THE EVALUATION OF THE STABILITY OF ACOUSTIC FEATURES IN AFFECTIVE CONVEYANCE ACROSS MULTIPLE EMOTIONAL DATABASES

A Thesis
Presented to
The Academic Faculty

by

Rui Sun

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of electrical and computer engineering

Georgia Institute of Technology
August 2013

THE EVALUATION OF THE STABILITY OF ACOUSTIC FEATURES IN AFFECTIVE CONVEYANCE ACROSS MULTIPLE EMOTIONAL DATABASES

Approved by:

Professor Hongwei Wu,
Committee Chair
Department of Electrical and
Computer Engineering
Georgia Institute of Technology

Professor Elliot II Moore, Advisor
Department of Electrical and
Computer Engineering
Georgia Institute of Technology

Professor Mark A. Clements
Department of Electrical and
Computer Engineering
Georgia Institute of Technology

Professor Allen R. Tannenbaum
Department of Electrical and
Computer Engineering
Georgia Institute of Technology

Professor Bruce N. Walker
School of Psychology and School of
Interactive of Computing
Georgia Institute of Technology

Date Approved: 16 May 2013

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Dr. Elliot Moore, whose vision and support makes this thesis possible. Without the guidance and patience from him, the research presented in this thesis can not be finished. I also feel thankful to Dr. Mark A. Clements, who always generously provided me the speech data, the ideas of cooperation work, and the answers of my questions. My heartfelt thanks go to Dr. Monson Hayes, without who I could not meet my advisor Dr. Moore and feel so encouraged to march towards my doctoral degree.

I would also like to thank other members of my committees, Dr. Allen R. Tannenbaum, Dr. Hongwei Wu, and Dr. Bruce Walker, for the time they devoted and the insightful comments offered. I deeply appreciate their extraordinary efforts to schedule my defense and make it happen smoothly. My great thank also goes to Dr. Shrikanth S. Narayanan, Dr. Sidney D'Mello, Dr. Bjoern Schuller, and the HUMAINE society for generously providing me the databases for my research.

During my graduate years in Georgia Tech, I was honored to work with a group of talented and passionate colleges, Winston Percybrooks and Juan F. Torres, who inspired and encouraged me. And my special thanks go to Ms. Mariliouse Mayko, Ms. Tasha Torrence, Ms. Pat Potter, and Ms. Jacqueline Trappier, who take care of the necessary administrative details to make my life much easier.

The credits of all my accomplishment should go to my paresents, Jifang Sun and Xiulan Qian and my big family. Without the endless support from them, I could not have such a beloved life and the work can not be done. I also thank all my friends in Savannah, Atlanta, San Jose, and China, for their love, accompany, help, and faith in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Speech is enriched by emotion. Emotion recognition in speech benefits human-computer interface to build more effective, user-friendly, and intelligent applications. While the majority of traditional research in emotional speech recognition has focused on the use of a single database for assessment, it is clear that the lack of large and diverse enough databases has presented a significant challenge in generalizing results for the purpose of building an emotion classification system yielding results for database with varieties. Recently, work has been reported on cross-training emotional databases to examine the consistency and reliability of acoustic measures in performing emotional assessment. However, the acoustic features that have been studied were limited to prosodic and spectral features, and the examination remained mainly on attempts on selecting prototypical data. The objective of the research presented in this thesis is to systematically investigate the computational structure for cross-database emotion recognition. The systematic research consists of evaluating the stability of acoustic features, particularly the glottal and Teager Energy based features, and investigating the critical procedures of normalization methods and data fusion techniques. In cross-database classification, challenge rises from the variation possessed by different databases, e.g., the naturalness of emotion expressed, and the recording conditions. To cope with the variation, three normalization methods are studied to show that normalization can improve the performance of cross-database classification even to a higher accuracy rate than the self cross-validation without normalization. Motivated by the lack of large and diverse enough emotional database to train the classifier, using multiple databases to train poses another challenge of data fusion. This thesis proposes two data fusion techniques, pre-classification SDS and post-classification ROVER to study the issue. The systematic computational structure proposed in this thesis improves the performance of cross-database binary-emotion recognition by up to 23% for neutral vs. emotional and 10% for positive vs. negative.

CHAPTER I

INTRODUCTION

Recognizing the emotional information embedded in speech has drawn much research effort [62, 13, 86, 45]. The characteristics of emotional databases are affected by many factors including the recording conditions, the naturalness of emotion expressed (acted or authentic), the language, and the basic information of speakers (e.g age, culture, and gender). Additionally, the lack of a single emotional database large and diverse enough to cover the variety of emotional expressions makes it difficult to train a system that can be generalized. As the direction of emotion recognition heads to more realistic use, the evaluation of emotion recognition engines in generalizing the performance on one data set to another data set is more and more important. Currently, the widely accepted emotion recognition method is to train the system on one data set and test it using the same data. However, this makes the system sensitive to the data's quality, based on which the system is built. Cross-database training and testing can contribute on building a generalized emotion recognition system by training it on one data set and testing it on another one. limited to prosodic and spectral features, and the examination remains mainly on attempts on selecting prototypical data. The goal of the research presented in this thesis is to systematically investigate the computational structure for cross-database emotion recognition. The systematic research consists of evaluating the stability of acoustic features, particularly the glottal and Teager Energy based features, and investigating the critical procedures of normalization methods and data fusion techniques.

Cross-database training and testing has many challenges, three of which are the main focus of this thesis. First, the features from the training data should possess the similar emotion distinguishing ability for the testing data. Therefore using the features exhibiting the stable affective conveyance capability across multiple databases is important for cross-database evaluation, which remains as a challenge. Second, the variation coming form different databases (e.g., the recording conditions, the naturalness of emotion, the gender of speaker) leads to the mismatch of features from different databases even expressing the same emotion. The third challenge rises from the availability of multiple training databases. The issue of the methodology to combine the information gained by each training database should be addressed. While cross-database training and testing has not traditionally been a heavily researched topic, the challenge of building a practical emotion recognition system from limited data has prompted several publications addressing some of these challenges. Previous research efforts focused on reporting the direct comparison between databases using thousands of prosodic and spectral features in a brute-force way, and the approaches of selecting the prototypical samples to train the recognition engine. An account of the past research efforts is presented in Chapter 3.

To address the first challenge of cross-database emotion recognition of the stability of features, the glottal and Teager energy based features are evaluated using multiple databases under different experimental conditions. The emotion recognition research has been conducted in several channels, e.g., facial expression, body gesture, and speech. From the speech standpoint, the effort of emotion detection was made using the conversational cues (e.g., lexicon, discourse) [19], and the acoustics (e.g., pitch, spectral) [8, 64]. In the acoustic channel, large set (more than 4000) of prosodic and spectral features have been studied [64]. Even using the pitch-related features solely, research [8] showed fairly good performance in emotion recognition. Research has shown that the glottal waveform dynamics and Teager energy features can play an important role in voice characterization [15, 92, 57, 58, 84, 9, 96, 37]. Therefore, this thesis investigates the stability of the glottal and Teager energy based features in cross-database emotion recognition without the information from other channels. The extraction and stability evaluation of the glottal and Teager energy features using multiple databases is presented in Chapter 5.

Another problem of cross-database evaluation comes from the mismatch between databases in the emotion labeling strategies. Emotion has attracted much psychologists' effort on studying the best way to describe. There are two prominent emotion description methods available and widely used, discrete lexicon of emotional words (e.g. "sad" and "interest") or a dimensional scale (e.g. [0.25, 0.75] and [-1,0.5]). In Chapter 2, the details of basic emotion theory is introduced as the fundamental knowledge of this thesis. In this thesis, six emotional databases are employed to reach the purpose of multiple databases study in this thesis and the detailed information for each database is described in Chapter 4. Each of the six database has its own emotion labeling strategy, e.g., in category (different choice of emotional status words) or in dimension (different resolution, steps, and number of dimensions). To solve the emotion label problem, the decision of mapping emotion labels to an uniform label is required. The decision of mapping emotion labels for cross-database experiments is discussed in Chapter 6.

Facing the variation among multiple databases (e.g., caused by the recording conditions, the naturalness of emotions, the gender of speakers), the necessity of normalization is investigated to address the second challenge of mismatch between databases. Three normalization methods, Speaker Normalization, Speaker Normalization with Reference, and the Neutral Reference Model, are evaluated and the performance is compared in this thesis. Due to the availability of multiple training databases, the methodology of combining the information gained from each database is investigated as the study for the third challenge of data fusion. Two data fusion techniques, Sequential Forward Database Selection (SDS) and Recognizer output voting error reduction (ROVER) are developed to solve the data fusion problem. In Chapter 7, the methodology and evaluation of normalization and data fusion is presented. The overall improvement brought by the proposed systematic computational structure for cross-database emotion recognition using six emotional databases is presented in Chapter 7. Chapter 8 summarizes the conducted research, draws the conclusion and suggests the future work.

CHAPTER II

EMOTION THEORY

The two most prominent means of emotion characterization have relied on either a discrete lexicon of emotional words (e.g. “sad” and “interest”) or a dimensional scale (e.g. [0.25, 0.75] and [-1, 0.5]) for estimating levels of affect generated by a speech. The discrete lexicon of emotional words convey specific meanings and intent with the most commonly studied being “happy”, “sad”, “angry”, “disgust”, “surprise” and “fear”. The advantage of this framework for analysis is that it helps to establish a controlled vocabulary for creating objective assessments. However, research literature does not always agree on a set lexicon of emotional words with there being hundreds of potential emotional labels that can be assigned to a particular human experience. This lack of agreement has produced conflicting research results and difficulties in comparing one set of research results to another given differences in emotion labeling strategies [13]. An alternative method of modeling affect is based on using a dimensional scale. The dimensional approach assumes that humans are in a perpetual affective state that can be represented along three dimensions:

- Valence- A dimension that represents the degree to which a particular stimulus is viewed as being *pleasant* or *unpleasant* (i.e., *positive* or *negative*). Valence may also be referred to as Evaluation.

- Arousal - A dimension that represents the degree of activity that is generated by a particular stimulus. It relates to the intensity of an emotional experience and may range from *energized* or *alert* to *calm* or *drowsy* (i.e., *active* or *passive*). Arousal may also be referred to as Activation.

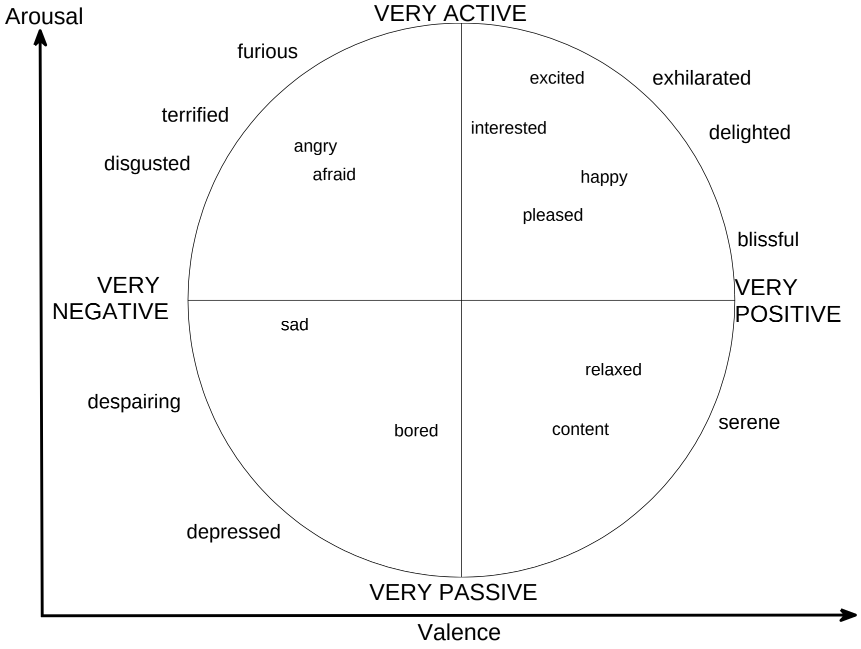


Figure 1: Two-dimensional model of emotion proposed in [14].

- Control - A dimension that relates to the degree of power a person feels over their emotional response to a particular stimulus. It is considered to be helpful in distinguishing between responses with similar arousal and valence. Control may also be referred to as Dominance or Power.

It has been shown that a simpler 2-dimensional model of valence and arousal is able to account for most of the necessary variation in emotional responses [32, 16]. An example of the 2-D valence/arousal (VA) model is shown in Fig. 1 (based on [14])along with the estimated locations of several discrete emotional words based on this scale. However, some researchers (for example [52, 34]) still utilize the full 3-dimensional space to evaluate emotion in speech. The example in Fig. 1 is meant as a reference and is based on the FEELTRACE model developed by Cowie et al. [14]. In this study, we utilize a square representation of the 2-dimensional valence and arousal space where the values are allowed to vary from [-1, +1] in both dimensions as shown in Fig. 2.

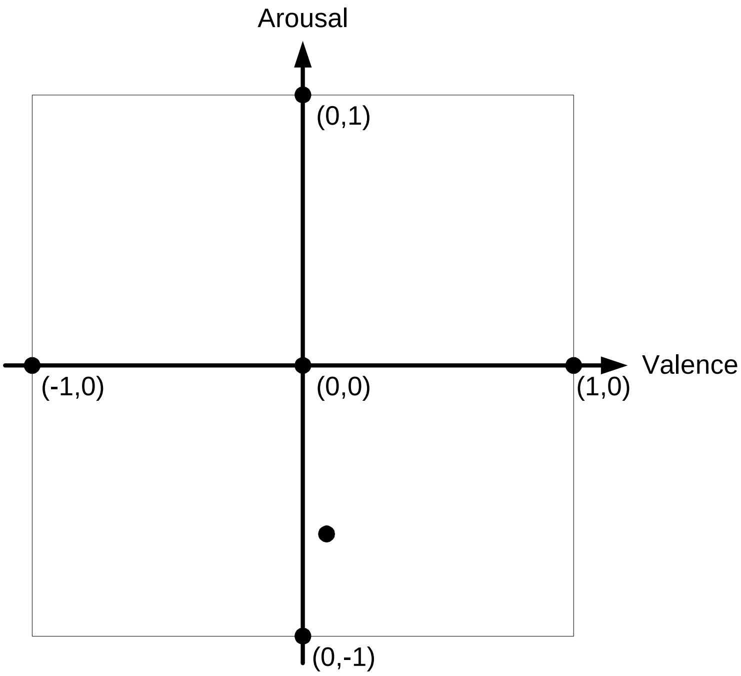


Figure 2: Two-dimensional model of emotion in this dissertation.

Both methods have value in creating labels of emotional intent for analysis. The discrete lexicon has the advantage of providing lexical terms that are commonly used in communication and generally well understood. However, the number of words in an emotional lexicon is great enough to cause some difficulty in finding the most reliably accurate set across all circumstances. The dimensional (VA) approach to emotion categorization has the advantage of creating a numeric representation that can track emotional changes over time and give more quantitative metrics regarding the degree to which an emotional expression is expressed. However, the mapping of the VA space to a specific set of acoustic cues is still a matter of study. Additionally, the concepts of arousal and valence are not generally used in public conversation to describe emotional state. Therefore, the process of collecting data from subjects using the VA scale requires greater care.

CHAPTER III

BACKGROUND

While cross-database training has not been a subject of extensive research, there are several instances of work that offer relevant background for this article. In [78], the emotion anger from three French databases was studied in terms of the averaged values of spectral and time domain features. By comparing the acoustics of anger from three databases, the difference could be observed. This observation motivates the research of this thesis on models and techniques to reduce the difference between databases representing the same emotion status, which is discussed in Chapter 7.1.1. What's more, the similarity between two databases has been studied in [5]. Sequential Forward Floating Search (SFFS) was used to select feature from each database. The selected features from different databases were compared to show the similarity. The similarity between databases is the indicator of possibility of cross-database classification.

In [25], four emotional databases were used for cross-database classifications (train on three databases and test on the other one) in four binary-class tasks in valence, i.e., negative vs. non-negative, positive vs. non-positive, neutral vs. emotional, and positive vs. negative. Support Vector Machine was used as the classifier and 2832 acoustic features were employed without feature selection. The feature set consists of prosodic features (pitch, energy), formant, jitter, shimmer, spectral, and MFCCs. The results showed positive vs. negative gave the best results while neutral vs. emotional was lower. In Chapter 7 when the cross-database training and testing is evaluated, we will start from recognizing neutral vs. emotional in cross-database to investigate the possibility of improving the performance.

In [66], the Euclidean distance from 6552 features between positive and negative in valence/arousal provided objective measurements to data. Selecting the prototypical (with larger Euclidean distance) data samples (a subset data) as the training set helped improving the accuracy rate to reach the highest accuracy rate of 68% for arousal and 54% for valence. In this thesis, since the focus is on the investigation of the stability of acoustic features not the selection of prototypical data, all data samples will be used without selecting only the prototypical subset. The study of [67] studied two methods of using multiple databases for training, i.e., uniting and voting, employing six emotional databases. Uniting is joint training with multiple databases and voting is the late fusion of classifiers trained on single databases. The results show that majority voting performs better. In this thesis, the two ideas of joint training and voting are both considered. Sequential Forward Database Selection (SDS) based on the joint training and Recognizer output voting error reduction (ROVER) as the later fusion are developed and presented in Chapter 7.1.2.

To cope with variances among databases (recording conditions, languages, types of observed emotions, etc.), [65] investigated four normalization conditions, the speaker-level normalization, the database-level normalization, the combination of the two, and no normalization. Using six databases, the cross-database classification was conducted in the way to training on single/combination of multiple databases and testing on another database. Results showed that the speaker-level normalization produced the highest accuracy rate. And the highest accuracy rate (around 66% for valence and 78% for arousal) was observed when testing on the GES database [7]. This thesis will keep using the speaker-level normalization but involving a reference neutral data, i.e., the neutral reference model. The neutral reference model, referred as NRM, [93, 8] is of interest because a methodology was proposed which may improve the generalization ability of features by transforming the acoustic features into *fitness measurements* using a neutral model trained on the reference neutral data. In [8], the pitch-related features was classified to detect neutral vs. emotional samples for four emotional databases in two ways, cross-validation on the combination of all emotional databases (referred as “joint analysis”) and cross-language tests (trained on one language and tested on another). The delivered results emphasized the generalization ability of the neutral reference model in cross-database binary emotion classification (neutral vs. emotional). In the research of this thesis, the neutral reference model is employed as one of the normalization methods. Together with the speaker-level scaling methods (with and without the reference data), three normalization strategies are evaluated and discussed in Chapter 7.1.1.

The studies completed so far usually suffered from the lack of systematic model to improve the performance of the cross-database classification. Most of them focus on reporting experiment results of prototypical sample selection, data agglomeration, and measuring database by direct comparing the statistics of acoustic features. What’s more, the acoustic features used so far are limited in the prosodic, spectral, and voice quality related features (e.g., jitter, shimmer), which have been well studied in emotion recognition. Therefore, the research of this thesis systematically studies the elements contributing to improving the performance of cross-database training and testing, from the emotion labeling strategy to data fusion techniques. The glottal and Teager energy related features are introduced into the feature sets, which makes up the lack of study of these two features in this field.

CHAPTER IV

EMOTIONAL DATABASES

Six emotional speech databases have been collected to be used in this thesis. The basic information is summarized in Table 1. More details and the work on each database is provided in this chapter.

Emotional Prosody Speech Transcripts (EPST) [49] (Publicly available):

- English; Acted; 7 speakers (4 female; 3 male); 275-422 utterance/ speaker (Numbers and dates)
- 15 Emotions (neutral, disgust, panic, anxiety, hot anger, cold anger, despair, sadness, elation, happy, interest, boredom, shame, pride, and contempt)
- Perceptual ratings done in our lab using 20 subjects provide discrete Valence/ Arousal data

An acoustic feature study conducted by Bitouk evaluated the combination of class-level/ utterance-level features and prosodic/ spectral features in emotion recognition using EPST and another acted database in German [4]. The results delivered showed that class-level MFCC statistics outperformed both prosodic and utterance-level spectral features in speaker independent emotion recognition. Another study [43] compared the recognition of negative emotional states using the acted speech in

Table 1: The basic information of six emotional databases.

	EPST	**EMA**	**GES**	**SEMAINE**	**VAM**	**UM**
Language	English	English	German	English	German	English
Speakers	7	3	10	7	47	30
Emotions	15c	4c	7c	4-dim	2-dim	7c
Sampling freq.	22.05kHz	16kHz	48kHz	48kHz	16kHz	22.05kHz

the EPST (using “hot anger” for negative and “neutral” for non-negative) database and real-world speech collected by attempts to provoke negative user reactions in several tasks in a smart home application. The emotional speech was broadly categorized into “negative” and “non-negative” based on subjective tests. While the classification results indicated a higher accuracy in distinguishing the acted emotional speech (i.e., EPST database), it was also noted that the real-world speech contained less definitive boundaries in emotional expression and that there was disagreement among listeners regarding the appropriate emotional labels for the real-world speech. Sethu, et al. [70] used the EPST to examine the impact of speaker-specific feature warping algorithms for improving accuracy of emotion recognition across multiple speakers and emotional classes within EPST. Wu, et al.[90] used the EPST database for testing feature modification strategies in speaker recognition that can compensate for the acoustical variations presented by emotional speech. Hirschberg et al. [38] investigated the completeness of the emotional labels for the EPST database by assigning multiple sets of emotional labels to a single utterance in the EPST database as opposed to each utterance having one emotional descriptor.

University of Memphis (UM) [31] (Not currently available publicly)

- English; Induced emotion through computer interaction in AutoTutor system dialog; 30 speakers (19 female; 11 male); 16-344 utterance / emotional expression

- 7 basic emotions (boredom, confusion, flow, frustration, delight, surprise, neutral)

- The speakers provide their own word labeling for the emotions they felt during recording (Fig. 1a)

This database was made available to our lab by the University of Memphis. The UM database (referred to as UM in this thesis) contains a record of students using a computer guided learning system known as Auto Tutor [31]. Work in [21] performed emotion recognition on subsets of 4 and 5 of the emotion categories using conversational cues, gross body language, and facial features. The recognition rate in these experiments was 0.47 and 0.51, respectively. Currently, no acoustic analysis has been reported.

Vera-Am-Mittag (VAM) [33] (Publicly available)

- German; Speech from TV talk shows; 47 speakers (36 females; 11 males); 947 utterances total
- Perceived evaluation range is [-1,1] for all three dimensions VAD
- Database is split into VAM I (rated by 17 subjects) and VAMII (rated by 6 subjects); perceptual labeling provides VA ratings

Work in [88, 89] proposed a long-term spectro-temporal (ST) representation and modulation spectral features (MSF) in research designed to correlate acoustic features with the valence, arousal, and dominance emotion labels in the database. Their results showed the highest correlations for the arousal and dominance labels based on the acoustic measures. Work in [34] utilized various machine learning techniques based on support vector regression, fuzzy k-Nearest Neighbor estimators, and Rule-based Fuzzy logic estimators to estimate the valence, arousal, and dominance values of the emotional labeling based on prosody and other spectral based features.

German Emotional Speech (GES) [7] (Publicly Available)

- German; Acted speech using set of sentences; 10 speakers (5 female; 5male); 535 total utterances
- 7 emotions (happy, angry, sad, fearful, bored, disgust, and neutral)
- Perceptual labeling by 20 subjects for emotional authenticity

Work in [63] used MFCC features extracted from the voiced sections of the speech to achieve an average recognition rate of 60.57%. Work in [39] combined prosodic, formant, and MFCC features to achieve recognition rates of 83.17% on this database. Other work on this database utilized features base on psychoacoustics [91] and empirical model decomposition [48] to achieve recognition rates around 83%. Wu proposed modulation spectral features (MSFs) combined with prosody to achieve an average recognition rate of 91.3% [89].

SEMAINE [54] (Publicly available)

- English; Acted and Induced speech where a speaker acts on a specific emotion in order to induce an emotion from a subject; 7 speakers (4 female, 3 male); 40 sessions (each 5 minutes long)
- Rated by 1-6 subjects for continuous ratings of valence and arousal on a range from [-1, 1].

As of the time of the writing of this proposal, no specific publications could be found that directly reference the acoustic analysis of the SEMAINE database in possession.

Electromagnetic Articulography (EMA) [46] (Publicly available)

- English; Acted speech based on set of sentences; 3 speakers (2 females; 1 male); 14 sentences for the male and 10 sentences per female
- 4 emotions (angry, happy, sad, and neutral)
- Perceptual labeling using emotional words by 3 subjects to validate emotion expressions.

The EMA database was created by the SAIL group at USC. While the data is publicly available, it's told by USC that this proposed work is the first to request access to it outside of their lab. Some of the previous studies on the data [34, 8] have focused on prosodic features as the primary source for distinguishing the emotion types present in the data. Work in [34] achieved an overall recognition rate of up to 83.5% on the data while Busso [8] used EMA in conjunction with other databases to build a recognition model that separates neutral and non-neutral emotion (i.e. a binary classification). Kim [42] used MFCCs and measures of the articulatory space in an isometric feature mapping (Isomap) experiment to compare emotion recognition on a reduced feature space.

CHAPTER V

ACOUSTIC FEATURES

The literature shows that previous work used prosodic features such as pitch/energy as the sole acoustic feature for emotion recognition [8]. In addition to pitch, this study also extracts features based on Teager energy and the glottal waveform as well as the Mel-cepstral coefficients. Research has shown that the glottal waveform dynamics and Teager energy features can play an important role in voice characterization [77, 72, 73, 15, 92, 57, 58, 84]. In this chapter, the extraction and evaluation of glottal and Teager Energy based features are presented and the emotion distinguishing ability is investigated using the databases introduced in Chapter 4.

5.1 Glottal Waveform Parameters

5.1.1 Extraction

The glottal waveform is considered as a representation of the volume velocity of airflow through the vocal folds during voiced speech. In this section (Section 5.1.1), the extraction of the glottal features is introduced. And in Sections 5.1.2-5.1.4, the performance of emotion recognition using the glottal features is evaluated on different databases. Fig. 3 shows an example of the glottal waveform (Fig. 3(b)) and glottal waveform derivative estimate (Fig. 3(c)) for one cycle of voiced speech (Fig. 3(a)). One total cycle (TC) consists of an open phase (O) and closed phase (C). The open phase is divided into an opening phase (OP) (i.e., abduction) and closing phase (CP) (i.e., adduction). The opening phase may sometimes be further divided into the length of the primary opening (T_{o1}, i.e. OP) and a secondary opening (T_{o2}). The distinction between T_{o1} and T_{o2} is marked by a an increase in the slope during the opening phase (i.e. smaller slope for (T_{o1}, larger slope for T_{o2}).

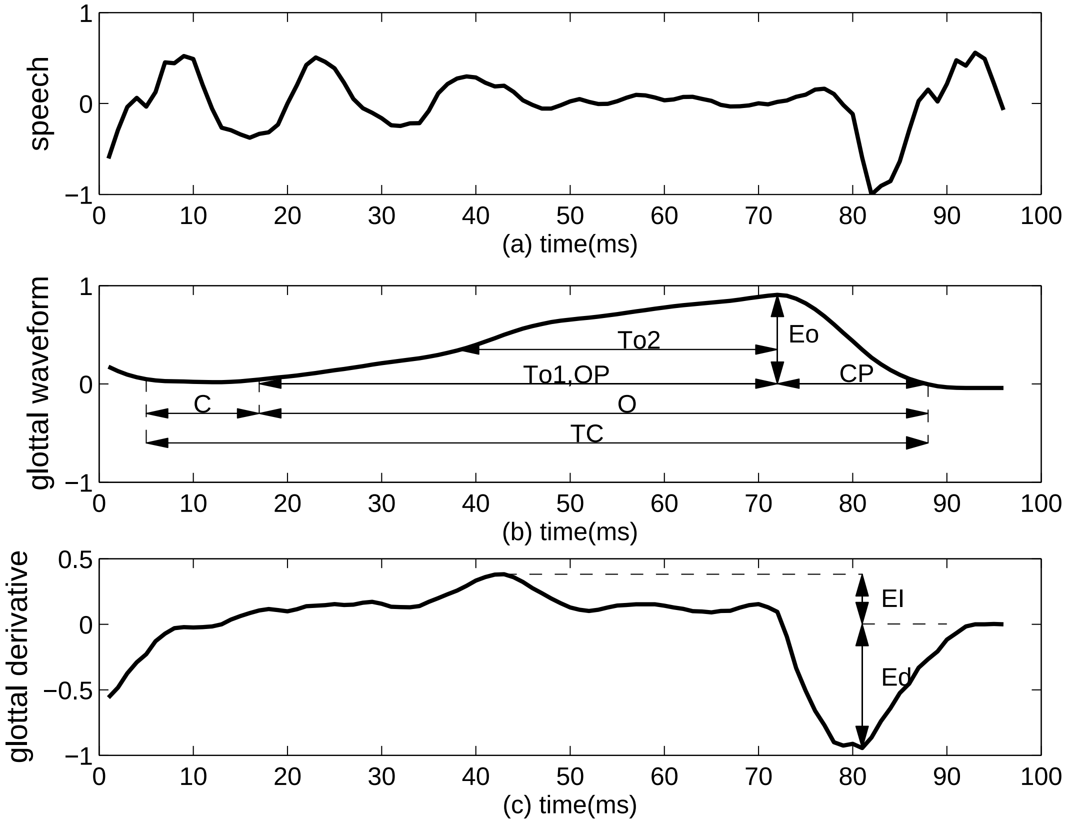


Figure 3: Example of the time-based parameters extracted from the glottal waveform during a single speech cycle. (a) One-pitch cycle of speech (b) Glottal waveform estimate (c) Glottal waveform derivative.

Table 2: Time-based glottal features extracted from the glottal waveform estimations (see Fig. 3).

Abbr.		Equation
OQ1	(open quotient, from primary glottal opening)	$OQ1 = \frac{T_{01}+CP}{TC}$
OQ2	(open quotient, from secondary glottal opening)	$OQ2 = \frac{T_{02}+CP}{TC}$
AQ	(amplitude quotient)	$AQ = \frac{Eo}{Ed}$
NAQ	(normalized amplitude quotient)	$NAQ = \frac{AQ}{TC}$
ClQ	(closing quotient)	$ClQ = \frac{CP}{TC}$
OQa	(open quotient based on Liljencrants-Fant model)	$OQa = \frac{E_o}{TC}(\frac{\pi}{2EI} + \frac{1}{E_d})$
SQ1	(speed quotient, from primary glottal opening)	$SQ_1 = \frac{T_{01}}{TC}$
SQ2	(speech quotient, from secondary glottal opening)	$SQ_2 = \frac{T_{02}}{TC}$
QOQ	(quasi-open quotient)	[44]

The glottal waveform may be parameterized using time-based features to quantify the shaping of the signal and spectral features. The extraction of the glottal features for each speech utterance was processed in four steps: (1) each utterance was divided into frames 4 pitch periods long (2) glottal closure instants (GCI's) were obtained using the DYPSA algorithm [61] on each frame (3) glottal waveform estimates were obtained for each frame using the Rank-Based Glottal Quality Assessment (RBGQA) [59], which iterates around approximate locations of GCI's to find the optimal analysis window position for deconvolution via the covariance method of linear predictive analysis (LPA) (an LPA order of 16 was used) (4) for each frame, the 11 glottal features were extracted using version 0.3.1 of the APARAT [1] program. Table 2 lists nine of the time-based features extracted from the glottal waveform based on the parameters shown in Fig. 3. In Fig. 3, E_o represents the amplitude of maximum glottal opening of glottal waveform, E_d is the absolute amplitude of minimum point in glottal derivative, and El is peak amplitude of glottal derivative. In the time-based features in Table 2, the Quasi-Open Quotient (QOQ) was calculated in APARAT using a variation on estimation of the Open Quotient (OQ) based on work in [44].

Spectral-based features calculated by APARAT on the glottal waveform included: *DH12* (the difference in the first and second glottal formants, in dB [83]) and *HRF* (the Harmonic Richness Factor, in dB [11]). These parameters were calculated as shown in Eq. 1 and Eq. 2, where *X(F0)* is the spectral amplitude at the fundamental frequency ($F0$), *X(2*F0)* is the amplitude at two times the fundamental frequency, and *X(f_i)* is the spectral amplitude in the i^{th} harmonics (f_1 is the fundamental frequency $F0$).

$$DH12 = 10 \log \frac{|X(F0)|^2}{|X(2 * F0)|^2}, \tag{1}$$

$$HRF = 10 \log \frac{\sum_{i>1} |X(f_i)|^2}{|X(f_1)|^2}. \tag{2}$$

5.1.2 Emotion categories with similar prosody

Fundamentally, automated emotion detection is the attempt to quantify an abstract interpretation into objectively measured components of recorded human interaction. A review of the study of emotion for human computer interaction in [14] shows that prosodics (e.g., pitch, energy, speaking rate, etc.) are the most common form of speech analysis in literature. Additionally, [14] shows support that the general prosodic tendencies in distinguishing between different emotion categories can be extremely qualitative, subtle, and likely speaker dependent. For example, a person who is happy may tend to raise their prosody (e.g., increased pitch, energy, speaking rate, etc.) from their neutral state but may also show similar tendencies when expressing anger or panic. Work on the use of glottal features (i.e., features extracted from the estimated signal representing the air-flow through the vocal folds) in classifying emotion [15, 26, 58] has shown that these features can provide valuable insight into distinguishing different types of emotional expression. The purpose of this section is to target speaker dependent expressions of emotional pairs that share statistically similar prosodic information and investigate a set of glottal features for their ability to find measurable differences in these expressions.

5.1.2.1 DATABASE

The speech used for this study was provided by the Emotional Prosody Speech and Transcripts (EPST)[49] database. The EPST database contains recordings of emotional expression on semantically neutral speech from 7 professional actors (4 females and 3 males) who are native speakers of standard American English. Each actor reads short (4-syllables) dates and numbers in 15 different emotional categories [2] (*"neutral", "disgust", "panic", "anxiety", "hot anger", "cold anger", "despair", "sadness", "elation", "happy", "interest","boredom", "shame", "pride", "contempt"*. The speech was recorded at a sampling frequency of 22.05 KHz with 2-channel interleaved 16-bit PCM format. The duration of each utterance varied from 1sec to 2sec. While it is in no way assumed that acted speech provides a complete picture of authentic emotion, the value of this information is that the actors adjusted their speech patterns to fit their *perception* of different emotions. These voice changes are objectively evaluated at this time without the need to explicitly determine the degree to which each utterance represents the intended emotion to an observer.

5.1.2.2 OBJECTIVE MEASURES

Pitch represents a high-level view of the motion of the vocal folds as it provides information on the rate at which air from the lungs is allowed into the vocal tract. The glottal waveform, on the other hand, provides a representation of the volume velocity of airflow through the vocal folds during voiced speech. While pitch information provides the rate, glottal features ideally provide a more detailed look at the phonatory process. This section used prosodic features of speech based on the mean pitch and energy. Pitch was obtained using the RAPT pitch estimation algorithm in VOICEBOX[6] using a 10 ms frame rate. Energy was calculated as the squared sum of the values within each frame across the voiced sections in each utterance as indicated by the pitch information. The glottal waveform provides a representation of the shaping of the volume velocity of airflow *through* the vocal folds during voiced speech. The glottal parameters used in this study is ClQ, NAQ, OQ, OQa, SQ1, DH12, and HRF from Table 2. The extraction of the glottal features for each speech utterance was processed in four steps as described in Section 5.1.1. All features were quantified using only 1^{st} order statistics (i.e, the mean) across all frames of an utterance. The use of higher order statistics was excluded from the study at this time as the goal was to study the basic discriminatory power of the features themselves and not to build a complex model for general classification.

5.1.2.3 METHODOLOGY

Because of the high number of discrete emotion categories, most research on emotion has focused on smaller subsets of emotion (such as happy, anger, fear, etc.). However, a pairwise comparison is conducted on 14 distinct emotional categories in an effort to identify which emotional pairs statistically share the same prosody information. Four actors (2 females (F1, F2) and 2 males (M1, M2)) were chosen from the EPST database based on the speakers with the highest total number of observations (i.e., utterances). Pitch, energy, and glottal features were extracted on a speaker-dependent basis as described earlier. The feature extraction procedure resulted in 13644 frames with a 9 dimensional feature vector (i.e., mean pitch, energy, and glottal features). The average number of frames per speaker was 3411 with an average of 227 frames per emotion. There were approximately 25 utterances per emotion for each speaker on average with no emotion allowed to have less than 20 utterances for inclusion in the study (this resulted in the exclusion of *neutral* utterances).

The purpose of this study was to evaluate the discrimination power of glottal features on emotional categories that share statistically similar prosodics. Therefore, the mean pitch values of each of the pairwise groups of emotions (91 pairs total) was subjected to a Kruskal-Wallis (KW) significance test. Pairwise groups that showed no statistical difference in their pitch distributions at a significance level of $p < 0.05$ were targeted for further analysis. The discrimination of these emotional pairs was then evaluated by finding the error rates from using each of the 9 single features as classifiers and the error rates from using a Sequential Feature Selection (SFS) algorithm for selecting any combination of features for classification. SFS starts with an empty feature set and sequentially adds features that have not yet been selected. Every feature combination set is evaluated 10-fold cross validation until there is no improvement in the criterion function. For this study, the criterion was set to the error rate from a quadratic discriminant computed as the number of incorrect classifications divided by the total number of observations. The SFS algorithm added features in an effort to reduce the error rate as much as possible.

5.1.2.4 RESULTS

Table 3 shows the emotional pairs that showed no significant difference ($p < 0.05$) in their pitch distributions after the Kruskal-Wallis test on a speaker-dependent basis. Intuitively, many of the emotional pairs reflect an expected similarity in prosodic tendencies. For example, many of the pairs reflect a confusion between two high (such as *elation* and *hot anger* for speaker F1) or low arousal states (such as *pride* and *sadness* for speaker M1). Additionally, there is very little overlap in the confused emotional states across actors, which reflects the highly speaker dependent nature of emotional interpretation and expression. For all of the emotion pairs listed in Table 3, at least one glottal feature showed a statistically significant difference and 19 out of 30 pairs had 4 or more of the 7 glottal features show statistical significance. Further evaluation was conducted by finding the error rate (ER) for each of the individual features in discriminating the emotional pairs using 10-Fold cross validation. The error rate was computed as the number of incorrect classifications divided by the total number of observations (i.e., utterances). The number of observations was approximately equal for each of the emotional pairs, making the chance error rate roughly equal to 50%. Due to the relatively small number of observations, the 10-Fold cross validation was repeated 50 times, where each iteration randomized the data in a way ensure that enough variations on the combinations of data observations for training and testing were used. Table 3 shows the mean of the error rate computed across all 50 runs of the 10-fold cross-validation. Only the best performing glottal feature is shown in the table. A lower error rate is achieved by a glottal feature in 24 out of the 30 pairs. Of the the 6 pairs where a glottal feature is not the best feature, energy has the lowest error rate in 4 pairs and pitch has the lowest error rate in 2 pairs. That pitch could have the lowest error rate (though slight) even though there was no statistically significant difference highlights the reasons for evaluating the classification performance of each feature.

Table 4 shows the resulting mean error rates from the SFS procedure along with the percentage change from the lowest error rate achieved for a single feature. The SFS was run on each subset of the 9 features 50 times using 10-fold cross validation to ensure enough randomization of training and testing combinations in the data. The '%Change' column indicates the percentage change in the error rate that resulted from using multiple features over the single feature with the lowest error rate shown in Table 3. Table 4 shows that pitch and energy continue to play an important role in emotional classification even when the emotional pairs are selected based on non-significant differences in pitch distributions. In only one instance (M1, $contempt, sadness$) was neither pitch nor energy selected for the classifier. For the females, 'hrf' feature was among the most prominent glottal features selected while for the males the 'oqa', 'clq', and 'naq' were the most prominent glottal features. The 'dh12' feature was a prominent feature across all speakers while the 'sq' feature showed little impact on discrimination and was rarely chosen. While the discrimination for most emotional

Table 3: Minimum error rate (ER) for emotional pairs using single features.

Actor	Emotional Pairs	Pitch	Energy	(Glottal, ER)
F1	*pride, anxiety*	0.46	0.24	(hrf,0.16)
	elation, hotanger	0.48	0.49	(hrf,0.21)
	boredom, coldanger	0.52	0.33	(oq,0.21)
	contempt, coldanger	0.43	0.47	(dh12,0.29)
	happy, sadness	0.36	0.38	(oqa,0.40)
	interest, sadness	0.20	0.29	(hrf,0.23)
	pride, interest	0.51	0.36	(hrf,0.13)
F2	*coldanger, disgust*	0.32	0.15	(hrf,0.30)
	sadness, disgust	0.32	0.23	(hrf,0.32)
	despair, panic	0.38	0.20	(oq,0.05)
	happy, panic	0.41	0.08	(hrf,0.06)
	despair, hotanger	0.40	0.34	(oq,0.21)
	elation, hotanger	0.32	0.57	(dh12,0.16)
	happy, hotanger	0.35	0.32	(oq,0.23)
	sadness, coldanger	0.36	0.39	(hrf,0.30)
	elation, despair	0.55	0.31	(oq,0.08)
	contempt, sadness	0.40	0.41	(oq,0.30)
	happy, elation	0.45	0.26	(hrf,0.08)
	contempt, boredom	0.40	0.36	(hrf,0.18)
M1	*shame, anxiety*	0.38	0.51	(oqa,0.32)
	elation, coldanger	0.46	0.38	(clq,0.30)
	interest, coldanger	0.46	0.40	(naq,0.31)
	pride, sadness	0.70	0.36	(dh12,0.30)
	contempt, sadness	0.39	0.60	(naq,0.25)
M2	*shame, disgust*	0.62	0.16	(oqa,0.34)
	happy, panic	0.68	0.47	(naq,0.28)
	despair, anxiety	0.40	0.43	(oqa,0.22)
	contempt, anxiety	0.51	0.36	(clq,0.37)
	interest, coldanger	0.45	0.46	(oqa,0.07)
	contempt, despair	0.40	0.42	(oqa,0.17)

pairs was greatly improved through the multiple feature classifier, the discrimination for speaker F2 with the emotional pairs $(sadness, disgust)$, $(happy, panic)$ and $(happy, elation)$ could not be improved over the performance of the single features of energy and harmonic richness factor, respectively.

Table 4: Minimum error rate (ER) for emotional pairs using SFS and the top five selected features

Actor	Emotional Pair	ER	%Change	(Feature, Selection Percentage(%))				
F1	$pride, anxiety$	0.10	-40%	(pitch,94),	(hrf,92),	(sq,6),	(dh12,4)	
	$elation, hotanger$	0.19	-7%	(hrf,100),	(sq,14),	(eng,10),	(naq,8),	(oq,4)
	$boredom, coldanger$	0.20	-3%	(oq,100),	(eng,20),	(dh12,6),	(hrf,6)	
	$contempt, coldanger$	0.17	-36%	(eng,96),	(dh12,92),	(hrf,34),	(oq,8),	(sq,8)
	$happy, sadness$	0.16	-54%	(pitch,98),	(naq,60),	(oqa,60),	(clq,42),	(hrf,38)
	$interest, sadness$	0.08	-59%	(pitch,96),	(oq,74),	(hrf,72),	(dh12,28),	(clq,16)
	$pride, interest$	0.09	-31%	(hrf,98),	(pitch,76),	(clq,40),	(oqa,10),	(dh12,4)
F2	$coldanger, disgust$	0.03	-80%	(pitch,100),	(eng,100),	(oq,44),	(naq,30),	(hrf,14)
	$sadness, disgust$	0.23	0%	(eng,100),	(pitch,6),	(clq,4),	(oq,4),	(oqa,4)
	$despair, panic$	0.04	-16%	(oq,70),	(eng,26),	(hrf,26),	(dh12,22),	(pitch,16)
	$happy, panic$	0.06	0%	(hrf,76),	(dh12,18),	(eng,4),	(oqa,2),	(sq,2)
	$despair, hotanger$	0.17	-20%	(hrf,74),	(oqa,36),	(oq,32),	(eng,24),	(pitch,22)
	$elation, hotanger$	0.09	-41%	(pitch,100),	(dh12,100),	(hrf,2),	(sq,2)	
	$happy, hotanger$	0.18	-23%	(oq,92),	(naq,60),	(oqa,48),	(eng,32),	(pitch,26)
	$sadness, coldanger$	0.13	-58%	(hrf,96),	(oq,90),	(oqa,70),	(naq,70),	(eng,46)
	$elation, despair$	0.04	-44%	(pitch,64),	(clq,60),	(dh12,30),	(eng,24),	(hrf,6)
	$contempt, sadness$	0.12	-60%	(oq,100),	(oq,76),	(eng,68),	(hrf,58),	(pitch,44)
	$happy, elation$	0.08	0%	(hrf,98),	(naq,98),	(pitch,6),	(dh12,2)	
	$contempt, boredom$	0.18	-3%	(hrf,84),	(dh12,18),	(sq,10),	(oqa,8),	(eng,4)
M1	$shame, anxiety$	0.28	-14%	(oqa,84),	(pitch,36),	(clq,32),	(naq24),	(hrf,18)
	$elation, coldanger$	0.23	-25%	(clq,100),	(pitch,90),	(eng,40),	(sq,26),	(hrf,14)
	$interest, coldanger$	0.24	-23%	(naq,96),	(eng,76),	(oq,36),	(dh12,24),	(oqa,20)
	$pride, sadness$	0.25	-16%	(dh12,90),	(eng,36),	(pitch,28),	(hrf,28),	(naq,24)
	$contempt, sadness$	0.11	-55%	(naq,98),	(clq,92),	(oqa,76),	(oq,62),	(dh12,34)
M2	$shame, disgust$	0.13	-18%	(eng,100),	(oq,42),	(clq,36),	(naq,8),	(oqa,4)
	$happy, panic$	0.22	-20%	(naq,100),	(clq,56),	(pitch,26),	(oqa,16),	(dh12,14)
	$despair, anxiety$	0.09	-57%	(oqa,100),	(pitch,90),	(dh12,78),	(clq,28),	(naq,18)
	$contempt, anxiety$	0.27	-25%	(eng,68),	(oqa,44),	(clq,32),	(dh12,22),	(pitch,14)
	$interest, coldanger$	0.06	-9%	(oqa,54),	(hrf,50),	(pitch,46)		
	$contempt, despair$	0.15	-13%	(oqa,98),	(oq,54),	(hrf,32),	(eng,4),	(dh12,4)

The results highlight a few critical points about emotion in speech. The first confirms that there are emotional pairs that carry subtle differences that can be difficult to express and interpret based on prosody alone. Additionally, while the types of emotional ambiguities are largely speaker dependent, there are subtleties that can exploited from features of the glottal flow to help resolve some of them. The presented work examined the ambiguities present in an actors' *intended* emotional expressions.

5.1.3 Authentic emotion: Auto Tutor

The interest of researchers in the field of human-computer interaction (HCI) has been developed to build more effective, user-friendly, and intelligent applications [51, 21, 28, 13, 20, 19, 54]. Computer tutoring system with user emotion detection is one of the focuses. Researchers extracted multiple cues to recognize users' emotion. One of the computer tutoring systems is ITSPOKE[51]. The ITSPOKE group collected features including the acoustic-prosodic (pitch related, energy related, duration, speaking rate, pause-duration, and number of internal silence) and the lexical (i.e., manually transcribed or recognized speech) features to predict three emotion states (negative, neutral, and positive) of the users. Their result, in general, showed the lexical features yielded higher predictive utility than acoustic-prosodic features [51].

Another computer tutoring system, Auto Tutor, developed by the University of Memphis involved multiple channels [21] to detect the learners' emotion, such as facial expression, body gesture [20], and speech [19]. In the speech channel, the features they used were the conversational cues, which consist of five aspects of information: temporal, response, answer quality, tutor directness, and tutor feedback (more details in Section 5.1.3.3). However, no work has been reported involving the features of acoustics from the speech channel. This motivated the study to investigate the performance of acoustic cues from speech in learner's emotion detection of Auto Tutor. We use the same speech data and methodology as that in the work using conversational cues for emotion detection of Auto Tutor [19] and compare the performance of acoustic cues working in emotion detection with that of the conversational cues. In addition, Sequential Floating Forward Selection (SFFS) is applied as extra study on the acoustics for feature selection and the selected features was evaluated and examined to provide more detailed analysis of the acoustic features in emotion detection of Auto Tutor HCI.

Table 5: Number of samples of UM database

Emotion	neutral	boredom	confusion	flow	frustration	delight	surprise
No.samples	277	268	319	348	204	78	17

5.1.3.1 Data

The speech data used for this study was provided by the Auto Tutor system from the University of Memphis [31] (UM data). The UM data contains the recordings of 30 users' (15 females and 15 males) dialog when they were learning with Auto Tutor. UM data covers seven emotions including 'neutral', 'boredom', 'confusion', 'flow', 'frustration', 'delight', and 'surprise'. 'Neutral' was defined as no emotion or feeling. 'Boredom' was defined as being weary or restless through lack of interest. 'Confusion' was defined as a noticeable lack of understanding. 'Flow' was a state of interest resulting from involvement in an activity. 'Frustration' was defined as dissatisfaction or annoyance. 'Delight' was a degree of satisfaction. And 'surprise' was wonder or amazement, especially from the unexpected [31]. These emotion categories were labeled by the user himself/herself (i.e., self-evaluation) through reviewing the recorded video of their learning interaction procedure after the learning session. The number of utterances with emotion labels is shown in Table 5.

To make a balanced classification and equivalent comparison, this study excluded emotions 'delight' and 'surprise' in the analysis with the following reasons: 1) the number of samples in emotion categories 'delight' and 'surprise' is considerably smaller than that of other emotions, which could cause a bias in classification evaluation; 2) the comparable work in [19] excluded the two emotions for the same reason. Therefore, the dataset in this analysis consists of 1416 samples from 30 speakers in five emotion categories: 'neutral', 'boredom', 'confusion', 'flow', and 'frustration'.

5.1.3.2 Features

The acoustic cues used in this study consist of two features sets, glottal waveform features and a feature set including the prosodic, spectral, and other voice related features (e.g., probability of voicing, jitter, and shimmer). The glottal features are the parameters shown in Section 5.1.1. Another set of energy related, spectral related, and other voice related features were extracted using the openSMILE toolkit [23]. The low-level descriptors (LLD) are listed in Table 6. Up to 39 functionals were applied to LLD to generate 4368 features [68].

Table 6: openSMILE low-level descriptors (LLD)[68]

Energy Related	Spectral Related	Other Voice Related
Sum of auditory spectrum	RASTA-style spectrum	F0
Sum of RASTA-style filtered spectrum	MFCC 1-12	Probability of voicing
RMS energy	Spectral energy	Jitter
Zero-crossing rate	Spectral roll off point	Shimmer
	Spectral statistics	

Based on the above, a 4445-dimensional acoustic feature set was created for each utterance sample for all five emotion categories (excluding delight and surprise) as listed in Table 5.

5.1.3.3 Methodology

In the comparison work [19], the emotion states of speech data (UM data) were labeled by four evaluators: the user himself/herself (i.e., self-evaluation), peer (other user), and two trained judges. Based on the evaluator, seven sets of emotion labels were generated: self-evaluation (the user evaluated himself/herself), peer-evaluation (another user's evaluation), two trained judges separately, agreement shared between the two trained judges, agreement shared between more than two evaluators (including self, peer, and the two trained judges), and agreement shared between more than three evaluators (including self, peer, and the two trained judges). Except for the self-evaluation label, other label sets are not available in this study, so the comparison will focus on the experiments using self-evaluation as the emotion label. The features used in the comparison work [19] are 17 conversational cues from five aspects: temporal (e.g., duration, number of topics, number of turns, etc.), response (number of words, number of chars, etc.), answer quality (similarity to an expectation, the change in the similarity, etc.), tutor directness (hint, prompt, correction, etc.), and tutor feedback (positive, neutral, negative, etc.). The 17 features was reduced in dimension using Principle Component Analysis (PCA). The reduced feature set were evaluated by a list of classifiers in the WEKA environment [35], among which Adaboost.M1 is the classifier yielded the highest accuracy rate for the self-evaluation label data. The classification was conducted in two sets of experiments: multiple-emotion classes and binary-emotion classes. The multiple-emotion classification including 5-classes (all five emotions) and 4-classes ('neutral' excluded); the binary classification is between two emotions: neutral and one from the other four emotions. For each classification task, the number of samples in each class was forced to be equal by randomly selecting N samples from every class (N is the smallest class's size). The classification for each experiment was repeated 10 times (trials) using the balanced samples. The averaged result over 10 trails was the representation of the classification experiment.

For the comparison purpose, this study adopted the methodology in [19]. Dimension reduction using PCA was conducted, followed by classification using Adaboost.M1. However, due to the large difference in the dimension of features (17 conversational cues vs. 4445 acoustic cues), additional feature selection (Sequential Floating Forward Selection - SFFS) was applied to the 4445-dimension acoustic feature set. The selected features were evaluated by classification tasks using Adaboost.M1 classifier as well. AdaBoost.M1 is a boosting algorithm. It improves the 'weak' learning algorithm by repeatedly applying it to different distributions or weighings of training samples and eventually forming a 'stronger' learning algorithm [29]. This classifier was chosen in this additional study because it was one of the classifiers yielded best classification result in D'Mello's work [19], and also it was evaluated in prior works to show more robust performance compared with other learning methods [51]. The software WEKA [35] was used for SFFS and Adaboost.M1 classification and MATLAB executed PCA to create new feature sets with the 95% variation represented.

It should be noticed that acoustic features are speaker-dependent because they capture the characteristics of speakers (e.g., gender and culture). To eliminate the acoustic difference from factors other than emotion, speaker normalization was applied to the data first. The equation for speaker normalization is shown in Equation (15):

$$\hat{f_{i,j}} = \frac{f_{i,j} - mean(f_{i,j})}{std(f_{i,j})}, \tag{3}$$

where $f_{i,j}$ is the i_{th} feature descriptor for speaker j across the samples of all emotions and std refers to the standard deviation.

5.1.3.4 Result

The classification result is shown in Table 7. The classification results using PCA (95% variation represented) are all approximately equal to the baseline (chance) as shown in the row of 'PCA95%variation'. While all classification accuracy rates using SFFS are above chance (i.e, the baseline). The higher accuracy rate using SFFS than PCA implies that feature selection is benefit for large dimension feature set. Therefore, the focus of result expression will be on comparing the results using SFFS (instead of PCA) with D'Mello's work [19]. From the multiple-emotion tasks shown in columns '5class' and '4class' of Table 7, the 5-class and 4-class ARs are lower than D'Mello's results using conversational cues. However, acoustic cues yielded higher (boredom and flow) or similar (confusion and frustration) accuracy rate in all binary-emotion classifications between neutral and emotional states (the rightmost four columns in Table 7). The statistics significant test was used to help the comparison. In Table 7, the one-sided one-sample t test was used to test the null hypothesis that the accuracy rates variable using acoustic features comes from the population with the mean of the comparison accuracy rate reported in [19] against the alternative hypothesis that the mean is greater than the comparison work. T test was chosen because the Jarque-Bera test results did not reject any of the null hypothesis that the accuracy rates fit the normal distribution. And one-sided (i.e., right) was chosen because, if the null hypothesis of the t test was rejected, instead of the general interpretation of not equaling to, the author was more interested in the detail about whether the mean is greater or smaller than the comparison work. The significant test showed that using acoustic features, the classification between neutral vs. boredom, neutral vs. confusion, and neutral vs. flow produced significantly higher accuracy rates than the comparison work.

Table 7: Comparison: Classification accuracy rate (AR) in %, (D'Mello: the comparison work [19], P: the presented work)

AR	5class	4class	neutral vs. boredom	neutral vs. confusion	neutral vs. flow	neutral vs. frustration
baseline (chance)	20.0	25.0	50.0	50.0	50.0	50.0
SFFS	24.9	31.6	69.1	61.8	66.6	64.6
PCA95%variation	20.1	24.9	47.3	54.2	50.5	52.6
D'Mello [19]	29.5	35.1	61.3	58.9	52.9	64.1

Following the analysis in [19], the F-measure scores in Table 8 were calculated by dividing the doubled number of correctly classified samples (i.e., true positive) belonging to one class by the total number of samples of false positive, false negative, and doubled true positive from the confusion matrix [87]. For multiple-emotion tasks (5-class and 4-class), Acoustic cues outperform conversational cues for flow, while conversational cues win in confusion and frustration for both 5/4-class tasks(significantly different tested by t test). Neutral and boredom are distinguished from other emotions with fairly similar scores using conversational cues and acoustics (26%, 25%

for neutral, and 32%, 32% for boredom) for 5-class. In the 4-class case, acoustics yield higher score in boredom (significantly different tested by t test). Considering 4 and 5 classes together, flow exhibits the most acoustic separation, while the score using conversational cues is the lowest. This could be explained that among the five emotions (including neutral), flow (i.e., interest, engaged) is mostly an arousal related emotion [13]. Previous work showed that arousal degree (involvement) were captured more using acoustic cues comparing using the degree valence (pleasant of emotion) [38]. Therefore, the emotion flow, describing the involvement, has the largest separation using acoustic cues and is captured by acoustics better than the conversational cues in this study. The F-measure scores for binary classification (the rightmost four columns in Table 8) support the accuracy rate results with significant test results (the rightmost four columns in Table 7). Acoustic cues yield higher (boredom and flow) or similar (confusion and frustration) scores than conversational cues. And the two scores in each binary classification is fairly balanced. The observation indicates the distinguishing capability of acoustic cues working in detecting 'flow' and 'boredom' of the computer tutoring system.

Table 8: Comparison: F-measure scores in % (D: the comparison work [19], P: the presented work)

F-measure	5class		4class		boredom		confusion		flow		frustration	
	D	P	D	P	D	P	D	P	D	P	D	P
neutral	26	25	na	na	57	71	63	63	59	66	66	63
boredom	32	32	35	51	67	68						
confusion	35	11	43	11			57	62				
flow	13	48	33	66					54	67		
frustration	35	6	41	0							63	66

To have a closer investigation on the performance of acoustic cues in distinguishing emotions, the binary classification between two emotional states was also conducted and the result is shown in Table 9. The best performance is from the pair of boredom and flow with the AR 71.4% while the lowest AR is 61.8% from the pair of confusion and frustration. This is possibly caused by the fact that the largest involvement separation is between the pair flow (i.e., interest) and boredom (i.e., lack of interest) and the involvement, as described above, is highly related with acoustics. The result in Table 9 shows that although confusion and frustration failed in multiple-emotions classification (i.e., all 5-class and 4-class), they can be recognized with a over 61.8% AR in pair-wise classification tasks. Confusion and frustration exhibit discriminant ability in acoustics.

Table 9: Binary classification accuracy rate (AR) between Emotional States using UM data(%)

AR	boredom	confusion	flow
confusion	67.7		
flow	71.4	65.8	
frustration	63.4	61.8	68.5

Examining the selected acoustic features provides the details of which aspect of acoustics works for emotion recognition in this study. The features selected in all 10 trials for each experiment are listed in Table 10. Pitch (i.e., F0) related feature was selected in all experiments. Comparing emotions boredom and flow with the other two, more shimmer and spectral related features were selected. The larger number of features selected for boredom and flow (mostly from shimmer and spectral) can contribute to the better performance of the binary classification of the two emotions.

Table 10: Features Selected using SFFS in All 10 Trials for UM data

	5class	4class	boredom	confusion	flow	frustration
F0	1	1	1	1	1	1
jitter		1	1		1	
shimmer			5	1		
spectral			4	1	4	
mfcc			1	1	1	1
energy					2	
voicing			1	1		
total	1	2	13	5	9	2

This section examined the acoustic cues in detecting learner's emotion of Auto Tutor system and compare the results with available work using conversational cues [19] from the same dataset. The binary-emotion classification result is better (for emotion boredom and flow) or comparable (confusion and frustration) to the comparison work in [19]. This result reveals that the emotions: flow and boredom are better captured in acoustics than conversational cues while conversational cues play a more important role in multiple-emotion classification. These results are related with the results delivered by the ITSPOKE group. Their group studied the acoustic-prosodic features (part of the acoustic cues in the present study) and lexical cues (part of the conversational cues in [19]) from the user's speech in the computer tutoring system ITSPOKE. Their results showed lexical cues outperformed acoustic-prosodic cues in distinguishing negative, neutral, and positive emotions of users [51]. The emotion states in their research are more distributed along the valence dimension (pleasant or unpleasant) than the arousal (active or passive). While in the present study, boredom and flow are emotions with extremely opposite degrees along the arousal dimension, for which the acoustic cues yield higher accuracy rate (AR).

The comparison work used Principle Component Analysis (PCA) for feature selection and extraction on their 17 conversational cues. However, the PCA did not work well on the 4445-dimension features set in this study (the AR was close to the rate by chance). This could be possibly explained that the new dimensions created by PCA representing the largest diversity of data, however, this diversity may come from the inner-class of emotion itself instead of between-classes. Also, the 95% variation PCA applied on 4445-dimension feature set still resulted in a large dimension of subset comparing with the size of sample, which is not suitable for classification. Therefore, the feature selection SFFS was required in this study and was the focus in the comparison result. What's more, the comparison between acoustic cues and conversational cues is available only using the self-judgement of emotion labels.

Table 11: Summary of three acted emotional databases and the number of samples for emotion categories.

	EPST	**EMA**	**GES**
Language	English	English	German
Speakers	7	3	10
Emotions	15	4	7
Sampling frequency	22.05kHz	16kHz	48kHz
duration/sample	1-2s	1-4.5s	1-9s
total samples	543	568	339
neutral	80	146	79
angry	139	141	127
sad	159	157	62
happy	165	124	71

5.1.4 Cross-Databases Emotion Recognition

Being unaware of any attempt to evaluate glottal features on a cross-database training platform as has been shown in the previous work in introduced as the background in Chapter 3, the goal of this section is present a preliminary report on the consistency of glottal-based features for cross-emotion database training. For comparison, an equivalent study on pitch-related features is additionally included. This section reports the performance of cross-databases 4-emotion recognition using the glottal-based features and compares it with the pitch-related features. Three databases are studied, two are spoken in English and one in German.

5.1.4.1 Data

The emotional speech data in this study involves three databases, the Emotional Prosody Speech and Transcripts database (EPST), the Electromagnetic Articulography database (EMA), and the German Emotional Speech database (GES). Details of the three databases are shown in Table 11. The EPST database [49] contains recordings of emotional and semantically neutral utterances of dates and numbers (e.g. "five hundred one"). The EMA database [46] consists of recordings of 14 sentences, most of which are semantically neutral (e.g.,"Your grandmother is on the phone"). The speech material of GES [7] is a set of 10 sentences with no semantically emotional bias covering everyday life content (e.g., “The cloth is lying on the fridge”). The sampling rates of the three databases are different, as shown in Table 11, “f_s”. The three databases are all down sampled to 16 kHz for this study. All databases are acted emotional speech with target emotion status varies from 4 to 15 in number. To provide better comparability among databases for cross-databases emotion recognition, the four emotion categories (neutral, angry, sad, and happy) consistently presenting in all three databases are investigated. The number of sample units for each emotion is shown in Table 11.

5.1.4.2 Feature

The extraction of the glottal features for each speech utterance was processed in four steps as described in Section 5.1.1. Seven statistics (i.e, the mean, median, minimum, maximum, standard deviation, range, and inter-quartile) were applied over frames to represent one speech sample, i.e., 77 glottal features in total. For comparison, the pitch-related features were included in this study. 84 pitch-related features were extracted using openSMILE toolkit [24] at the 25ms frame with the step of 10ms. Instead of seven, 14 statistics were applied on six the frame-level base pitch features (i.e., pitch, pitch envelop and their 1^{st} and 2^{nd} derivatives) for the following reasons: 1) to make the comparison between glottal and pitch features more equivalent using more balanced size of features, and 2) to represent more information about pitch contour.

5.1.4.3 Methodology

To eliminate the acoustic difference from factors other than emotion (e.g., gender, language, culture), speaker normalization was applied to the extracted features first.

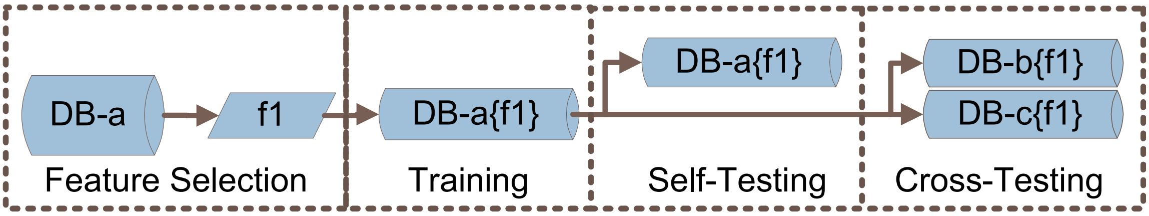


Figure 4: The block diagram of the self-evaluation and the cross-database training and testing methodology.

The processing of speaker normalization is shown in Eq. 15:

$$\hat{f_{i,j}} = \frac{f_{i,j} - mean(f_{i,j})}{std(f_{i,j})}, \tag{4}$$

where $f_{i,j}$ is the i^{th} feature descriptor for speaker j across the samples of all four emotions and std refers to the standard deviation. The normalization was conducted within database.

Given the normalized features, as shown in Fig. 4, the analysis in this study consists of *feature selection*, *self-testing* and *cross-testing*. Feature selection was applied to the normalized data using Sequential Floating Forward Selection (SFFS) for each database by 10-fold cross-validation, individually. The evaluator of SFFS was the consideration of the predictive ability along with the degree of redundancy. Wrapper was not utilized for the purpose of eliminating the effect of classifier-dependence. Features selected at least in one fold were considered as the selected feature for the database. In Fig 4, $f1$ represents the feature subset selected from Database-a("DB-a"). Then, the selected features were first used to build a 10-fold cross-validation classification trained on the database (e.g., $DB - a\{f1\}$ in Fig. 4) and tested on the same database (e.g., $DB - a\{f1\}$), referred to as "self-testing". The features selected using one database were then trained on this database (e.g., $DB - a\{f1\}$) and tested on the other two databases (e.g., $DB - b\{f1\}$ and $DB - c\{f1\}$), referred to as "cross-testing". The procedure was repeated by starting from all three databases. The software WEKA [35] was used for SFFS and the 4-emotion classifiers were built using LibSVM [10] with the RBF kernel.

5.1.4.4 Results

The results are shown in Table 12. It could be read in this way: the first block "Feature Selection" reports the number of the selected feature for each database. The following three blocks represent the 4-emotion classification trained using the selected features for three databases, individually. Inside each block, the sub-block with $*$ in bold represent "self-testing" results by 10-fold cross-validation while other entry is the result of "cross-testing" tested on the other two databases. The results are shown in accuracy rate ("AR"), precision ("prec"), and recall("reca") using glottal ("GLO") and pitch ("PCH") features, separately.

From the block of "Feature Selection" of Table 12, the number of selected features of glottal and pitch are approximately equivalent for three databases. The acknowledged by this is that the difference between the following classification results using glottal and using pitch features is not caused by the different the numbers of the selected features. From the "AR" rows of Table 12, overall, the observations are that 1) the ARs of "self-testing" are higher than "cross-testing", 2) the ARs using glottal features ("GLO") are higher than pitch features ("PCH"). For "self-testing", comparing the three databases, both glottal and pitch features exhibit higher ARs for EMA (79.7%G, 66.4%P) and GES (82.6%G, 70.5%P) while relatively lower for EPST (64.3%G, 44.0%P) in the 4-emotion classification.

For "cross-testing", training using EPST produced up to 26.8% AR (i.e., tested on GES using glottal features). These ARs are around or lower than the rate by chance. This indicates the model trained using EPST has difficulty in working on EMA and GES, which also reveals the difference of both glottal and pitch features between EPST and the other two databases in emotion expression. From the block of "train:EMA" of Table 12, the model trained on EMA performed better when tested on GES than EPST. Although EMA and EPST are in English, while GES is in German, the language is not the issue to affect cross-testing in glottal and pitch features for

Table 12: The results of accuracy rate (AR), precision, and recall for four emotions, ("*" indicates the self-tested experiments using 10-fold cross validation, the self-tested accuracy rates are in bold).

Feature Selection							
No.feat		EPST		EMA		GES	
		GLO 23	PCH 27	GLO 30	PCH 24	GLO 28	PCH 37

train: EPST							
test:		***EPST***		EMA		GES	
		GLO	PCH	GLO	PCH	GLO	PCH
accuracy rate		64.3	44.0	24.6	19.7	26.8	14.7
precision	Happy	0.61	0.44	0.30	0.22	0.36	0.19
	Angry	0.71	0.39	0.16	0.13	0.36	0.15
	Sad	0.57	0.51	0.26	0.06	0.15	0.02
	Neutral	0.80	0.47	0.24	0.44	0.19	0.53
recall	Happy	0.65	0.70	0.40	0.66	0.39	0.35
	Angry	0.57	0.38	0.14	0.11	0.31	0.11
	Sad	0.66	0.35	0.33	0.02	0.31	0.03
	Neutral	0.73	0.19	0.14	0.08	0.06	0.11

train: EMA							
test:		EPST		***EMA***		GES	
		GLO	PCH	GLO	PCH	GLO	PCH
AR		26.0	20.1	79.7	66.4	64.9	49.3
precision	Happy	0.37	0.27	0.74	0.68	0.52	0.39
	Angry	0.20	0.13	0.72	0.64	0.76	0.64
	Sad	0.25	0.17	0.84	0.62	0.57	0.41
	Neutral	0.22	0.21	0.88	0.72	0.75	0.60
recall	Happy	0.32	0.26	0.72	0.78	0.56	0.55
	Angry	0.22	0.11	0.72	0.45	0.66	0.39
	Sad	0.20	0.18	0.89	0.64	0.89	0.61
	Neutral	0.34	0.30	0.84	0.80	0.52	0.51

train: GES							
test:		EPST		EMA		***GES***	
		GLO	PCH	GLO	PCH	GLO	PCH
AR		28.2	20.4	56.9	33.3	82.6	70.5
precision	Happy	0.40	0.12	0.55	0.47	0.74	0.71
	Angry	0.25	0.22	0.52	0.31	0.79	0.66
	Sad	0.08	0.15	0.76	0.19	0.92	0.84
	Neutral	0.26	0.21	0.56	0.41	0.89	0.71
recall	Happy	0.42	0.01	0.64	0.06	0.55	0.07
	Angry	0.34	0.48	0.70	0.72	0.90	0.94
	Sad	0.03	0.05	0.34	0.07	0.90	0.74
	Neutral	0.40	0.44	0.63	0.48	0.90	0.87

this study. It can be further observed from the block of "train:GES" since the AR results tested on EMA are higher than EPST, too. Based on the above observation, glottal features, in the cross-databases way, could be trained and tested better using the pair of EMA and GES (with AR up to 64.9%) than other combinations involving EPST. This pattern can also be observed in pitch features, but with lower ARs.

Table 13: The demonstration of precision and recall.

		actual class	
		positive	negative
predicted	positive	tp (true positive)	fp (false positive)
class	negative	fn (false negative)	tn (true negative)

More details of the classification results are shown in Table 12 in terms of *precision* and *recall*, given by Eqs.(5) and(6),

$$precision = \frac{tp}{tp + fp}, \tag{5}$$

$$recall = \frac{tp}{tp + fn}, \tag{6}$$

where tp, fp, and fn are shown in Table 13. Examining the precision and recall provides the details of performance of single emotion for "cross-testing". Due to the low values of EPST, the focus will be on EMA and GES in this discussion. When trained on EMA and tested on GES (block "train:EMA", column "GES"), the smaller precision (0.57) and larger recall (0.89) of "sad" using glottal features indicate that other three emotions are likely to be classified as "sad". "Sad" is also the reason for the performance of the model with the training and testing sets exchanged (block "train:GES", column "EMA"). "Sad" is classified as other emotions than itself from the larger precision (0.76) and smaller recall (0.34). However, for pitch, the distribution of the error of classification is widely spread.

It should be pointed out that the results show better performance for both "self-testing" and "cross-testing" in all databases using glottal than pitch. However, the conclusion should be drawn with caution because glottal features include the acoustic information not only in the time-domain but also the spectral while pitch features capture mainly the time-domain of speech. From the results, up to 64.9% could be achieved using glottal features only. More similarity are shared by EMA and GES in terms of emotion expression in glottal features than EPST. And this pattern could be observed by pitch features as well. This serves the goal of investigating the glottal features in cross-database emotion recognition of this study.

This section reports the study of cross-databases emotion recognition in four emotion states (neutral, angry, happy, sad) using glottal-based features and compares it with pitch-related features. Three acted databases (two in English, one in German) are studied from the perspective of "self-testing" (i.e, trained on one database and tested on the same one) and "cross-testing" (i.e., trained on one database and tested on others) for 4-emotion recognition. In the results, Using glottal features only, up to 64.9% could be achieved for cross-database emotion recognition. Difference in performance between training/testing database pairs could be observed and the baseline of multiple database emotion recognition is provided for future work.

In this study, all the databases were recorded in the lab environment at last 16kHz. Considering the potential application of detecting emotions in recordings at the telephone quality level, e.g., Call Center speech, the challenge arose that the glottal features may fail in the emotion recognition tests in the telephone quality speech (with the sampling frequency of 8kHz). To study the emotion distinguishing ability of the glottal features extracted form the telephone quality speech in the cross-database experiments, comparable study was conducted and presented in ChapterA.

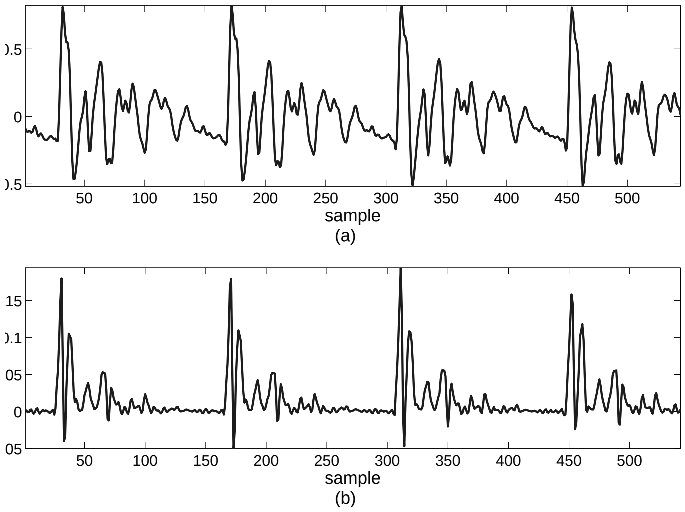


Figure 5: Example of the TEO derived from one frame of speech signal. (a) One frame of speech signal, (b) The TEO derived from (a) signal.

5.2 Teager Energy Operator (TEO)

5.2.1 Extraction

Another domain of acoustic features in this thesis is Teager Energy Operator (TEO) based features. TEO was motivated by the experiments in speech and hearing by Teager and Teager [79, 80, 81]. The results of the experiments showed that the vocal fold model is nonlinear, under the effect of vortex action. The vortex action caused the modulation in speech signal. Teager developed the energy measurement motivated by the experiments based on the modulation pattern. Then the calculation of TEO $\Psi[x(n)]$ of discrete-time speech signal $x(n)$ was formulated by by Kaiser [41] as shown in Eq. 7.

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1). \tag{7}$$

Given the TEO, three sets of measurements of TEO [96] were calculated as the Low-level Descriptors in the feature set:

(1) **Variation of FM component (TEO-FM-Var)**, the FM demodulation

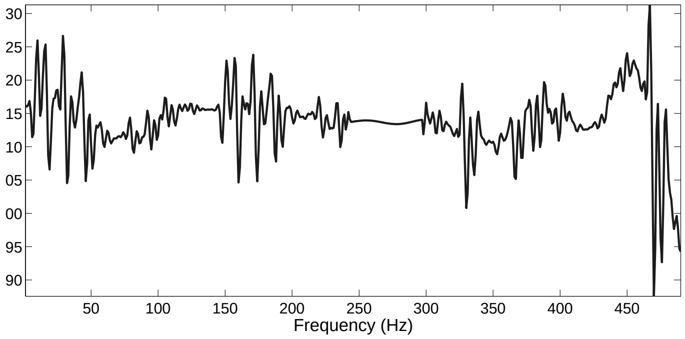


Figure 6: Example of one frame of FM component of TEO.

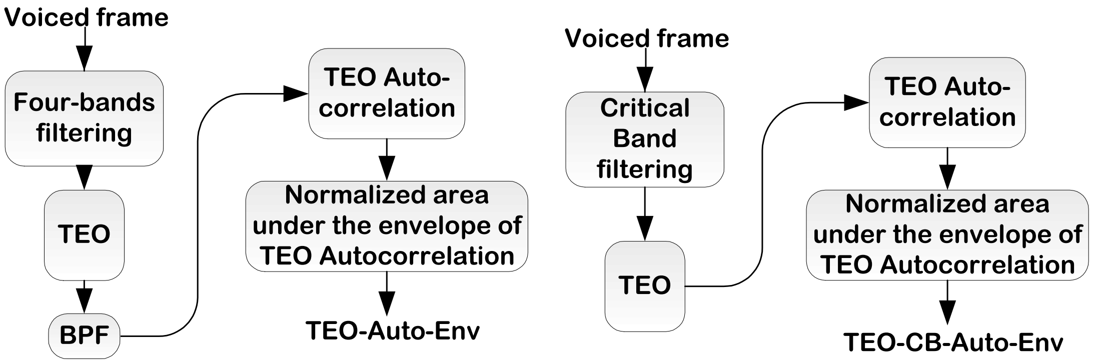
component was obtained by Eq. 8 [53, 36], and eight statistics (mean, minimum, maximum, range, log-range, standard deviation, median, and inter-quartile) were calculated for each frame to form FM-Var feature set,

$$\omega(n) = arcsin\sqrt{\frac{\Psi[x(n+1)] - \Psi[x(n-1)]}{4 * \Psi[x(n)]}}. \tag{8}$$

(2) **Normalized TEO autocorrelation envelope area (TEO-Auto-Env)**, TEO was applied on each of the four-band filtered (0-2kHz, 2-4kHz, 4-6kHz, 6-8kHz) speech and the normalized area under envelope of TEO autocorrelation was computed for each band, respectively, as shown in Fig. 7(a). The four band was chosen by equally dividing half of the sampling frequency of the speech (i.e., 16Hz);

(3) **Critical band based TEO autocorrelation envelope (TEO-CB-Auto-Env)**, 16-critical band filterbank [50] was applied to the voiced speech and TEO was calculated in each of the band. The normalized area under the envelope of TEO autocorrelation was computed for each band, respectively, as shown in Fig. 7(b).

Besides the above measurements of TEO, Teager Energy Cepstrum Coefficient (TECC) was calculated as another TEO based feature set. TECC was proposed with the motivation of the processing of MFCC feature and TEO [18]. The extraction procedure of TECC was similar to MFCC but using Teager Energy $TEO[x(n)]$ instead of the squared energy $[x(n)]^2$ as the primary difference (as shown in Eq. 7). The voiced speech was segmented into frame with length approximating four times of



(a) TEO-Auto-Env

(b) TEO-CB-Auto-Env

Figure 7: The block diagram of the extraction of (a)the normalized TEO autocorrelation envelope area (TEO-Auto-Env) and (b) the critical band based TEO autocorrelation envelope (TEO-CB-Auto-Env).

the pitch period with a step size of 10ms. The extraction on one frame of signal was demonstrated in Figure 8. The Gammatone filter is given by Eq. 9 in the time domain,

$$g(t) = At^{n-1}exp(-2\pi ERB(f_c)t)cos(2\pi f_c t), \tag{9}$$

where A, b, and n are Gammatone filter design parameters and f_c is the center frequency. According to [40, 18], the parameters are set as $b = 1.019$ and $n = 4$. Equivalent Rectangular Bandwidth (ERB) represents the bandwidth of filters, which is given by Eq. 10,

$$ERB(f) = 6.23(f/1000)^2 + 93.39(f/1000) + 28.52. \tag{10}$$

where f is the center frequency in Hz. And the filter gain A is set under the consideration that the frequency response at the center frequency equals to one. The filter placing is in Bark-scale (critical filterbank) [97] and the number of filterbank in this study is 25. More details of the extraction of TECC and the evaluation compared with MFCC is presented in the following section. The motivation and evaluation of TECC is presented in Section 5.2.2.

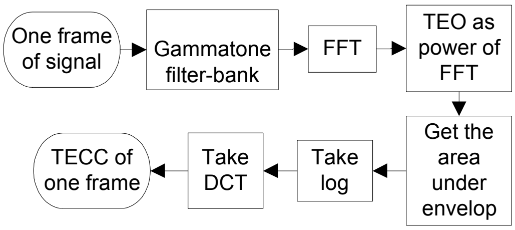


Figure 8: The block diagram of TECC extraction algorithm on one frame.

5.2.2 The noise analysis of TECC compared with MFCC

Automated emotion detection is the attempt to quantify an abstract interpretation into objectively measured components of recorded human interaction. Emotion recognition in a noisy condition remains a challenging problem. The literature shows that Mel-Frequency Cepstrum Coefficients (MFCCs) exhibit robust performance in speech analysis in noisy environment, especially for speech recognition [71, 18, 56, 60, 95, 94, 12, 47, 85, 55]. On the other hand, work on the use of Teager Energy Operator (TEO) [9, 96, 37, 72] in classifying emotion has shown that these features can provide valuable insight into distinguishing different types of emotional expression. This motivates the study of emotion recognition using features combining the advantage of both MFCC and TEO, which has not been reported much yet. Teager Energy Cepstrum Coefficient (TECC) was first proposed by Dimitriadis and his colleges and studied to show its robust performance in speech recognition [18, 17]. In this section, the robustness of TECC in emotion recognition was investigated and compared with MFCC at different noise levels for three databases, two of which are English emotional databases and one is in German.

5.2.2.1 Data

The emotional speech data in this study involves three databases, the Emotional Prosody Speech and Transcripts database (EPST), the Electromagnetic Articulography database (EMA), and the German Emotional Speech database (GES). To provide better comparability among databases, the four emotion categories (neutral, angry, sad, and happy) consistently presenting in all three databases are investigated. The number of sample units for each is shown in Table 11.

5.2.2.2 Feature

The statistics of MFCC and TECC features and their derivatives (ΔMFCC and ΔTECC) were extracted and calculated to form the feature set in this study.

MFCCs were computed from the log-squared-energy in frequency bands distributed over a Mel-scale. The extraction of MFCC features for each speech sample was processed in five steps: (1) marked the voiced section of speech, (2) divided the voiced section into frames approximating four pitch periods in length with a 10ms step, (3) took the Fourier Transform on each frame, (4) mapped the power spectrums on to a Mel-scale, (5) took the log of the power at each Mel-scale band, (6) took the Discrete Cosine Transform (DCT) of Mel-log powers. The amplitude of the resulting spectrum was MFCC. The ΔMFCC feature was calculated by Eq. 11,

$$\Delta MFCC_j(i) = MFCC_j(i+1) - MFCC_j(i), \tag{11}$$

where $MFCC_j(i)$ is the j^{th} coefficient of MFCC from the i^{th} frame. The number of coefficients of MFCC used in this study is 12.

TECC was proposed with the motivation of the processing of MFCC feature and Teager Energy Operator [18]. Teager Energy was proposed by Teager based on his nonlinear model of the true source of sound production, which is actually the vortex-flow interactions [79, 82]. He developed the Teager Energy Operator supporting the observation that hearing is the process of detecting the energy. The TEO of discrete-time speech signal $x(n)$ can be calculated following Eq. 7 derived by Kaiser [41] as shown in Section 5.2.1. The ΔTECC feature is calculated by Eq. 12,

$$\Delta TECC_j(i) = TECC_j(i+1) - TECC_j(i), \tag{12}$$

where $TECC_j(i)$ is the j^{th} coefficient of TECC from the i_{th} frame. All features were quantified using seven statistics (i.e, the mean, median, minimum, maximum, standard deviation, range, and inter-quartile) across all frames of a sample to form the representation of an utterance. The feature extraction produced 168 MFCC features and 350 TECC features.

5.2.2.3 Methodology

The purpose of this study was to evaluate the robustness of the discrimination ability of TECC features in noisy conditions. Investigating the relationship between the robustness of features and the noise degree of speech requires emotional speech whose noise level is quantified and measurable. Therefore, five sets of data were created by adding White Gaussian noise to the "clean" speech dataset at five Signal Noise Ratio (SNR) levels from 20dB to 0dB with the step of 5dB. In total, six datasets (including the clean data) were available for each database (i.e.,five noisy and one clean). The White Gaussian noise was chosen as the additive noise because based on the research of [18], the White Gaussian noise produced the largest difference between the features extracted from the clean speech and the noisy speech, comparing with Babble, Pink, and Car noise.

It has been shown that acoustic features are speaker-dependent because they capture the characteristics of speakers (e.g., gender, language, culture) [22]. To eliminate the acoustic difference from factors other than emotion, speaker normalization was applied to the extracted features first. The processing of speaker normalization is shown in Eq. (15):

$$\hat{f_{i,j}} = \frac{f_{i,j} - mean(f_{i,j})}{std(f_{i,j})}, \tag{13}$$

where $f_{i,j}$ is the i^{th} feature descriptor for speaker j across the samples of all four emotions and std refers to the standard deviation. The normalization was conducted within database.

Given the normalized features, the normalized mean squared error (NSME) [18, 94] for MFCC and TECC was calculated at each noise level and compared. NSME is the measurement on the distance between feature of the noisy and clean speech from the same signal segment. The calculation of NMSE is shown in Eq. (14). It's defined as the average Euclidean distance between the "clean" and "noisy" features divided by the mean of "clean" feature vector norm [18],

$$NMSE = \frac{mean(D(f_{i,clean}, f_{i,noisy})}{mean(|f_{i,clean}|)}, \tag{14}$$

where $D(f_{i,clean}, f_{i,noisy})$ is the Euclidean distance between the i^{th} feature in feature set of the clean speech and the noisy speech, and $|f_{i,clean}|$ is the vector norm of the i^{th} feature of the clean speech . The interpretation of NSME is that a smaller NSME value implies more robustness the feature possesses (i.e., NSME value is zero for the clean speech).

The robustness of the discrimination ability of features were evaluated in emotion recognition experiments. Using 5-fold cross-validation, the experiment built a four-class classifier on four subsets of data using a Support Vector Machine (SVM) and tested it on the other subset(using LibSVM tool [10] in MATLAB, linear kernel). This procedure was repeated using another choice of training and testing sets till all sets has been tested. This classification was repeated 10 times for randomization. The further study was carried out as the discrimination ability in pair-wise emotion classification task. Four emotion categories formed six emotion pairs. One classifier was built for each pair using the liner kernel SVM with 5-fold cross-validation and the whole analysis was repeated 10 time as well.

5.2.2.4 Results

In this section, the normalized mean squared errors of MFCC and TECC on six datasets are reported. Then the classification results of multi-emotion and pair-wise emotion tasks are presented.

Mean Squared Error Analysis

Table 14: The normalized mean squared error (NSME) of MFCC/TECC at five SNR levels for three databases. The smaller value between MFCC and TECC under the same noisy condition using the same data is shown in **bold.**

SNR	**EPST**		**EMA**		**GES**	
	MFCC	TECC	MFCC	TECC	MFCC	TECC
0	0.58	**0.39**	0.52	**0.24**	0.56	**0.33**
5	0.50	**0.33**	0.44	**0.20**	0.47	**0.29**
10	0.41	**0.28**	0.35	**0.16**	0.38	**0.25**
15	0.32	**0.24**	0.27	**0.12**	0.30	**0.21**
20	0.24	**0.20**	0.20	**0.09**	0.22	**0.17**

Table 14 lists the NSME values at five SNR levels for three databases. It could be observed that, for all three databases, the values of MFCC and TECC decreases while the noise is reduced. It indicates the reliability of values in Table 14 according to the interpretation of NSME. It should be noticed that, at all noise levels of all three databases, the value of TECC is smaller than MFCC (value in bold). Especially for EMA, the all-level TECC values are less than half of MFCC. Based on the observation, the conclusion could be reached that TECC is more robust than MFCC facing additive noise in emotional speech. To further investigate the robustness of emotion-distinguishing ability of TECC, multi-emotion and pairwise-emotion classification experiments were conducted.

Emotion Recognition experiments

In emotion recognition experiment, both TECC and MFCC features were applied in the multi-emotion classification (four-emotion) task first. The four-emotion classifier was built using LibSVM [10] with the liner kernel. The accuracy rate (AR) was calculated as the average of those from 10 repetitions of classifications (5-fold cross-validation).

The accuracy rates at different noise levels are shown in Figure 9. From Figure 9 it's clear that the accuracy rates at all noise level using TECC are equal to or higher than MFCC for all three databases (i.e., AR using TECC is up to 71% in EPST, up to 89% in EMA, and up to 85% in GES for the four-emotion classification). When SNR equals to zero, TECC and MFCC performed equally. As the noise is reduced, using TECC improved the AR up to 38% for EPST, 9% for EMA, and 8% for GES. Overall, the ARs of EPST are relatively lower than EMA and GES. The possible explanation is that EPST contains 15 emotion categories while EMA has 4, and GES covers 7. The wider variety of emotion categories led to less acoustic difference between emotions for speech in EPST than the other two. Moreover, the robustness of emotion-distinguishing ability of TECC and MFCC is shown in the relationship between the variation of ARs with the change of noise levels.

For a better evaluation of the variation of ARs,the standard deviation of ARs at six noisy conditions using MFCC/TECC for each database is shown in Table 15. From Table 15, the standard deviation of MFCC and TECC is approximating equal. But for EPST, the variation of TECC is larger than MFCC. The reason for the larger variation of TECC is the increase in Figure 9(a). The conclusion could be reached that, both MFCC and TECC exhibit robustness in emotion recognition in noisy conditions while the overall AR of TECC is relatively higher. The larger variation of AR using TECC (in EPST) is caused by the performance improvement of ARs with the reduction of noise than MFCC. The significance of the difference between the accuracy rates using TECC and MFCC was tested by the Kruskal-Wallis test. The reason for choose this test was because using Jarque-Bera test, the accuracy rates of the classification using the MFCC from the “clean” (no additive noise) EPST data did not fit the normal distribution. Using the speech of three databases (EPST, EMA, GES) at six noise levels (lab-quality/clean, SNR=0, 5, 10, 15, 20), 18 experiments were conducted. The significant test showed that, in four-emotion recognition, expect for SNR equaling to zero for EPST and GES, the classification results (the other 16 experiments) at all six different noise levels for three databases exhibit statistically significant difference between using TECC and MFCC.

Table 15: The mean and standard deviation (std) of ARs over six noise levels for 4-emotion classification using MFCC and TECC separately for three databases.

	EPST		**EMA**		**GES**	
	MFCC	TECC	MFCC	TECC	MFCC	TECC
mean	52.3	61.1	79.1	84.3	79.6	84.1
std	1.11	7.69	3.97	4.01	1.55	2.05

The hypothesis is that the performance of TECC and MFCC is not the same to all emotion categories. To test it, a pair-wise emotion classification experiment was conducted and tested by the Kruskal-Wallis test. The results are shown in Table 16. Since this experiment contains six emotion pairs at six noise levels for three databases. The resulting number of classification will be 108. Similar as the multi-classification task, for each classification, the accuracy rate was obtained as the average from 10 repetitions of 5-fold cross-validation classifications.

From the mean ARs row of Table 16, the AR using TECC only reaches 74-85% for EPST, 86-99% for EMA, and 94-99% for GES in pair-wise emotion classification. From the standard deviation row of Table 16, the variation of ARs is quite small comparing with their mean values (up to 5% for EPST, 5% for EMA, and 4% for GES). This indicates the little effect from noise on the distinguishing ability of both TECC and MFCC. Comparing TECC with MFCC, the AR of TECC is increased by up to 15% for EPST (neutral-angry), 8% for EMA (angry-happy), and 5% for GES (angry-happy) than MFCC. In pair-wise emotion recognition, 108 experiments were conducted using TECC and MFCC separately. The significant test shows that for 80.6% of EPST and 86.1% of EMA experiments exhibit the significant difference between the accuracy rates using TECC and MFCC. For GES, only 33.3% of the experiments exhibit the significant difference, most of which are from the classification of neutral vs. angry, neutral vs. sad, angry vs. sad, and sad vs. happy, the accuracy rates of which are relatively higher for both TECC and MFCC.

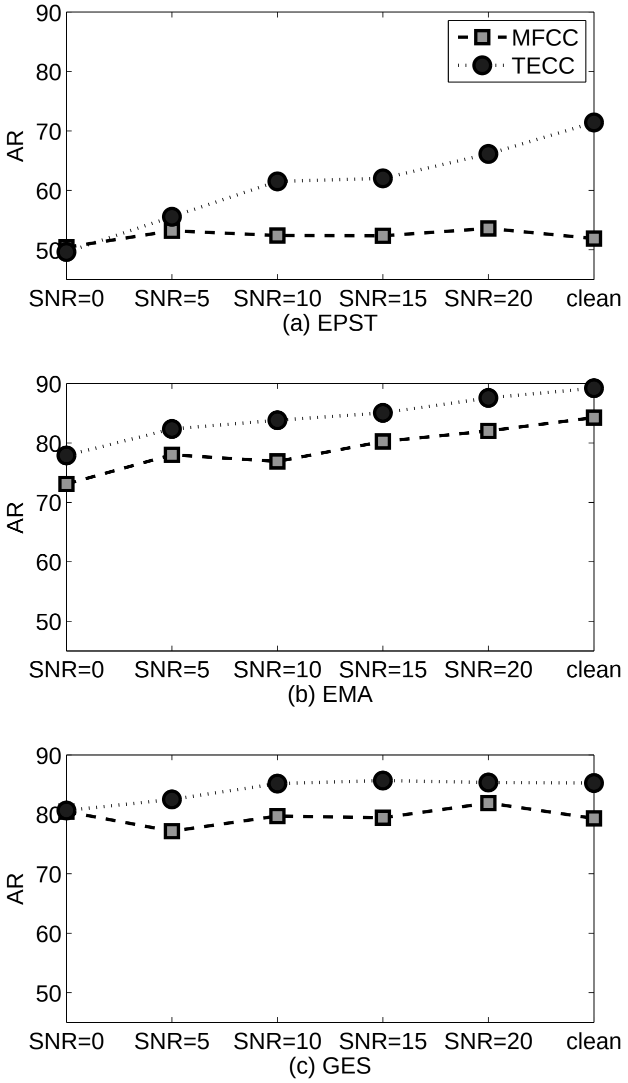


Figure 9: The accuracy rate (AR in %) of 4-emotion classification using MFCC and TECC separately for three databases, (a)EPST, (b)EMA, and (c)GES.

Table 16: The mean, and standard deviation (std) of accuracy rate from the pairwise emotion (N:nutral, A:angry, S:sad, H:happy) classification of six noise levels using MFCC and TECC features for three databases.

		N-A		N-S		N-H		A-S		A-H		S-H	
		MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC	MFCC	TECC
EPST	mean	69.9	80.3	76.8	84.4	79.4	86.1	67.5	74.2	72.0	77.5	71.7	74.0
	std	2.1	4.3	2.5	4.4	1.4	4.6	3.3	2.6	1.4	3.1	1.8	6.8
EMA	mean	96.6	98.6	85.6	88.1	96.0	97.5	97.9	98.5	79.6	85.9	94.3	97.9
	std	1.5	0.4	2.3	3.9	0.9	0.3	0.9	0.7	3.1	4.1	2.1	1.0
GES	mean	97.8	98.5	94.2	94.2	93.0	96.4	99.3	99.1	75.6	78.8	99.3	99.4
	std	0.8	0.2	1.1	1.4	2.9	1.0	0.3	0.2	2.1	3.0	0.5	0.2

Overall, the ARs from EPST are lower than EMA and GES, which has been observed and explained in the multi-emotion task. Even though, the ARs for all three databases using TECC are fairly high, especially for EMA and GES (up tp 99%). Among six emotion pairs, the pair angry and happy possesses relatively lower ARs than other pairs for EMA and GES of both features. This observation could be explained by the conclusion that emotions with valence difference could be less captured by acoustics than arousal difference, which has been studied. This observation is not obvious in EPST data. The reason for this is that we chose “hot anger” in EPST, in which “cold anger” also uttered. Therefore, the angry in EPST was supposed to exhibit more difference in arousal than angry in other databases. As discussed with Table 16, the highest improvement of TECC than MFCC happens in the neutral and angry pair for EMA and GES. This emphasized the performance of TECC when less acoustic difference exists.

This study investigates the robustness of TECC in emotion recognition facing additive noise at different levels. The results from three databases (two in English, one in German) highlight the robust emotion-discrimination ability of both TECC and MFCC. But the higher accuracy rate is achieved by TECC than MFCC. For the condition when SNR equals to zeros, TECC and MFCC performed similarly. While the noise level is reduced ($SNR = 5 \sim +\infty$), TECC outperformed MFCC in all emotion recognition tasks. Overall, using TECC features only, the up to 89% for four-emotion classification and 99% for pair-wise emotion classification accuracy rate could be achieved. Future work will involve the study using the real-life authentic emotional speech data in different speech quality conditions.

5.3 Emotion Distinguishing Capability of Glottal and TEO

5.3.1 Continuous Emotion labels with Glottal and TEO

Real-life application is the goal of emotion research. Therefore, it's interesting to examine the discrimination ability of glottal parameters and TEO using *real* emotional speech data in more dimensions. The purpose of the study in this section is to evaluate the performance of glottal waveform parameters and TEO in distinguishing binary classes in four emotion dimensions (activation, expectation, power, and valence) using the authentic emotional speech from SEMAINE corpus.

5.3.1.1 Data and Feature

The speech data used in this study is from the SEMAINE corpus [54]. The speech is the recording of conversations between humans (the user) and artificially intelligent agents (the operators). The emotion labels of the speech are provided in four emotion dimensions: activation, expectation, power, and valence. In each dimension, a binary label 1/0 represents the emotion possessing a higher/lower degree than the averaged. The provided corpus is divided into three parts: the training (for train model), the development (to test model), and the test (the challenge data, see [69]). More information about this data is available in [69]. Because of the large size of data, the speech data was down-sampled from 48kHz to 16kHz in this study.

This section studied three sets of acoustic features, the given openSMILE feature set [23, 24], glottal waveform parameters, and Teager Energy Operators (TEO). The given openSMILE features consist of prosodic related, spectral related, and other voice related (e.g., probability of voicing, jitter, and shimmer) features extracted using the openSMILE toolkit [23, 24] and provided with the speech data. The 1941-dimensional

Table 17: Number of words in the training dataset ('Train') and the development dataset ('Develop') in total and in each binary class ('class-1/0').

	Train				**Develop**			
	activate	expect	power	valence	activate	expect	power	valence
total	15307	15307	15307	15307	12663	12663	12663	12663
class-1	7695	6253	8611	8357	7275	4240	8595	8131
class-0	7612	9054	6696	6950	5388	8423	4068	4532

acoustic feature set was created at the word-level (more information in [69]).

The second acoustic set used in this study is the glottal waveform parameter. The feature extraction following the instruction in Section 5.1.1 produced 77 glottal waveform features. The third domain of acoustic features in this study is the measurements of Teager Energy Operator (TEO) [9] in Section 5.2.1. All TEO features were quantified using seven statistics (i.e, the mean, median, minimum, maximum, standard deviation, range, and inter-quartile) across all frames of a word to obtain 196 TEO features. Together with the first two sets, 2214-dimension acoustic features were extracted at the word-level. Due to the algorithm for feature extraction, some words did not produce valid glottal parameters or TEO (e.g., impulsive phoneme). Therefore, the number of words with all three features sets available is smaller than the given speech data. The number of words in the training dataset and development dataset is summarized in Tabel 17. These datasets were used in the following analysis including the recalculated baseline of openSMILE (instead of the baseline given in [69]).

5.3.1.2 Methodology

The purpose of this study was to evaluate the discrimination ability of glottal and TEO features on binary categories in four emotion dimensions using authentic emotion speech SEMAINE corpus.

First, the feature was subjected to a Kruskal-Wallis (KW) significance test individually for a fundamental sight of the distinguishing possibility. The significant test was conducted using the combination of the training data and the test data. Number of features showing statistically significant difference (i.e., $p < 0.01$) in each feature set was counted and the percentage of feature showing significant difference was calculated. The purpose of this test was to assess the individual discrimination ability of features in each feature set.

Although features showing statistically significant difference individually have been selected in the above test, features in subset may exhibit more discrimination ability than individually. To compare the classification results between three feature sets in subset, this experiment built a model on training data using a Support Vector Machine (SVM) and tested it on the development data using one set of features at a time (using LibSVM tool [10] in WEKA [35]). Due to the large number of features in each feature set, feature selection using Sequential Forward Floating Selection (SFFS) was applied to each feature set using WEKA [35] before classification. The evaluator of SFFS was to select the features possessing higher correlation with the emotion labels and lower intercorrelation ('CfsSubsetEval' option in WEKA), which was not classifier related.

Finally, the discrimination ability of feature sets was compared by using a Sequential Feature Selection (SFS) algorithm for selecting any subset of features out of the combination of three feature sets together. SFS starts with an empty feature set and sequentially adds features that have not yet been selected. Every feature combination set is evaluated until there is no improvement in the criterion function. For this study, the criterion was set to the accuracy rate from a quadratic discriminant computed as the number of correct classifications divided by the total number of observations (i.e., words). The SFS algorithm added features in an effort to increase the accuracy rate as much as possible. Due to the large size of feature dimension (2214), this study 1) chose SFS instead of SFFS for feature selection, and 2) SFS was applied onto features showing statistically significant difference obtained in the Kruskal-Wallis test instead

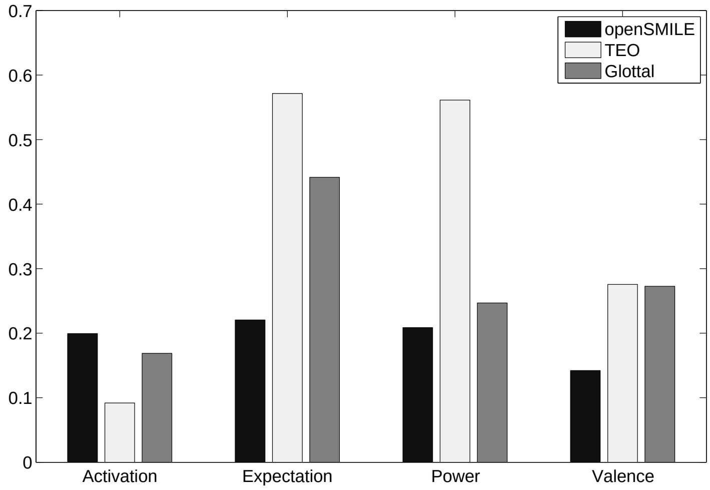


Figure 10: Percentage of features with statistically significant difference between two classes in each feature set for SEMAINE.

of the full set (2214 dimensions). The SFS was run on the feature subset 20 times (20 trials) to ensure enough randomization.

5.3.1.3 Result

The percentage of features with statistically significant difference between the binary classes for four dimensions is shown in Fig. 10. Considering the percentage for each feature set across the four dimensions, the percentage of openSMILE is fairly similar across all dimensions. TEO possesses a much higher percentage in expectation and power than activation. For glottal parameters, the percentage for expectation is highest followed by valence. This observation indicates the discrimination power of TEO in expectation and power and glottal parameters in expectation. This discrimination power is further evaluated by classification experiments using the three feature sets individually.

The binary classification experiments were conducted using one feature set at a time. The accuracy rate (AR) and F-measure [3] of binary classification is calculated and shown in Table 18. The highest AR (in bold) is from TEO in activation and

Table 18: The accuracy rate (AR) and F-measure (F-1/0) of classification using three sets of features separately (%) (op: openSMILE as the *baseline*, glo: glottal parameters, F-1/0: F-measure for class 1/0).

	activation			**expectation**			**power**			**valence**		
	op	TEO	glo	op	TEO	glo	op	TEO	glo	op	TEO	glo
AR	*55.0*	**59.3**	56.5	*66.0*	**66.3**	**66.3**	*66.3*	**67.4**	66.8	*62.9*	63.7	**64.1**
F-1	51.2	46.4	56.8	79.3	79.7	79.7	8.5	17.2	8.8	9.3	4.0	0.2
F-0	58.3	67.2	56.2	5.3	1.0	1.3	79.4	79.7	79.7	76.7	77.6	78.1

power, and glottal parameters in valence, while TEO and glottal features yielded equal ARs in expectation. This result is reasonable for the higher percentage of TEO showing significant difference in expectation and power and glottal in expectation and valence than other dimensions (in Fig. 10). Except for activation, the F-measures of other dimensions show the bias of classification results (i.e., expectation-1, power and valence-0). This could be the reason to cause the AR of activation lower than other three dimensions.

To get a closer look at the performance of glottal parameters and TEO comparing with openSMILE, the feature selection result using SFS on the combination of three feature sets was conducted. Because of the large dimension of feature set, the original 2241 features were represented by a subset consisting of features showing significant difference in the KW test. Table 19 shows the resulting mean accuracy rates from the SFS procedure along with the number of features selected over 20 trials. From Table 19, fewer features were selected in valence than others. This result supports the previous work that valence is less captured by acoustics than other dimensions (e.g., arousal [38]). Comparing the three sets, more TEO features were selected for activation and power, and more glottal parameters were selected for expectation and valence. This indicates the stronger discrimination ability of TEO in activation and power, and of glottal parameters in expectation and valence. This also explains the highest AR in Table 18 achieved by TEO in activation and power and by glottal in expectation and valence. The AR by using SFS shows that combining three feature

Table 19: The feature selection result: the averaged accuracy rate (AR %) over 20 trials and the number of features selected in all 20 trials.

	activation	**expectation**	**power**	**valence**
AR	64.3	62.3	61.2	58.6
No. OpenSMILE	4	5	2	1
No. TEO	0	3	2	0
No. glottal	0	1	0	1

sets, the accuracy rate of activation is increased. The AR of other three dimensions is lower. This could be explained that the bias of classification results is reduced for the other three dimensions.

This section (Section 5.3.1) examined the performance of glottal waveform parameters and TEO in distinguishing binary classes in four emotion dimensions (activation, expectation, power, and valence) using authentic emotional speech SEMAINE corpus. The result highlights the discrimination ability of TEO in emotion dimensions activation and power, and glottal parameters in expectation and valence for the authentic speech data. In the same classification experiment, TEO and glottal parameter outperformed or performed similarly to the prosodic, spectral and other voicing related features (i.e., the feature set obtained using openSMILE).

CHAPTER VI

PRE-PROCESSING OF SIX DATABASES

In the effort to perform cross-database training and testing, one challenge is about the emotion labels. Each of the six databases investigated in this thesis has its own emotion labeling strategy, e.g., in category (different choice of emotional status words) or in dimension (different resolution, steps, and number of dimensions). To solve the emotion label problem, Section 6.1 describes the methodology of mapping all the labels from different databases to three categories in valence/arousal, i.e., positive, neutral, and negative in valence/arousal. In the following chapters, the emotion recognition is conducted in binary-class between neutral vs. emotional (positive and negative), and positive vs. negative (neutral excluded) in valence and arousal. Having studied different sets of acoustic features, especially the glottal and TEO based features, Section 6.2 summarizes the acoustic feature sets used in the following chapters. Features were selected by Sequential Forward Floating Selection (SFFS) for each database to provide a preliminary similarity comparison between databases in terms of acoustic features. The features shared by two or more databases are presented in Section 6.2 while the detailed feature selection results are shown in Appendix C.

6.1 Emotion Labeling Decision

Among the six emotional databases studied, four (EPST, EMA, GES, and UM) are labeled with categories and two (SEMAINE and VAM) are in dimensions. Since there is no widely accepted method to map emotion categories onto dimensions, to reach generalized result, no assumption was made to assign each category the dimension values. Instead, three classes of emotions (valence/arousal) are decided to represent emotion status in this study, positive/high, neutral, and negative/low. The emotion

Table 20: Emotion labeling strategy of emotional databases.

DBs	**Valence+**	**Valence0**	**Valence-**	**Arousal+**	**Arousal0**	**Arousal-**
EPST	Elation, happy, interest, pride	Neutral	Hot anger, panic, anxiety, disgust, contempt, cold anger, sadness, shame, despair, boredom	Hot anger, elation, happy, interest, pride, panic, anxiety, disgust, contempt, cold anger	Neutral	Sadness, shame, despair, boredom
EMA	Happy	Neutral	Angry, sad	Happy, angry	Neutral	Sad
GES	Happy	Neutral	Angry, disgust, boredom, sad, fear	Happy, angry, fear, disgust	Neutral	Sad, boredom
semaine	(0.2, 1]	[-0.2, 0.2]	[-1, -0.2)	(0.2, 1]	[-0.2, 0.2]	[-1, -0.2)
VAM	0.5, 1	0	-0.5, -1	0.5, 1	0	-0.5, -1
UM	Delight, flow, surprise	Neutral	frustration, confusion, boredom	frustration, surprise, delight, confusion, flow	Neutral	Boredom

labeling strategy in this study is shown in Table 20.

In Table 20, the four databases with categorical labels are classed into the three classes along valence and arousal based on the study of [14] (and common sense). The labels of VAM are numeric values [-1, -0.5, 0, 0.5, 1]. Therefore, 0 is considered as neutral, positive values are positive/high in valence/arousal, and negative a negative/low.

The process of emotion labels of SEMAINE is described as follows. The speech sample unit of SEMAINE is “turn”. One example of two turns is given below.

```
00:00:100 00:02:800 User "Oh, I’d like to speak to Poppy please".
00:02:900 00:06:100 Poppy "Hi. I’m poppy. Have we met before?"
```

Table 21: Sample selection strategy for SEMAINE.

Number of rators	The least number of rators with agreement (N)
2	2
3	2
5	3
6	4
7	5
8	6

The time stamp shown at the beginning of each turn is used to segment the speech into turns. The "user" speech was selected to analyze. Each emotional speech was evaluated by multiple raters. The number of raters vary from two to eight as shown in Table 21. The decision was made to select samples with N raters agreed, which is shown in Table 21 as well. The emotional labels of SEMAINE was continuous (with the step of 0.002). Therefore, the mapping procedure from values to three classes is:

1. Set up a threshold value, within which the VA values are considered as "neutral".

2. Along each dimension (valence/arousal), values larger than the *threshold* is considered as positive/high and the lower is negative/low.

3. Account the resulted number of samples in each classes, if the portion of neutral samples is smaller than 20% or larger than 40% (much unbalanced data), repeat from step one.

Finally, the *threshold* was set as 0.02 considering balancing the number of samples in each category.

Following the emotion mapping decision in Table 20, the number of samples for each category and database is shown in Table 22. The column of "Intersect neutral" is the number of samples labelled as neutral in both valence and arousal. Although valence and arousal are orthogonal as shown in Fig. 1 and 2, the samples considered as neutral in one dimension are regarded as neutral in another one. The categories are

Table 22: Number of samples for each emotional category.

DBs	Valence				Arousal				Intersect
	+	0	-	total	+	0	-	total	neutral
EPST	653	80	1630	2363	1624	80	659	2363	80
EMA	124	146	298	568	265	146	157	568	146
GES	71	79	385	535	313	79	143	535	79
SEMAINE	2361	1022	950	4333	2529	1023	781	4333	1022
VAM	19	522	364	905	256	472	177	905	349
UM	420	267	774	1461	927	267	267	1461	267

not balanced in either dimension with much less neutral samples. The number of the positive and negative categories in valence is more bias than arousal. This unbalance issue in binary classification is addressed by randomly selecting equal number of samples from each category and repeating the classification to gain the averaged output as the result.

6.2 Feature selection for each database

Five sets of features are used in the evaluation section, glottal, TEO, TECC, MFCC, and pitch related features. The glottal and TEO, TECC, and MFCC features were introduced in Chapter 5. The pitch features were obtained using the voicebox library[6]. The frame length is four times of the length of the averaged pitch cycle with a 10ms update step. As shown in Table 23, the 21 statistics including the derivative were applied to the features to generate a 1596-dimension feature set.

Feature selection was conducted for each database not only because it can reduce the dimension size, but also the selected features are of interest to examine. The features selected by each database could be considered as the acoustic representation of that database. The comparison between databases using the selected features is one means to show the similarity between databases. The feature selection was implemented by Sequential Forward Floating Selection (SFFS) of WEKA [35] for each database by 10-fold cross-validation self-training and testing. The evaluator of SFFS was to select the features possessing higher correlation with the emotion labels

Table 23: Summary of acoustic features, Low-Level Descriptors (LLDs) and statistics applied in hte evaluation of normalization with reference.

Feature			
Group	LLD	Abbr.	No.
glottal	time-based freq-based	see Table 2	189 42
TEO	(1)FM	fm1(mean),fm2(min), fm3(max), fm4(range),fm5(log-range), fm6(std), fm7(median), fm8(irq)	168
	(2)Env	evn1(0-2kHz),evn2(2-4kHz), evn3(4-6kHz), env4(6-8kHz)	84
	(3)CB-Env	cb$i$ (the i_{th} critical band)	336
TECC	24 coefficients	i (the i_{th} coefficient)	504
MFCC	12 coefficients	i (the i_{th} coefficient)	252
pitch			21

Statistics
mean, median, min(minimum), max(maximum), range std(standard deviation), irq(inter-quartile), q25(25%-quartile), q75(75%-quartile), fit1(slope), fit2(curvature), fit3(inflexion), Δmean, Δstd, Δmax, Δmin, Δrange, Δmedian, Δiqr, Δq25(Δ25%-quartile, Δq75(Δ75%-quartile)

and lower inter-correlation, which was not classifier related. The feature selection was conducted for each database by self-evaluation using 10-fold cross-validation. In the cross-validation, selecting features which were selected in more than 80% of the tests of 10-fold cross-validation were considered as the feature selection criterion used here and in the following chapters. In the following chapters, the same feature selection methodology was applied to each database after specific processing, e.g., after normalization in Chapter 7 Section 7.2.1. The detailed feature selection results are provided in Appendix C.

CHAPTER VII

CROSS-TRAINING AND TESTING WITH NORMALIZATION METHODS AND DATA FUSION TECHNIQUES

7.1 Methodology

7.1.1 Normalization Methods

Combining multiple databases can pose significant challenges for classifiers dependent on statistical training. Recording mismatches can make it difficult to determine whether observed statistical differences are caused by the variable of interest or anomalies of the environmental differences. Additionally, acoustic features capture the natural characteristics of speakers related to differences in gender, language, and culture [65, 22]. All of these factors make speaker normalization an important and necessary element of any cross-corpus study. Three methods of normalization are employed in two groups: the scaling, i.e., speaker normalization (SN), and normalization with the reference, i.e., the speaker normalization with reference (SR), and the neutral reference model (NRM). The flowchart of the three methods is shown in Fig. 11.

7.1.1.1 Scaling: Speaker Normalization

The processing of speaker normalization (SN) is shown in Eq. (15):

$$\hat{f_j^i} = \frac{f_j^i - mean(f_j^i)}{std(f_j^i)}, \tag{15}$$

where f_j^i is the j^{th} feature descriptor for speaker i across the samples of all four emotions and *std* refers to the standard deviation. The normalization was conducted within database. In this way, features from the same speaker will have the mean of

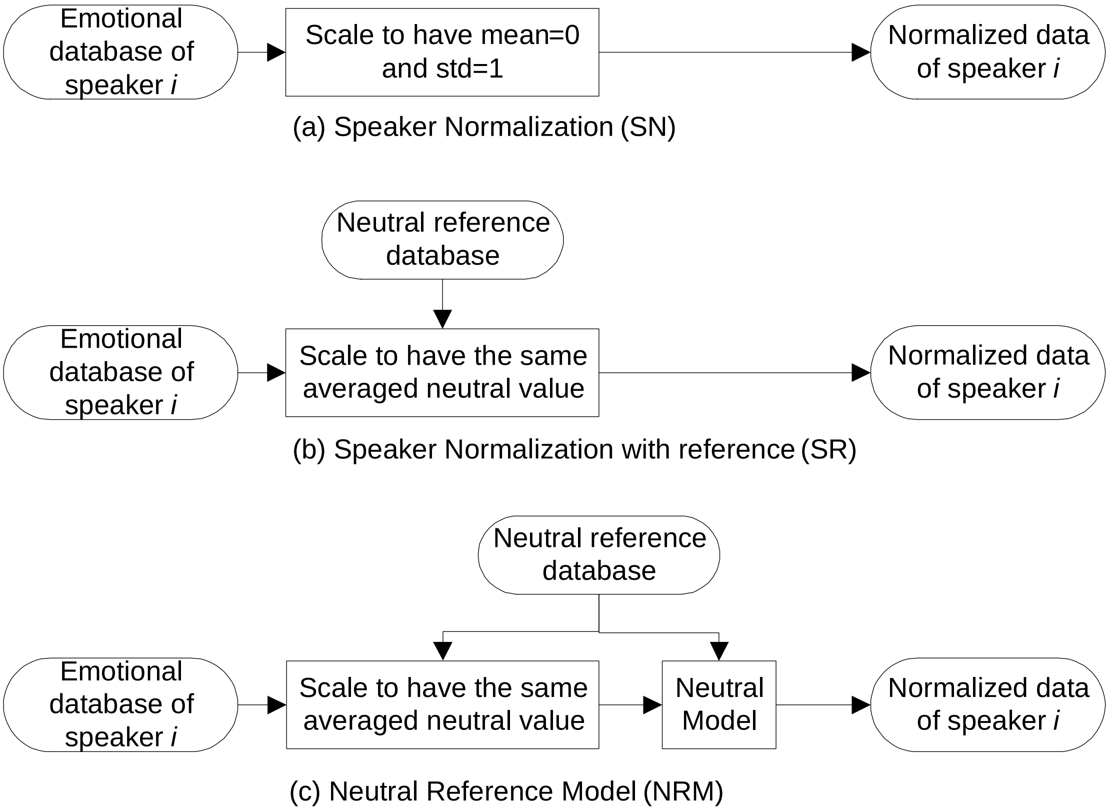

Emotional database of speaker *i*
Scale to have mean=0 and std=1
Normalized data of speaker *i*
(a) Speaker Normalization (SN)
Neutral reference database
Emotional database of speaker *i*
Scale to have the same averaged neutral value
Normalized data of speaker *i*
(b) Speaker Normalization with reference (SR)
Neutral reference database
Emotional database of speaker *i*
Scale to have the same averaged neutral value
Neutral Model
Normalized data of speaker *i*
(c) Neutral Reference Model (NRM)

Figure 11: The framework of (a) speaker normalization, (b) speaker normalization with reference, and (c) the neutral reference model.

zero and standard deviation of one. This normalization method has been used in previous chapters in this thesis and widely used in the literature [65, 77, 72, 73].

7.1.1.2 speaker normalization with reference

The equation of speaker normalization with reference (SR) is shown in Eq. (16):

$$\hat{f_j^i} = \frac{mean(f_{ref,j})}{mean(f_{neu,j}^i)} \cdot f_j^i, \tag{16}$$

where f_j^i is the j^{th} feature vector for speaker i across the samples of all emotions, $f_{neu,j}^i$ refers to the neutral samples of this speaker, $f_{ref,j}$ is the feature j of the reference neutral samples, and $mean$ represents the mean value. The goal of normalization process is to have the neutral samples of emotional data set scaled as the equal mean value as the reference data. This method was used in the work of [8]. It's still based on scaling at the speaker-level but involving the reference data.

As shown in Fig.11(b), the SR method of emotional recognition is to first extract and select the acoustic features from the same speaker in the emotional database, then scale the data from this speaker to make the neutral samples of the same speaker have the equal value to the mean of the reference neutral data. The scaling method introduced in Section 7.1.1.1 does not employ the reference data. The features are scaled to a mean of zero and standard deviation of one [65] just as described in Section 7.1.1.1. The SR method is considered as the intermediate stage between SN and the neutral reference model (NRM).

7.1.1.3 Neutral reference model

The neutral reference model (NRM) investigated in this study was originally proposed in [93] to analyze the emotional modulation observed in expressive speech. As shown in Fig. 11(c), first of all, the acoustic feature extracted from the emotional data was scaled by speaker normalization with reference (SR) as described in Section 7.1.1.2. For each speaker, the mean value of the neutral samples from this speaker was scaled to equal to the mean value of the neutral reference data. Then the neutral model (based on GMM in [8]) was trained with the neutral reference data set as the core of this methodology. The normalized acoustic feature from the emotional database was transformed to *fitness measurement* by calculating the likelihood score when applied to the trained neutral model. Finally, the *fitness measurements* calculated using the original features individually were classified instead of the acoustic features. In their follow-up studies involving this framework [8], GMM with two mixtures was set as the neutral model after the experimental comparison. In this thesis, the GMM with two mixtures are used, too.

7.1.2 Data Fusion Techniques

The methodology of data fusion could be considered in two directions: pre-fusion, i.e., before the classification, and post-fusion, i.e., after the classification. The difficulty of cross-database training and testing is the availability of "perfect" training data, which has enough diversity to represent all possible emotions in the testing data. Using as many as possible databases on hand to train the classifier may address the issue, however, will introduce another problem: data redundancy and unnecessary computation load. Therefore, an algorithm to select the "necessary" databases as the training data is proposed, named as Sequential Database Selection (SDS). In this way, a subset of databases from all candidates will be selected by SDS to train one classifier for the testing data, which will achieve the highest accuracy rate compared other combination of training sets. On the other way, one classifier was trained using each of the available training database. The final classification result is determined by all the outputs using each of the training data. The methodology to deal with the multiple output is developed from ROVER (Recognizer output voting error reduction). Other than simply majority voting, ROVER considers more factors such as confidence measurement of each classification and weights of each training databases together.

7.1.2.1 Pre-fusion: SDS

The algorithm of Sequential Database Selection (SDS) is similar to the Sequential forward feature selection (SFS). Starting with an empty set, SDS adds one database at a time after evaluating all candidate databases. The evaluator is the accuracy rate of Support vector machine (SVM) as the classifier. The database, by adding which can produce the highest accuracy rate, is added into the selected set. It's repeated till 1) all candidates are added into the selected set, or 2) the accuracy rate is not increased by adding any of the candidate. It's noticeable that the choice of the starting candidate could affect the selection results. Therefore, for each testing data, the SDS was conducted by starting with all candidate, respectively, to search the best selected training set.

1. Start with empty database set $D^i = \emptyset$, $i = 0$,
2. Select the additional database d_n which can maximize the criterion of $\Gamma(D^i)$, where $D^i = \{D^{i-1}, d_n\}$,
3. Update $D^i = \{D^{i-1}, d_n\}$, $i = i + 1$,
4. If (criterion is not increased) or ($i ==$ the number of candidate databases), End; else go to 2.

7.1.2.2 Post-fusion: ROVER

The method used to combine the different classification results from different training databases is based on the ROVER "Recognizer output voting error reduction" technique used in automatic speech recognition (ASR). ROVER [27] is used to combine the independent word results from different engines to reach a composite output in ASR. First, ROVER performed word alignment over different word recognized from different ASR engines. Then the final decision was made based on the combination of the frequency of the word recognized and the confidence of this word. In this paper, the word alignment processing can be skipped, and the score of each emotion candidate is calculated by Eq. 17,

$$Score(e_n) = \alpha \frac{N(e_n)}{Ns} + (1 - \alpha) Conf(e_n), \tag{17}$$

given Ns outputs of emotions and the corresponding confidence scores (i.e., the probability of the output class) from Ns classifications, $N(e_n)/Ns$ represents the frequency of candidate emotion e_n from all Ns outputs and $Conf(e_n)$ is the confidence measurement of the emotion candidate e_n. The mean value of all confidence scores of emotion (e_n) is used as the confidence measurement. The confidence score from each classifier is the probability estimation in support vector machine (SVM), which varies from 0 to 1. The parameter α is the weight to balance the frequency (i.e., majority voting) and the confidence. It's obtained by exhaustively searching from 0.1 to 0.9

with the step of 0.1. Finally, the emotion candidate e_n with the highest $Score(e_n)$ is decided as the label.

7.2 Evaluation

This section evaluates the normalization methods and data fusion techniques proposed in Section 7.1.1 and 7.1.2. The acoustic feature sets consist of pitch, glottal, TEO, TECC, and MFCC features as introduced in Section 6.2. Because the testing data in the "real-world" will be "blind", the feature set selected by the training data (the combination feature set if more than one training database) was used as the feature set for all the classifications. Section 7.2.1 presents the classification results from the use of the three normalization methods described in Section 7.1.1 (with the addition of classification results without the use of normalization) on cross-validation studies involving training/testing with a single database as well as pairwise training and testing (i.e., training performed on a different database than the one tested). In Section 7.2.2, the improvement of performance in cross-database emotion recognition brought by the systematic computational structure studied in this thesis will be presented. All the training and testing data in this evaluation section (Section 7.2) was randomly selected to have equal number of samples (i.e., rate by chance is 50%). The classification using randomly selected samples was repeated 50 times and the average value was the value listed in the tables in this section. A Kruskal-Wallis (K-W) test was conducted on every combination of classification results to highlight differences with a significance level of 5% ($p < 0.05$). The K-W test was chosen over ANOVA because a Jarque-Bera test at a significance level of 5% ($p < 0.05$) indicated the classification results did not follow a normal distribution. The results of the Jarque-Bera tests for normality are shown in Table 24. Bonferroni correction was used for all K-W tests involving more than two factors.

Table 24: The classifications whose accuracy rate variables reject the null hypothesis of Jarque-Bera test with the significant level of 5% ($p < 0.05$)(i.e., not fit the normal distribution).

Emotions	Dim	training data type	testing data	normalization	p-value
neutral vs. emotional	VA	self-testing	EMA	SR	0.0480
	V	best other	EPST	NA	0.0136
	A	best other	EPST	SR	0.0001
positive vs. negative	V	best other	GES	SR	0.0174
	A	best other	VAM	SN	0.0235
	A	best other	UM	SR	0.0301

7.2.1 Normalization methods

The goal of this section is to evaluate the three normalization methods by comparing with no-normalization. By the results from this section, the question of why cross-database emotion recognition can benefit from normalization will be answered. Also, comparing the results using the three methods described in Section 7.1.1 will show the performance of each method in particular application (i.e., emotion choice and database choice). The experiments were conducted in binary classification of neutral vs. emotional (in valence and arousal) and positive vs. negative (in valence and arousal). All the training and testing data was randomly selected to have equal number of samples (i.e., rate by chance is 50%). The classification using randomly selected samples was repeated 50 times and the average value was the value listed in the tables in this section. In this section, the TIMIT database was used as the neutral reference data for (b) SR and (c) NRM in Fig. 11. TIMIT [30] is the widely used read-speech database in English. It recorded 4620 sentences from 460 speakers with the sampling frequency 16kHz.

Tables 25-28 shows the results of recognizing neutral vs. emotional and positive vs. negative in two experiments: “self” and “best other”. “Self” represents training and testing on the same database by 10-fold cross-validation. “Best other” lists the highest accuracy rate for each testing database that could be achieved by training another database. More details of the “best other” training data are presented in Appendix B.

Table 25: The accuracy rate of **neutral vs. emotional (Valence)** classification using four normalization methods (NA represents no normalization). For each testing database, the **bold** values are the highest values with statistically significant difference ($p < 0.05$) from the un-bold values in "self" and "best other" experiments separately.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
self: train and test on the same data by 10-fold cross-validation						
NA	83.4	92.3	84.8	64.8	62.5	57.4
SN	84.8	95.0	89.0	58.7	61.6	60.3
SR	86.6	**96.2**	92.7	71.0	67.9	**63.6**
NRM	**88.4**	**96.1**	**96.2**	**76.1**	**69.4**	**62.1**
best other: the highest AR when train on another database						
NA	66.9	70.9	73.3	**58.6**	56.1	50.5
SN	67.1	73.5	82.5	52.5	55.4	51.5
SR	73.2	72.9	78.4	54.7	54.3	57.2
NRM	**83.7**	**76.7**	**90.2**	57.3	**57.4**	**59.6**

The entry in bold represents the highest rate that is statistically significant from the others utilizing the K-W test at a significant level of $p < 0.05$. For each testing database, the comparison includes four accuracy rates. The null hypothesis for the significance test (K-W test, 5% significant level) is there is no difference among the four accuracy rates (the three normalization methods + no-normalization), however, if the null hypothesis is rejected, the interpretation is too general. Therefore, a post hoc test were used. With Bonferroni correction, the multiple comparison results were given in a pair-wise manner. And the significant test results were interpreted in the following way. In Tables 25-28, among the four accuracy rates, the bold values inside one column are the highest values with the statistically significant difference with the un-bold values. When more than one values are bold in one column, it means all the bold values are statistically the highest accuracy rates (no significant difference between the bold values). Between the un-bold values, the significant different may or may not exhibit, which is not the focus of this thesis because they are not the higher accuracy rates.

When differentiating emotional status from neutral, most entries in Table 25 and 26 are the same. The reason for this is that the samples labeled as neutral/emotional

Table 26: The accuracy rate of **neutral vs. emotional (Arousal)** classification using four normalization methods (NA represents no normalization). For each testing database, the **bold** values are the highest values with statistically significant difference ($p < 0.05$) from the un-bold values in “self” and “best other” experiments separately.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
self: train and test on the same data by 10-fold cross-validation						
NA	83.5	93.4	86.5	64.5	64.8	57.4
SN	85.3	93.1	90.9	58.7	64.6	60.1
SR	86.7	**96.6**	91.3	71.1	**73.0**	**63.7**
NRM	**87.9**	**96.8**	**96.1**	**76.4**	**73.5**	**62.2**
best other: the highest AR when train on another database						
NA	65.5	67.9	72.0	57.7	53.7	50.8
SN	67.5	73.9	81.7	52.5	55.2	51.2
SR	73.3	72.2	77.9	53.3	56.3	55.0
NRM	**83.8**	**77.3**	**89.7**	**59.0**	**61.4**	**58.0**

in valence is also neutral/emotional in arousal (refer to Table 20 and 22). From Tables 25-26, NRM gives the highest accuracy rates for at least five out of six databases in both “self” and “best other” tests. Especially when the training and testing databases are different (i.e., “best other”), using NRM improves the accuracy rate significantly than no-normalization (NA).

Tables 27-28 shows the accuracy rate of classification between positive and negative emotions in valence and arousal. Different from neutral vs. emotional, the normalization method achieving the highest accuracy rate depends on the testing database. Comparing using the results with and without normalization, in most cases, classification using at least one normalization method can produce statistically significant higher accuracy rates than without normalization.

Based on the results comparing normalization methods with no-normalization, both the self-testing and the cross-database training and testing can achieve significantly higher accuracy rate using at least one normalization method for most testing databases. In cross-database test, when the accuracy rate is lower than self-testing, using normalization can improve the performance to as good as that of self-testing

Table 27: The accuracy rate of **positive vs. negative (Valence)** classification using four normalization methods (NA represents no normalization). For each testing database, the **bold** values are the highest values with statistically significant difference ($p < 0.05$) from the un-bold values in "self" and "best other" experiments separately.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
self: train and test on the same data by 10-fold cross-validation						
NA	**67.4**	**91.7**	78.7	77.8	70.3	59.6
SN	**67.5**	**92.5**	**82.1**	67.7	66.4	61.8
SR	**67.2**	90.2	77.7	86.9	73.7	65.4
NRM	66.7	91.3	78.5	**90.7**	**80.2**	**66.7**
best other: the highest AR when train on another database						
NA	59.6	72.5	72.9	**60.7**	55.8	48.2
SN	**61.3**	**75.7**	**76.5**	51.6	**59.6**	52.7
SR	59.8	70.0	70.7	55.7	58.8	**53.8**
NRM	56.2	71.7	70.1	53.7	58.3	51.6

Table 28: The accuracy rate of **positive vs. negative (Arousal)** classification using four normalization methods (NA represents no normalization). For each testing database, the **bold** values are the highest values with statistically significant difference ($p < 0.05$) from the un-bold values in "self" and "best other" experiments separately.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
self: train and test on the same data by 10-fold cross-validation						
NA	75.5	98.9	97.7	66.7	**92.5**	55.6
SN	**76.6**	96.8	**98.7**	56.1	88.2	60.8
SR	75.1	**99.7**	97.5	76.5	**92.2**	62.8
NRM	71.8	97.8	96.8	**81.0**	88.8	**67.4**
best other: the highest AR when train on another database						
NA	69.5	**92.2**	**90.5**	53.6	**79.9**	50.8
SN	68.8	**91.0**	**92.4**	51.3	76.4	**55.0**
SR	**69.9**	85.8	90.8	**55.5**	74.5	50.0
NRM	64.2	82.9	89.7	53.7	74.6	52.2

(no-normalization) for some databases (e.g., EPST in Table 29 and 30, UM in Table 32) or even higher (e.g., GES and UM in Table 29 and 30). This results indicate the importance of normalization in cross-database emotion recognition. Comparing the three normalization methods, NRM helps improve the accuracy rate for both self-testing and cross-database classification when differentiating between neutral and emotional. To detect positive and negative motions, SN was highlighted based on the frequency to provide the highest accuracy rates for each testing database.

7.2.2 Improvement by combining normalization and data fusion

In this section, the two data fusion techniques are included to evaluate the system using acoustic features summarized in Section 6.2 as well as the three normalization methods discussed in Section 7.2.1. Combining all the efforts (i.e., acoustic features, normalization and data fusion), the highest accuracy rate the system can achieve will be compared with the baseline (the “best other” without normalization) and shown in Tables 29-32. Same as the evaluation of the normalization methods, the experiments in this section are conducted in binary classification of neutral vs. emotional (in valence and arousal) and positive vs. negative (in valence and arousal). The Kruskal-Wallis test ($p < 0.05$) is used to test the statistically significance of the differences in classification methods.

Three data fusion techniques were evaluated in this section, ROVER with database (Rover-db), ROVER with database and feature set (Rover-db*f), and SDS. “Rover-db” represents the results using ROVER based on databases using all feature sets together. “Rover-db*feat” is calculated using ROVER based on the combination of databases and each of the five feature sets. For example, for one test set, “Rover-db” evaluates five recognition engines from the other five databases, and “Rover-db*feat” evaluates 30 engines resulted from the combination of five other databases and six feature sets (five sets + all five together). “SDS” means the results using SDS as

Table 29: The improvement of the accuracy rate for **neutral vs. emotional (Valence)** classification combining normalization methods and data fusion techniques. All Δs are statistically significant by $p < 0.05$.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
baseline	65.5	67.9	72.0	57.7	53.7	50.8
highest AR	89.2	91.3	92.8	62.1	64.7	60.9
normalization	NRM	NRM	NRM	SN	NRM	NRM
datafusion	R-db*f	R-db*f	R-db*f	R-db*f	R-db*f	R-db*f
Δ	23.7	23.4	20.8	4.4	11.0	10.1

Table 30: The improvement of the accuracy rate for **neutral vs. emotional (Arousal)** classification combining normalization methods and data fusion techniques. All Δs are statistically significant by $p < 0.05$.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
baseline	66.9	70.9	73.3	58.6	56.1	50.5
highest AR	89.2	91.3	92.8	62.1	64.7	60.9
normalization	NRM	NRM	NRM	SN	NRM	NRM
datafusion	R-db*f	R-db*f	R-db*f	R-db*f	R-db*f	R-db*f
Δ	22.3	20.4	19.5	3.5	8.6	10.4

the data fusion method. The bottom row of "Δ" is the increase of accuracy rate from the baseline "best other" in "NA" to the highest accuracy rate achieved. The detailed classification results with fully combination of normalization and data fusion techniques are available in Appendix B. In this section, the baseline and the best performance information is abstracted from Appendix B to make the explicit present of the improvement brought by the proposed systematic computational structure.

Table 31: The improvement of the accuracy rate for **positive vs. negative (Valence)** classification combining normalization methods and data fusion techniques. All Δs are statistically significant by $p < 0.05$.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
baseline	59.6	72.5	72.9	60.7	55.8	48.2
highest AR	65.1	77.8	77.7	62.2	66.1	56.4
normalization	SN	SN	SN	NA	SN	SN
datafusion	R-db*f	SDS	R-db*f	SDS	SDS	SDS
Δ	5.5	5.3	4.8	1.5	10.3	8.2

Table 32: The improvement of the accuracy rate for **positive vs. negative (Arousal)** classification combining normalization methods and data fusion techniques. All Δs are statistically significant by $p < 0.05$.

TEST:	EPST	EMA	GES	SEMAINE	VAM	UM
baseline	69.5	92.2	90.5	53.6	79.9	50.8
highest AR	71.5	93.0	94.2	56.1	84.9	54.6
normalization	SN	SN	SN	SR	SR	SN
datafusion	SDS	SDS	R-db*f	R-db*f	R-db*f	SDS
Δ	2.0	0.8	3.7	2.5	5.0	3.8

In Table 29-30, the highest accuracy rates are achieved by ROVER with the combination of database and feature ("Rover-db*f") for the EPST, EMA, GES, VAM, and UM databases with NRM. The increase of accuracy rates ranges from 4.4% to 23.7% for valence and 3.5% to 22.3% for arousal, which is all statistically significant by Kruskal-Wallis test($p < 0.05$). This observation highlights the improvement introduced by ROVER for all databases and the neutral reference model normalization for most databases in neutral vs. emotional recognition.

When the two categories are positive and negative (i.e., no neutral involved), the results are shown in Table 31 for valence and Table 32 for arousal. By acoustic features only, valence is more difficult than arousal to recognize. Examining the values in Table 31 for valence. The overall accuracy rates for valence are relatively lower than arousal in Table 32. The highest values are achieved by using SN for five databases and using SDS for four databases. The improvement of accuracy rate is up to 10.3%.

Comparing with valence, the accuracy rates of arousal shown in Table 32 are higher. Although the values are higher, the pattern is similar. The best performance is obtained by speaker normalization (SN) for four databases except for SEMAINE and VAM, which exhibit the best performance by the speaker normalization with reference (SR). Using SDS, half databases gain the highest accuracy rates with SN.

Another half benefit from the ROVER with database and feature method to combination information from multiple training data sets. The increase of accuracy rate ranges from 0.8%-5.0%. One reason for this is that the baseline rate is fairly high (EMA and GES are over 90%, VAM is close to 80%) so less space left for the improvement. Another reason can be detected from the comparison of the evaluation between the normalization methods in Section 7.2.1. Unlike the experiments of neutral vs. emotional, the NRM method does not increase the results much for positive vs. negative. This indicates NRM, which gained much improvement in neutral vs. emotional, can hardly represent the acoustic difference between positive and negative status as well as in neutral vs. emotional.

To sum up, ROVER with the combination of databases and feature sets improved the performance of cross-database training and testing when detecting the emotional samples from neutral by up to 23.7% for valence and 22.3% for arousal. Using SDS, the higher accuracy rate than training on any single database ("best other") could be guaranteed. SDS only outperformed ROVER in the recognition between the positive and negative emotions for a set of databases. When detecting the emotional from the neutral, NRM gained much improvement by transferring the features into the fitness measurements based on the neutral GMM. However, the advantage of NRM did not exhibit in the classification between positive and negative. Although the fitness measurement also represented the distance from one emotional sample to neutral, the category inside the emotional group (i.e., positive and negative) could not be captured by the measurements obtained by NRM using the neutral reference data. For detecting positive and negative, the normalization method producing highest accuracy rates for the most testing databases was scaling the samples from one speaker to have zero mean and the standard deviation of one (SN).

CHAPTER VIII

CONCLUSION

8.1 Research Summary

The objective of the research presented in this thesis was to systematically investigate the computational structure for cross-database emotion recognition. The research consisted of evaluating the stability of acoustic features, particularly the glottal and Teager Energy based features, and investigating three normalization methods and two data fusion techniques. One of the challenges of cross-database training and testing is accounting for the potential variation in the types of emotions expressed as well as the recording conditions. In an attempt to alleviate the impact of these types of variations, three normalization methods on the acoustic data were studied. The results showed that in cross-database test, using normalization improved the performance to as good as that of self-testing (no-normalization) for some databases or even higher. Comparing the three normalization methods, NRM helped improving the accuracy rate when differentiating between neutral and emotional and SN was highlighted for positive vs. negative classification based on the frequency to provide the highest accuracy rates for each testing database. Motivated by the lack of large and diverse enough emotional database to train the classifier, using multiple databases to train posed another challenge: data fusion. This thesis proposed two data fusion techniques, pre-classification SDS and post-classification ROVER to study the issue. Using the glottal, TEO and TECC features, of which the stability of emotion distinguishing ability has been highlighted on multiple databases, the systematic computational structure proposed in this thesis could improve the performance of cross-database binary-emotion recognition by up to 23% for neutral vs. emotional and 10% for positive vs. negative. The details of research conducted were discussed as follows.

Unlike the glottal and Teager Energy based features, the prosodic features (i.e., pitch, energy, etc.) are the widly used cues of speech analysis in literature [14, 8]. However, certain pairs of emotions remain difficult to discriminate due to the similar displayed tendencies in prosodic statistics. Thirty emotion pairs showing no statistically significant difference by prosodics from the EPST database (15 acted emotions in total) was targeted by the Kruskal-Wallis test [77]. The statistically significant difference could be observed from at least one glottal feature for all 30 emotion pairs and 19 out of 30 pairs had four or more of the seven glottal features (i.e., ClQ, NAQ, OQ, OQa, SQ, HRF, and DH12). Evaluation involving classification between emotion pairs using the quadratic classifier and one feature per classification showed a lower error rate could be achieved by a glottal feature in 24 out of 30 pairs. Combining the prosodic and glottal features together, a further study using SFS (Sequential Forward Selection) with the same quadratic classifier as the evaluator listed the top five features for each emotion pair based on the frequency of selection in 10-fold cross-validation. It was found that HRF was the most prominent feature for female speakers while for the male OQa, ClQ, and NAQ were the most prominent glottal features.

While it's not assumed that acted speech provides a complete picture of authentic emotion, the aforementioned study using the EPST database represents the emotions which the actors adjust to fit their own perception of emotions. The emotion distinguishing ability of the glottal features was evaluated using authentic emotional data recorded when students were interacting with an automatic computer tutoring system, named Auto Tutor (referred to as UM database in this thesis). Multimodal emotion recognition study has been taken about Auto Tutor system, e.g., facial expression, body gesture, and speech. However, in the speech channel, the features for emotion detection is conversational cues, which consisted of five aspects of information: temporal, response, answer quality, tutor directness, and tutor feedback. The lack of examining acoustic cues motivated the study of [72]. More glottal measurements (AQ, OQ2, SQ2, and QOQ) were added to make 11 glottal features in total. Another acoustic set was obtained by openSMILE [24], which consisted of energy, spectral, and voice quality related features. To detect five emotions in UM database, the comparison showed that *flow* and *boredom* were better captured in acoustics than conversational cues.

As introduced in Chapter 2, emotion categories (e.g., happy, sad) and emotion dimensions (e.g., 0.5 in valence/arousal) are the two emotion description methods mainly used. The aforementioned study of the EPST and UM data focused on emotion categories, emotions in dimensions were examined using SEMAINE data in four dimensions: valence, arousal, expectation, and power [73]. The acoustic feature set consisted of the Teager Energy Operator (TEO) based features as well as the glottal measurements and the features by openSMILE. Three sets of features were derived based on TEO: the variation of FM (Frequency Modulation) component, the normalized autocorrelation envelope area, and the critical band based autocorrelation envelope area. Using the three sets of TEO, glottal, and openSMILE features separately, the binary-classification was conducted in four emotion dimensions by SVM (linear kernel) and the output accuracy rates were compared. The comparison results showed that using the same classification methodology, TEO and glottal features outperformed or performed similarly to the openSMILE set. The result also highlighted the discrimination ability of TEO in emotion dimensions arousal and power comparing with the other two dimensions.

An additional study was conducted to investigate the acoustic features from the speech with noise. Motivated by the robust performance of Mel-Frequency Cepstrum

Coefficients (MFCCs) in speech analysis under noisy condition, Teager Energy Cepstrum Coefficients (TECCs) were developed combining the concepts of TEO and MFCCs [74]. The extraction of TECC was similar to MFCC but using TEO instead of the squared energy as the primary difference. The experiments involved three databases, EPST, EMA, and GES, all of which were recorded in a lab environment to minimize noise (referred to as "clean" speech). The four emotions consistently presented in all three databases were studied, i.e., neutral, happy, sad, and angry. The noisy speech was obtained by adding white Gaussian noise to the"clean" speech at five Signal Noise Ratio (SNR) levels from 20dB to 0dB with the step of 5dB, which leaded to six data set with different noise levels including the "clean" set from each emotional database. TECCs were extracted from all data sets, individually. The performance of TECC was evaluated and compared with MFCC in two ways: the Normalized mean square error (NMSE) and recognizing four emotion status. NMSE measures the average difference between the feature from noisy speech to from the clean speech. The smaller the value of NMSE is, the more robust the feature possess. The results showed that almost at all noisy levels, the NMSE value of TECC was only half of the corresponding MFCC. The conclusion was reached that TECC was more robust than MFCC facing additive noise in emotional speech. Furthermore, to evaluate the emotion distinguishing ability of TECC, emotion recognition experiments were conducted. In the four-emotion classification using SVM (linear kernel), TECC achieved higher averaged accuracy rate (across different noise levels) than MFCC for all three databases. The averaged accuracy rates of TECC were 61.1% for EPST (52.3%-MFCC), 84.3% for EMA (79.1%-MFCC), and 84.1% for GES (79.6%-MFCC) in four-emotion classification. In the pair-wise classification, the average accuracy rate could reach 99% using TECC only (angry-sad, and happy-sad in GES).

In the interest of the knowledge of how much difference in accuracy rate could be observed between cross-database and self-cross-validation, a preliminary study [75] involving three emotional databases (EPST, EMA, and GES) was conducted using glottal and also pitch-related features for the comparison purpose. Four-emotion classification was implemented in two ways: training on one database and test on another one (referred to as "cross-testing"), and train and test on the same data by cross-validation (referred to as "self"). The results [75] suggested that the glottal features were more stable in the 4-emotion classification system over three databases and were able to perform well above chance for several of the cross-testing experiments. Overall, the difference in the accuracy rates between cross-testing (lower) and self-testing (higher) was observed using both the glottal and pitch features. However, when the cross-testing occurred between EMA and GES, the accuracy rate was higher (0.60) than other cases (i.e., close to the rate by chance in most cases, 0.30).

In the effort to perform cross-database training and testing, one problem is about the emotion labels. Six emotional databases are employed for cross-database emotion study, each of which has its own emotion labeling strategy, in categories (different choice of emotional status words) or in dimensions (different resolution, steps, and number of dimensions). To solve the emotion label problem, the decision was made to map all labels to a three-class way in valence/arousal, i.e., positive, neutral, and negative in valence/arousal.

To establish the generalized result involving multiple databases, the fact should be faced that the speech recordings are affected by many factors, such as the recording conditions, the naturalness of emotion expressed (acted or authentic), the language, and the basic information of speakers (e.g. age, culture, and gender). Recording mismatches can make it difficult to determine whether observed statistical differences are caused by the variable of interest or anomalies of the environmental and speaker differences. All of these factors make speaker normalization an important and necessary element of any cross-corpus study. Three methods of normalization were employed in this study: speaker normalization (SN) (i.e.,scaling to make the samples belonging to one speaker have zero mean and one standard deviation), the speaker normalization with neutral reference (SR) (scaling to make the neutral samples have the same mean value as the reference neutral samples from another database), and the neutral reference model (NRM)[76] . The neutral model was trained with the neutral reference data set. The pre-normalized acoustic features from emotional databases were transformed to *fitness measurements* by calculating the likelihood score when applied to the trained neutral model. Finally, the *fitness measurements* were classified instead of the acoustic features. Gaussian Mixture Model (GMM) with two mixtures was chosen as the neutral model. Acoustic feature sets consisted of pitch, MFCC, TECC, TEO, and glottal based features. Results showed that, in most cases of comparing the acoustic feature sets, the NRM outperformed SN and SR for classification between neutral vs. emotional. However, for recognizing status between positive vs. negative, SN showed the best results when tested on most databases. Overall, using normalization improved the performance of cross-database classification even to a higher accuracy rate than the self cross-validation without normalization. This results indicated the importance of normalization in cross-database emotion recognition.

The final phase of this research was to investigate the methodology of how to combine the information obtained from different training sets, referred to as the data fusion techniques. The methodology of data fusion could be considered in two directions: pre-fusion, i.e., before the classification, and post-fusion, i.e., after the classification. The difficulty of cross-database training and testing is the availability of a large enough training data, which has enough diversity to represent all possible emotions in the testing data. Using as many as possible databases on hand to train the classifier may address the issue, however, introduces another problem: data redundancy and unnecessary computation load. Therefore, an algorithm to select the "necessary" databases as the training data was developed, named Sequential Database Selection (SDS). In this way, a subset of databases from all candidates was selected by SDS to train one classifier for the testing data, which achieved the highest accuracy rate compared with any other combination of the training sets. The pre-fusion reduced the data redundancy and the computational load by reducing the data size. One possible issue caused by the reduction was that the loss of data might lower the highest accuracy rate which could be possibly reached with all the available data. To avoid the data loss, the post-fusion technique ROVER was developed to combine the information gained from each database. For post-fusion, the final classification result was determined by all the outputs of classifications using each of the training data (and feature sets). The methodology to deal with the multiple classification outputs was ROVER. ROVER (Recognizer output voting error reduction) considered more factors such as confidence measurement of each classification and weights of each training databases. For the binary classification between neutral and emotional, comparing all the normalization methods and data fusion techniques, the highest accuracy rates were achieved by ROVER with NRM for EPST, EMA, VAM, and UM databases, ROVER with SR for GES, and ROVER without normalization for SEMAINE. This observation showed the improvement caused by ROVER for all databases and the neutral reference model normalization in most cases. Examining the values in positive vs. negative, the highest values were achieved by both ROVER (for 2 testing databases in valence and 3 testing databases in arousal) and SDS (for 4 testing databases in valence and 3 in arousal). However, speaker normalization (SN) produced the highest accuracy rates for in most cases. Overall, combining the effort from normalization and data fusion techniques, the accuracy rate of cross-database training and testing was improved 3-23% (absolute value here and in the following text) for neutral vs. emotional in valence and arousal (varies based on the tested database), up to 10% for positive vs. negative in valence, and up to 5% for positive vs. negative in arousal.

8.2 Contributions and Future Work

The available publication on the cross-database emotion recognition research focuses on comparing the statistics of acoustic features between databases directly, reporting the benchmark results of cross-database classification, or selecting the prototypical data samples as the training data. This thesis systematically studied the computational structure of cross-database emotion recognition focusing on investigating three main challenges, the stability of acoustic features, the normalization methods, and the data fusion techniques. The contribution of this thesis consists of the following:

- *Evaluating the stability of emotion distinguishing ability of the glottal and Teager Energy based features using multiple databases.* The report of the glottal and Teager energy features in the field of cross-database evaluation was very limited. This thesis studied the stability of the two sets of features and showed the stable emotion distinguishing performance across multiple databases in multiple emotion recognition experiments.

- *The investigation of three normalization methodologies at the speaker-level.* In emotion recognition using multiple databases, normalization was considered of interest and has been studied in the manner of self-scaling at different levels, e.g., speaker-level, database-level. This thesis investigated three normalization methods, two of which involving a reference neutral data. The comparison results showed that using normalization could significantly improve the cross-database classification performance.

- *The development of two data fusion techniques.* This thesis proposed the Sequential Forward Database Selection (SDS) and developed the Recognizer output voting error reduction (ROVER) approaches to combine information gained from different training databases due to the lack of large and diverse enough emotional database for training. Using SDS as the pre-classification technique to choose the combination of training databases can guarantee the same or better performance than training on any single database. Using ROVER as the post-classification technique to efficiently combine the classification results using multiple training databases can achieve better performance than simply majority voting (the most commonly used method for data fusion).

- *The improvement on cross-database classification performance.* Combining all the efforts made in this thesis to investigate the challenges, the systematic computational structure proposed in this thesis could improve the accuracy rates of the cross-database classification significantly by up to 23% in detecting neutral vs. emotional and up to 10% in positive vs. negative. When testing the GES database, the similar research results could be found in [66] of 68% for arousal and 54% for valence in positive vs. negative classification. The accuracy rates using the systematic structure in this thesis are 93% for arousal and 78% for valence. And when GES was tested for neutral vs. emotional, [8] reported the accuracy rate of 80.2%, while the performance of the systematic structure in this thesis is 92.8%.

Future work will include the development of normalization in a speaker-independent way because the it will require less information from the speech data and make the emotion recognition engine have more generalized real-life applications. Furthermore, the normalization model to process positive vs. negative emotion need to be studied since the current neutral reference model works much better for detecting the emotional samples from the neutral. Training the GMM using emotional data instead of the neutral data could be a direction to address this issue. However, the lack of emotional database with the emotion at the controlled degree arises as the challenge for it.

APPENDIX A

THE GLOTTAL FEATURES FROM TELEPHONE QUALITY SPEECH WITH CODEC G.729

The glottal features were investigated in Chapter 5.1.4 to show better performance in cross-database emotion recognition compared with the pitch-related features. The three databases used were EPST, EMA, and GES. Results showed that EMA and GES possessed higher accuracy rates. However, all the databases were recorded in the lab environment at last 16kHz. Considering the potential application of detecting emotions in recordings at the telephone quality level, e.g., Call Center Speech, the study of the emotion distinguishing ability of the glottal features was extended to conduct the cross-database experiments using the telephone quality speech.

For comparison reason, EMA and GES were employed in this study (EPST was excluded due to its relatively lower accuracy rate). The telephone quality speech was obtained by encoding the original speech to bitstream and then decoding the bitstream to the speech. The codec to generate the telephone quality speech was G.729. Given the telephone quality speech, following the same methodology introduced in Chapter 5.1.4, 4-emotion classification was conducted using the glottal features extracted form the telephone quality speech and the results are shown in Table 33.

In Table 33, comparing the performance of the glottal features extracted from the telephone quality speech (“G729-GLO”) and the original speech (“GLO”), the accuracy rate of G729 speech is slightly lower than the original. The accuracy rates are fairly close for the cross-database experiments. The patterns observed from precision and recall are similar as well. Comparing “G729-GLO” with the pitch-related features extracted from the original speech (“PCH”), the accuracy rates using the glottal

Table 33: The results of 4-class emotion categorization using the glottal features extracted from the telephone quality speech ("G729-GLO"), and the glottal("GLO") and pitch ("PCH") features from the original speech (as seen in Table 12).("*" indicates the self-tested experiments using 10-fold cross validation, the self-tested accuracy rates are in **bold**).

Train: EMA							
Test:		***EMA***			GES		
		G729-GLO	GLO	PCH	G729-GLO	GLO	PCH
accuracy rate		**0.73**	**0.80**	**0.66**	0.63	0.65	0.49
precision	Happy	0.70	0.74	0.68	0.44	0.52	0.39
	Angry	0.65	0.72	0.64	0.72	0.76	0.64
	Sad	0.77	0.84	0.62	0.58	0.57	0.41
	Neutral	0.78	0.88	0.72	0.73	0.75	0.60
recall	Happy	0.62	0.72	0.78	0.49	0.56	0.55
	Angry	0.63	0.72	0.45	0.63	0.66	0.39
	Sad	0.89	0.89	0.64	0.78	0.89	0.61
	Neutral	0.74	0.84	0.80	0.64	0.52	0.51

Train: GES							
Test:		***GES***			EMA		
		G729-GLO	GLO	PCH	G729-GLO	GLO	PCH
accuracy rate		**0.79**	**0.83**	**0.71**	0.58	0.57	0.33
precision	Happy	0.63	0.74	0.71	0.55	0.55	0.47
	Angry	0.75	0.79	0.66	0.52	0.52	0.31
	Sad	0.95	0.92	0.84	0.83	0.76	0.19
	Neutral	0.84	0.89	0.71	0.57	0.56	0.41
recall	Happy	0.46	0.55	0.07	0.49	0.64	0.06
	Angry	0.85	0.90	0.94	0.74	0.70	0.72
	Sad	0.88	0.90	0.74	0.38	0.34	0.07
	Neutral	0.92	0.90	0.87	0.73	0.63	0.48

Table 34: The correlation coefficients of features extracted from speech and from the corresponding telephone quality speech with codec G.729.

	pitch	TEO	TECC	MFCC	glottal
EMA	0.961	0.963	0.931	0.944	0.961
GES	0.970	0.955	0.917	0.928	0.953

features from the telephone quality speech is higher than using the pitch features from the original speech, for both self-test and cross-database cases. The observation form Table 33 strengthens the emotion distinguishing ability of the glottal features extracted from the telephone quality speech. It indicates that the application of detecting emotion in telephone recordings using the glottal features is feasible.

As a slight extension of the aforementioned study, the correlation coefficients between the acoustic features extracted from the telephone quality speech and the original speech were calculated and listed in Table 34. The correlation coefficients were calculated using Eq. 18,

$$\rho(\nu_1, \nu_2) = \frac{cov(\nu_1, \nu_2)}{\sqrt{cov(\nu_1, \nu_1) \cdot cov(\nu_2, \nu_2)}}. \tag{18}$$

where *v1*, *v2* represent feature sets from the original sample and the corresponding telephone quality speech sample. $cov(v1, v2)$ is the covariance between the two. For each database, the averaged value across all the samples was listed in Table 34.

In Table 34, among the five feature sets, the pitch, TEO, and glottal features exhibit the larger correlation coefficients while TECC has the lowest for both databases. Even though, the correlation coefficients are all above 0.90, which indicates that the two sets are strongly correlated.

APPENDIX B

THE DETAILED RESULTS FOR CROSS-DATABASE TRAINING AND TESTING

The classification results using data without any normalization or data fusion techniques are the baseline to compare the performance of the proposed system (marked as "best other" in the block of "NA" in Tables 35 to 38). Each table contains three blocks, "NA", "SN", "SR", and "NRM", representing normalization processing of none, speaker normalization, speaker normalization with reference, and neutral reference model, in order. Inside each block, "self" is the classification trained and tested on the same data by 10-fold cross-validation. "Best other" represents the highest accuracy rate achieved by training the classifier on another database, the ID of which is listed in "best db". "Rover-db" shows the results using ROVER based on databases and all features together and "Rover-db*feat" is calculated using ROVER based on the combination of databases and each of the five feature sets. For example, for one test set, "Rover-db" evaluates five recognition engines from the other five databases, and "Rover-db*feat" evaluates 30 engines resulted from the combination of five other databases and six feature sets (five sets + all five together). "SDS" shows the results using SDS as the data fusion method, and "dbs" is the ID of databases selected by SDS to train the classifier. The last row of "Δ" is the increase of accuracy rate from the baseline "best other" in "NA" to the highest accuracy rate achieved. Tables 35 to 38 represents the binary emotion recognition in valence and arousal between neutral vs. emotional (Tables 35 for valence, Table 36 for arousal) and between positive vs. negative (Tables 37 for valence and Table 38 for arousal). In each table, the highest accuracy rate is marked in bold and the baseline to compare the performance is underlined. The classifier to implement all the experiments is SVM with liner kernel and the two classes in each binary classification was randomly chosen to have equal number of samples to make the accuracy rate 50% by chance. The values shown in tables are the mean values over 100 repeated classifications.

Table 35: Accuracy rates of binary classifications (**neutral vs. emotional in valence**) tested on each emotional database when the training data is the test data itself ("self"), one of the other database producing the highest accuracy rate ("best other"), and combing all other databases by ROVER and SDS. Three normalization methods are included. The baseline performance to compare with is underlined.

neutral vs. emotional in valence						
NA	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	83.4	92.3	84.8	64.8	62.5	57.4
best other	66.9	70.9	73.3	58.6	56.1	50.5
best db	4	4	2	5	1	3
Rover-db	67.9	78.0	74.8	60.8	59.3	46.0
Rover-db*feat	71.4	72.9	75.3	62.1	60.0	44.2
SDS	67.3	72.9	72.8	58.7	61.1	50.6
dbs	[2,5]	[4,5]	[2]	[5]	[4,6,1]	[3]
SN	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	84.8	95.0	89.0	58.7	61.6	60.3
best other	67.1	73.5	82.5	52.5	55.4	51.5
best db	2	3	2	2	2	2
Rover-db	73.1	78.2	82.5	53.0	59.3	50.0
Rover-db*feat	77.2	79.1	83.5	54.0	59.0	50.2
SDS	71.7	75.9	82.1	53.8	59.0	52.3
dbs	[3,2]	[3,1]	[2]	[5,2,6,3]	[4,6,3,2]	[1,2]
SR	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	86.6	96.2	92.7	71.0	67.9	63.6
best other	73.2	72.9	78.4	54.7	54.3	57.2
best db	3	3	2	5	6	5
Rover-db	78.7	81.1	88.6	54.7	61.6	55.5
Rover-db*feat	87.6	87.7	92.7	57.9	63.2	59.5
SDS	79.1	76.6	82.4	55.8	63.4	57.2
dbs	[3,2]	[5,3]	[2,6]	[2,6]	[2,6]	[5]
NRM	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	88.4	96.1	96.2	76.1	69.4	62.1
best other	83.7	76.7	90.2	57.3	57.4	59.6
best db	2	3	2	2	2	5
Rover-db	87.5	88.9	90.8	61.4	64.1	60.0
Rover-db*feat	89.2	91.3	92.8	60.7	64.7	60.9
SDS	83.8	82.2	89.9	61.3	60.1	59.7
dbs	[3,2]	[3,1]	[2]	[2,6]	[1,6]	[5]
Δ	22.3	20.4	19.5	3.5	8.6	10.4

Table 36: Accuracy rates of binary classifications (**neutral vs. emotional in arousal**) tested on each emotional database when the training data is the test data itself ("self"), one of the other database producing the highest accuracy rate ("best other"), and combing all other databases by ROVER and SDS. Three normalization methods are included. The baseline performance to compare with is underlined.

neutral vs. emotional in arousal						
NA	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	83.5	93.4	86.5	64.5	64.8	57.4
best other	65.5	67.9	72.0	57.7	53.7	50.8
best db	4	4	2	1	3	3
Rover-db	67.9	78.0	74.8	60.8	59.3	46.0
Rover-db*feat	71.4	72.9	75.3	62.1	60.0	44.2
SDS	67.4	73.3	73.8	58.5	55.9	50.4
dbs	[2,5]	[4,5]	[5,4,2]	[1,3]	[6,3]	[3]
SN	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	85.3	93.1	90.9	58.7	64.6	60.1
best other	67.5	73.9	81.7	52.5	55.2	51.2
best db	2	3	2	2	3	2
Rover-db	73.1	78.2	82.5	53.0	59.3	50.0
Rover-db*feat	77.2	79.1	83.5	54.0	59.0	50.2
SDS	72.4	75.7	82.1	53.5	58.4	51.4
dbs	[3,2]	[3,1]	[2]	[2,6,5,3]	[3,4]	[2]
SR	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	86.7	96.6	91.3	71.1	73.0	63.7
best other	73.3	72.2	77.9	53.3	56.3	55.0
best db	3	3	2	5	3	5
Rover-db	78.7	81.1	88.6	54.7	61.6	55.5
Rover-db*feat	87.6	87.7	92.7	57.9	63.2	59.5
SDS	79.2	76.7	83.2	55.3	65.6	56.5
dbs	[3,2]	[5,3]	[2,6]	[2,6]	[2,6]	[5,4,2]
NRM	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	87.9	96.8	96.1	76.4	73.5	62.2
best other	83.8	77.3	89.7	59.0	61.4	58.0
best db	2	3	2	5	2	5
Rover-db	87.5	88.9	90.8	61.4	64.1	60.0
Rover-db*feat	89.2	91.3	92.8	60.7	64.7	60.9
SDS	84.3	85.5	90.4	61.0	64.4	58.1
dbs	[2]	[5,3]	[2]	[2,6]	[1,6,2]	[5,2]
Δ	23.7	23.4	20.8	4.4	11.0	10.1

Table 37: Accuracy rates of binary classifications (**positive vs. negative in valence**) tested on each emotional database when the training data is the test data itself ("self"), one of the other database producing the highest accuracy rate ("best other"), and combing all other databases by ROVER and SDS. Three normalization methods are included. The baseline performance to compare with is underlined.

positive vs. negative in valence						
NA	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	67.4	91.7	78.7	77.8	70.3	59.6
best other	59.6	72.5	72.9	60.7	55.8	48.2
best db	3	1	2	1	4	2
Rover-db	61.2	72.6	72.7	50.8	42.1	43.6
Rover-db*feat	64.4	70.6	74.6	59.3	36.8	46.1
SDS	59.8	72.6	73.0	62.2	58.8	54.9
dbs	[3]	[1]	[1,2]	[3,2]	[6,4]	[3,5]
SN	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	67.5	92.5	82.1	67.7	66.4	61.8
best other	61.3	75.7	76.5	51.6	59.6	52.7
best db	2	1	2	3	2	4
Rover-db	63.5	68.4	74.1	50.7	58.0	50.5
Rover-db*feat	65.1	77.3	77.7	50.3	50.7	51.7
SDS	61.2	77.8	76.5	51.9	66.1	56.4
dbs	[2]	[1,3]	[2]	[1,5]	[4,2,1]	[4,1]
SR	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	67.2	90.2	77.7	86.9	73.7	65.4
best other	59.8	70.0	70.7	55.7	58.8	53.8
best db	2	3	1	3	6	2
Rover-db	57.1	61.7	62.9	52.2	51.7	52.9
Rover-db*feat	60.0	70.0	72.0	51.8	49.6	52.0
SDS	60.1	70.9	70.5	54.8	61.8	55.0
dbs	[2,3]	[1,3]	[1,2]	[3]	[3,4,2]	[5,3]
NRM	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	66.7	91.3	78.5	90.7	80.2	66.7
best other	56.2	71.7	70.1	53.7	58.3	51.6
best db	2	1	1	2	1	5
Rover-db	55.9	57.2	66.5	55.9	54.1	48.0
Rover-db*feat	56.9	66.4	69.1	55.8	48.0	49.1
SDS	60.1	70.9	70.5	54.8	61.8	55.0
dbs	[2,3]	[1,3]	[1,2]	[3]	[3,4,2]	[5,3]
Δ	5.5	5.3	4.8	1.5	10.3	8.2

Table 38: Accuracy rates of binary classifications (**neutral vs. emotional in arousal**) tested on each emotional database when the training data is the test data itself ("self"), one of the other database producing the highest accuracy rate ("best other"), and combing all other databases by ROVER and SDS. Three normalization methods are included. The baseline performance to compare with is underlined.

positive vs. negative in arousal						
NA	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	75.5	98.9	97.7	66.7	92.5	55.6
best other	69.5	92.2	90.5	53.6	79.9	50.8
best db	2	5	1	1	2	3
Rover-db	67.9	89.6	90.6	52.1	81.0	51.0
Rover-db*feat	68.3	91.9	92.3	51.5	83.1	48.5
SDS	69.6	92.2	90.3	55.0	81.5	52.1
dbs	[2]	[5]	[5]	[3,1]	[2,3]	[5,4]
SN	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	76.6	96.8	98.7	56.1	88.2	60.8
best other	68.8	91.0	92.4	51.3	76.4	55.0
best db	2	5	1	6	2	4
Rover-db	69.4	91.7	94.0	50.5	78.9	51.5
Rover-db*feat	71.3	92.9	94.2	50.2	81.2	50.8
SDS	71.5	93.0	92.5	51.5	78.1	54.6
dbs	[3,2]	[5,1]	[1]	[5,6,3,1]	[3,2]	[4,1]
SR	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	75.1	99.7	97.5	76.5	92.2	62.8
best other	69.9	85.8	90.8	55.5	74.5	50.0
best db	5	1	5	1	3	2
Rover-db	69.3	83.3	91.3	55.5	83.2	50.3
Rover-db*feat	68.5	82.3	93.0	56.1	84.9	50.9
SDS	70.1	86.0	90.5	55.7	81.9	52.2
dbs	[5]	[1]	[5,2]	[1,2]	[3,2]	[2,1]
NRM	1.EPST	2.EMA	3.GES	4.SEMAINE	5.VAM	6.UM
self	71.8	97.8	96.8	81.0	88.8	67.4
best other	64.2	82.9	89.7	53.7	74.6	52.2
best db	5	1	5	2	3	5
Rover-db	64.1	84.7	90.3	51.0	79.5	50.5
Rover-db*feat	64.7	86.7	93.0	50.7	78.9	49.3
SDS	70.1	86.0	90.5	55.7	81.9	52.2
dbs	[5]	[1]	[5,2]	[1,2]	[3,2]	[2,1]
Δ	2.0	0.8	3.7	2.5	5.0	3.8

APPENDIX C

FEATURE SELECTION RESULT FOR EACH DATABASE

Due to the limitation of space, only features selected by all 10 tests of the 10-fold cross-validation are listed.

Table 39: the features selected by all the 10 folds in self-testing by 10-fold cross validation in neutral vs. emotional.

EPST	EMA	GES	SEMAINE	VAM	UM
neutral vs. emotional in valence					
TECC6-Δmean TECC11-mean HRF-std HRF-iqr	pch-fit3 TEOfm2-range TEOcb2-fit1 TEOcb16-fit2 TECC1-mean TECC22-Δq75 MFCC12-std MFCC12-range QOQ-fit1 SQ2-median	TEOcb1-fit2 TEOcb11-median TEOcb11-q75 TEOcb14-q75 MFCC2-fit3	pch-q75 TEOfm4-median TEOevn4-median TEOevn4-q25 TECC15-min TECC15-max TECC18-max	TEOevn1-mean TEOcb13-q75 HRF-std	TEOcb14-std TECC13-mean TECC13-q75 TECC14-Δmean
neutral vs. emotional in arousal					
TECC6-Δmean TECC11-mean HRF-std HRF-iqr	pch-fit3 TEOfm2-range TEOcb2-fit1 TEOcb16-fit2 TECC1-mean TECC22-Δq75 MFCC12-std MFCC12-range QOQ-fit1 SQ2-median	TEOcb1-fit2 TEOcb11-median TEOcb11-q75 TEOcb14-q75 MFCC2-fit3	pch-q75 TEOfm4-median TEOevn4-median TEOevn4-q25 TECC15-min TECC15-max	pch-Δiqr TEOcb13-Δstd TEOcb16-q75	TEOcb14-std TECC13-mean TECC13-q75 TECC14-Δmean

Table 40: the features selected by all the 10 folds in self-testing by 10-fold cross validation in positive vs. negative.

EPST	EMA	GES	SEMAINE	VAM	UM
positive vs. negative in valence					
pch-Δstd pch-Δq75 pch-Δiqr TEOcb4-std MFCC9-Δstd MFCC11-Δstd MFCC12-std DH12-Δstd DH12-Δq75	pch-std pch-iqr TEOfm2-iqr TEOcb9-q75 TECC8-q25 MFCC3-iqr MFCC12-max ClQ-iqr	TEOcb9-iqr	TECC13-q25 TECC15-mean TECC15-max TECC15-q75 TECC18-max TECC20-max QOQ-max	TEOcb12-Δq25 TEOcb16-Δmedian TECC4-range MFCC3-Δstd	TECC10-max TECC10-Δmax TECC18-min
positive vs. negative in arousal					
pch-Δiqr TEOcb8-fit2 TEOcb8-Δmax TECC1-std TECC1-Δstd TECC2-std TECC2-Δq75 TECC4-Δiqr MFCC1-q25 MFCC5-min	pch-iqr pch-Δmedian pch-Δiqr TEOfm2-iqr TEOcb8-std TECC1-range TECC1-Δiqr TECC2-Δq75 TECC6-std TECC8-max TECC8-Δstd TECC11-Δstd TECC15-Δstd TECC16-std TECC18-std TECC22-Δstd MFCC3-std MFCC3-iqr MFCC5-mean MFCC5-min MFCC5-q25 MFCC9-min MFCC12-std OQ1-median AQ-q25 ClQ-iqr	pch-fit2 pch-Δmedian pch-Δq25 pch-Δq75 pch-Δiqr TEOfm2-Δstd TEOcb5-fit1 TEOcb10-Δiqr TEOcb15-Δmedian TECC2-mean TECC2-q25 TECC2-Δq75 TECC5-fit1 TECC5-Δiqr TECC6-Δstd TECC7-Δstd TECC8-Δstd TECC10-q75 TECC11-Δstd TECC15-fit2 TECC16-std TECC16-Δstd TECC17-Δstd MFCC2-Δiqr MFCC3-Δmean MFCC12-fit1 OQ1-median AQ-fit1	TEOfm1-q25 TEOfm4-min TEOevn4-q25 TEOcb16-min TECC1-max TECC1-q75 TECC4-q25 TECC9-min TECC9-q25	pch-Δq25 pch-Δiqr TEOevn1-iqr TEOcb11-q75 TECC1-q75 TECC2-std TECC2-q25 TECC2-iqr TECC10-q75 TECC14-std TECC17-std TECC18-std TECC18-Δstd MFCC1-Δq75 MFCC8-std MFCC8-q75 MFCC11-std OQ1-Δq75 OQa-std QOQ-q25	TECC8-fit2

REFERENCES

[1] Airas, M., Pulakka, H., Backstrom, T., and Alku, P., "A toolkit for voice inverse filtering and parametrisation," in *INTERSPEECH-2005*, (Lisbon, Portugal), pp. 2145–2148, 2005.

[2] Banse, R. and Scherer, K., "Acoustic profiles in vocal emotion expression," *Journal of ersonality and Social Psychology*, vol. 70, pp. 614–636, 1996.

[3] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V., "Combining efforts for improving automatic classification of emotional user states," in *Proc. IS-LTC 2006, Ljubliana*, pp. 240–245, 2006.

[4] Bitouk, D., Verma, R., and Nenkova, A., "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[5] Brendel, M., Zaccarelli, R., Shuller, B., and Devillers, L., "Towards measuring similarity between emotional corpora," 2010.

[6] Brookes, M., "Voicebox: Speech processing toolbox for matlab," in *http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 2007.

[7] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B., "A database of german emotional speech," *in Proceedings of Interspeech, Lissabon*, pp. 1517–1520, 2005.

[8] Busso, C., Sungbok, L., and Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[9] Caims, D. and Hansen, J. H. L., "Nonlinear analysis and classification of speech under stressed conditions," *The Journal of the Acoustical Society of America*, vol. 96, no. 6, pp. 3392–3399, 1994.

[10] Chang, C.-C. and Lin, C.-J., "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, p-p. 27:1–27:27, 2011.

[11] Childers, D. G., "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, 1995.

[12] Chu, W. and Champagne, B., "A noise-robust fft-based auditory spectrum with application in audio classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 137 –150, 2008.

[13] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[14] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M., "Facetrace: An instrument for recording perceived emotion in real time," 2000.

[15] Cummings, K. E. and Clements, M. A., "Analysis of the glottal excitation of emotionally styled and stressed speech," *The Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 88–98, 1995.

[16] Dietz, R. B., "Effective agents: Effects of agent affect on arousal, attention, liking and learning," in *3rd International Conference of Cognivite*, 1999.

[17] Dimitriadis, D., Maragos, P., and Potamianos, A., "On the effects of filterbank design and energy computation on robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, p-p. 1504–1516, 2011.

[18] Dimitriadis, D., Maragos, P., and Potamianos, A., "Auditory teager energy cepstrum coefficients for robust speech recognition," in *Interspeech'2005 - Eurospeech*, (Lisbon, Portugal), 2005.

[19] D'Mello, S., Craig, S. D., Witherspoon, A., McDaniel, B., and Graesser, A., "Automatic detection of learners affect from conversational cues," *User Modeling and User-Adapted Interaction*, vol. 18, no. 1-2, pp. 45–80, 2008.

[20] D'Mello, S. and Graesser, A., "Automatic detection of learner's affect from gross body language," *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.

[21] D'Mello, S. and Graesser, A., "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Modeling and User-Adapted Interaction*, vol. 20, no. 2, pp. 147–187, 2010.

[22] Dromey, C., Silveira, J., and Sandor, P., "Recognition of affective prosody by speakers of english as a first or foreign language," *Speech Communication*, vol. 47, no. 3, pp. 351–359, 2005.

[23] Eyben, F., Wollmer, M., and Schuller, B., "Openear - introducing the munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–6, 2009.

[24] Eyben, F., Wollmer, M., and Schuller, B., "opensmile-the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia (MM)*, (Florence, Italy), pp. 1459–1462, 2010.

[25] Eyben, F., Batliner, A., Schuller, B., Seppi, D., and Steidl, S., "Cross-corpus classification of realistic emotions - some pilot experiments," in *Third international workshop on EMOTION*, 2010.

[26] Fernandez, R. and Picard, R., "Classical and novel discriminant features for affect recognition from speech," in *INTERSPEECH*, pp. 473–476, 2005.

[27] Fiscus, J. G., "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997*, pp. 347–354, 1997.

[28] Fragopanagos, N. and Taylor, J. G., "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[29] Freund, Y. and Schapire, R. E., "Experiments with a new boosting algorithm," in *the Thirteenth Interantional Conference on machine Learning*, 1996.

[30] Garofolo, J. S. and Lamel, L. F., "Timit acoustic-phonetic continous speech corpus," 1993.

[31] Graesser, A., D'Mello, S., Chipman, P., King, B., and McDaniel, B., "Exploring relationship between affect and learning with autotutor," in *the 13th international conference on artificial intelligence in education*, pp. 16–23, 2007.

[32] Greenwald, M. K., Cook, E. W., and Lang, P. J., "Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli," *Journal of Pyschophysiology*, vol. 3, pp. 51–64, 1989.

[33] Grimm, M., Kroschel, K., and Narayanan, S., "The vera am mittag german audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, pp. 865–868, 2008.

[34] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S., "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, 2007.

[35] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[36] Hanson, H. M., Maragos, P., and Potamianos, A., "A system for finding speech formants and modulations via energy separation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 436–443, 1994.

[37] He, L., Lech, M., Maddage, N., Memon, S., and Allen, N., "Emotion recognition in spontaneous speech within work and family environments," in *Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*, pp. 1–4, 2009.

[38] Hirschberg, J., Liscombe, J., and Venditti, J., "Experiments in emotional speech," in *ISCA and IEEE Workshop on Spontanous Speech Processing and Recognition*, (Tokyo, Japan), pp. 119–125, 2003.

[39] Iliou, T. and Anagnostopoulos, C. N., "Statistical evaluation of speech features for emotion recognition," in *Digital Telecommunications, 2009. ICDT '09. Fourth International Conference on*, pp. 121–126, 2009.

[40] Irino, T. and Patterson, R. D., "A time-domain, level-dependent auditory filter: The gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, pp. 412–419, 1997.

[41] Kaiser, J. F., "On a simple algorithm to calculate the 'energy' of a signal," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 381–384 vol.1, 1990.

[42] Kim, J., Lee, S., and Narayanan, S. S., "An exploratory study of manifolds of emotional speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 5142–5145, 2010.

[43] Kostoulas, T., Ganchev, T., Mporas, I., and Fakotakis, N., "Detection of negative emotional states in real-world scenario," in *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, vol. 2, pp. 502–509, 2007.

[44] Laukkanen, A.-M., Vilkman, E., Alku, P., and Oksanen, H., "Physical variations related to stress and emotional state: a preliminary study," *Journal of Phonetics*, vol. 24, no. 3, pp. 313–335, 1996.

[45] Lee, C. M. and Narayanan, S., "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[46] Lee, S., Yildirim, S., Kazemzadeh, A., and Narayanan, S., "An articulatory study of emotional speech production," *INTERSPEECH-2005*, pp. 497–500, 2005.

[47] Li, Q. and Huang, Y., "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1791 –1801, aug. 2011.

[48] Li, X., Li, X., Zheng, X., and Zhang, D., "Emd-teo based speech emotion recognition," *Life System Modeling and Intelligent Computing*, vol. 6329, pp. 180–189, 2010.

[49] Liberman, M., K., D., Grossman, M., Martey, N., and Bell, J., "Emotional prosody speech and transcripts," 2002.

[50] Lippmann, R., Martin, E., and Paul, D., "Multi-style training for robust isolated-word speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, vol. 12, pp. 705–708, 1987.

[51] Litman, D. J. and Forbes-Riley, K., "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," *Speech Communication*, vol. 48, no. 5, pp. 559–590, 2006.

[52] Lugger, M. and Yang, B., "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *Acoustics, Speech and Signal Processing, (ICASSP 2008). IEEE International Conference on*, pp. 4945–4948, 2008.

[53] Maragos, P., Kaiser, J. F., and Quatieri, T. F., "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[54] McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M., "The semaine corpus of emotionally coloured character interactions," in *(ICME 2010)*, pp. 1079–1084, 2010.

[55] Milner, B. and Darch, J., "Robust acoustic speech feature prediction from noisy mel-frequency cepstral coefficients," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 338 –347, feb. 2011.

[56] Milner, B., Darch, J., and Vaseghi, S., "Applying noise compensation methods to robustly predict acoustic speech features from mfcc vectors in noise," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3945 –3948, 31 2008-april 4 2008.

[57] Moore, E., Clements, M., Peifer, J., and Weisser, L., "Investigating the role of glottal features in classifying clinical depression," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 3, pp. 2849–2852 Vol.3, 2003.

[58] Moore, E., Clements, M. A., Peifer, J. W., and Weisser, L., "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, 2008.

[59] Moore, E. and Torres, J., "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Communication*, vol. 50, no. 1, pp. 56–66, 2008.

[60] Muralishankar, R. and O'Shaughnessy, D., "A comparative analysis of noise robust speech features extracted from all-pass based warping with mfcc in a noisy phoneme recognition," in *Digital Telecommunications, 2008. ICDT '08. The Third International Conference on*, pp. 180 –185, 29 2008-july 5 2008.

[61] Patrick, A. N., Anastasis, K., Jon, G., and Mike, B., "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[62] Picard, R., *Affective Computing.* MIT Press, 2000.

[63] Ringeval, F. and Chetouani, M., "Exploiting a vowel based approach for acted emotion recognition," *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, vol. 5042, pp. 93–110, 2008.

[64] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A., "Acoustic emotion recognition: A benchmark comparison of performances," in *ASRU 2009.*, pp. 552–557, 2009.

[65] Schuller, B., Vlasenko, B., Eyben, F., Wo, x, llmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G., "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[66] Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G., "Selecting training data for cross-corpus speech emotion recognition: prototypicality vs. generalization," in *AFEKA-AVIOS*, 2011.

[67] Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G., "Using multiple databases for training in emotion recognition: to unite or to vote?," in *INTERSPEECH 2011*, 2011.

[68] Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J., "The interspeech 2011 speaker state challenge," in *Interspeech*, (Italy), 2011.

[69] Schuller, B., Valstar, M. F., Eyben, F., McKeown, G., Cowie, R., and Pantic, M., "Avec 2011-the first international audio/visual emotion challenge," in *First International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) held in conjunction with the International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2011 (ACII 2011)*, (Memphis, Tennessee), Springer LNCS, 2011.

[70] Sethu, V., Ambikairajah, E., and Epps, J., "Speaker normalisation for speech-based emotion detection," in *Digital Signal Processing, 2007 15th International Conference on*, pp. 611–614, 2007.

[71] Skowronski, M. D. and Harris, J. G., "Increased mfcc filter bandwidth for noise-robust phoneme recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–801 –I–804, may 2002.

[72] Sun, R. and Moore, E. I., "Investigating acoustic cues in automatic detection of learners' emotion from auto tutor," in *the fourth International Conference of Affective Computing and Intelligent Interaction (ACII)*, vol. 6975, (Memphis, TN), pp. 91–100, 2011.

[73] Sun, R. and Moore, E. I., "Investigating glottal parameters and teager energy operators in emotion recognition," in *the 4th Affective Computing and Intelligent Interaction*, vol. 6975, pp. 425–434, 2011.

[74] Sun, R. and Moore, E. I., "Investigating the robustness of teager energy cepstrum coefficients for emotion recognition in noisy conditions," in *(FLAIRS-25)*, (Macro Island, FL), 2012.

[75] Sun, R. and Moore, E. I., "A preliminary study on cross-databases emotion recognition using the glottal features in speech," in *INTERSPEECH*, (Portland, OR), 2012.

[76] Sun, R. and Moore, E. I., "Evaluation of the neutral reference model in cross-database emotion recognition," *IEEE Transactions on Affective Computing*, vol. (in revision), 2013.

[77] Sun, R., Moore, E. I., and Torres, J., "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, (Taipei, Taiwan), 2009.

[78] Tahon, M. and Devillers, L., "Acoustic measures characterizing anger across corpora collected in artificial or natural context," in *the Fifth International Conference on Speech Prosody*, (Chicago, Illinois), 2010.

[79] Teager, H., "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[80] Teager, H. and Teager, S., *The effects of separated air flow on vocalization.* Vocal Fold Physiology: Contemporary Research and Clinical Issues, 1981.

[81] Teager, H. and Teager, S., *Active fluid dynamic voice production models, or there is a unicorn in the garden.* Vocal Fold Physiology, 1983.

[82] Teager, H. and Teager, S., "A phenomenological model for vowel production in the vocal tract," *Speech Science: Recent Advances*, pp. 73–109, 1983.

[83] Titze, I. R. and Sundberg, J., "Vocal intensity in speakers and singers," *The Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[84] Torres, J. F., Moore, E., and Bryant, E., "A study of glottal waveform features for deceptive speech classification," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4489–4492, 2008.

[85] Varela, O., San-Segundo, R., and Hernandez, L., "Robust speech detection for noisy environments," *Aerospace and Electronic Systems Magazine, IEEE*, vol. 26, pp. 16 –23, nov. 2011.

[86] Ververidis, D. and Kotropoulos, C., "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[87] Witten, I. H. and Freank, E., *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.

[88] Wu, S., Falk, T. H., and Chan, W.-Y., "Automatic recognition of speech emotion using long-term spectro-temporal features," in *Digital Signal Processing, 2009 16th International Conference on*, pp. 1–6, 2009.

[89] Wu, S., Falk, T. H., and Chan, W.-Y., "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[90] Wu, Z., Li, D., and Yang, Y., "Rules based feature modification for affective speaker recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. 661–664, 2006.

[91] Yang, B. and Lugger, M., "Emotion recognition from speech signals using new harmony features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010.

[92] Yanguas, L. R. and Quatieri, T. F., "Implications of glottal source for speaker and dialect identification," in *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, vol. 2, pp. 813–816 vol.2, 1999.

[93] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Shrikanth, N., "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing*, (Jeju Island, Korea), pp. 2193–2196, 2004.

[94] Yu, D., Deng, L., Wu, J., Gong, Y., and Acero, A., "Improvements on mel-frequency cepstrum minimum-mean-square-error noise suppressor for robust speech recognition," in *Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on*, pp. 1 –4, dec. 2008.

[95] Zhang, J., Li, G.-l., Zheng, Y.-z., and Liu, X.-y., "A novel noise-robust speech recognition system based on adaptively enhanced bark wavelet mfcc," in *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on*, vol. 4, pp. 443 –447, aug. 2009.

[96] Zhou, G., Hansen, J. H. L., and Kaiser, J. F., "Nonlinear feature based classification of speech under stress," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 201–216, 2001.

[97] Zwicker, E. and Terhardt, E., "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

VITA

Rui Sun was born in the Northeast of China in a family full of love, where both her father and mother were engineers in power and Smart Grid. She spent her childhood painting, reading, and exploring the outdoor nature. She attended Shanghai Jiaotong University in Shanghai, China in 2002 as the first time left her hometown. In 2006 right after she earned her bachelor degree, she started the graduate life in Georgia Tech. After two years as a master student, she joined the Voice and Audio Lab of Dr Elliot Moore and moved to the beautiful city Savannah, GA. She earned her PhD in 2013 at Georgia Institute of Technology, where she was working as a graduate research assistant focusing on emotion recognition and speech analysis.