

Speech-Based Emotion Recognition

A Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment
of the requirements for the degree

Master of Science


by

Ye Gao

May 2019

APPROVAL SHEET

This Thesis
is submitted in partial fulfillment of the requirements
for the degree of
Master of Science

Author Signature: 

This Thesis has been read and approved by the examining committee:

Advisor: John Stankovic

Committee Member: Alfred Weaver

Committee Member: John Lach

Committee Member: _____

Committee Member: _____

Committee Member: _____

Accepted for the School of Engineering and Applied Science:



Craig H. Benson, School of Engineering and Applied Science

May 2019

To my parents.

ABSTRACT

Many algorithms on speech-based emotion detection that utilize machine learning are published. They are often trained and tested on datasets that consist of audio clips in which the speaker emulates emotion such as anger, happiness, neutrality, and sadness. Despite the high accuracy that the algorithms have achieved, they are not suitable for real-life deployment for two reasons. First, the datasets are often times collected in strictly controlled environments where noises are minimum and the microphone is placed very close to the speaker, which is not representative of real-life environments in which background noises are present and people are not expected to be adjacent to the acoustic sensor(s) all the time. Second, each audio clip is usually uttered by an actor, and labeled with the emotion that the actor attempts to simulate. However, research indicates no evidence that the acoustic features of acted emotion are representative of the acoustic features of authentic, spontaneous emotions. As a result, algorithms trained on acted speech may not achieve the same excellent performance when deployed in real-life environments to detect emotions in people's speech. This thesis explores different approaches to address the problem that high-performing machine learning classifiers on speech-based emotion recognition may not be fit for use in real-life deployment, and proposes an acoustical classifier for emotion detection fit for real-life deployment. The classifier is intended to be part of a smart healthcare system to monitor the users' emotions.

TABLE OF CONTENTS

Abstract	v
Chapter 1: Introduction	1
Chapter 2: Related Works	4
Chapter 3: Approach	6
3.1 Datasets	6
3.1.1 Danish Emotional Speech Datasbase (DES)	6
3.1.2 Berlin Database of Emotional Speech (EMO-DB)	7
3.1.3 Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)	7
3.1.4 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	8
3.1.5 Surrey Audio-Visual Expressed Emotion Database (SAVEE)	8
3.1.6 Electromagnetic Articulography Database (EMA)	9
3.1.7 Toronto Emotional Speech Set (TESS)	9
3.2 Acoustic system overview	10
3.3 Feature selection	11
3.4 Dataset compensation	12
3.5 Audio signal pre-processing	13

3.6	Emotion classifier	14
Chapter 4: Microphone Selection		17
Chapter 5: Evaluation on the emotion classifier		21
5.1	The downfall of playing out audio clips to the microphone via an electronic speaker	21
5.2	Trained on clean datasets	23
5.3	Trained on the synthetic dataset	29
5.4	Comparison of the performance of the classifiers trained on the clean dataset and synthetic dataset	30
5.4.1	Validation accuracy	30
5.5	Accuracy when evaluated on TESS	31
5.6	Accuracy when evaluated on live speech	33
Chapter 6: Conclusion		37
Chapter 7: Future works		38
References		41

CHAPTER 1

INTRODUCTION

Emotion plays an essential role in one's physical and psychological well-being. Negative emotions, such as sadness and anger, are correlated to underlying mental health issues which, if left untreated, may result in serious complications such as self-harm or even suicide. In cases when the negative emotions are not caused by mental health problems, they can be unpleasant at the present moment and negatively influence one's daily life and interaction with others. Emotion can be picked up by one's voice. With more and more acoustic sensors being connected to the Internet, we see a surge of interest in speech-based emotion detection in the field of affective computing.

This thesis presents a speech-based emotion detection classifier. When deployed, this classifier keeps track of each individual's emotion in home environments. Its output provides detailed record of each individual's emotion through time, which can prove useful to behavioral studies: For example, family eating dynamics (FED), consisting of dining environment and time at home, are potentially crucial to the diners' dietary intake. Family dining environment consists of the dinners' emotion [1]. Therefore, the detailed understanding will provide better insights into FED and problems related to dietary intake, such as obesity. Likewise, research into maintaining healthy and pleasant patient-caregiver relationship can benefit from the output of the pipeline. Taking care of patients suffering from chronic diseases such as dementia is a demanding and exhausting task. The detailed understanding will detect the onset of negative emotions from the caregivers. Based on the knowledge, negative emotions could be managed via a recommendation system that recommends a technique to control negative emotions. Therefore, the detailed understanding will be crucial to emotionally healthy environments in which caregivers tend patients.

The main contributions of the proposed acoustic system are:

- **Identified major limitation in published speech-based emotion detection approaches.** The datasets are often collected from actors, and research [2] suggests that the evidence that acted emotional speech can be used in the place of authentic, natural emotional speech is lacking. Even when trained and evaluated on emotional speech representative of the actual emotions, state-of-the-art approaches assume ideal sound-collecting environments in which reverberation and background noises are minimal, because many datasets are collected in strictly controlled studio environments in which the effect of reverberation and background noises are kept minimal with professional noise reduction and reverberation reduction sound devices. When training and evaluating solely on the datasets, the classifiers are not adaptive to the realistic scenario in which reverberation and background noises are constantly present.
- **Designed for realistic environment.** Having identified the major limitation of state-of-the-art approach, the speech-based emotion recognition classifier is suitable for deployment in realistic environments where background noises as well as reverberation are present. The microphone used to pick up sounds is distance-resistant. The transcription accuracy of the speech it captures does not degrade even if it is placed 5 meters away from the sound source, which is about the width of an average-sized room. A state-of-the-art noise-filtering algorithm [3] is also used to filter out noise in the audios captured by the microphone, before the audio samples are passed to the emotion classifier. The emotion classifier is trained on a synthetic dataset, which are obtained from randomly amplifying/de-amplifying and adding reverberation and background noises to the original audio clips from emotional speech datasets to imitate realistic environments.
- **Designed for rapid deployment.** The acoustic system only requires a computer and a specific microphone, both of which are available on the market. The computer does

not have to be an expensive high-end product with best computing powers, and the microphone is available at a reasonable price, despite its exceptional performance in preserving acoustic features at different distances. The microphone is powered via a USB-cable, so it can be directly connected to the computer. Batteries and additional work to forward audio clips captured by the microphone to the emotion detection classifier are unnecessary. The classifier is pre-trained, so there is no need to gather acoustic samples from the environment in which it is about to be deployed. Because of the availability and low maintenance of the components in the acoustic system, the system is fit for rapid deployment.

- **Wide range of potential applications.** Given that it is designed for realistic environments, the acoustic pipeline has a wide range of potential applications, such as an analysis tool for the development or progression of one's emotional states over time. Combining the analysis of the emotional progression of more than one individuals in a household, analysis can be made on family eating dynamics (FED) and patient-caregiver relationship.

CHAPTER 2

RELATED WORKS

Most state-of-the-art algorithms are trained and evaluated on datasets of emotional speech. Two popular datasets of emotional speech are the Danish Emotional Speech Database (DES) [4] and Berlin Database of Emotional Speech (EMO-DB) [5]. A study [6] provides a list of machine learning classifiers, such as [7], [8], [9], and [10], trained on DES and a list of machine learning classifiers trained on EMO-DB. In the former lists, approaches include GentleBoost [11], Bayes classifier [12], instance based learning [7], and vector quantification [10]. In the latter list, approaches include SVM [13], GentleBoost [11], linear discriminant analyses [14], and Two-stage neural network [15]. [16], a work not in the previous lists, achieves a very high accuracy (88.9%) on the EMO-DB dataset using CNN and LSTM. However, there are two issues with training and evaluating solely on datasets of emotional speech.

- **The emotion in some clips from datasets of emotional speech is not typical enough for humans to discern.** Although emotions are highly subjective, it is not difficult for human beings to pick up the emotion from one's speech when that emotion is typical. In some datasets of emotional speech, the emotion in the audio clips are not typical enough and humans cannot discern these emotions with satisfying accuracy. [6] compares the accuracy of machine learning classifiers with human evaluation. Human evaluation on DES only yields an accuracy of 0.67. There is no evidence suggesting that classifiers trained on DES will be able to yield a reasonable accuracy when tested on speech that conveys typical emotions. Thus, classifiers trained on speech datasets of atypical emotions are not usable in real-life deployment.
- **The datasets of emotional speech are collected in laboratory environments.** There

are datasets on which human evaluation yields reasonable accuracy; the human evaluation on EMO-DB results in an accuracy of 0.86 [6]. However, most available datasets of emotional speech, such as [4], [5], [17], [18], [19], are collected in laboratory environments in which the acoustic sensors are closed to the speaker's mouth and background noises are kept minimal with professional acoustic devices, the evidence suggesting that they will yield the same accuracy when applied in real-life scenarios is lacking.

A recent work [20] focuses on developing an automatic emotion recognition classifier that will work in a realistic scenario in which the speakers are not necessarily near the acoustic sensor. It interprets speech as the progression of acoustic states over time. An audio clip is segmented into small, overlapping frames, and each frame is represented by an acoustic state. Each acoustic state is associated with a word that describes the state, and each word is represented by the low level descriptors (LLD) extracted from the small frame. Using k-means clustering, [20] identifies the centroids of the clusters as words in the codebook. Each small frame is represented by the the set of centroids (words) to which it has the shortest Euclidean distances. Then, [20] uses the Emo2vec model to train the classifier based on the idea that if two audio clips are often observed in similar context (surrounding small frames), words associated with the small frames in both clips must be similar. Otherwise, the words associated with the small frames in both clips must be different. During the training, basic vector addition is applied to pull two vectors closer if they tend to show up in similar context, and basic vector subtraction is used to pull them apart if otherwise. It claims to be the first work that addresses realistic scenarios in which the speaker is not adjacent to the microphone, the likely scenario of a real-life deployment. However, only using two relatively small datasets of emotional speech [21], [22], the evidence that their approach is robust enough to handle noises and reverberation present in a realistic environment is lacking.

CHAPTER 3

APPROACH

3.1 Datasets

This section describes several publicly available datasets of emotional speech that are frequently used to develop automatic emotion recognition modules. While it is crucial to know that the validation accuracy of the classifiers trained on these datasets does not necessarily translate well into the accuracy of them being deployed in real-life, some of the datasets are good as training sets that allow features indicative of emotions to be pinpointed or correctly identified because these datasets are collected in idealistic environment in which background noise and reverberation effects are minimized. In other words, the most relevant features extracted from them will be accurate because the original audio samples are not distorted and presented at their best quality.

3.1.1 Danish Emotional Speech Datasbase (DES)

Danish Emotional Speech Datasbase (DES) consists of 260 samples of emotional speech in the Danish language obtained from two male and two female actors. There are five categories in the dataset: happiness, anger, neutrality, sad, and surprise. [4] claims that the actors are believed to convey the emotions realistically. Based on verification process that [4] uses, 67% of samples are correctly predicted by human evaluators. Danish and English have distinct linguistic features. Since the emotion recognition classifier proposed in this thesis targets the English-speaking population, DES is not a good candidate for the training of the proposed classifier.

3.1.2 Berlin Database of Emotional Speech (EMO-DB)

Berlin Database of Emotional Speech (EMO-DB) consists of 535 samples of emotional speech in the German language obtained from five male and five female speakers. There are seven categories in the dataset: anger, happiness, sadness, neutrality, boredom, disgust, fear. The speech samples are collected in a lab with high-end acoustic devices, while the microphone is placed in front of the speaker. Human evaluators can recognize 86% of the emotions in the samples correctly [6], indicating that the emotions in the emotional speech are typical enough for humans to recognize. EMO-DB appears to be a good candidate training set for speech processing classifiers in German, but it is unclear if its usability translates into English. Therefore, EMO-DB is not chosen as a training set for the proposed classifier.

3.1.3 Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)

Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) consists of 7442 samples of emotional speech obtained from 91 actors of different ethnicity. There are six categories in the dataset: anger, happiness, sadness, neutrality, fear, and disgust. Several human evaluators are asked to label the clips with the most likely emotion. The accuracy of human evaluation on the angry, happy, neutral, and sad audio samples are 68.2%, 62.4%, 67.2%, 54.1% [17]. The accuracy of human evaluation on the angry, happy, neutral, and sad audio-video samples are 76.1%, 89%, 71.7%, 59.9% [17]. The difference between the human evaluation accuracy on the audio-only samples and the human evaluation accuracy on the audio-video samples indicates that these emotions in the audio samples are still recognizable, although humans need visual cues help them recognize emotions more accurately. The audio clips are used as a part of the training set for the classifier proposed in this thesis.

3.1.4 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) consists of happy, angry, neutral, sad, calm, surprised, disgusted, and fearful emotional utterances of North American English obtained from a gender-balanced group of 24 professional actors. For all emotion classes except neutrality, there exist two types of intensity: normal and strong. On samples of happy speech with strong intensity, human evaluation yields an accuracy of 44%. On happy speech with normal intensity, the accuracy is 29%. While the accuracy seems low, the authors also provide the accuracy when the audios are paired with videos, and the accuracy for happiness with strong intensity becomes 84%, and happiness with normal intensity becomes 80%. On angry speech with strong intensity, the audio-only accuracy is 91%. On angry speech with normal intensity, the accuracy is 59%. When paired with videos, the accuracy for anger with strong and normal intensity is improved significantly to 94% and 75% respectively. On audio-only sad clips with strong intensity, the accuracy is 62%. On audio-only sad clips with normal intensity, the accuracy is 34%. Paired with videos, the accuracy for sadness with strong and normal intensity becomes 81% and 56%. The accuracy of human evaluation on audio-only neutral clips is 91%. The comparison between the accuracy of the audio-only samples and the accuracy of the audio clips paired with videos suggest that these emotions are indeed emulated with high accuracy, but voice alone may not be sufficient enough to correctly identify the emotions for humans. The audio clips from this dataset are used to train the classifier proposed in this thesis.

3.1.5 Surrey Audio-Visual Expressed Emotion Database (SAVEE)

Surrey Audio-Visual Expressed Emotion Database (SAVEE) consists of happy, angry, neutral, sad, disgusted, fearful, surprised samples of emotional speech. The dataset is obtained from four male actors and the audio samples are spoken in British English. There are 480 utterances of emotional speech in total. The speech samples are evaluated by ten human

evaluators, five of whom are native English speakers and the other have stayed in England for at least a year. For the recognition of all seven classes, the accuracy achieved by human evaluation is 67% when the evaluators are only given audio samples. The accuracy rises to 92% when the evaluators are given audio-video samples. The audio clips from this dataset are used to train the classifier proposed in this thesis.

3.1.6 Electromagnetic Articulography Database (EMA)

Electromagnetic Articulography Database (EMA) [22],[23] consists of happy, angry, neutral, and sad samples of emotional speech, produced by one male and two female speakers. The male speaker produces 70 sentences for each emotion and 280 emotional utterances in total. Each of the two female speakers produce 200 samples of emotional speech. Each of the samples of emotional speech is rated numerically at least three times by different human evaluators from different linguistic and cultural backgrounds. The accuracy on human evaluation is not explicitly spoken, but the authors provide a list of best utterances. A sample will be rated as a best utterance if it achieves high numerical result on its target emotion and low numerical result on other emotions. The dataset is used to train the classifier proposed in this thesis.

3.1.7 Toronto Emotional Speech Set (TESS)

Toronto Emotional Speech Set (TESS) [19] consists of happy, angry, neutral, sad, disgusted, fearful, and surprised samples of emotional speech, produced by two female actors aged 26 and 64 years. Both actresses have received proper musical training. For each of the emotion classes, the two actresses produce 100 samples together. 56 undergraduate students whose first language is English are recruited to evaluate the emotional utterances. The average accuracy for all the samples of emotional speech is 82%.

In this thesis, there are two training sets. The first training set is the clean dataset, consisting of the original, unaltered audio samples from CREMA-D, RAVDESS, SAVEE, and

EMA. The second training set is the synthetic dataset, consisting of the original audio samples from CREMA-D, RAVDESS, SAVEE, and EMA, as well as the altered versions of the samples - background noise, amplification/de-amplification, and reverberation are added to them. The performance of the emotion classifiers trained on the clean dataset and the synthetic dataset is compared. The details of the classifiers are described in Section 2.6.

Along with live person speech described in Chapter 5.6, TESS is used to test the classifier due to the following reason: As the dataset in English that yields the highest accuracy on human evaluation when the human evaluators are given audio samples only. In other words, this dataset consists of speech samples with the most acoustically recognizable emotions, indicating that the speech samples are most similar to spontaneous emotional speech when compared to the acted emotional speech samples in other English datasets. TESS, like the other datasets listed above, is collected in a strictly controlled lab environment where the microphone is placed adjacent to the speaker. Therefore, TESS can be seen as the set of clean, spontaneous emotional speech, with "cleanness" indicating that the speech clips are collected when background noise and reverberation effects are minimized. The classifier's performance when it is evaluated on TESS after training is indicative of its performance when it is deployed in an idealistic environment where a speaker is speaking spontaneously to a microphone placed near his or her mouth (distance between the speaker and the microphone is almost 0 meter) and the background noise is minimal (assuming optimal performance from the noise reduction algorithm in the pre-processing unit).

3.2 Acoustic system overview

The microphone used in the acoustic pipeline has a sampling rate between 44.1kHz - 48kHz and frequency response between 40Hz - 16kHz [24]. The evaluation (Chapter 5) indicates that it is distance resistant, capable of picking up sounds from 15 feet away with 100% transcription accuracy paired with a state-of-the-art transcription tool [37]. When the system starts running, the microphone begins capturing continuous speech signals. The continu-

ous speech signal is sliced into 5-second sound windows for further processing. A silence filter is applied to each of the sound windows. If more than an adjustable percentage of the sound window is silence, the entire sound window is regarded as silence and discarded. Otherwise, the sound signal is passed to the noise filter, then to the overlapped speech filter to get rid of the overlapped speech. After all the above pre-processing steps, the sound window is passed to the emotion recognition classifier.

3.3 Feature selection

There are two groups of features that can be extracted from a speech clip. The first group consists of features that can be extracted from small time frames in a clip, such as loudness and Mel-frequency cepstral coefficients. These features are low-level descriptors (LLDs). The second group consists of features that must be extracted from the LLDs obtained from all small time frames in the entire audio clip, such as skewness, flatness, standard deviation and quartile. These descriptors are global descriptors [20]. Since emotion in speech clips is represented as a progression of states and states are extracted from small time frames in the entire speech clip [20], v this thesis proposes to take small, overlapping small time frames from each speech clip and obtain LLDs from each of the small frames. In total, there are 272 LLDs associated with emotion recognition. Table 3.1 describes the LLD features.

Table 3.1: The 272 low-level descriptor features.

Low-Level descriptor features	Amount
Mel-Frequency cepstral coefficients (MFCC) 1-13	104
Delta coefficients for MFCC 1-13	104
Zero-crossing rates	8
Delta coefficients for zero-crossing rates	8
Continued on next page	

Table 3.1 – continued from previous page

Low-Level descriptor features	Amount selected to train the CNN
Root-mean-square signal frame energy	8
Delta coefficients for root-mean-square signal frame energy	8
Spectral centroid features	8
Delta coefficients for the spectral centroid related features	8
Pitch-related features	8
Delta coefficients for the pitch-related features	8
Total amounts	272

3.4 Dataset compensation

Since available datasets are collected in strictly controlled laboratory environments in which noises and reverberation are minimal and the microphone is placed adjacent to the speakers, modules trained on these datasets will not be robust enough to yield similar accuracy when deployed in real-life scenarios due to the presence of noises and reverberation. In order to compensate this problem, synthetic datasets are created by adding reverberation, amplification, and background noises to the original datasets. The original datasets of emotional speech used for training are CREMA-D [17], SAVEE [21], RAVDESS [18], and EMA [22]. In total, there are 1693 samples of happy speech, 1693 samples of angry speech, 1693 samples of sad speech, and 1473 samples of neutral speech from the original datasets. For each sample from the four classes from the original datasets, a copy of the sample is made, and reverberation effect is added to the copy. Both the clean samples and the copies of the clean samples that are contaminated with reverberation are kept. Then, the clean samples and contaminated samples are randomly amplified or de-amplified, and subsequently mixed with 136 samples of random background noise. As a result, the synthetic dataset

consists of 6772 samples for happy speech, 6772 samples for angry speech, 6772 samples for sad speech, and 5892 samples for neutral speech.

- Reason to add background noise to create the synthetic training set: Although an audio signal pre-processing unit (Section 3.5) is added in the system to filter out noise, we cannot assume that the noise will be completely eliminated. Portions of noisy frames will remain in a sound window even after it is pre-processed by this unit. Therefore, in the synthetic training set, background noise is fabricated into the original audio samples to better imitate actual signals that the acoustic system will receive.
- Reason to add reverberation to create the synthetic training set: The pre-processing unit does not reduce the reverberation effect from its input.

3.5 Audio signal pre-processing

When the system starts running, the microphone will always be on and continuously capturing acoustic signals. However, we only want single-person speeches to be passed to the mood classifier. Therefore, we need to (1) filter out silence, (2) get rid of noises that are both within and outside of the human vocal range, and (3) disregard the acoustic segments in which more than one speaker are speaking at the same time.

- **Silence filter** [3]. If the energy signal of an acoustic segment is below a predetermined threshold, this acoustic segment is treated as a silent segment and discarded. In other words, it will not be passed to the classifier.
- **Noise filter** [3]. The first step is to filter out noises outside of the human vocal range. A Butterworth band-pass filter will eliminate such noises from an acoustic segment. The third order of band-pass range (100Hz to 3500 Hz) is used. The second step is to use the standard spectral subtraction to detect noises that are inside the human vocal

range. The spectral profile of the background noise is analyzed and a fingerprint for the profile is generated. The input audio is sliced into several sub-segments. For any sub-segment, if its frequency spectrum is lower than the mean of the fingerprint, this sub-segment is replaced by silence. Non-speech segments are also treated as noise as they are not desirable candidate input for the emotion recognition classifier. Long-term spectral divergence (LTSD) VAD filter [25] is used to filter out such sounds. After calculating the LTSD between actual speech and noises, the filter will decide if this segment is speech or non-speech based on decision rule [3].

- **Overlapping speech filter** [20]. Proposed by Salekin et al., the classifier that detects overlapping speech is a binary neural network that categorizes an input sound signal into either single-speaker speech or multiple-speaker speech. If a sound segment is classified as a multiple-speaker speech, the entire segment is disregarded and will not be passed to the other classifiers in the system.

3.6 Emotion classifier

Figure 3.1 is an overview of the emotion classifier. The emotion classifier is a hierarchical structure consists of three sub-classifiers for mood, the top classifier that distinguishes happy/angry clips from neutral/sad clips, the happy/angry classifier that distinguishes happy clips from angry clips, and the neutral/sad clips that distinguishes neutral clips from sad clips.

Distinguishing happy voice samples from angry ones are hard, because happy and angry sound clips have similar affective activation labels (activation denotes the level of energy [26]). Meanwhile, distinguishing sad voice samples from neutral ones experiences the same difficulty, as samples from the two categories also have similar affective activation labels [26]. Therefore, a hierarchy of three classifiers are used. Each of the classifiers are convolutional neural networks.

The three classifiers (the top classifier, the Happy/Angry classifier, and the Neutral/Sad

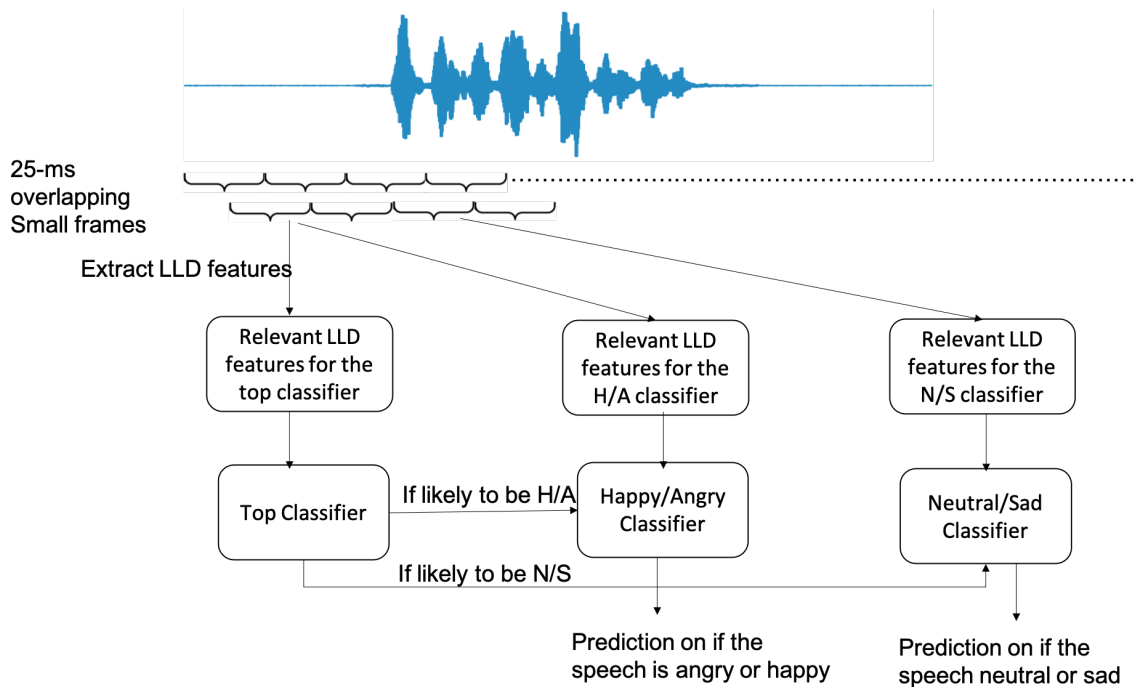


Figure 3.1: Overview of the emotion classifier

classifier) are convolutional neural networks with the same structure, as described in Figure 3.2. Such structure contains 3 CNN layers and each layer has 100 filters. Each layer is processed using max pooling, with window of size 2 and stride of size 2. The dropout rate for each layer is 20%. After the 3 CNN layers with max pooling, 1 dense layer of 200 neurons is added, and the dropout rate is set as 20%. The output layer is a dense layer with 1 neuron.

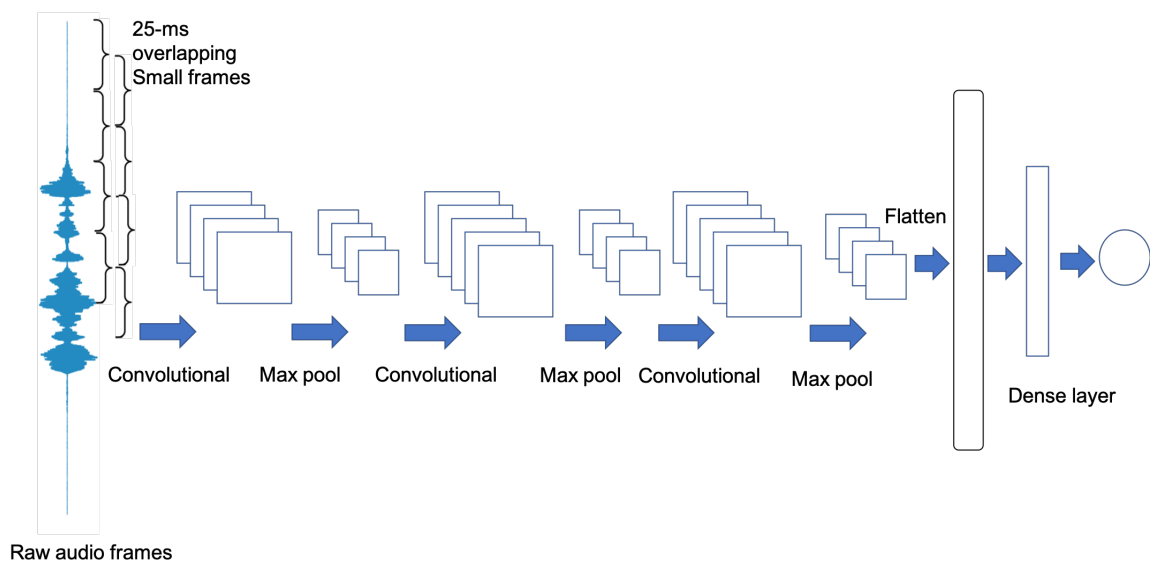


Figure 3.2: Structure of the CNN classifier. The top classifier, the Happy/Angry classifier, and the Neutral/Sad classifier share the same structure. The only difference among them is that they use different LLD features.

CHAPTER 4

MICROPHONE SELECTION

The microphone (MXL AC-404) used in the acoustic pipeline has a sampling rate between 44.1kHz - 48kHz and frequency response between 40Hz - 16kHz [24]. A comparative analysis is performed on MXL and other two microphones, Google Home Mini and Vaddio EasyMic. A person is speaking while standing at distances varying from 3 to 25 feet from the microphones.

The following paragraph describes the design of the testing environment. Vaddio, attached to the ceiling, is 8 feet away from the ground. The perpendicular line from Vaddio to the floor intersects the floor at point O. Google Home Mini, point O, and MXL are placed in a straight line on the floor, with point O in the middle, Google Home Mini on the left, and MXL on the right. Google Home Mini is placed 4 feet away from O on the left, and MXL is placed 4 feet away from O on the right. A another line is drawn on the floor that is perpendicular to the line consisting of Google Home Mini, point O, and MXL. The line is marked at points that are at different distances to the point O, starting at 3 feet away and ending at 25 feet away.

During the testing, both three microphones are turned on and spontaneously capturing acoustic signals. This is to make sure that the three microphones' performance is obtained when they process the same acoustic signal. The speaker stands and the marked points and speak sentences to the microphones. At each marked point, the speaker will speak spontaneously a sentence. The longest sentence consists of 23 words, while the shortest consists of only 7 words. The performance of the microphones are measured by transcription accuracy.

One complication needs to be addressed, as the transcribed sentences from Google Home Mini can not be directly obtained. The device needs to be invoked with the phrase

”Okay Google”, followed by the command ”repeat after me.” Then, the device will ask via its electronic speaker ”what would you like me to repeat.” Then the speaker will proceed to say the sentence. After the speaker finishes speaking the sentence, Google Home Mini will announce via its electronic speaker the transcribed sentence. In the meantime, the other two microphones are keep listening to the environment, and they will capture the aforementioned audible interaction between Google Home Mini and the speaker. For example, if the speaker intends that the three microphones should transcribe the sentence ”congratulations on your sister’s recent graduation from college”, the interaction between Google Home Mini and the speaker will be as follows. Note that, in the described scenario, Google Home Mini transcribes the sentence with 100% accuracy.

- Speaker: Okay Google, repeat after me.
- Google Home Mini: What would you like me to repeat?
- Speaker: Congratulations on your sister’s recent graduation from college.
- Google Home: Congratulations on your sister’s recent graduation from college.

Since the interaction conversation between Google Home Mini and the speaker is audible, the acoustic signal received by the other two microphones will be the entire conversation, instead of the sentence that is intended to be transcribed. The accuracy of the other two microphone’s transcription is only based on how the sentence intended to be transcribed is transcribed, while the transcription in the other part of the conversation is ignored.

A microphone that are suitable for real-life deployment must satisfy all three requirements below.

- Distance-resistant. All three microphones achieve excellent transcription accuracy when placed within 4.57 meters or 15 feet from the speaker. MXL and Google Home Mini achieve accuracy about 90 % when placed within 7.62 meters or 25 feet from the speaker. MXL and Google Home Mini can capture speech in a very large room and the speech can still be transcribed with high accuracy.

Distance	Vaddio	MXL	Google
3 ft	1	1	1
4 ft	0.9285	0.9285	1
5 ft	0.8888	0.8888	0.8888
6 ft	1	1	1
7 ft	1	1	1
8 ft	1	1	1
9 ft	1	1	1
10 ft	1	1	0.9090
11 ft	1	1	1
12 ft	1	1	1
13 ft	1	1	1
14 ft	1	1	1
15 ft	1	1	1
16 ft	1	1	1
17 ft	1	1	1
18 ft	0.9565	0.4782	0.9565
19 ft	1	1	1
20 ft	0	0.7058	0.7058
21 ft	0	1	1
22 ft	1	0.9	0.9
23 ft	0.7	0.7	1
24 ft	1	1	1
25 ft	1	1	1
average	0.8901	0.9391	0.9721

Table 4.1: Comparative analysis on microphones. The columns indicate the transcription accuracy of the acoustic signals collected by the microphones when the speaker stands at different distances from the microphones and speaks a sentence.

- Fit for rapid development. In other words, the microphone must be easily available for purchase at an affordable price to average users. It must be easily maintainable. It must not require additional efforts to pass the audio clips it captures to the speaker identification module and mood classifier. MXL and Google Home Mini are obtainable at affordable prices, available to online purchases. However, it is impossible to directly obtain the captured sound signals from Google Home.
- Comfortable to users. In other words, the microphone should not be too visually intrusive or aesthetically unpleasant. Vaddio is a pair of ceiling microphones that must be attached to the ceiling, and the pair of microphones must be connected to a mixer/amplifier device via wires. Some users may not be pleased with either the wires.

All three microphones are distance-resistant enough to be deployed in an average sized room, with MXL and Google Home Mini having a better performance in a large room. Since Vaddio is too visually intrusive and Google Home Mini requires extra effort to obtain the captured audio clips, MXL is used as the microphone in the acoustical pipeline.

CHAPTER 5

EVALUATION ON THE EMOTION CLASSIFIER

5.1 The downfall of playing out audio clips to the microphone via an electronic speaker

Since the acoustic pipeline is designed to be deployed in users' houses and the users are not expected to be always near the microphone, an intuitive approach to test the emotion classifier is to have actual human volunteers speak when the system is running. However, this approach requires volunteers who can speak in a way that carries out the emotions faithfully. It is hard for people who have not been trained in acting to act out different emotions accurately.

A second approach is to play out the audio clips with an electronic speaker to the microphone to test the mood classifier. However, the result produced by this approach may be influenced by the electronic speaker that are used to play out the samples of emotional speech. In order to investigate if the electronic speaker will indeed distort the audio signal and impact, an experiment is performed. A lay person who has not received any acting training is speaking spontaneously to the microphone that is placed 0.5 meter away; 211 samples of speech are produced, each of which is 5-second long. Then, these samples are stored in a laptop placed 0.5 meter away from the microphone and the samples are played out via the electronic speaker of the laptop. Both the original samples and the samples obtained from the electronic speaker are passed to five early versions of the top classifier (the classifier that distinguishes happy and angry samples from neutral and sad samples).

Original samples are the audio samples directly obtained by the microphone from the speaker. Played-out samples are audio samples obtained when the microphone captures the acoustic output from the speaker of the laptop when the original samples are being played

out. If the electronic speaker does not significantly distort the acoustic signal, the accuracy of the classifiers should not be significantly different. Since the human speaker who provides the speech samples is untrained and not emotionally simulated when the samples are collected, all the samples are labeled, by the speaker themselves, as neutral speech. Since a person knows what emotion he or she feels at the current moment, the labeling is accurate. When passing each of the samples to the top classifier, the top classifier will either classify it as happy or angry samples, or neutral or sad samples. In this case, since all samples are labeled as neutral samples, the accuracy of the top classifier is calculated as the number of samples predicted as neutral/sad divided by the number of all the samples.

In addition to further test that playing out audio clips from an electronic speaker and using a microphone to capture the output acoustic signal of the electronic speaker will distort the original audio clips, another evaluation is performed. 78 speech samples consisting of evenly distributed happy, angry, neutral, and sad speech samples from TESS are played out by a the same laptop using the same speaker. The laptop is placed 0.5 meter away from the microphone that captures the acoustic output. Since most audio samples in TESS are 2 to 3-second long, and the emotion classifier requires its input to be exact 5-second long, each of the captured samples is padded with silence. The 78 played-out samples are passed to the same early version of the top classifier. As a control group, the entire TESS dataset that consists of 100 happy audio samples, 100 angry audio samples, 100 neutral audio samples, and 100 sad audio samples, are directly passed to the same classifier after they are padded with silence. In this case, the accuracy of the top classifier is calculated as

$$accuracy = \frac{c}{sum} * 100\% \quad (5.1)$$

where sum is the overall amount of samples and c is the sum of the number of happy and angry clips that are predicted as happy or angry clips, and the sum of neutral and sad clips that are predicted as neutral or sad clips.

Table 5.1 is the result of this experiment. The performance of the top classifier varies

Table 5.1: Analysis on the effect of an electronic speaker.

Audio	Random Forest + CNN
Original live speech	0.98
Played-out live speech	0.38
Original TESS	0.83
Played-out TESS	0.42

significantly in both evaluations in this experiment. When an actual person is speaking directly to the microphone, the top classifier classifies 98 % of the speech samples as neutral or sad samples, which correspond to the fact that the person is feeling not happy, angry, or sad when the spontaneous speech was recorded. However, when the same speech clips are played out by an electronic speaker and re-captured by the same microphone, the top classifiers can only identify 38% of them as neutral or sad. The audio clips in TESS are collected when the human speakers (actors) are speaking directly to the microphone when the microphone is placed close by. However, when those audios are played out via an external electronic speaker and re-captured by a microphone, the performance of the classifier also changes dramatically. Therefore, playing out audio signals to the microphone is not a reliable way to conduct evaluation on the classifier.

5.2 Trained on clean datasets

It is unrealistic to have untrained volunteers demonstrate authentic happiness, anger and sadness for the classifier evaluation to be done, and playing out audio samples from datasets of emotional speech to the microphone requires an electronic speaker which distorts the original acoustic signal, but the set of features that results in the highest cross-validation accuracy are accurately indicative of the target emotion when the classifier is trained on the clean dataset, because the audio samples are collected in idealistic environments and subject to minimal distortion. Training and cross validation are performed on the combined dataset of CREMA-D, EMA, SAVEE, and RAVDESS. In other words, training and performing cross-validation on the clean dataset for the three emotion classifiers are to obtain

the feature set for each of them.

The audio clips in the datasets are only padded with silence if they are less than 5-second long. There is no other alternation to the datasets than silence padding. There are 272 LLD features associated with emotions, out of which we use random forest for feature selection. The random forest consists of 100 trees, and the sizes of the trees range from 1 to 7 layers. The random forest provides a ranking for all of the 272 features.

After obtaining ranking of the 272 features, different amounts of feature, ranging from 48 to 272, are selected to train the CNN for each of the Happy/Angry, Neutral/Sad, and top classifier in order to select the best amount of features. For example, when the amount of feature is specified as 50, the top 50 features from the feature ranking will be used to train the CNN. The following figures and tables illustrate the top 100 LLD features for the Top classifier, the top 100 LLD features for the Happy/Angry classifier, and the top 90 features for the Neutral/Sad classifier, as the CNNs trained on the top 100 LLD features for the top classifier, the top 100 LLD features for the Happy/Angry classifier, and the top 90 LLD features for the Neutral/Sad classifier yield the best accuracy on their respective validation sets.

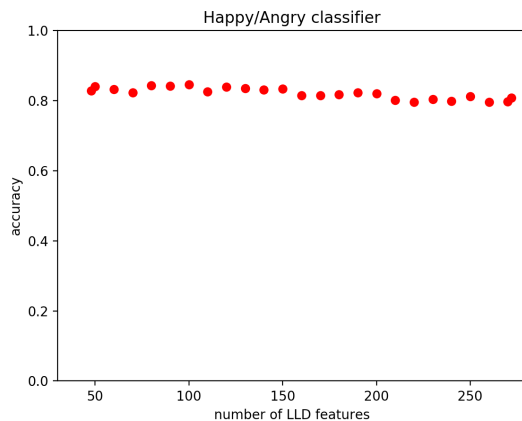


Figure 5.1: Validation accuracy of the Happy/Angry classifier over different LLD features.

Table 5.2: Top 100 relevant LLD features for the Happy/Angry classifier.

Low-Level descriptor features	Amount selected to train the CNN
Mel-Frequency cepstral coefficients (MFCC) 1-13	62
Delta coefficients for MFCC 1-13	6
Zero-crossing rates	6
Delta coefficients for zero-crossing rates	1
Root-mean-square signal frame energy	6
Delta coefficients for root-mean-square signal frame energy	4
Spectral centroid features	6
Delta coefficients for the spectral centroid related features	1
Pitch-related features	6
Delta coefficients for the pitch-related features	2
Total amounts	100

Figure 5.1 describes the cross validation accuracy of the Happy/Angry classifier with different features. The x-axis indicates the top n features based on the feature ranking produced by a random forest feature selection (100 trees, and the sizes of the trees range from 1 to 7 layers), and the y-axis describes the classifier's accuracy on the validation set when trained on the top n features. Figure 3 indicates that there is no significant difference on the validation accuracy of classifiers when trained on different features, as all of them providing satisfactory performance, although the validation accuracy is highest when the top 100 feature are selected. Figure 5.1 provides insights into feature reduction - since there is no significant improvement of the classifier when different top features are chosen, as small as a set of 48 features may prove sufficient if a smaller feature set can reduce the complexity

during training.

Table 5.2 describes what the top 100 features are. The top 100 features are used as the selected features when the same classifier is trained on the synthetic dataset.

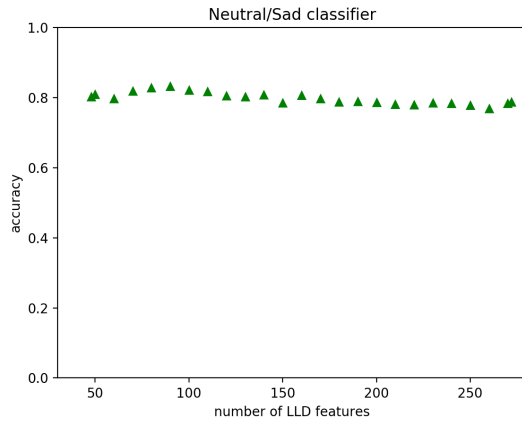


Figure 5.2: Validation accuracy of the Neutral/Sad classifier over different LLD features

Table 5.3: Top 90 relevant LLD features for the Neutral/Sad classifier.

Low-Level descriptor features	Amount selected to train the CNN
Mel-Frequency cepstral coefficients (MFCC) 1-13	60
Delta coefficients for MFCC 1-13	7
Zero-crossing rates	4
Delta coefficients for zero-crossing rates	0
Root-mean-square signal frame energy	6
Delta coefficients for root-mean-square signal frame energy	4
Spectral centroid features	6
Delta coefficients for the spectral centroid related features	1
Pitch-related features	2
Delta coefficients for the pitch-related features	0

Continued on next page

Table 5.3 – continued from previous page

Low-Level descriptor features	Amount selected to train the CNN
Total amounts	90

Figure 5.2 describes the cross validation accuracy of the Neutral/Sad classifier with different features. Again, the x-axis indicates the top n features based on the feature ranking produced by the same random forest feature selection. The y-axis describes the classifier's accuracy on the validation set when trained on the top n features. Figure 5.2 indicates that there is no significant difference on the validation accuracy of classifiers when trained on different features, as all of them providing satisfactory performance, although the validation accuracy is highest when the top 90 feature are selected. Since there is no significant improvement of the classifier when different top features are chosen, as small as a set of 48 features may prove sufficient if a smaller feature set can reduce the complexity during training.

Table 5.3 describes what the top 90 features are. The top 90 features are used as the selected features when the same classifier is trained on the synthetic dataset.

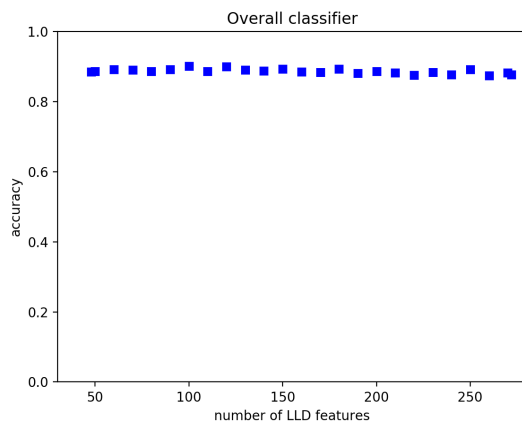


Figure 5.3: Validation accuracy of the top classifier over different LLD features

Table 5.4: Top 100 relevant LLD features for the top classifier.

Low-Level descriptor features	Amount selected to train the CNN
Mel-Frequency cepstral coefficients (MFCC) 1-13	60
Delta coefficients for MFCC 1-13	9
Zero-crossing rates	6
Delta coefficients for zero-crossing rates	1
Root-mean-square signal frame energy	7
Delta coefficients for root-mean-square signal frame energy	6
Spectral centroid features	6
Delta coefficients for the spectral centroid related features	1
Pitch-related features	4
Delta coefficients for the pitch-related features	0
Total amounts	100

Figure 5.3 describes the cross validation accuracy of the top classifier with different features. Again, the x-axis indicates the top n features based on the feature ranking produced by the same random forest feature selection. The y-axis describes the classifier's accuracy on the validation set when trained on the top n features. There is no significant difference on the validation accuracy of classifiers when trained on different features, as all of them providing satisfactory performance, although the validation accuracy is highest when the top 100 feature are selected. Since there is no significant improvement of the classifier when different top features are chosen, as small as a set of 48 features may prove sufficient if a smaller feature set can reduce the complexity during training.

Table 5.4 describes what the top 100 features are. The top 100 features are used as the

selected features when the same classifier is trained on the synthetic dataset.

Over the training dataset, the top classifier yields an accuracy of 0.9013 with 100 LLD features, the happy/angry classifier yields an accuracy of 0.8461 with 100 LLD features, and the neutral/sad classifier yields an accuracy of 0.8322, with 90 LLD features. The reason why the evaluation starts at 48 features is that [20] summarizes there at 48 emotion-related LLD features that distorts less than 50% when the microphone is placed away from the sound source at different distances. In the future work, we plan to take the intersection between the 48 distance-agnostic features and the set of features that yield the highest accuracy for each classifier (the top classifier, the happy/angry classifier, and the neutral/sad classifier).

5.3 Trained on the synthetic dataset

In order to serve as a comparison to the classifier trained on the clean dataset discussed in the previous subsection, a classifier sharing the same structure (3 convolutional neural networks with the same parameters) is trained on the synthetic dataset. As discussed in the Dataset Compensation section, the synthetic dataset is obtained from adding reverberation effect, background noise, and random amplification/de-amplification effect to the original audio clips in the clean datasets.

The same features selected that yields the highest validation accuracy for each emotion classifier is used as the set of features for the same classifier trained on the synthetic dataset. The optimal amount of relevant LLD features for the top classifier, the happy/angry classifier, and the neutral/sad classifier are determined as 100, 100, and 90 respectively. Based on the knowledge of the most relevant LLD features for each classifier, the three classifiers are trained. When trained on the synthetic datasets, the accuracy of the top classifier becomes 0.8759, the accuracy of the happy/angry classifier becomes 0.8357, and the accuracy of the neutral/sad classifier becomes 0.8180.

5.4 Comparison of the performance of the classifiers trained on the clean dataset and synthetic dataset

5.4.1 Validation accuracy

Table 5.5 is a comparison on the accuracy over the validation set of the classifiers trained on the synthetic dataset and the clean dataset. The validation accuracy of the Happy/Angry classifier when trained on synthetic and clean dataset are very similar. The validation accuracy of the Neutral/Sad classifier is slightly higher when it is trained on clean dataset, while the validation accuracy of the Happy/Angry classifier is slightly lower when it is trained on clean dataset.

Table 5.5: Validation accuracy obtained when the classifiers are trained on synthetic and clean datasets.

Classifier	Training set	Validation accuracy
Happy/Angry	Clean	0.8461
Happy/Angry	Synthetic	0.8357
Neutral/Sad	Clean	0.8322
Neutral/Sad	Synthetic	0.8180
Top	Clean	0.9013
Top	Synthetic	0.8759

The validation accuracy obtained from classifiers trained on synthetic datasets are very similar to the validation accuracy obtained from classifiers with the same structure trained on clean datasets. The highest validation accuracy is 90.13% given by the top classifier when it is trained on clean datasets, and the lowest validation accuracy is 81.80%, given by the Neutral/Sad classifier trained on synthetic datasets. The classifier's validation accuracy

falls in the range between 81.80% and 90.13%, indicating that the classifiers are not subject to the risk of over-fitting and consequently not biased on the datasets.

5.5 Accuracy when evaluated on TESS

As discussed before, TESS can be interpreted and used as spontaneous emotional speech that is collected in idealistic environment in which the speaker is speaking closely to the microphone and background noise and reverberation is minimum.

Table 5.6 illustrates the performance of the top classifier, the happy/angry classifier, and the neutral/sad classifier trained on clean dataset. The performance is indicated by the accuracy that these classifiers achieve on the clean TESS dataset. “Clean” means that the audio clips in TESS are padded with silence if they are less than 5-second to make it longer, and there is no further modification. When trained on clean datasets, the top classifier can accurately identifies 87% of happy speech samples, 99% of angry speech sample, and 84% of sad speech samples, yet it can only correctly classify 52% of neutral speech samples into the neutral/sad category. The performance of the Happy/Angry and Neutral/Sad classifiers are lacking, with the former only yielding an accuracy of 60.5%, while the latter’s accuracy is very similar to random guessing. This again confirms that distinguishing emotions that are dramatically different is easy, while distinguishing emotions that are subtly different from each other requires more than CNNs with a standard structure.

Table 5.6: Performance of the three classifiers on the TESS dataset when the classifiers are trained on the clean dataset. Row: Emotional utterances in TESS. Column:Classifier accuracy

	Top classifier	Happy/Angry	Neutral/Sad
Happy	0.87	0.49	N/A
Angry	0.99	0.72	N/A
Neutral	0.52	N/A	0.08
Sad	0.84	N/A	0.93
Overall	0.805	0.605	0.505

Table 5.7 illustrates the performance of the top classifier and the neutral/sad classifier trained on synthetic datasets. The performance is indicated by the accuracy that these classifiers achieve on the clean TESS dataset. The top classifier’s performance on each emotional category in TESS remain almost identical to its performance when it is trained on the clean speech. The accuracy of the happy/angry classifier and the neutral/sad classifier drops by 5.5% and 0.5%.

Table 5.7: Performance of the three classifiers on the TESS dataset when the classifiers are trained on the synthetic dataset. Row: Emotional utterances in TESS. Column:Classifier accuracy

	Top classifier	Happy/Angry	Neutral/Sad
Happy	0.87	0.6	N/A
Angry	0.99	0.5	N/A
Neutral	0.55	N/A	0.01
Sad	0.84	N/A	0.99
Overall	0.81	0.55	0.50

When trained on both synthetic and clean datasets, the top classifier, the Happy/Angry classifier, and the Neutral/Sad classifier all yield satisfactory validation accuracy. When tested on TESS, a dataset previously unseen by the classifiers and resembling a semi-realistic environment in which the impact of noise and reverberation is at its minimum, the top classifier trained on the clean dataset and its counter-part trained on synthetic trained on the synthetic dataset both generalized well. This indicates that, despite TESS and other aforementioned datasets are collected with the intention to minimize background noise and reverberation effect, background noise and reverberation effect is not eliminated. As a result, the top classifier trained on the synthetic dataset performs slightly better than its counterpart trained on the clean set.

5.6 Accuracy when evaluated on live speech

A person is speaking spontaneously to the microphone when standing at different distances (0.5, 1.5, 3, and 6 meters) from the microphone. The captured audio clips are passed to the pre-processing component to filter out silence, noise, and overlapping speech. The person is in a calm and slightly joyful mood when the samples are collected. Intuitively, the clips should be labeled as neutral because the speaker is not overly stimulated, but the neutral speech samples from the datasets tend to be more similar to disinterest to boredom. As a result, the audio clips are labeled as happy samples. Although the speech samples are collected in a lab environment, the environment is realistic because there is no external acoustical equipment to assure the minimum of background noise and reverberation effect, unlike the idealistic environments where such equipment is present. In other words, the environment resembles the home environments in which the acoustic system will be deployed, and this evaluation is indicative of the evaluation of the classifiers on the samples collected in those home environments.

This paragraph describes the details of the collection of the audio samples when a live person is speaking. The microphone used in the acoustic pipeline has a sampling rate between 44.1kHz - 48kHz and frequency response between 40Hz - 16kHz [24]. The evaluation of the microphone indicates that it is distance resistant, capable of picking up sounds from 15 feet away with 100% transcription accuracy paired with a state-of-the-art transcription tool [37]. When the system starts running, the microphone begins capturing continuous speech signals. The continuous speech signal is sliced into 5-second sound windows for further processing. A silence filter is applied to each of the sound windows. If more less than 25% of the sound window is speech, the entire sound window is regarded as silence and discarded. Otherwise, the sound signal is passed to the noise filter, then to the overlapped speech filter to get rid of the overlapped speech. After all the above pre-processing steps, the sound window is passed to the mood classifier.

Given the results from Table 5.6 and Table 5.7, the top classifier generalizes well on a semi-realistic environment. Therefore, comparing its performance on the live person speech and its performance on TESS will provide insights into if how the effects introduced by a realistic environment affect the features. Since the feature sets of all three classifiers are all subsets from the same 272 features, information obtained from how the top classifier’s features are influenced by a realistic environment does translate to the other two classifiers.

Table 5.8: Accuracy obtained when the classifiers are trained on synthetic datasets tested in a realistic environment

Classifier	Distance	Accuracy
Top	0.5 meter	0.7254
Top	1.5 meters	0.84
Top	3 meters	0.74
Top	6 meters	0.5283

Table 5.8 describes the accuracy obtained when the classifiers are trained on synthetic datasets tested in a realistic environment with a live person speaking to the microphone at different distances. When the speaker is 3 meters away from the microphone, the average accuracy of the top classifier is 74%, in contrast to its cross validation accuracy that is 87.59%. We observe a significant drop of 21.17% when the speaker moves from being 3 meters away to 6 meters away to the microphone. This suggests that, although the microphone is distance-resistant when it comes to transcription and the synthetic dataset is created to mitigate the effect introduced by the environment in which the acoustic signals are collected, distance still impacts the performance of the classifier.

In Table 5.8, the classifier’s accuracy from 0.5 meter away is lower than the classifier’s accuracy from 1.5 meters away. This is because of the assumption of the ground truth - all the samples produced by the speaker is labeled as happy. However, when a person is happy, his or her speech samples will consist of a large portion of happy samples and perhaps a smaller portion of neutral samples.

Table 5.9: Accuracy obtained when the classifiers are trained on clean datasets and tested in a realistic environment

Classifier	Distance	Accuracy
Top	0.5 meters	0.6470
Top	1.5 meters	0.56
Top	3 meters	0.3
Top	6 meters	0.3018

Table 5.9 describes the classifier’s performance when the speaker is at different distances from the microphone and when the classifier is trained on the clean dataset. Classifiers trained on clean datasets do not generalize well on audio clips from live speech, despite the fact that noise-filtering techniques are applied to the captured audio clips. Similar to when the classifier’s performance when it is trained on the synthetic dataset, the accuracy decreases over different distances. However, the classifier trained on the synthetic dataset always significantly outperforms the classifier trained on the clean dataset. This is because (1) the pre-processing procedure, despite having noise filters, does not mitigate the effect of reverberation, (2) Despite the best effort of the pre-processing component, the background noise can only be reduced, instead of completely eliminated, because of this, the top classifier trained on the clean dataset is not adaptive to the effect introduced by the environment,

(3) the microphone is distance-resistant.

The goal of testing the classifier in a realistic setting is to see the effect of the synthetic dataset. By comparing Table 5.8 and Table 5.9, we see that the accuracy of the classifier trained on the clean dataset is lower than the accuracy of the classifier trained on the synthetic dataset at each of the distances. Therefore, the synthetic dataset helps the classifier adapt to a realistic environment.

The top classifier performs well when a live person is speaking and the speech passes through the pre-processing components, suggesting that the pre-processing components function well in filtering out silence and reducing noise and results in audio samples that are similar to those collected in idealistic or strictly controlled lab environments.

CHAPTER 6

CONCLUSION

This thesis proposes an acoustic system consisting of audio pre-processing algorithms and a speech-based emotion recognition classifier. This thesis identifies several limitations in published speech-based emotion detection approaches: they are trained on datasets of emotional speech collected in strictly controlled laboratory environment in which noise and reverberation are minimal. As a result, the evidence to prove that these approaches will still yield similar accuracy when deployed in real-life environments is lacking. The emotion recognition classifier in this thesis is trained on the synthetic dataset obtained from adding background noise, amplification and de-amplification, and reverberation effects to imitate the actual environments in which speech is taking place, instead of the idealistic and strictly controlled environments in labs. The emotion recognition classifier is a hierarchical classifier consisting of three components: the top classifier that separates happy/angry speech samples from neutral/sad speech samples. The happy/angry classifier will determine if the speech sample is happy or angry if the top classifier decides that the speech sample is either happy or angry. Likewise, the neutral/sad classifier will determine if the speech sample is neutral or sad if the top classifier decides that the speech sample is either neutral or sad. Top classifier performs well on the clean TESS dataset which can be seen as a semi-realistic scenario in which a live person is speaking. It's semi-realistic due to two factors: first, the speaker is adjacent to the microphone; second, there is minimal background noise and reverberation effect. Its performance is also satisfactory when the noise and reverberation effect are nullified from the speech sample, based on its evaluation when a live person is speaking to the microphone. In summary, the speech-based emotion recognition system can identify if the speaker is happy/angry or neutral/sad despite the speaker's distance from the microphone.

CHAPTER 7

FUTURE WORKS

It is shown that the top classifier's performance degrades gradually when the speaker is moving away from the microphone. This is because not all of the LLD features it uses are distance-agnostic. In fact, [20] identifies that many LLD features distort over distances. A future step to improve the accuracy is to take the intersection of the top LLD features and the set of distance-agnostic features identified in [20] and retrain classifiers with them.

REFERENCES

- [1] D. Spruijt-Metz, K. de la Haye, J. Lach, and J. A. Stankovic, “M2fed: Monitoring and modeling family eating dynamics: Poster abstract,” in *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM*, ser. SenSys '16, Stanford, CA, USA: ACM, 2016, pp. 352–353, ISBN: 978-1-4503-4263-6.
- [2] N. Campbell, “Databases of emotional speech,” *SpeechEmotion-2000*, 2000, pp. 34–38.
- [3] A. S. Z. Chen M. Ahmed and J. A. Stankovic, “Arasid: Artificial reverberation-adjusted indoor speaker identification dealing with variable distances,” 2019.
- [4] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, “Design, recording and verification of a danish emotional speech database,” in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in *Ninth European Conference on Speech Communication and Technology*, 2005.
- [6] T. Danisman and A. Alpkocak, “Emotion classification of audio signals using ensemble of support vector machines,” in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Springer, 2008, pp. 205–216.
- [7] P. Zervas, I. Mporas, N. Fakotakis, and G. Kokkinakis, “Employing fujisakis intonation model parameters for emotion recognition,” in *Hellenic Conference on Artificial Intelligence*, Springer, 2006, pp. 443–453.
- [8] M. Shami and W. Verhelst, “An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech,” *Speech Communication*, vol. 49, no. 3, pp. 201–212, 2007.
- [9] M. H. Sedaaghi, C. Kotropoulos, and D. Ververidis, “Using adaptive genetic algorithms to improve speech emotion recognition,” in *2007 IEEE 9th Workshop on Multimedia Signal Processing*, IEEE, 2007, pp. 461–464.
- [10] L. X. Hung, G. Quot, and E. Castelli, *Speaker-dependent emotion recognition for audio document indexing*.

- [11] D. Datcu and L. J. M. Rothkrantz, “Facial expression recognition with relevance vector machines,” in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 193–196.
- [12] Z. Hammal, B. Bozkurt, L. Couvreur, D. Unay, A. Caplier, and T. Dutoit, “Passive versus active: Vocal classification system,” in *2005 13th European Signal Processing Conference*, 2005, pp. 1–4.
- [13] H. Altun and G. Polat, “New frameworks to boost feature selection algorithms in emotion detection for improved human-computer interaction,” in *Advances in Brain, Vision, and Artificial Intelligence*, F. Mele, G. Ramella, S. Santillo, and F. Ventriglia, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 533–541.
- [14] M. Lugger and B. Yang, “Classification of different speaking groups by means of voice quality parameters,” *Proceedings of ITG-Sprach-Kommunikation*, 2006.
- [15] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, “Two-stage classification of emotional speech,” in *International Conference on Digital Telecommunications (ICDT’06)*, IEEE, 2006.
- [16] J. Kim and R. A. Saurous, “Emotion recognition from human speech using temporal information and deep learning,” Sep. 2018, pp. 937–940.
- [17] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [18] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PLoS one*, vol. 13, no. 5, e0196391, 2018.
- [19] K. Dupuis and M. K. Pichora-Fuller, *Toronto emotional speech set (tess)*. University of Toronto, Psychology Department, 2010.
- [20] A. Salekin, Z. Chen, M. Y. Ahmed, J. Lach, D. Metz, K. De La Haye, B. Bell, and J. A. Stankovic, “Distant emotion recognition,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, 96:1–96:25, Sep. 2017.
- [21] S. Haq, P. J. Jackson, and J. Edge, “Audio-visual feature selection and reduction for emotion classification,” in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP’08), Tangalooma, Australia*, 2008.
- [22] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, “An articulatory study of emotional speech production,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

- [23] *Electromagnetic articulography (ema) database*, https://sail.usc.edu/ema_web/ema_info.htm.
- [24] *Mxl ac-404 usb-powered microphone*, <http://www.mxlmicro.com/microphones/web-conferencing/AC-404/>, Accessed: 2019-04-14.
- [25] J. Ramirez, J. C. Segura, C. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.
- [26] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.