Speech Emotion Recognition using Connectionist Models
in a Tandem System

by

Mary Najafi

A thesis
submitted to College of Engineering and Computing at
Florida Institute of Technology
in partial fulfillment of the requirements
for the degree of

Master of Science
in
Computer Engineering

Melbourne, Florida
May, 2018
© Copyright 2018 Mary Najafi
All Rights Reserved

The author grants permission to make single copies.
We the undersigned committee
hereby approve the attached thesis

Speech Emotion Recognition using Connectionist Models
in a Tandem System by Mary Najafi

Dr. V. Kepuska, Ph.D.
Associate Professor
Electrical and Computer Engineering
Committee Chair

Dr. M. Silaghi, Ph.D.
Associate Professor
Computer Sciences and Cybersecurity
Outside Committee Member

Dr. S. Kozaitis, Ph.D.
Professor
Electrical and Computer Engineering
Committee Member

Dr. G. Anagnostopoulos, Ph.D.
Associate Professor
Electrical and Computer Engineering
Committee Member

Dr. S. Kozaitis, Ph.D.
Professor and Department Head
Electrical and Computer Engineering
ABSTRACT

Title:

Speech Emotion Recognition using Connectionist Models

in a Tandem System

Author:

Mary Najafi

Major Advisor:

Dr. V. Kepuska, Ph.D.

This article presents a tandem Speech Emotion Recognition (SER) system by which 8 archetypal emotions can be differentiated based upon two different types of acoustic features as inputs to Artificial Neural Network (ANN) models. The two types of features that are fed into the classifiers reveal the degree of excitement and pleasantness of speech. The author has implied the time-based characteristics of speech in the feature extraction method by monitoring the trend of local features through time. Thus, two global features are proposed that are derived from Teager Energy Operator (TEO)-based ($TEO_g$) and spectral features, Mel-Frequency Cepstral Coefficients ($MFCC_g$). In this study we established a tandem system of two hierarchies that follows a cognitive model to separate emotions based on the amount of stress in the voice Teager Energy Operator-Critical Band-Autocorrelation-Envelope ($TEO-CB-Auto-Env$) and the pleasantness of the emotion ($MFCC$). In this research, we proposed a baseline measurement of the recognition based on the current feature vectors and make an analogy between the baseline and the tandem system to demonstrate the superiority of the proposed tandem system against the non-hierarchical systems. Moreover, we compared our results with the recognition rates from some of the cited articles. Additionally,
inspired by the cognitive model, the author defined a hybrid tandem system in which the first hierarchy gets the $TEO_g$ as input to the classifier and two models in the second hierarchy get the $MFCC_g$ features for their input layers. This system will be compared to a tandem system with only $MFCC_g$ feature vectors in the hierarchies in terms of the effectiveness and efficiency. Based on our experiments, it turns out that the former system returns a higher degree of efficiency whereas the latter tandem system gives a higher recognition rate. In our system, we made use of a binary-class Multi-Layer Perceptron (MLP) and two multi-class MLPs for the first and the second hierarchies, respectively. Considering only the audio part, the classification is performed on three emotion-based datasets: Surrey Audio-Visual Expressed Emotion (SAVEE), Berlin Database of Emotional Speech (Emo-DB), and eNTERFACE Audio-Visual Emotion Database (eNTERFACE). The systems are considered speaker- and gender-independent. We have used Unweighted Accuracy (UW) accuracy to evaluate our methods. Our tandem system at its best given only $MFCC_g$ returns prediction rates as 77.26, 71.42 and 66.49 on the Emo-DB, SAVEE and eNTERFACE datasets, respectively. Whereas this measurement using a hybrid feature (second best) are 75.067, 67.596 and 65.197.
# Table of Contents

Abstract iii

List of Figures vii

List of Tables ix

Acknowledgments x

Dedication xi

1 Introduction 1

2 Related Work 3

3 Approach 7

3.1 Feature Extraction 8

3.1.1 Spectral Features 8

3.1.2 $MFCC_g$ Features 12

3.1.3 TEO-based Features 12

3.1.4 $TEO_g$ Features 14

3.2 Classification: Artificial Neural Networks 15

3.2.1 Hierarchy I 16
3.2.2 Hierarchy II ................................................. 16

4 Experimental Evaluation ...................................... 26

4.1 Procedures .................................................. 26

4.2 Datasets ..................................................... 27

4.2.1 SAVEE .................................................. 28

4.2.2 EMO-DB .................................................. 28

4.2.3 eENTERFACE .......................................... 29

4.3 Evaluation .................................................. 29

4.4 Result ........................................................ 30

4.5 Analysis ...................................................... 34

5 Conclusion ..................................................... 36

References ....................................................... 37
List of Figures

2.1 The Cognitive Emotion Model that spots emotions based on their arousal and valence measurement. ........................................... 4

3.1 An Example of a Tandem Classification System using Hybrid Features. 8
3.2 The Schema of the \textit{MFCC} \textsubscript{g} Feature Extraction from the Mel-Frequency Cepstral Coefficients ................................................... 10
3.3 The Schema of the \textit{TEO} \textsubscript{g} Feature Extraction from TEO-CB-Auto-Env. ................................................................. 18
3.4 The Original Audio File Showing Two Active (a) and Passive (b) Emotions. ................................................................. 19
3.5 The Extracted TEO Feature for the First Critical Band of an Anger (Active Emotion) ............................................................... 20
3.6 The Extracted TEO Feature for the Critical Band 6 of an Anger (Active Emotion) ............................................................... 21
3.7 The Extracted TEO Feature for the Critical Band 16 of an Anger (Active Emotion) ............................................................... 22
3.8 The Extracted TEO Feature for the First Critical Band of Boredom (Passive Emotion) ......................................................... 23
3.9 The Extracted TEO Feature for the Critical Band 6 of Boredom (Passive Emotion) ......................................................... 24
3.10 The Extracted TEO Feature for the Critical Band 16 of Boredom (Passive Emotion) .................................................. 25

4.1 The Result of the PCA for Spectral and TEO-based Features for the Emo-DB Corpus. ............................................. 30

4.2 The Trend of Accuracy for Different Number of Hidden Units ... 32
List of Tables

3.1 Critical Band Frequency Information in Hz for 16 Sub-Bands . . . . 13

4.1 Emotions that Exist in Datasets . . . . . . . . . . . . . . . . . . . 28
4.2 The Comparison between the Tandem Architecture and the Baseline 32
4.3 UW Accuracy for Different Features . . . . . . . . . . . . . . . . . 33
4.4 The Elapsed Time for Classifying Emotions using Different Features in Tandem System . . . . . . . . . . . . . . . . . . . . . . . . . . 34
I would like to express my gratitude to my advisor, Dr. Kepuska for his continuous support and patience during this research. I wish to express my sincere thanks to Dr. Chan, Dr. Anagnostopoulos and Dr. Kozaitis, whose guidance, knowledge and support across my Master’s degree have been a major factor to my academic success.
Dedication

I would like to dedicate this thesis work to my dearly loved husband, Ebad, whose positive spirit, unconditional support, encouragement and academic help have been a great motivation over the course of my research.
Chapter 1

Introduction

There are unified characteristics in human speech that reflect behavioral and emotional states of speaker, independent of the content of speech. These features could be perceived differently by different audience, and thus they are assumed to be subjective in terms of the way that the expressed emotion is interpreted. This paper proposes an idea that the machine mimics a natural way of human perception of the expressed emotion in speech.

Inspired by the fact that the process of perceiving one’s emotion in speech is likely a time-based process, the author approaches the Speech Emotion Recognition (SER) problem by implying the time-based nature of speech in their acoustic feature extraction. Using a cognitive emotion model, the author presents a tandem system where the machine learns how to interpret the emotional characteristics of human speech by measuring the arousal of an utterance followed by its valence. The proposed tandem system is compared to a non-hierarchical approach and the paper provides results such that a tandem system performs effectively comparing to the other proposed approaches in this paper.
In this article, the system is broken down into two major tasks, each of which is being executed by a connectionist model utilizing their own specific types of acoustic features. At the first hierarchy of the system, a non-sequential ANN model has utilized prosodic features which represent global characteristics of TEO-CB-Auto-Env feature vectors across audio frames ($TEO_g$). Also at the second level of the hierarchy two different ANN models are studied to monitor the trend of MFCC spectral features through time across audio frames ($MFCC_g$). The result of our predictive analysis based on this hierarchy is then compared to a different approach with the same model architecture except that the features refer to the local characteristics of audio files across their frames.
Chapter 2

Related Work

Speech recognition has been extensively studied using a variety of connectionist and non-connectionist models, among which some are sequence-based models that follow the fact that speech has a time-based feature space in which each event at a given time depends on both the preceding and following time frames. The speech recognition systems have been based on both generative and discriminative models. For the former the model describes the distribution of data as a joint probability \( p(x, y) \); where \( x \) and \( y \) are features and target pairs) such as Gaussian Mixture Model/Hidden Markov Model (GMM/HMM) approach in speech recognition [11] whereas in the latter model, it spits out the posterior probability with no prior assumption. According to the Bayesian rule, these models directly calculate a conditional probability \( p(y|x) \); where \( x \) and \( y \) are features and target pairs) of an event occurrence in data given no a-priori, and with only the target value (e.g. Support Vector Machine (SVM)). Nowadays, variants of Neural Networks as discriminative models such as Recurrent Neural Networks (RNN) or Long Short Term Memory/RNN (LSTM/RNN) have become popular in many different speech
recognition applications; however one of the drawbacks of using these algorithms is the lack of model visibility or readability [7] [3].

According to the emotion/cognitive model [16] in the Cartesian coordinate system (Figure 2.1), this study is planned based upon a crucial concept that differentiates emotions based on their altitude on the vertical axis (arousal) and positivity on the horizontal axis (valence).

![Figure 2.1: The Cognitive Emotion Model that spots emotions based on their arousal and valence measurement.](image)

The number of emotions is as great as 300 but researchers agreed with a limited number of emotions as stated in the Palette Theory [4]. According to the palette theory all existing emotions are derived from combinations of seven primary emotions plus neutral that are named archetypal emotions: happiness, anger, sadness, fear, disgust, surprise, boredom, and neutral. Due to the uncertainty of the precise locations of each of these emotions on the cognitive model, we tackle the classification problem at its primary stage by simplifying the problem which divides the emotions into two categories of active and passive classes. According to [1], happi-
ness, surprise, anger, fear and disgust are five emotions that by nature are grouped together so they are designated as an active class whereas sadness, boredom and neutral are marked among passive emotions in terms of their arousal.

Generally, the challenge still remains in speech processing for specifying the most significant and powerful features to discriminate speech tokens. Acoustic features are not the only types of features that carry speech information. For Automated Speech Recognition (ASR) problems, among those features that play significant role to recognize phonemes one can name spectral features such as MFCC, Linear Prediction Cepstral Coefficients (LPCC), or Linear Predictive Codes (LPC) \[4\]. In Speech Emotion Recognition (SER), aforementioned features are considered standard features that in conjunction with Prosodic features have provided comparative results \[14\]. Prosodic features are continuous features that monitor the pattern of one particular characteristic of utterance throughout the speech. These patterns manifest the rhythm and intonation of produced speech. Along with acoustic and prosodic features the discourse and linguistic features are also extensively studied in speech and image modality \[4\]. An important group of features that measures the excitation of speech is named Teager Energy Operator (TEO)-based features. TEO-based features measure the non-linearity of the spoken utterance through time, and thus they provide information about the level of stress and thereby arousal of the speech.

There are other features that reveal the excitement of speech such as paralinguistic features that are related to glottal pulses. The characteristics of the paralinguistic features are called the Voice Quality Parameters (VQP) and they refer to the physical characteristics of the vocal tract and glottal movement \[10\]. Generally speaking, the vocal tract characteristics are appealing to phoneme recog-
ition [11]. Therefore for SER problems researchers are interested to compensate
the vocal tract effects on the produced utterance and keep the glottal pulses. There
have been studies that show VQP which represent physical characteristics of glot-
tal movement that can be estimated with non-invasive methods using different
harmonics in signals [10].

In 2015, Wang et. al. presented a feature extraction method as a combination
of the spectral features and Fourier Parameters (FP) of the speech signal. They
made use of an SVM model on the first 20 harmonics of the signal. In their paper,
they applied global features using statistics such as mean, maximum, minimum
and standard deviation of FPs [15]. In the same year, using Deep Neural Networks
(DNN) models, Fayek et. al. have introduced a model that utilizes the raw speech
spectrograms. They claimed that the chance of lowering the accuracy of the ma-
chine learning algorithm and the lack of model generalization is relatively high in
the presence of a load of preprocessing on the input signal [5].

The following sections discuss feature extraction methods and classification
techniques. The article continues with the experiments on three emotional databases:
SAVEE, EMO-DB and eNTERANCE and their results are studied and tabulated
under the experimental evaluation section. Finally, we will discuss the remarks in
the conclusion section.
Chapter 3

Approach

In this chapter, we show different feature extraction techniques to improve the effectiveness of the classifiers in different emotional speech data sets. We have made use of a single-layer ANN model for TEO-based global features to discriminate the arousal of the speech (active/passive classes). This structure is followed by two other ANN models using the MFCC global features to measure the valence of each prediction from the previous classifier. In the sense that each of two groups of emotions gets fed to one of these two classifiers separately. These steps are deployed as illustrated in Figure 3.1. Every speech processing and recognition has at least two main phases in it: 1) feature extraction techniques and 2) learning models. In the following two subsections we will discuss the methodologies to extract the aforementioned features and the techniques to build the tandem classification system consisting of two feature extraction methods and one classification technique.
Figure 3.1: An Example of a Tandem Classification System using Hybrid Features.

### 3.1 Feature Extraction

#### 3.1.1 Spectral Features

As it can be interpreted from their names, spectral features picture a spectrum of a signal in a range of a certain time interval (e.g. 25 milliseconds). Mel-Frequency Cepstral Coefficient (MFCC) features are derived from spectral features. Cepstral features are used to efficiently represent input signal. This feature has been used for many different applications in speech recognition. In our approach, one MFCC vector indicates a single time-frame of the input signal. Not only do we calculate the energy over the range of frequency for each frame but also we monitor the
spectrum of each of these frames with respect to their previous ones. A positive impact of using MFCC features is that unlike TEO-based features we can detect the mimiced emotion via MFCC features [14]. However, there are potential problems using this approach for our particular model that are reserved to be discussed later on in chapter 4.

This feature calculates the Discrete Cosine Transform (DCT) transform of the mel-scaled power spectrum of a signal. Before extracting this feature, we should pre-process the input signal. First we digitize an analog signal and transform it into a digital signal using Analog-To-Digital Codec (ADC). Second we do sampling from the digitized vector from the previous step given that the sampling rate must obey the Nyquist theorem. This operation is followed by scalar quantization.

According to the Figure 3.2 (with our contribution in dark blocks), pre-emphasis is the first step in processing the actual digital signal. Using pre-emphasis on audio signal we remove the spectral tilt problem by boosting up higher frequencies because they naturally decay due to the radiation loss at the lips. The speech is a modality with non-stationary data, therefore a framed-based speech processing is required. Hence, the signal is broken up into segments of roughly stationary frames of equal length (e.g. 25 milliseconds). A slight delay or gap between every frame pair is called frame shift that is at most half the frame length (e.g. <= 12.5 milliseconds). This process is called windowing. There are different windowing functions that could be applied on each frame. Since a rectangular window will cause discontinuity and spikes in its boundaries, we apply a bell shaped window (hamming function) to have each frame smooth on both ends. In order to get the power spectrum of each frame we get the signal transformed from time-domain to a frequency domain with the Fast Fourier Transform (FFT). Due to the non-linear
characteristics of the human auditory system and its insensitivity toward higher frequencies, we compute filter bank energies by mapping the magnitude frequen-
cies onto a mel-scale filter bank (a Hertz scale pitch transformed to a mel scaled pitch). Taking the log of each mel-scaled feature, we take the feature back to the time domain by taking the Discrete Cosine Transform (DCT). Per MFCC feature vector we derive a 13-D of features consisting of 12 cepstral features plus the energy of the frame which is a summation over squared value of each sample through time.

\[ \text{Energy} = \sum_{t=1}^{N} x^2[t] \quad (3.1) \]

where \( N \) refers to the length of the frame in terms of samples. At this moment, we apply a Sinusoidal liftering filter on every feature vector per frame with the liftering parameter \( L = 22 \) which represents the length of the Sinusoidal filter; the number of samples. This filter is a high-pass filter to re-emphasize the higher frequencies.

The MFCC feature extraction process is finalized by adding the first and second derivatives of each feature vector followed by their energy. These two vectors are \( MFCC_\delta \) (velocity feature) and \( MFCC_{\delta\delta} \) (acceleration feature) respectively [11] and they give us a better understanding of cepstral feature trends per frame. To calculate the derivatives, we have used stencil point of 5 for all experiments to get a more accurate derivative for deltas by calculating the slope between the current frame and 5 points preceding and following frames.

At this point, the \( MFCC_{\delta\delta} \) feature vectors are extracted for the entire audio file. For simplicity sake we refer to \( MFCC_{\delta\delta} \) as MFCC. Due to unbalance datasets inevitably we get different number of feature vectors for different emotion classes. However, we stratify data points in both train and test sets such that they have the same distribution of emotions.
3.1.2 \(MFCC_g\) Features

In \(MFCC_g\), \(g\) stands for global features. To get the \(MFCC_g\) derived from the 39-D MFCC vectors per frame, this paper proposed \(S = 7\) statistics including mean, max, min, median, standard deviation, kurtosis and skewness by which we can monitor signal’s behavior over time. The skewness reveals the symmetry of the sound in terms of the distribution curve of MFCC vectors over time, and the kurtosis shows the sharpness of this curve. As the result of this layout, we will have a feature vector of length \(39 \times S = 273\) for multiple frames per audio file.

3.1.3 TEO-based Features

Teager Energy Operator (TEO) features inspect the characteristics of speech where some amount of stress exists in the utterance. TEO features measure the non-liwith nearity of the utterance by processing its behavior through both frequency and time domain. The Figure 3.3 (our contribution in dark blocks) shows the process of TEO-CB-Auto-Env feature extraction. In this paper, we extract TEO-CB-Auto-Env features according to [17]. We show the feature extraction process for two different emotions: anger and boredom which are respectively classified as active and passive emotions with respect to their arousal value (Figure 3.4). TEO feature extraction method gets an input signal passed through a filter bank comprising with \(M = 16\) Band-Pass Filters (BPF) (e.g. Butterworth) that are called Critical Bands (CB). The boundaries for these CBs are tabulated in Table 3.1. Each one of these CBs represents a sub-band energy which is an energy distribution of a certain frequency range that could be regarded as the signal behavior in the frequency domain.
<table>
<thead>
<tr>
<th>Band Number</th>
<th>Critical Band Frequency (Hz)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Center</td>
<td>Upper</td>
<td>Bandwidth</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td>300</td>
<td>350</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>400</td>
<td>450</td>
<td>510</td>
<td>110</td>
</tr>
<tr>
<td>5</td>
<td>510</td>
<td>570</td>
<td>630</td>
<td>120</td>
</tr>
<tr>
<td>6</td>
<td>630</td>
<td>700</td>
<td>770</td>
<td>140</td>
</tr>
<tr>
<td>7</td>
<td>770</td>
<td>840</td>
<td>920</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>920</td>
<td>1000</td>
<td>1080</td>
<td>160</td>
</tr>
<tr>
<td>9</td>
<td>1080</td>
<td>1170</td>
<td>1270</td>
<td>190</td>
</tr>
<tr>
<td>10</td>
<td>1270</td>
<td>1370</td>
<td>1480</td>
<td>210</td>
</tr>
<tr>
<td>11</td>
<td>1480</td>
<td>1600</td>
<td>1720</td>
<td>240</td>
</tr>
<tr>
<td>12</td>
<td>1720</td>
<td>1600</td>
<td>2000</td>
<td>280</td>
</tr>
<tr>
<td>13</td>
<td>2000</td>
<td>1850</td>
<td>2320</td>
<td>320</td>
</tr>
<tr>
<td>14</td>
<td>2320</td>
<td>2150</td>
<td>2700</td>
<td>380</td>
</tr>
<tr>
<td>15</td>
<td>2700</td>
<td>2500</td>
<td>3150</td>
<td>450</td>
</tr>
<tr>
<td>16</td>
<td>3150</td>
<td>3400</td>
<td>3700</td>
<td>550</td>
</tr>
</tbody>
</table>

Table 3.1: Critical Band Frequency Information in Hz for 16 Sub-Bands.

Back to the time domain, each output stream from CBs will go through TEO profile estimation process as shown in Figure 3.5b. For the TEO-profile estimation, each of \( M \) signal output is segmented into frames of an equal length (e.g. 25 milliseconds with 10 milliseconds of frame-shift); where \( M \) is the number of critical bands, and \( f \) is the frame number for which the TEO profile is being extracted.

\[
\Psi_M[x_f[t]] = (x_f[t])^2 - (x_f[t - 1]x_f[t + 1])
\]  

Next, we extract envelope area parameters of \( M \) normalized Auto-Correlation Functions (ACF) for each time frame. According to a proposed method by Cheveigne and Kawahara in [2], given a signal \( x \) an ACF of each time-frame is denoted by
where \( l \) stands for the lag time or delay.

\[
r_x[l] = \sum_{n=0}^{n=N-1-l} (x[n]x[n+1])^2; l = 0, ..., N - 1
\]  \hspace{1cm} (3.3)

Accordingly, we calculate the normalized ACF for every frame. Subsequently, for every result an amplitude of the analytic signal (envelope) will be computed by calculating the absolute values of the Hilbert transform of each ACF. We have approximated the area under envelope using the Simpson's rule to avoid sharp edges around each subinterval. Then we normalize the area under envelope per time frame. In Figure 3.5c the broken line shows an unnormalized amplitude analytic signal. Eventually for each frame we derive \( M \) scalar values representing TEO-CB-Auto-Env features. Note that the area under envelope is scaled between 0 and 1. The maximum area value is then equal to the half the length of each frame \((N/2; \text{where } N \text{ refers to the length of the frame.})\) which is a straight descending line [13]. The TEO-CB-Auto-Env features show no presence of excitement if the area under envelope is very close to \( N/2 \). Hence, for the active-passive classifier, closer to \( N/2 \) means closer to the passive class.

### 3.1.4 TEO\(g\) Features

In this feature, \( g \) refers to global type of features. In order to derive the \( TEOg \) features from TEO-CB-Auto-Env features, we have calculated 7 statistics to exploit the speech behavior over time. Similar to the MFCC-\(g\) features, we extract the kurtosis, skewness, standard deviation, median, mean, max and min to monitor this feature's trend across frames. Eventually, we come up with a feature vector of length \( M \times S = 112 \) per audio file.
Additionally, we expanded the tandem system by defining it as the system that utilizes the $TEO_g$ features in its first hierarchy while $MFCC_g$ features in the second hierarchy and we denote it as the hybrid feature. The $TEO_g$ feature at the first hierarchy supports the fact that our system determines the amount of stress (arousal) followed by the recognition of the positivity (valence) of the spoken utterance (Figure 2.1).

### 3.2 Classification: Artificial Neural Networks

The Artificial Neural Network (ANN) model is among classic Machine Learning (ML) models that is superior for its fast evaluation of the learned target function [12] and for being robust to noisy or erroneous data set which seems to be very common specifically in speech-related classification problems.

The input layer of each ANN in our study depends on the level of hierarchy in the tandem system that will be discussed individually under the following two sub-sections. Regarding the architecture of the models, the size of the hidden layer is 165 units. As a non-linear and differentiable activation function, a Sigmoid unit is designated in the intermediate layer (Equation 3.4).

\[
f(z_i) = \frac{1}{1 + e^{-z_i}} \quad (3.4)
\]

A softmax function is used for the output layer (Equation 3.5).

\[
f(z_i) = \frac{e^{z_i}}{\sum_{k=1}^{N} e^{z_k}} \quad (3.5)
\]

To speed up the convergence and avoid getting stuck in the local minima we have
used the momentum coefficient equal to .9 in our experiments. The Limited-Memory Quasi-Newton Method (L-BFGS) is used as the line search algorithm in all of our model trainings. In spite of the basic Newton optimization algorithm which works perfectly on small datasets we have made use of the L-BFGS algorithm for the sake of the scalability of the application for which the Hessian matrices are costly to be computed. Regarding the overfitting problem while training the model, we have implemented a 10-Fold Cross Validation with 10% of the dataset as the hold-out set. The weight difference threshold is set to 1e-5.

In the next chapter we will discuss the experiments in further details.

3.2.1 Hierarchy I

In this article, we use a binary classifier architecture of Multi-Layer Perceptron (MLP) whose input layer is $TEO_g$ features that have already been discussed in this chapter. As mentioned before, the length of the feature vector for a single example that gets fed into the ANN model is 112 that represents $S \times M$ features including $S = 7$ statistics of $M = 16$ critical bands.

The first hierarchy is to classify the data into two active and passive emotions with respect to their arousal. The classified samples from this phase will play the role as input samples to the next hierarchy in the sense that the classified active and passive samples will then form two other hyperplanes separately.

3.2.2 Hierarchy II

In the second hierarchy, two ANN models are formed to discriminate a group of five emotions in the active and at most three emotions in the passive category. Per ANN model we have utilized a multi-class MLP model with a single layer as their
hidden layer and $MFCC_g$ feature vectors as their input of the length of 273 scalar values (See section 3.1.2).

As an alternative model in this step, an RNN/LSTM [6] was designated for this step to take into account the sequence-based characteristics of speech in the model. However, due to the relatively fewer number of data points with respect to the higher number of updating weights and the time-step in the RNN/LSTM model, but we used an ANN model but included the time-based characteristics of speech in the feature extraction phase.
Figure 3.3: The Schema of the $TEO_g$ Feature Extraction from TEO-CB-Auto-Env.
Figure 3.4: The Original Audio File Showing Two Active (a) and Passive (b) Emotions.
Figure 3.5: The Extracted TEO Feature for the First Critical Band of an Anger (Active Emotion)
Figure 3.6: The Extracted TEO Feature for the Critical Band 6 of an Anger (Active Emotion)
Figure 3.7: The Extracted TEO Feature for the Critical Band 16 of an Anger (Active Emotion)
Figure 3.8: The Extracted TEO Feature for the First Critical Band of Boredom (Passive Emotion)
Figure 3.9: The Extracted TEO Feature for the Critical Band 6 of Boredom (Passive Emotion)
Figure 3.10: The Extracted TEO Feature for the Critical Band 16 of Boredom (Passive Emotion)
Chapter 4

Experimental Evaluation

In this section we discuss a tandem system consisting of three connectionist models (ANNs) over three emotion-based speech databases leads to a reasonable performance. Two categories of features that are fed into these models are TEO-based ($TEO$ and $TEO_g$) and spectral features ($MFCC$ and $MFCC_g$). In chapter 3 we have discussed the steps toward each feature extraction technique. In this chapter we will discuss the results out of the classification evaluations on each one of the acoustic features. Particularly we claim that our tandem system is superior versus a non-hierarchical classification setting. Then in the tandem architecture, we will discuss the trade-offs of using different features.

4.1 Procedures

This article provides an application to demonstrate the performance of the SER system in the tandem setting. The entire system is built and executed in the Linux Ubuntu 64-bit OS platform installed on a host machine with the processor Intel
Core i7-7500U CPU @ 2.70GHz, 12 GiB of memory and Intel HD Graphics 620 of the graphic card. The $MFCC_g$ and $TEO_g$ features have been extracted via Python 3.5 whereas the classifier is implemented and tested in PySpark 2.2.1 with Python 2.7.12 for the sake of the system scalability.

Given the previously extracted features from our Python program, the data visualization and Principal Component Analysis (PCA) are performed in PySpark merely using the stored Comma Separated Values (CSV) files containing the features.

### 4.2 Datasets

We have performed our tandem-system classification approach on three datasets: the Surrey Audio-Visual Expressed Emotion (SAVEE) Databases, Berlin Dataset of Emotional Speech (EMO-DB) and eNTERFACE Audio-Visual Emotion Database (eNTERFACE). The time duration of the audio files are not necessarily of the same length, however it is assumed that each file is monophonic and represents one single emotion. Moreover, the emotions in each data set are not uniformly distributed. Hence, the model training technique is constrained to unbalance data classification. In our experiments, every audio file is segmented into frames of length 25 milliseconds (i.e. 1102 samples) with 10 milliseconds of frame shift length. In all three datasets no gender-based model is determined to be learned. No silence removal has been applied on these audio files, therefore inevitably the Signal-to-Noise Ratio (SNR) of the data sets is lower than when the silence-removal process is applied.

Prior to going through detail of each dataset, first we outline the list of emotions
that exist in our datasets. Table 4.1 shows emotions in each of three datasets, categorized based on the cognitive model from the second chapter. As seen in the table, the datasets have missing some passive emotions in significant level.

Table 4.1: Emotions that Exist in Datasets

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Active</th>
<th></th>
<th></th>
<th>Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>happy</td>
<td>surprise</td>
<td>anger</td>
<td>fear</td>
</tr>
<tr>
<td>Emo-DB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SAVEE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>eNTERFACE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

4.2.1 SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) data set contains 7 archetypal emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. Each audio file is named according to the dominant expressed emotion during the speech. We have modified the names of the audio files such that the name of the emotion is preceded by either "active" or "passive" term, depending on the class to which they are assigned. The audio files are recorded with the sampling frequency of 44.1 KHz, mono-channel and digitized with the PCM modulation of 16-bit depth resolution. The four speakers in this data set are male speakers with the British accent. The SAVEE data set has 2 of its expressed emotions in the passive category: sadness and neutral.

4.2.2 EMO-DB

The Berlin Dataset of Emotional Speech (EMO-DB) with 536 audio files is a German data set that consists of 7 archetypal emotions including anger, boredom,
disgust, fear, happiness, sadness and neutral. In terms of the gender of speakers Emo-DB has an equal number of 5 male and female speakers. The audio files originally were sampled with the rate of 16 KHz with 16-bit per sample. In our article, we have upscaled the audio files to the sampling rate of 44.1 KHz for the sake of consistency with the other data sets during the experiments. We have 3 emotions in this dataset in the passive category: boredom, sadness and neutral.

4.2.3 eNTERFACE

The eNTERFACE Audio-Visual Emotion Database is a non-native English spoken data set with the Belgian accent. This database consists of 1257 audio-visual files of 6 archetypal emotions: anger, disgust, fear, happiness, surprise and sadness. The audio files are consistent with two other datasets and thus they are digitized with 16-bit depth PCM modulation and sampled with 44.1 KHz rate. There are 43 speakers among whom 8 speakers are female. In this data set there is only one emotion (sadness) in the passive class.

4.3 Evaluation

To evaluate the performance of the system, we take an unweighted (UW) average accuracy over the classifiers in the tandem system. The process of measuring the effectiveness of the model is simply the calculation of the arithmetic mean of the accuracy of the first hierarchy and two other accuracies of the second hierarchy.
4.4 Result

In this section we will tabulate and plot the collected results. First, we have visually compared the MFCC feature vector to the $MFCC_g$ and the TEO feature vector to the $TEO_g$ by plotting them in Figure 4.1. The high dimensional features undergo a PCA dimensionality reduction method so that we will be able to visualize the significant difference between our features versus the $MFCC$ and $TEO$ features.

![Figure 4.1: The Result of the PCA for Spectral and TEO-based Features for the Emo-DB Corpus.](image)

The ANN models follow the single-layer architecture. According to our experiments increasing the hidden layers does not make improvements in our results.
Thus, we have our ANN model configured as a single-hidden layer MLP with 165 hidden units. The method to pick this certain number of hidden units is depicted in Figure 4.2. This figure shows three curves across different number of hidden units. In x-axis of the figure every slot represents a particular number of hidden units starting from 5 to 200. To get three curves computed, first we have executed the algorithm for the sub-categories (active and passive sets) of all datasets, and thus in total we have 10 accuracy values (in the scale of 0 to 1) for every slot. Given these 10 curves, the blue curve is the average of these accuracies across hidden units. Subsequently, the red and yellow curves will show the maximum and minimum accuracy values for each slot. The green histogram, the Standard Deviation, illustrates how sparse are the results of the execution per hidden unit. The appealing number of hidden units would be where the average (blue curve) is at its highest while the standard deviation (green histogram) is the minimum. The lower hidden units are not of our interest because of high sparsity in the accuracies (histogram) and low accuracy (blue curve). Thereby, to some degree, the hidden units in the middle would be appealing. Ultimately, we have trained all of our ANN models with 165 hidden units due to the lowest sample deviation and the highest accuracy.

The tandem system has been reflecting a higher effectiveness than the baseline. We assume that the baseline classification is a non-hierarchical system in which only one classifier is being trained over all emotions. Follow Table 4.2 for the comparison between $MFCC_g$ and $TEO_g$ in a tandem system with the same pair in a non-hierarchical system.

The tabulated result in Table 4.3 depicts the accuracies (in percent) of different features for three emotion-based datasets. Note that all the executions are based
Figure 4.2: The Trend of Accuracy for Different Number of Hidden Units

Table 4.2: The Comparison between the Tandem Architecture and the Baseline

<table>
<thead>
<tr>
<th>Feature</th>
<th>Emo-DB</th>
<th>SAVEE</th>
<th>eNTERANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC&lt;sub&gt;g&lt;/sub&gt; – Tandem</td>
<td>77.26</td>
<td>71.42</td>
<td>66.49</td>
</tr>
<tr>
<td>MFCC&lt;sub&gt;g&lt;/sub&gt; – baseline</td>
<td>63.66</td>
<td>51.56</td>
<td>46.7</td>
</tr>
<tr>
<td>TEO&lt;sub&gt;g&lt;/sub&gt; – Tandem</td>
<td>73.36</td>
<td>60.42</td>
<td>55.35</td>
</tr>
<tr>
<td>TEO&lt;sub&gt;g&lt;/sub&gt; – baseline</td>
<td>55.77</td>
<td>35.16</td>
<td>24.8</td>
</tr>
</tbody>
</table>

on 2-hierarchy system for the sake of consistency in our experiments. The final performance of the entire system is calculated using an average accuracy of the three trained models where the first model recognizes the excitement of the voice and the second and third models are responsible to measure the pleasantness of an active and passive classes separately. As it can be seen from this table, the tandem system outperforms the baseline system in all tests.

In Table 4.3, we compare the performance of each feature in all three corpora. According to the result, among last five features that are implemented in our study
the prediction result for the $MFCC_g$ is the highest for Emo-DB with 77.26% comparing to the rest of the features with 71.42% and 66.48% for SAVVE and eNTERFACE corpora respectively. The second best among our proposed features is the hybrid feature with a slightly different prediction rate of 75.06%, 67.59% and 65.19% for Emo-DB, SAVVE and eNTERFACE corpora respectively.

To compare our result with the previous works, the accuracy of the Hybrid system for both SAVVE and eNTERFACE datasets are higher than the results from [5] for both $MFCC_g$ and Hybrid features. However regarding the Emo-DB dataset the accuracy that Wang et. al. provided in [15] using their (global) FP + MFCC features on an SVM model is %2.25 and $4.44$ higher than our $MFCC_g$ and Hybrid method, respectively.

Table 4.3: UW Accuracy for Different Features

<table>
<thead>
<tr>
<th>Accuracy (%)</th>
<th>Emo-DB</th>
<th>SAVVE</th>
<th>eNTERANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fayek et. al.</td>
<td>-</td>
<td>59.7</td>
<td>60.53</td>
</tr>
<tr>
<td>Wang et. al.</td>
<td>79.51</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$MFCC$</td>
<td>65.321</td>
<td>57.227</td>
<td>57.947</td>
</tr>
<tr>
<td>$TEO$</td>
<td>48.245</td>
<td>48.981</td>
<td>52.77</td>
</tr>
<tr>
<td>$MFCC_g$</td>
<td>77.264</td>
<td><strong>71.423</strong></td>
<td><strong>66.488</strong></td>
</tr>
<tr>
<td>$TEO_g$</td>
<td>73.358</td>
<td>60.424</td>
<td>55.349</td>
</tr>
<tr>
<td>Hybrid</td>
<td>75.067</td>
<td>67.596</td>
<td>65.197</td>
</tr>
</tbody>
</table>

Although the hybrid feature is the second best with a slight difference in terms of the effectiveness, the experiments show that the tandem system with the hybrid feature is more efficient than the system with the $MFCC_g$ feature vectors. According to Table 4.4, we have calculated the arithmetic mean of 10 attempts of training and testing over each aforementioned system with the 10-fold cross validation. The results return the elapsed run time of the system with the hybrid feature in all three datasets shorter than the one with only $MFCC_g$ feature by at least 10
seconds. Note that the features for all datasets are extracted separately in advance, and thus this report of the time consumption regards only the model training and testing phase. The hybrid feature in the tandem system for the Emo-DB dataset is 1.29 times faster than the same system with $MFCC_g$. Similarly, the system is 1.22 times faster with the hybrid feature in the SAVEE data set compared to the $MFCC_g$ feature vectors. Finally, the eNTERFACE dataset runs 1.11 times faster with the hybrid feature in a tandem system than the $MFCC_g$ features.

Table 4.4: The Elapsed Time for Classifying Emotions using Different Features in Tandem System

<table>
<thead>
<tr>
<th></th>
<th>Emo-DB</th>
<th>SAVEE</th>
<th>eNTERANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MFCC_g$</td>
<td>59.91</td>
<td>56.34</td>
<td>112.57</td>
</tr>
<tr>
<td>Hybrid</td>
<td>46.44</td>
<td>46.36</td>
<td>101.83</td>
</tr>
</tbody>
</table>

4.5 Analysis

In our study, the $MFCC_g$ spectral features have returned reasonable degree of separability in our trained model more than the $TEO_g$ features because of the fact that the former features are of a higher dimension in spite of the latter ones. This statement suggests that models trained based on $MFCC_g$ features are able to discriminate data with more complexity. Thus, according to the results from running the tandem system for only $MFCC_g$ in both hierarchies versus the Hybrid features ($TEO_g$ and $MFCC_g$ for the first and second hierarchies, respectively), we showed that the spectral features surpassed the Hybrid features with a slightly better performance.

On the other side, a potential draw-back with feeding MFCC features to the
recognition system is due to the DCT transformation that is a linear transform while the speech is non-linear by nature and thus the recognizers accuracy is generally alleviated in average.

The tandem system based on cognitive model is debatable for including the stress detection followed by the pleasetness of emotion in speech. There is a significant trade-off between including $TEO_g$ in the system and excluding it. In spite of the $MFCC_g$, the $TEO_g$ has the privilege of prompting a faster recognition from our classifiers. Thereby, although our hybrid system stands as the secound best in terms of the effectiveness, it has a higher efficiency than the $MFCC_g$ system.

Applying gender-based models in our system seems to be a potential improvement particularly in the prediction analysis based on $TEO_g$ features. There is a possibility that our methodology for the $TEO_g$ feature extraction could be improved by performing the extraction based on slightly different critical sub-bands for different genders.
Chapter 5

Conclusion

We showed that a tandem system comprising with three ANN models forming two hierarchies gives higher recognition rates for SAVEE and eNTERFACE datasets and a slightly lower recognition rate for the Emo-db dataset. Based on the cognitive model, we have divided the system such that it performs classification using two types of classifiers: binary-class MLP for the arousal detection and multi-class MLP to spot the valence (pleasentness). We defined a hybrid system that comprises with a binary-class classifier that takes the $TEO_g$ as its input layer in its first hierarchy, and two different multi-class MLPs that get $MFCC_g$ features as for their input layers for the second hierarchy of the hybrid system. Our speaker- and gender-independent emotion recognition method to classify 8 architypal emotions returned a better recognition rate for SAVEE and eNTERFACE datasets in comparison with the baseline system and recognition systems in [5]. However a combination of the Fourier Parameters and MFCC features with an SVM recognizer in [15] has returned an accuracy of %79.51 which is %4.4 higher than our Hybrid system and %2.2 higher than our tandem system that uses only $MFCC_g$. 
Bibliography


