 to 90 is between 10.96% and 18.90% which indicates that the model has a good performance for these deciles. The error in the last decile is 20.32% which indicates an acceptable performance and the first decile is about 107.07% which indicates that the model performance on the predictions is extremely poor [260]. To investigate the result further Figure 6.17 is plotted which shows the MAP variances in each decile that the neural network is aiming to learn. In the first decile the standard deviations shown by the blue line correspond to extremely sharp distributions where the loss gradient is low. The percentage of frames with low standard deviation is far greater for valence than arousal as seen in Figure 6.14.
It is worth noting that the most common measure of the central tendency of ambiguity in existing literature for continuous emotion prediction tasks is the mean of the predicted distribution. Since means and consequently standard deviations are influenced by extreme observations [265], these numerical descriptions of central tendency and spread would be more relevant in the contexts of symmetrical distributions without extreme outliers. Most research incorporating ambiguity modeling as mentioned before employs mean-based measures which are relevant if the predicted distribution is roughly symmetric. A beta distribution though is not constrained to be symmetric. Therefore, mean-based measures of spread were applied to enable comparisons with previous studies. However, it was of interest to investigate if the assumption of a roughly symmetric distribution still applies to the emotion data being modeled with beta distributions. The skewness \( \tilde{\mu}_3 \) which is a measure of the magnitude and direction of asymmetry of the probability distribution of a random variable about its mean [266] of each beta distribution was calculated. A skewness \( \tilde{\mu}_3 = 0 \) indicates that the distribution is perfectly symmetrical e.g. the Gaussian distribution. Generally, \( \tilde{\mu}_3 \in [-0.5, 0.5] \), the distribution can be assumed to be roughly symmetric, if \( \tilde{\mu}_3 \in (-1, -0.5) \cup (0.5, 1) \), then the distribution is moderately skewed and highly skewed when \( \tilde{\mu}_3 \in ((-\infty, -1) \cup (1, \infty)) \) [267]. Analysis of the beta MLE and MAP estimates on the RECOLA train partition revealed that the majority of the estimates were roughly symmetric but 1.5% of the valence MLE and 2% of the valence MAP estimates.
were moderately skewed. This may indicate that there is a need to further explore the ambiguity over the labels of emotion data that other measures of central tendencies such as median or mode may be considered for future evaluations of ambiguity predictions rather than the mean.
6.5 Summary

This chapter reports the proposed emotion prediction framework based on time-varying emotion ambiguity modeled with beta distributions for both arousal and valence. Experiments indicated that with the alternate representation of beta distribution parameters, namely mode-concentration instead of mean-variance, it is possible to enhance the loss function. The reported results (TABLE 6.3 and 6.4) of the proposed framework compared with the existing frameworks indicate that the proposed system is able to predict the spread of ambiguity besides the mean. The framework is able to predict the mean of less ambiguous labels more accurately, which is consistent with previous research, and also

Figure 6.14: Frame-wise standard deviation of arousal MAP estimates shaded differently for each decile.
Figure 6.15: The Relative Error per frame of valence ratings is exhibited with a * that is color-coded according to the decile it belongs to. The x-axis is the standard deviation of the MAP estimate and the y-axis is the RE percentage. The RRMSE is calculated collectively over all frames within a decile and is shown with black *.

able to provide an overall more accurate prediction on the level of ambiguity (both high and low) than existing systems. Moreover, the results from the different beta distribution parameterizations highlight the problems with certain intuitive parameter representations and the analyses on the log-likelihood loss reveal the problems with learning sharp distributions. Furthermore, the experiments demonstrated that incorporating prior beliefs while modeling ambiguity especially improves ambiguity prediction on valence ratings and provides insight into why a probabilistic framework may perform poorly while predicting a series of valence ambiguity distributions which is the nature of log-likelihood or KL divergence-based losses for narrow distributions as well as the concentration of extremely
narrow maximum likelihood estimates being unrealistic for real-world emotion prediction problems which improves by incorporating a prior belief on the distributions’ parameter space. The proposed framework reported in this chapter can not only track the level of ambiguity and mean emotion of the labels over time especially accurately within regions of low ambiguity but, is also simplistic, based on a simple unimodal distribution model which outperforms far more complex ambiguity models such as 8 Gaussian Mixtures. Moreover, the skewness of the MAP and MLE distributions showed that the ground truth distributions for a fraction of frames were skewed which may indicate the need to explore central tendency measures other than the mean rating. In the future, it is worth exploring methods of comparison between two time series of distributions that may indicate the

Figure 6.16: Frame-wise standard deviation of valence MAP estimates shaded differently for each decile.

![Standard deviation of Ratings per Decile](image)
consequence of deviations of the parameters of the predicted distribution from the MAP estimates, thereby not only enhancing the interpretability of the results but also making a pathway to improving loss functions tailored for distribution predictions.
Chapter 7

Conclusion and Future Directions

7.1 Conclusion

This thesis has described a series of investigations into various aspects of continuous emotion prediction systems, specifically, exploring alternative acoustic features discriminatively encoded, examining the robustness of the auto-regressive exogenous linear back-ends in a cross-language scenario, investigating the appropriateness of modeling ambiguity with different distribution families and finally predicting ambiguity modeled with beta distributions. Additionally, there were investigations made into the parameter spaces of the label distributions, and inferences were made to improve system performance. The aims and contributions of these investigations may be listed as follows: 1) examining the effect of discriminative encoding of acoustic features obtained from the proposed novel framework that learns from emotion labels while encoding acoustic feature distributions by investigating feature cluster positioning within the feature space; 2) examining the effect of inter-language labels distributions and investigating the robustness of a linear back-end due to the virtue of its simplicity and linearity for cross-language emotion prediction systems; 3) developing techniques to compare different distribution families to determine the most appropriate distribution family to model ambiguity and finally; 4) propose a framework to predict ambiguity modeled with beta distributions and investigate the parameter
distribution space to obtain insights into label distributions.

7.1.1 Effect of Discriminative Learning of Acoustic Feature Encoding

Chapter 3 explores marginal distributions of LLDs within an emotion window, indicating that the unimodal distribution parameters distribution may be insufficiently representing the feature distribution for a substantial number of windows and joint feature distribution-based representations especially learned based on emotion label information may enhance the feature representation for emotion prediction tasks. The Bag-of-Optimised-Clusters framework is proposed to learn such a feature representation by capturing the statistical information of the LLD distributions and encoding it based on the dynamics of emotion by joint training. The framework has been an effective deep learning framework in allowing the LLD feature space to be partitioned such that the statistical information is captured by the front-end in the form of histograms of cluster memberships. Comparisons with other systems that individually employ similar bag-of-words type features, LSTM back-ends, and end-to-end systems demonstrate that the proposed framework achieves similar state-of-the-art results while having all three of these characteristics.

A key insight into the effect of incorporating label information in the joint feature space partitioning as revealed in Chapter 3 is that the position of certain partitions in the feature space where there are more unusual feature distributions is key in distinguishing between emotion dimension intensities since there is a significant migration of the optimized codewords from their start points in the outer edges of the feature space representation. Moreover, the distance of migration when the codewords were initialized as centers obtained via k-means++ is significantly less during the training epochs than when the centers are randomly initialized thereby, clearly demonstrating that the supervised training of the codewords has an appreciable effect on them.
7.1.2 Investigation into Cross-Language Affect with a Linear Back-end

Chapter 4 presents investigations into continuous emotion prediction across languages which showed similarities in label distribution within languages and dissimilarities across languages when the subjects were prompted with similar tasks. Novel frameworks based on linear regressor ARX back-ends were proposed for cross-language affect based on the hypothesis that although the model parameters can incorporate temporal dynamics perhaps the model being a linear regressor would be robust across languages by not overfitting on a single language. The performance of the proposed model at times outperformed or gave similar performances to the complex LSTM-based systems. Furthermore, label normalization across languages was proven to be effective while joint training the emotion models with data from different languages. Within a culture, the language-dependent models which capture the specific language information were shown to be beneficial in predicting the unseen test languages with superior performance. The performance in valence prediction in Chinese culture could not outperform the baseline system, which may indicate that valence could be a more complex task when transferring among different cultures which cannot be well captured by a linear model.

7.1.3 Impact of Modelling Ambiguity with Parametric Distributions

Chapter 5 explores the limitations of employing normal distribution assumptions commonly applied in modeling the distribution of numerical emotion ratings and proposes measures comparing families of distributions in the suitability of modeling emotion ambiguity and investigated an exhaustive list of the most well-known families of bounded distributions as alternatives to model emotion ambiguity in a time instant. The Beta distribution family was found to be the most suitable to model the ambiguity of numerical emotion labels, consistently proving to be a better fit for the emotion ratings, across time, in the RECOLA dataset. Measures were proposed to enable comparisons between distribution families and mathematical reasoning was provided for selecting measures to compare distribution families. Furthermore, while the RECOLA dataset was chosen for
CHAPTER 7. CONCLUSION AND FUTURE DIRECTIONS

this study since it provides the highest number of label annotations amongst all the commonly employed datasets providing continuous affect annotations, the characteristics of continuous affect labels do not change with context. Therefore, the analyses and findings in this chapter are expected to be relevant in all similar scenarios. With the increasing trend towards modeling humans’ affective states being sampled from distributions rather than hard labels, it is reasonable to expect that the choice of distribution to model affect labels would play an increasingly crucial role.

7.1.4 Predicting Ambiguity Modeled with Beta Distributions

Chapter 6 proposes an emotion prediction framework based on time-varying emotion ambiguity modeled with beta distributions for both arousal and valence. The Concordance correlation coefficient between the predicted ambiguity and the emotion ambiguity from ratings (computed as the standard deviation among six raters) was adopted as the evaluation metric to quantify system performance. The results of the proposed framework compared with the existing frameworks indicate that the proposed system can predict the spread of ambiguity besides maintaining a performance comparable to state-of-the-art systems on mean emotion prediction, suggesting the proposed framework is a promising approach. Especially worth mentioning are the insights into valence ambiguity prediction which were obtained from the analysis of the prediction results with the proposed system. The chapter reports the importance and consequences of incorporating prior beliefs while modeling ambiguity for valence and that a framework can be developed to predict the ambiguity of valence ratings as a series of distributions and also consider the limitations of loss functions. The proposed framework can not only track the level of ambiguity and mean emotion of the labels over time especially accurately within regions of low ambiguity but, is also simplistic, based on a simple unimodal distribution model which outperforms the far more complex ambiguity models such as 8 Gaussian Mixtures, thereby opening new avenues of investigation with more complex ambiguity models on their limitations in ambiguity prediction. Since the framework predicts the mean of less ambiguous labels more accurately, it is consistent with recent research and additionally provides an overall
more accurate prediction on the level of ambiguity (both high and low) than existing systems. Therefore this work provides a path for using ambiguity to improve conventional emotion prediction systems.

7.2 Future Work

Keeping in mind that the ultimate goal of this and emotion-related research is to deploy the models and developed systems for real-world applications, the performance of proposed developments may as future direction, be incorporated into realistic scenarios such as ascertaining the affective state in call centers or in classroom settings where the ambiguity of the predicted emotion is likely to play an important role in delivering considerate responses. Besides this suggestion, several specific avenues for extending the work presented in this thesis are listed as follows.

About the BoOC framework introduced in chapter it may be worth exploring alternative distance metrics employed in the supervised learning of codewords rather than simply employing Euclidean (or other) distance metrics or their functions. Additionally, other approaches to representing statistical information of low-level descriptors within similar deep learning approaches are also worth exploring.

While the measures developed in this thesis to find an appropriate model for emotion ambiguity may be extended to finding models of soft labels within any research field that suffers from defining the ground truth as hard labels, the ambiguity models here have been computed over six ratings for each given time instant since that was the highest number of label annotations available amongst all the commonly employed datasets providing continuous affect annotations, i.e. from the RECOLA dataset. It would be useful to have more ratings per speech utterance in emotion databases for more precise distribution modeling. Moreover, ambiguity models could be further improved by computing uncertainty from various sources such as epistemic uncertainty which may be measured from the spread of the model parameters, and aleatoric uncertainty which may arise due
to the randomness of ratings and may be harder to measure. With the growing interest in considering emotion ambiguity and the development of ambiguity-based frameworks in affective computing systems, it is reasonable to expect that the choice of distribution to model affect labels would play an increasingly crucial role and therefore another avenue worth exploring is evaluating more complex mixtures of distributions with the measures proposed here to ascertain even better models for ambiguity.

Additionally, the ambiguity models presented in this thesis were based on individual emotion ratings compensated with a constant delay approximating the reaction lag from all raters to be equal. However, it is reasonable to expect that in a realistic scenario, the individual reaction lags are distinct and that realigning them will improve the ambiguity distribution estimation. Consequently, it is worth exploring alternative models to model the reaction lags of raters individually which may be due to a combination of perception lags and lag in operating the equipment which currently remains a confounding factor in emotion prediction systems coupled with the possibility of the lag varying throughout the utterance. It may be worth investigating if the perception lag of the raters is dependent on emotional ambiguity.

There is a pressing need for the development of appropriate evaluation techniques to evaluate emotion ambiguity prediction performance. The performance of the proposed framework predicting beta distribution is measured based on the concordance correlation coefficient of the mean and spread of the predicted distribution with the underlying label distributions. The most popular metric for comparing distributions is the KL divergence, which outputs a number that does not help infer if the distributions are misaligned in concerning the location, the spread, or their directions and therefore gives no insights into the predicted ambiguity. Although the ambiguity modeled is in terms of the distribution as a whole in this study the performance of the system was measured by the standard deviation of the predicted distribution which is only one of the important aspects of the modeled ambiguity. Therefore the avenue for future investigation that looks promising is exploring methods of comparison between two time series of distributions that would be able to lend insights into the consequence of deviations of the parameters of the predicted
distribution from the underlying distribution estimates, thereby not only enhancing the interpretability of the results but also making a pathway to improving loss functions for distribution predictions.

Finally, although the LSTM back-end can capture the temporal dynamics in the time series of the mean emotion ratings and consequently the mean of the ambiguity distributions, with the rapid developments in deep learning, a multitude of deep neural networks may be investigated in the future for improving both the central tendency and spread predictions of emotional ambiguity.
References


