Speech Emotion Recognition using Deep Neural Networks

by

Raiymbek Mustazhapov

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

July 2021

Author .....
Department of Computer Science
July 31, 2021

Certified by .....
Adnan Yazici, Muhammed Fatih Demirci
Professors
Thesis Supervisor

Accepted by .....
Vassilios D. Tourassis
Dean, School of Science and Technology

Speech Emotion Recognition using Deep Neural Networks

by

Raiymbek Mustazhapov

Submitted to the Department of Computer Science
on July 31, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

There is an apparent evolving interest in speech emotion recognition (SER), one of the particular cases of a broader problem of multimedia pattern recognition. SER is considered to possess the capability to enhance the communication efficiency between human and artificial intelligence providing an emotional context to the machine. The field has been developing fast with the emergence and increase in accessibility of deep learning techniques recently. This potential critical benefit and novel techniques have drawn the attention of many specialists in the field and generated a great number of research papers that furnish diverse intricate methods. One of such methods involving various data augmentation techniques has demonstrated high performance in this field. This paper performs an analysis of various simple augmentation methods to attempt to improve existing models. Particularly, this research focuses on state-of-the-art CNN models for RAVDESS, EMO-DB, and IEMOCAP datasets, and exploits temporal, spatial, and spectral transformations of sound as an underlying method for augmentation. As a result of exploiting simple augmentations, we achieved an increase in performance for IEMOCAP model and positive effects comparable to their complex counterparts for other datasets.

Thesis Supervisor: Adnan Yazici, Muhammed Fatih Demirci
Title: Professors

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisors Prof. Adnan Yazici and Prof. Fatih Demirci for the continuous support of my Master's study and research. I offer my sincere appreciation for the patience, knowledge that they shared, and motivation that helped me to progress in my studies. My completion of this project would not be possible without their guidance that helped me in all the time of research and writing of this work.

Besides my advisors, I would like to thank the committee members for their keen interest in my thesis project and insightful feedback. Especially, I thank Dias Issa, an expert in the field, who was kind enough to accept the membership in a committee.

My sincere thanks also go to my academic advisor Prof. Jean Marie Guillaume Gerard de Nivelle who greatly helped me with academic affairs at my weakest state during the historical pandemic times.

Finally, I offer my deepest gratitude to my caring and supportive parents. I am extremely thankful for their love, prayers, and sacrifices for my future. Also, I express my thanks to my friends that supported and believed in me and my success.

Contents

List of Figures

Chapter 1

Introduction

One of the most rapidly developing subfields of multimedia pattern recognition and human-computer interaction is speech emotion recognition (SER). SER is an interdisciplinary field that adopts practices from other realms of computer science such as machine learning, multimedia processing, and acoustic spectral analysis. The ultimate objective of SER is to derive an efficient and accurate model that can process the recorded voice of a person, extract and classify a person's "mood" into one of the emotion categories. In essence, speech emotion recognition can be considered as a computational problem that involves two major parts: feature selection/extraction and basic machine learning classification prediction. Therefore, fundamental questions that are concerned here are: Which acoustic feature to extract? Which acoustic feature possesses the most descriptive power in terms of emotions? Which combination of features performs best? Which classification method is most robust? Each question has its specific challenges.

Nevertheless, researchers in the field have generated numerous various models which are assembled out of different combinations of acoustic features (MFCC, LPCC, MSFs, etc.) and classifiers (NN, RNN, CNN, GMM, HMM, SVM, DBN, etc.) available in the field[1, 27, 16, 21], especially including hybrid models which involve several features and/or several classifiers combined[58, 62, 19]. The availability of such an abundant amount of choices is probably one of the leading factors that caused the SER field to develop fast recently. Furthermore, there is a clear correlation between the

field of deep learning and SER as an exploration of a new state-of-the-art classification approach in machine learning can directly lead to the creation of novel emotion recognition models that outperform previous ones. It will take a great amount of effort and time, however, until all possible combinations to be experimented with and reported not mentioning the challenges that reside out of the two major parts (extraction and classification).

There are various challenges in the development of a competent SER model. The existence of various languages, accents within those languages, speaking styles, sentence structures, speakers themselves adds a layer of difficulty to the problem of speech emotion recognition. One of the reasons is that "to be classified" speech utterances precisely depend on these characteristics. Therefore, usually, a model that demonstrates high performance on emotion recognition in one language may not achieve the same degree of accuracy in another language. The other challenge is the limited choice of data sets (i.e. labeled audio recordings of emotion). Those data sets are used to train classifiers and allocate a significant portion of the development process of emotion recognition models. One of the approaches to oppose the latter challenge is data augmentation.

Data augmentation, a label-invariant transformation of inputs, plays a significant role in the development of any field of machine learning by providing enrichment and balance to existing data. It has been a hot topic in recent years in speech emotion recognition, particularly for a reason that the data scarcity and imbalance of classes inside datasets are common in this field. Most remarkably, it has proven its decent capabilities in image recognition tasks improving the performance of models trained on both small and large image databases. In speech emotion recognition, however, researchers have to work with audio which in essence and digital representation has a different structure. Various researches apply different data augmentation techniques ranging from simple transformations to sophisticated models and algorithms. Pitch shifting, speed alteration, amplitude adjustments remain the most common simple augmentation methods currently in literature [39]. While the application of simple augmentations yields an increase in performance, some studies deploy more complex

augmentation steps on top of it. Despite the fact that these complex methods facilitate increment in performance, a thorough study of augmentation mechanisms needs to be performed. Thus, the central objective of this research is to conduct a comprehensive study of augmentation methods with respect to each class member in a given database and perform comparative analysis across existing speech datasets using quantitative metrics. We chose a state-of-the-art model as a baseline model for three different databases and applied simple data augmentations based on speed and pitch alterations. As a result of a total of twelve experiments, we were able to improve the existing model for one dataset in terms of accuracy, and obtain an increase in performance, as a result of augmentation, comparable if not higher than complex augmentation techniques, yet with a simpler approach for other datasets. Finally, we determined the effects (beneficial and detrimental) of each method for each class which can be utilized for selective augmentation in further studies in the area. In terms of structure, the next chapter gives an overview of SER, problems, and motivation behind the efforts of solving these problems. Chapter 3 describes the research methods used in this work, followed by Chapter 4 where corresponding results are presented. A detailed discussion of results is introduced in Chapter 5. Chapter 6 includes overall conclusions and Chapter 7 discusses future research directions.

Chapter 2

Literature Review

Speech is the fastest and most efficient paradigm of communication developed by humanity through millions of years of evolution. It has been developed to the extent that not only information and knowledge but also the emotional context of the speaker can be transmitted and perceived. This capacity power of speech has motivated specialists to exploit speech as a natural and efficient way of human-computer interaction. Although researchers achieved outstanding performance in speech recognition, there is a great difference between human-computer and natural human-human speech interactions. One of the primary reasons is that a computer is not capable of perceiving the human emotional context infused in speech. Speech emotion recognition (SER) aims to solve this by recognizing emotion automatically analyzing the voice of a user.

Fundamentally, human emotion in speech exists in different dimensions simultaneously. Its propagation in those dimensions can be represented with corresponding features. These are prosodic (sound), visual, and semantic features. Prosodic features are features that involve vocal or audio features of speech. These include pitch, loudness, tempo, and rhythm, and other low-level features of sound. Visual representation of emotion corresponds to facial expression, gestures, and movements during the exhibition of emotion. While both previous sets of features are the representation of some physical phenomenon, semantic features, on the other hand, do not refer to what exists physically. The underlying meaning of the words, sentence structure, and different concepts that exist within the semantic dimension of spoken words and

clauses (irony, sarcasm, etc) are the constructs of semantic features. Human is able to perceive and analyze all three channels of features, but most of the SER models are limited to a particular set of features. While there exist models that attempt a modular approach to adopt all three sets of features, this paper focuses on prosodic (audio) features of the speech.

Speech emotion recognition is capable of enhancing any process that involves a human, computer, and their essential interaction by providing an additional dimension of information. Such processes occur in computer tutorials where user's emotion can play a role of a feedback-response mechanism[41], embedded systems in cars where awareness of the emotional state of a driver can improve safety measures and decisions chosen by a system[41], automatic diagnostic systems as symptoms of various diseases affect the mental state of a patient[15] and any other fields where the realization of person's state can be beneficial.

2.1 Feature selection/extraction

As explained previously in the paper the problem of speech emotion recognition consist of two major sub-problems: feature extraction and classification of emotion based on those features. Features in the speech recordings, as the input is an audio file, refer to different acoustic characteristics of a sound. There are two methods of extracting these characteristics in the literature. First, extracting features using a machine learning approach. In this method, the neural network learns features and then passes them to the classification stage. Often this method is addressed as "end-to-end" or "black box" as the outside observer does not have knowledge on which features are extracted[14, 51, 52, 59]. The second is extracting known features. In this method handcrafted, previously studied features are extracted and fed to the classifier. Various studies propose distinct features to extract recent years: Mel-frequency cepstral coefficients (MFCCs)[60, 40], linear predictive cepstral coefficients (LPCCs)[2], Teager energy operators(TEO) based features[3], spectral features, prosodic features (i.e. pitch, rhythm, etc.)[28, 33] and handcrafted group of features (e.g. GEMAP)[46, 10].

There is still a debate, however, on which particular feature contains the most distinctive information about emotion. Some researchers fuse features with another one before feeding to a classifier to increase accuracy[28, 55], while others fix on a single feature and focus on improving the classifier[3]. Both methods, however, have common challenges.

The feature selection and extraction is not a trivial process and has its challenges associated with the selection of efficient approaches. Ayadi et. al. have classified these challenges into four distinct issues[11]. First is the choice between local features and global features. Former refers to the approach where speech recording is divided into time intervals (i.e. frames) and features of each frame extracted separately, called local features, whereas the latter attributes to the method where global characteristics are extracted from the entire speech recording. The next issue is choosing between different types of features. Each class of features (e.g. pitch, energy, frequency, zero crossings, etc.) behave differently in terms of emotion classification, since each characteristic describes different aspects of sound, in some cases mutually exclusive. Third, what are the effects of post-processing, noise removal, and other audio enhancing methods on the overall accuracy of the classification model? Finally, the last question is whether only acoustic features are enough for emotion classification or should speech information, linguistic features, and other speech-specific characteristics be combined with acoustic to achieve improved performance of a classifier.

2.2 Data Augmentation

One of the challenging tasks of machine learning is to increase the generalization ability of the neural model. Generalization refers to a degree of performance of the model on previously unseen data with respect to data that the model is trained on. The inability to generalize is a result of overfitting of the model to the training data and it usually happens with small datasets. In other words, the model memorizes the limited training data instead of learning the patterns that are common in the whole population. As a result, the model shows high accuracy on the training set but

fails to perform well on test data or unseen input in general. There are techniques to counter such undesirable behavior involving regularization techniques (dropout layers, batch normalization) [45, 49], pretraining [12], transfer learning [53, 43], and other architecture modifications. Data augmentation, however, is a method that tackles the problem in the root - the data.

Data augmentation is rather a set of techniques that increases the number of samples in the dataset exploiting the existing ones. Those techniques are usually transformations that dependant on the nature of the samples themselves. As a result, a dataset is populated applying those transformations and creating new data samples, as a result inflating the set with new synthetic samples. Evidently, the only transformations that are eligible for data augmentation are the ones that preserve the label of the sample. In other words, the sample must correspond to the previous label after transformation. For example, in the case of audio, the audio recording of a person laughing labeled as 'happy' should also be able to recognize as 'happy' after the transformation. Applying these techniques is capable of greatly reinforcing the generalization of a model.

Image recognition is a perfect example of the remarkable effectiveness of data augmentation techniques. This success comes from the fact, that image can be transformed in a plethora of ways and shapes. It showed efficiency and effectiveness in many image processing tasks such as classification of diseases from medical images (MRI, radiographs, etc) [34, 30] where the limited occurrence of disease are common, face recognition [32], handwriting classification [54], and even marine organisms classification [20]. However, the data in speech emotion recognition is audio which is fundamentally different from image. Most common augmentation techniques include simple transformations such as pitch shifting, speed alteration, amplitude adjustments. Apart from simple augmentations, complex techniques are developed and utilized in the literature. These include window slicing techniques [37], vocal tract length perturbation (VTLP), a normalization technique to reduce variance between speakers [13, 23], segment repetition based on the high amplitude (SRHA), augmentation of existing audio with the segment with the highest amplitude [39], and generative

adversarial networks (GAN) based models to generate samples [4, 8, 57, 56].

2.3 Classification

The second sub-task of speech emotion recognition is classification. It is the core stage where emotions are recognized and labeled to one of the defined categories (i.e. anger, happiness, sadness, etc.) depending on the model and database. There are various methods of solving this problem and the majority are handled by machine learning neural networks. Some researchers propose deep learning-based methods that exploit Convolutions Neural Networks (CNN) that proved its efficiency with image processing[60, 27, 1], others propose Recurrent Neural Networks (RNN) that is often used to work with time series and sequences since audio recording is, in essence, time-frequency (or time-energy) series[27, 35]. The third most offered approach is adopting Deep Belief Networks (DBN) as a classifier[33, 55]. Finally, there are classical approaches that use Support Vector Machines (SVM), k-st Nearest Neighbour (k-NN)[2, 47], Hidden Markov Model (HMM) or Gaussian Mixture Model (GMM)[3]. As with the feature extraction methods, one can combine approaches and design fused models that exploit several machine learning techniques. Some researchers combine CNN with Long-Short Term Memory (LSTM) networks [60], others RNN with LSTM [48, 31], CNN with SVM[25, 19], k-NN with SVM[2, 36] and other possible combinations to acquire increased accuracy. With multiple models, however, comes the multiple challenges associated with these techniques. Nevertheless, machine learning techniques currently proved to be a sufficient classification tool in the problem of speech emotion recognition.

2.4 Databases

The classification task requires large datasets for the training process as the primary tool for the classification step of speech recognition is machine learning models. These databases usually contain audio-video recordings labeled with particular emotion cat-

egories (anger, sadness, neutral, etc.). These are mostly used in the studies: IEMOCAP - multimodal database of English dyadic conversations containing both fixed speech and free, spontaneous speech[6]; Berlin Emotional Database (also called EmoDB on some studies) - different scripted emotional utterances of 10 actors[5]; Surrey Audio-Visual Expressed Emotion Database(SAVEE) - 4 male actors in 7 different emotions; 480 British English utterances in total[44]; CHEAVD - a Chinese natural emotional audio-visual database[26]; CASIA - utterance of 4 professional speakers, 12,000 sentences in total[http://shachi.org/resources/27], RAVDESS - speech and song recordings of 24 actors (12 male, 12 female), 1440 samples total [29] and others.

2.5 Other challenges

Some issues exist outside of feature extraction and classification areas. One such issue is the existence of different languages. A model that was designed for one particular language will not probably show the identical performance on other languages. The existence of languages such as Chinese where the tone and pitch are more defined and frequent creates an additional layer of complexity in designing universal multilingual SER models. The other challenge is the post-processing of recording. Post-processing is a technique of enhancing and altering the audio recording. Some researchers provide post-processing in their implementations, while others believe such techniques probably change patterns that contain information about the emotion of a speaker.

Chapter 3

Research Methods

Our research methodology is based on training and testing neural network models on RAVDESS, EMO-DB, and IEMOCAP datasets with incremental data augmentation. For this paper, we focused on diverse data transformations and retained architecture consistent across experiments. The overall setup was consistent throughout the experiments. Further, we will explain all the processes and techniques applied in all the experiments.

3.1 Architecture

We used state of the art model presented by Issa et. al. [22] as baseline architecture. It is a convolutional neural network that consists of six convolutional layers. First, the first convolution layer receives the input from the input layer, and feeds the output to the batch normalization layer, to the output of which rectified linear unit activation function is applied. Next, the output is fed to the second convolution layer, then again ReLU activation, then dropout is applied. Next comes batch normalization and after it ReLU activation. Further, the third and fourth convolution layers are feed sequentially with ReLU activation after each layer. Then the output is fed to the fifth convolution layer, batch normalization, ReLU, and dropout again sequentially. The sixth and last convolutional layer receives the output and ReLU is applied to the output of the convolutional layer. The fully connected layer receives

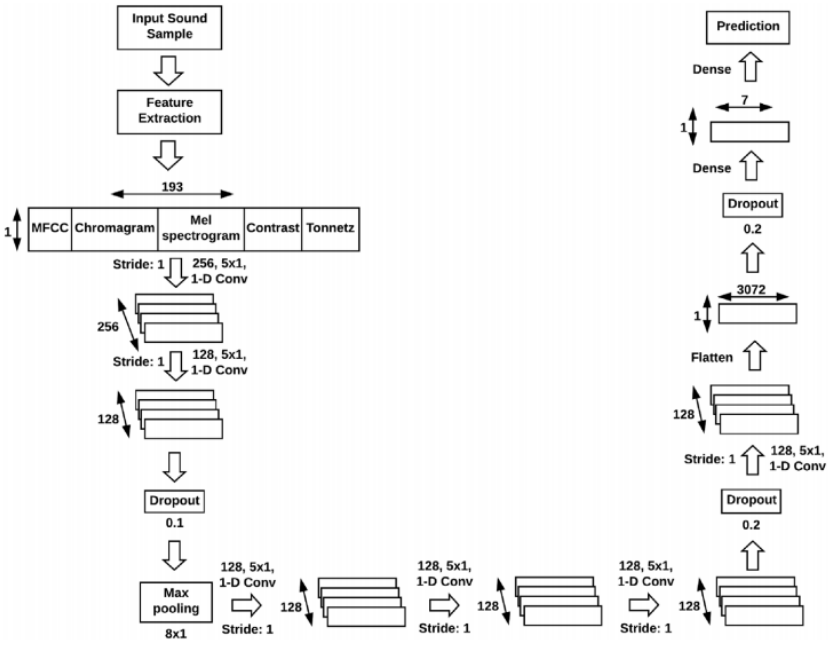

Figure 3-1: Model topology described in Issa et. al. [22]

```
BaseNet(
  (conv1): Conv1d(1, 256, kernel_size=(5,), stride=(1,), padding=(2,))
  (conv2): Conv1d(256, 128, kernel_size=(5,), stride=(1,), padding=(2,))
  (conv3): Conv1d(128, 128, kernel_size=(5,), stride=(1,), padding=(2,))
  (conv4): Conv1d(128, 128, kernel_size=(5,), stride=(1,), padding=(2,))
  (conv5): Conv1d(128, 128, kernel_size=(5,), stride=(1,), padding=(2,))
  (conv6): Conv1d(128, 128, kernel_size=(5,), stride=(1,), padding=(2,))
  (fc1): Linear(in_features=3072, out_features=8, bias=True)
  (batch_norm1): BatchNorm1d(256, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (batch_norm2): BatchNorm1d(128, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (batch_norm3): BatchNorm1d(8, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (pool): MaxPool1d(kernel_size=8, stride=8, padding=0, dilation=1, ceil_mode=False)
  (dropout1): Dropout(p=0.1, inplace=False)
  (dropout2): Dropout(p=0.2, inplace=False)
  (softmax): Softmax(dim=1)
)
```

Figure 3-2: Configuration of model implemented in PyTorch

the flattened output of the last convolution layer as input and feeds its output to the batch normalization layer. Finally, a softmax layer is applied at the end. The model is described robustly in research work by Issa et. al. [22]. The Figure 3-1 shows the model topology and Figure 3-2 shows the same model implemented in PyTorch for this work.

3.2 Dataset and preprocessing

The model is trained and tested on three different emotional audio databases: RAVDESS, EMO-DB, and IEMOCAP. Particularly, only the audio recordings from these databases were used for the experiments. Similar preprocessing steps were applied for all three datasets including dividing them into folds.

3.2.1 RAVDESS

The model is trained and tested on an only audio subset of The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset [29]. The subset that we use contains labeled audio recordings of 24 actors (12 male, 12 female) who express 8 different emotions, ['neutral', 'calm', 'happy', 'sad', 'angry', 'fearful', 'disgust', 'surprised']. Overall, it contains 1440 speech recordings in English. The classes are well balanced except the 'neural' class that is represented two times less than others. We used the k-fold validation method with k=5 and divided the dataset into five sets. We first shuffled the whole dataset with a constant randomizer seed to make results consistent. Then equally distributed the classes to retain the distribution of the whole dataset. The Figure 3-3(a) shows the distribution across folds.

3.2.2 EMO-DB

Berlin Emotional Database (EMO-DB) contains audio recordings of scripted speeches of 10 (5 males, 5 females) professional speakers [5]. Overall, it contains 535 utterances labeled in 7 emotion classes, ['neutral', 'boredom', 'happiness', 'sad', 'anger', 'fear', 'disgust']. The database is slightly imbalanced with proportions of classes falling in the range between 9.0% and 23.4% (Figure 3-3(b)). We divided the dataset into 5 folds in the same fashion as with the RAVDESS dataset. The language of the speeches in the dataset is German.

a)

Class	neutral	calm	happy	sad	angry	fearful	disgust	surprised
Instances	20	39	39	39	39	39	39	39
Percentage	6.8%	13.3%	13.3%	13.3%	13.3%	13.3%	13.3%	13.3%

b)

Class	neutral	boredom	happiness	sad	anger	fear	disgust
Instances	16	17	15	13	26	14	10
Percentage	14.4%	15.3%	13.5%	11.7%	23.4%	12.6%	9.0%

c)

Class	neu	sad	ang	hap
Instances	223	104	62	33
Percentage	52.8%	24.6%	14.7%	7.8%

Figure 3-3: Distribution of folds: a) RAVDESS, b) EMO-DB, c) IEMOCAP

3.2.3 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) is a dataset of audio, video, and motion capture recordings of acted English speech utterances [6]. Only the audio section was used for the work that contains scripted and improvised speech recordings of 10 (5 males, 5 females) actors. We used only improvised recordings and only 4 classes, ['neutral', 'sadness', 'anger', 'happiness'], which resulted in a total of 2280 samples, which is still the largest set we worked with. The dataset was also divided into 5 folds, yet the division was conducted accordingly to five sessions inside the database. The dataset (or the subset in our case) is poorly balanced with the proportions spread between 7.8% and 52.8% (Figure 3-3(c)).

3.3 Feature extraction

The features that are extracted for the experiments are consistent with ones presented in Issa et. al. [22] as the model is the same. These are the five audio features:

1. Mel Frequency Cepstral Coefficients. MFCC is one of the widely used features not only in SER, but in speech recognition itself. Furthermore, MFCC is widely used in speaker/voice identification tasks[18], speaker/voice reconstruction[9]

and other sound-related classification tasks. MFCC is low-level feature which means that it describes rather physical characteristics of a sound.

2. Chromagram. The chroma feature, a feature of music that represents a chromatic scale (or musical notes/pitches). Chromagram is frequently used in music genre classification [7, 42].

3. Mel Spectrogram. A spectrogram of a sound is a visual representation of an audio signal portrayed in the time coordinate. Mel Spectrogram is the same as Spectrogram with little difference. In the Spectrogram, the x-axis is time, and the y-axis is the frequency in the Hertz scale, whereas x is time and y is a frequency in the Mel scale in the case of Mel Spectrogram.

4. Spectral Contrast. Spectral Contrast was first introduced by Jiang et. al in 2002 [24]. It is a feature that represents a relative distribution of frequencies in the sound. It is also widely used in music recognition as its filters are based on an octave scale.

5. Tonal Centroids (tonnetz). Tonnetz is also a handcrafted feature presented by Harte et. al. in 2006 [17]. This feature is used to detect a change in harmonics by measuring the distance between major and minor thirds and fifths in the sound. It is useful for emotion classification as humans tend to associate major and minor scales with 'happy' and 'sad' feelings respectively.

All the above features are extracted and fused sequentially in a one-dimensional vector of length 193 (40 mfccs, 12 chromas, 128 mels, 7 contrasts, and 6 tonnetzs). This feature vector is fed into the input layer of the model. Further, we will describe each experiment conducted on this model.

3.4 Experiment Setup

We conducted an identical set of experiments for each dataset. We maintained consistency in parameters throughout all experiments in order to positively identify correlations if there are any. Overall, we conducted four different experiments for each speech database:

3.4.1 Experiment 1 - Baseline

The first experiment aimed to reproduce the experiment described by Issa et. al. [22] and validate the results. PyTorch library is used for model definition, training, and validation. We used RMSProp as optimizer with parameters of learning rate of 0.0001 and decay of 1e-6, and CrossEntropyLoss as loss function as described in the paper. The training was conducted in 700 epochs. This setup was consistent overall experiments conducted further. For this experiment, we trained and tested on all three datasets without any augmentation or transformations of samples.

3.4.2 Experiment 2 - Pitch Only

In the next experiment, we populated samples with its pitch-shifted versions. Two new samples were generated from one sample: one with pitch-shifted to 4.5 steps up and the other 4.5 steps down. As a result, the number of samples tripled. The value of 4.5 was chosen as changing up and down with this value voice of a human changed the most and naturally, i.e. woman voice with 4.5 steps up resembled child's voice. This is an indication of that the voice of adults (between 120-200 Hz) are transformed into the range of average child frequency range (around 300 HZ) [38]. This augmentation procedure (and the following ones) was conducted within the folds to avoid common samples between folds, even transformed ones. The training procedure remained the same, while testing was done on both augmented (i.e. with the same augmented samples) and original (i.e. without additional augmented samples) folds. In other words, in the latter case, the model would train on augmented folds and be tested on the other fold, except its version before augmentation.

3.4.3 Experiment 3 - Speed Only

In this experiment, we populated original samples with speed-shifted versions. Analogously, two new samples were generated from one sample: one sped up to 125% and the other slowed down to 75% relative to their original versions. Augmentation and validation methods are consistent with the previous experiment: augmentation

within the folds and validation on both augmented and original folds.

3.4.4 Experiment 4 - Full Augmentation

For this experiment, we combined all the transformations. These are the transformations that have been applied for each sample:

- Pitch up for 1.5, 3.0, and 4.5 steps
- Pitch down for 1.5, 3.0, and 4.5 steps
- Speed up to 125%
- Slow down to 75%
- Trim silence before and after the speech

We added more pitch transitions in between for the model to learn. Despite the fact that newly added pitch shifts do not transform voice to completely resemble another person for the human ear they might add robustness for the model to train on. We also added samples with removed leading and trailing silences. This composite augmentation increases the sample size 10 times since new nine samples are generated from single one. Similarly, augmentation and validation methods are consistent with previous experiments.

Chapter 4

Results

4.1 RAVDESS

Experiment 1 - Baseline. No data augmentation or transformation of samples were applied for this experiment. Our results show an accuracy of 66.6%, with a maximum of 70.8% and a minimum of 60.7%. The accuracy across folds is presented in Figure 4-1. The table can be interpreted as the accuracy of the trained model on a particular fold. For example, a number under column 'fold1' represents an accuracy of the model tested on 'fold1' but trained on a combination of all other folds. Confusion matrix of the experiment is shown in Figure 4-2.

Experiment 2 - Pitch Only. The results of the first experiment with pitch data augmentation are shown in Figure 4-3. The average accuracy for augmented and original test sets are 65.3% and 65.0% respectively. The maximum accuracy for the original set is 67.4%, and the minimum is 60.3%. Confusion matrix for both augmented and original test sets are presented in Figure 4-4.

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	70.8%	68.1%	60.7%	65.8%	67.7%	66.6%

Figure 4-1: RAVDESS: Accuracy across folds of Experiment 1

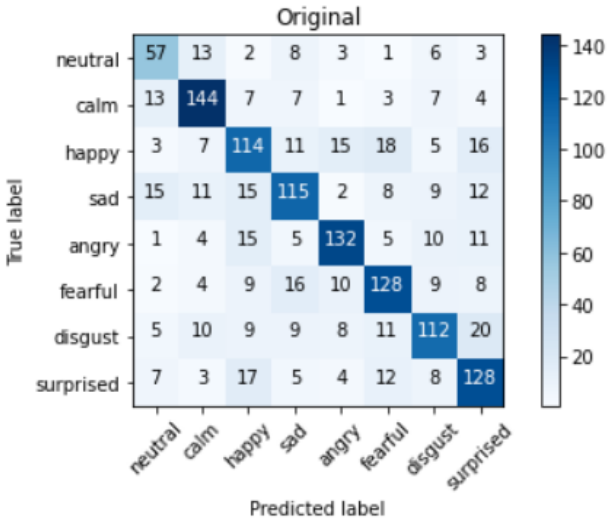

Figure 4-2: RAVDESS: Confusion matrix of Experiment 1

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	65.7%	68.6%	62.4%	64.1%	65.9%	65.3%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	67.4%	68.4%	60.3%	63.6%	65.4%	65.0%

Figure 4-3: RAVDESS: Accuracy across augmented and original folds of Experiment 2

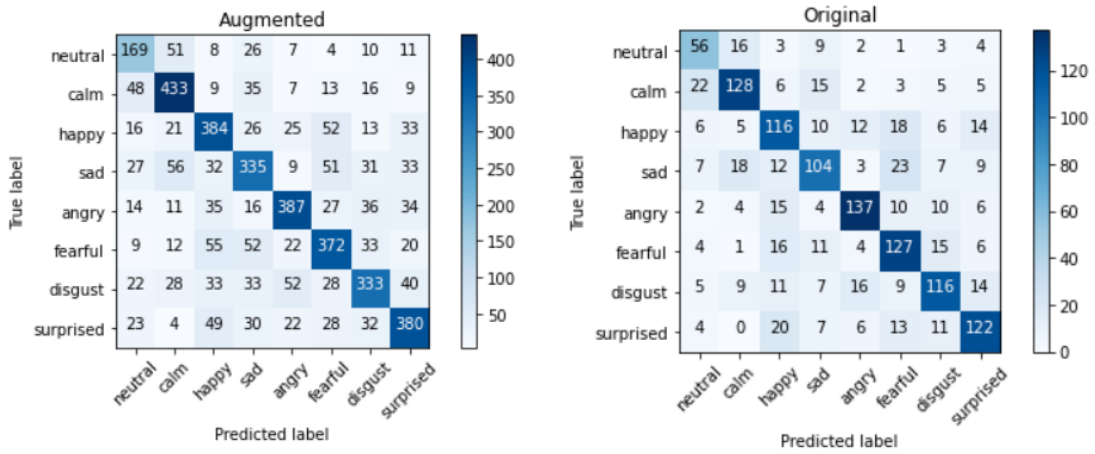

Figure 4-4: RAVDESS: Confusion matrices for augmented and original test sets of Experiment 2

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	70.7%	72.0%	64.9%	68.3%	70.3%	69.2%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	72.2%	72.6%	67.7%	69.1%	69.8%	70.3%

Figure 4-5: RAVDESS: Accuracy across augmented and original folds of Experiment 3

Experiment 3 - Speed Only. In the third experiment we augmented the dataset with speed modified samples. Results can be viewed in Figure 4-5. The average accuracy for augmented and original test sets are 69.2% and 70.3% respectively. With a maximum of 72.6% and a minimum of 67.7% for the original test set. Figure 4-6 show confusion matrices for augmented and original test sets.

Experiment 4 - Full Augmentation. In the final experiment we combined and added more transformations. Figure 4-7 presents results for both test sets. The average accuracy for augmented and original sets was reported to be 67.9% and 72.9% respectively. The maximum and minimum accuracy for the original set are 69.6% and 77.1% respectively. Confusion matrices for augmented and original test sets can be

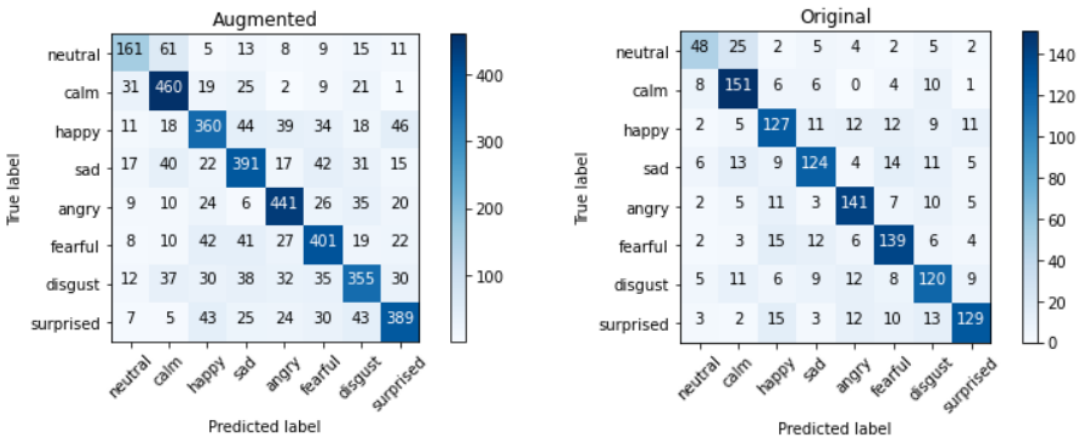

Figure 4-6: RAVDESS: Confusion matrices for augmented and original test sets of Experiment 3

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	69.6%	69.1%	65.2%	66.8%	68.6%	67.9%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	72.2%	77.1%	71.3%	71.0%	72.8%	72.9%

Figure 4-7: RAVDESS: Accuracy across augmented and original folds of Experiment 4

found in Figure 4-8.

4.2 EMO-DB

Experiment 1 - Baseline. Model trained on original dataset yielded an mean accuracy of 78.3% across folds, with the 81.2% highest and 71.9% lowest. Complete table of accuracy can be seen in Figure 4-9. Confusion matrix is demonstrated in Figure 4-10

Experiment 2 - Pitch Only. Pitch only transformations of samples resulted in a little higher average accuracy of 80.4% on the original test set with 72.9% and 87.5% as lowest and highest records across folds. Complete results for augmented and

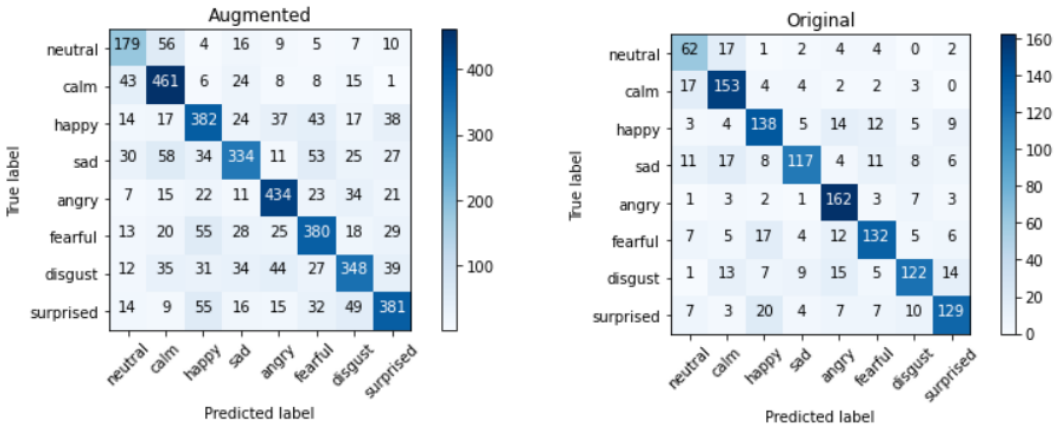

Figure 4-8: RAVDESS: Confusion matrices for augmented and original test sets of Experiment 4

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	79.2%	78.1%	81.2%	71.9%	81.2%	78.3%

Figure 4-9: EMO-DB: Accuracy across folds of Experiment 1

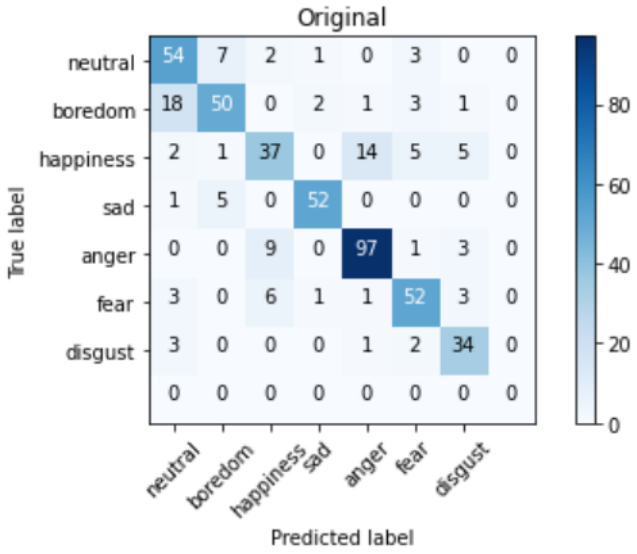

Figure 4-10: EMO-DB: Confusion matrix of Experiment 1

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	72.8%	75.6%	81.6%	75.7%	76.6%	76.5%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	72.9%	79.2%	87.5%	80.2%	82.3%	80.4%

Figure 4-11: EMO-DB: Accuracy across augmented and original folds of Experiment 2

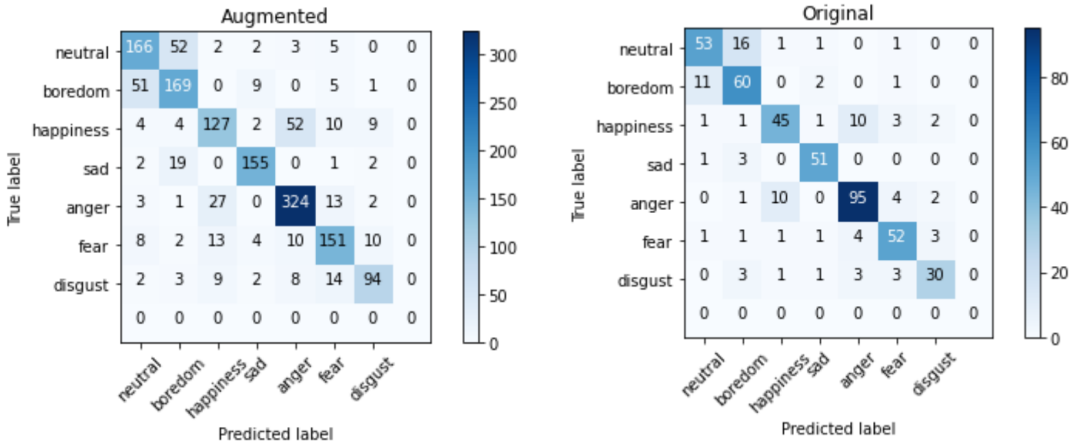

Figure 4-12: EMO-DB: Confusion matrices for augmented and original test sets of Experiment 2

original test sets can be observed in Figure 4-11, and Figure 4-12 shows confusion matrix for both augmented and original test sets.

Experiment 3 - Speed Only. An average accuracy of 80.8% was reached with speed only data augmentation on samples of EMO-DB dataset on the original testing set. Figure 4-13 shows that accuracy across folds range between 79.2% and 83.3%. Confusion matrices are shown in Figure 4-14.

Experiment 4 - Full Augmentation. Data augmentation with both speed and pitch transformations, and silence trimming recognized 82.3% of all samples from the original test set. This model fluctuates between 76.0% and 86.5% accuracy across all five folds. Accuracy and confusion matrices are demonstrated in Figures 4-15 and

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	76.2%	79.4%	82.9%	81.2%	81.2%	80.2%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	79.2%	80.2%	82.3%	83.3%	79.2%	80.8%

Figure 4-13: EMO-DB: Accuracy across augmented and original folds of Experiment 3

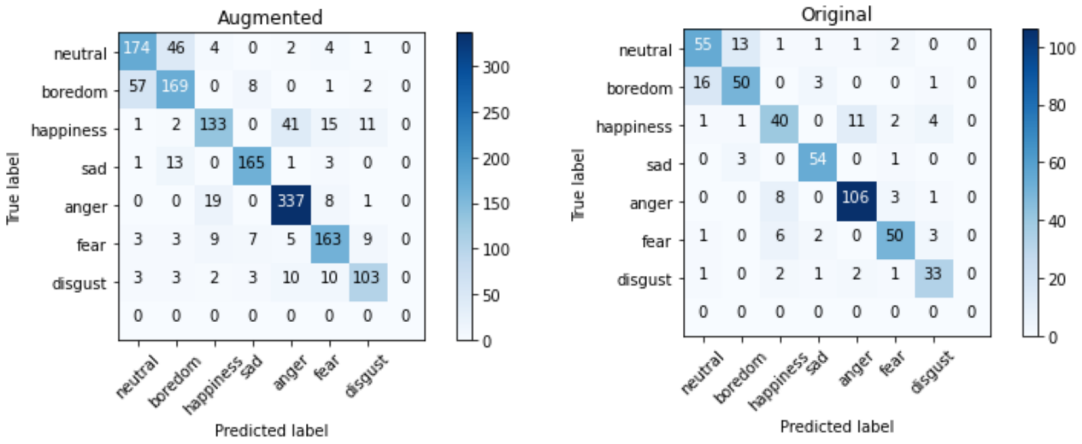

Figure 4-14: EMO-DB: Confusion matrices for augmented and original test sets of Experiment 3

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	74.7%	76.8%	82.1%	83.0%	79.2%	79.2%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	76.0%	82.3%	86.5%	85.4%	81.2%	82.3%

Figure 4-15: EMO-DB: Accuracy across augmented and original folds of Experiment 4

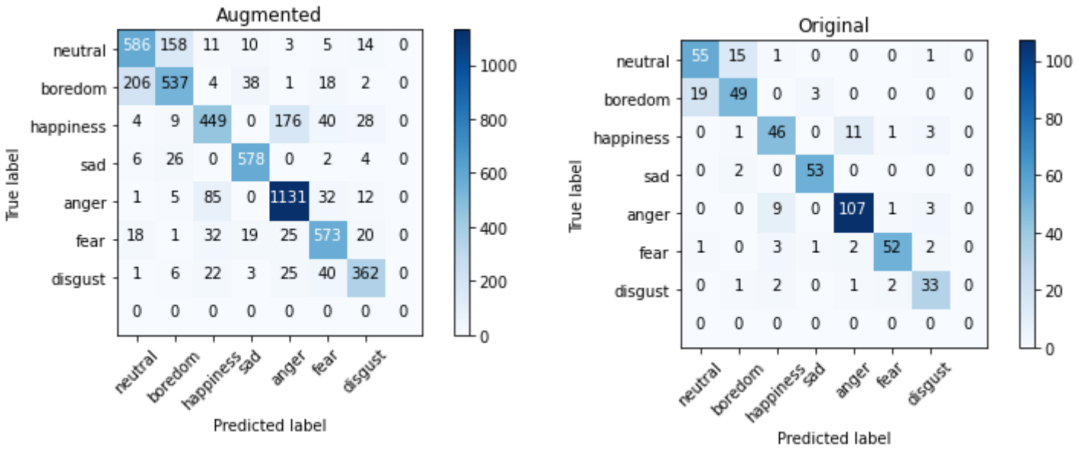

Figure 4-16: EMO-DB: Confusion matrices for augmented and original test sets of Experiment 4

4-16 respectively.

4.3 IEMOCAP

Experiment 1 - Baseline. Results of training and testing on original dataset can be discovered in Figure 4-17. Model reached average accuracy of 54.5% on largest and least balanced set. Accuracy across folds is distributed between 51.4% and 56.2%. Figure 4-18 presents confusion matrix of Experiment 1.

Experiment 2 - Pitch Only. Pitch shifting of samples in IEMOCAP set resulted in a change of mean accuracy to 56.5%, while its spread between folds fell in the range

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	51.4%	56.0%	56.1%	56.2%	52.9%	54.5%

Figure 4-17: IEMOCAP: Accuracy across folds of Experiment 1

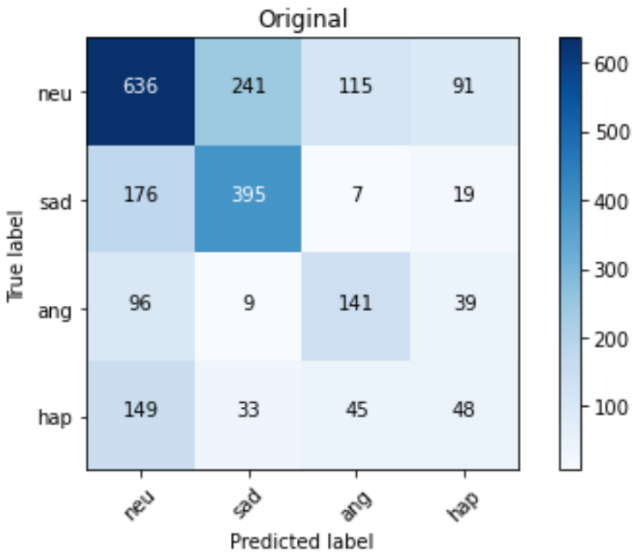

Figure 4-18: IEMOCAP: Confusion matrix of Experiment 1

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	53.0%	61.8%	58.1%	60.1%	48.3%	56.3%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	52.9%	63.0%	56.6%	57.6%	52.2%	56.5%

Figure 4-19: IEMOCAP: Accuracy across augmented and original folds of Experiment 2

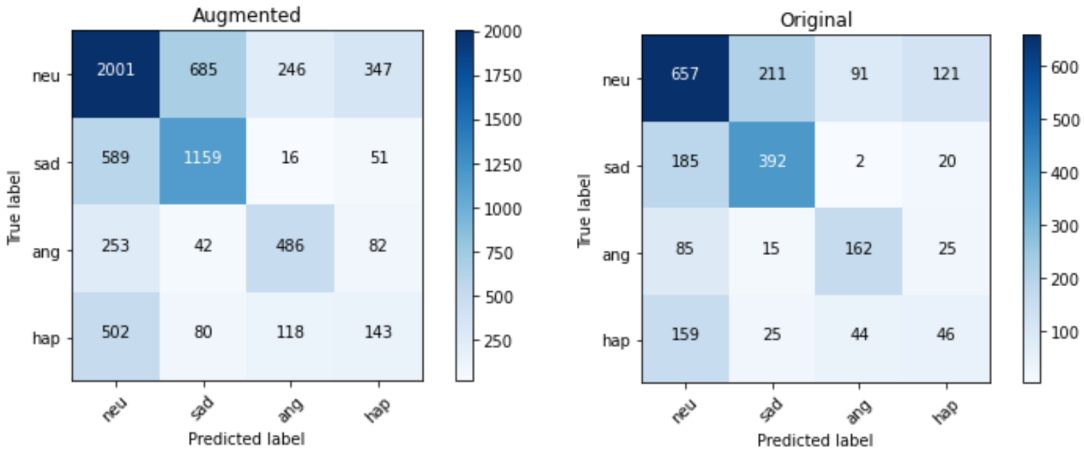

Figure 4-20: IEMOCAP: Confusion matrices for augmented and original test sets of Experiment 2

52.2% and 63.0%. Detailed list of results for each fold can be found in Figure 4-19, and Figure 4-20 shows the corresponding confusion matrix.

Experiment 3 - Speed Only. The results of the testing model that trained on a set with speed transformations is shown in Figure 4-21. It indicates average accuracy as high as 58.0% on the original set and accuracy of 55.3% and 59.9% correspond to lowest and highest values across folds respectively. Figure 4-22 presents confusion matrix of current experiment.

Experiment 4 - Full Augmentation. Final experiment with composite data augmentation led the model to reach the level of accuracy of almost 56.0%. The model demonstrated the lowest accuracy of 52.6% and highest accuracy of exactly

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	57.1%	61.2%	57.5%	58.2%	53.5%	57.5%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	58.9%	59.9%	58.5%	57.3%	55.3%	58.0%

Figure 4-21: IEMOCAP: Accuracy across augmented and original folds of Experiment 3

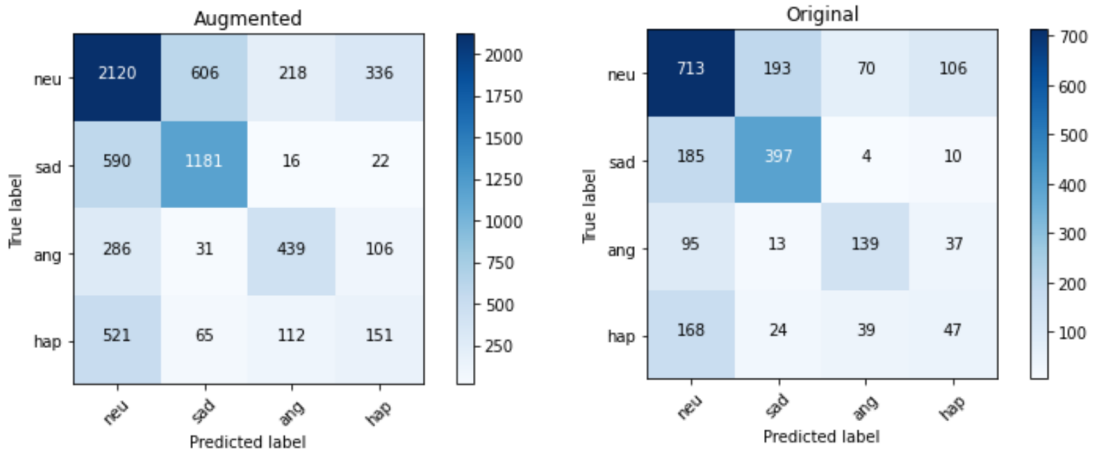

Figure 4-22: IEMOCAP: Confusion matrices for augmented and original test sets of Experiment 3

Augmented:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	52.8%	61.2%	52.8%	56.9%	50.3%	54.8%

Original:

Model	fold1	fold2	fold3	fold4	fold5	Average
Accuracy	54.3%	63.0%	53.4%	56.2%	52.6%	55.9%

Figure 4-23: IEMOCAP: Accuracy across augmented and original folds of Experiment 4

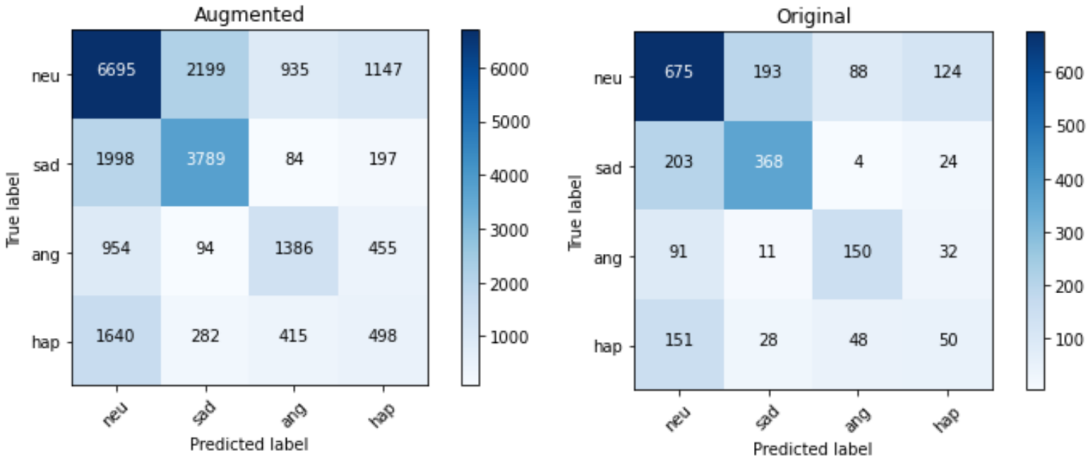

Figure 4-24: IEMOCAP: Confusion matrices for augmented and original test sets of Experiment 4

63.0%. Detailed list of values across folds is illustrated in Figure 4-23, while Figure 4-24 shows confusion matrices of the same results.

Chapter 5

Discussion

Discussion includes analysis of confusion matrices and dynamics of True Positive Rate. While confusion matrices of augmented test sets is valuable, our discussion is based on confusion matrices of original test sets. True Positive Rate is a widely used metric that describes the proportion of correctly predicted classes out of all true classes. It is useful for our analysis as we can quantitatively measure the performance of the model on the specific class. This way we can have insight into transformations and their effect on each class. TPR is calculated by the formula shown in Figure 5-3. In the formula, TP denotes True Positive outcomes which are located diagonally in a confusion matrix and FN denotes False Negative outcomes which correspond to rows of the confusion matrix (excluding TP). In this paper, we interchange the TPR with the terms 'sensitivity' and 'recall' which are different terms for the same concept.

5.1 RAVDESS

The first experiment on the original set without data augmentation has shown to be lower than expected. From the confusion matrix, it is clear that the most confusing emotions are neutral (to calm), disgust (to surprised), and happy (to fearful). Disgust class has the least true positive rate (TPR) of 60.3% followed by neutral and happy with 61.2% and 61.9% sensitivity respectively. The accuracy and confusion of happy class comply with the results of Issa et. all [22] experiment, even though

Classes	Baseline (Ex-1)	Pitch (Ex-2)	Speed (Ex-3)	Full (Ex-4)
neutral	61.29	59.57	51.61	67.39
calm	77.42	68.82	81.18	82.70
happy	60.32	62.03	67.20	72.63
sad	61.50	56.83	66.67	64.29
angry	76.30	72.87	76.63	89.01
fearful	68.82	69.02	74.33	70.21
disgust	60.87	62.03	66.67	65.59
surprised	69.57	66.67	68.98	68.98

Figure 5-1: RAVDESS: True Positive Rates (TPR) of classes (in %)

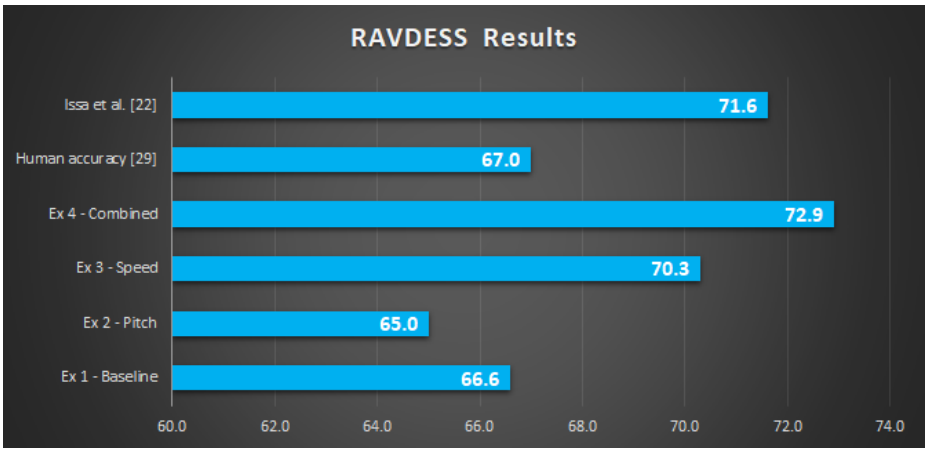

Figure 5-2: RAVDESS: Comparison of our results with literature [29, 22]

our implementation and testing of the identical model have some different results. We speculate that usage of different libraries and randomness of shuffles during the training process may cause such diversity.

Our first data augmentation, pitch transformation, had detrimental effects on the performance of the model. The average accuracy on the original test set dropped to 65.0% from 66.6% compared to the results without any data augmentation. Data augmentation with pitch modification affected positively on happy, angry and disgust with minor improvements of accuracy, while drastically decreasing performance in other classes. The angry class has become the most accurate class to recognize with the TPR of 72.8%, but achieving such property rather by an overall decrease in other classes rather than an improvement on this particular class. Pitch transformations also changed the most confused classes, neutral still retaining the most confused (to calm) class, for being most underrepresented class, followed by sad (to fearful) and

$$TPR = \frac{TP}{TP + FN}$$

Figure 5-3: True Positive Rate equation

calm (to neutral).

In the next experiment with speed transformations, we received a higher performance in comparison to the original experiment (Experiment 1). Specifically, 3.7% increase overall in average accuracy for the original test set. Accuracy of recognition of each emotion is increased with the only exception of the neutral class. Neutral class still remains the most confused class, followed by fearful and surprised both mostly confused with happy class. Speed transformation augmentation increased the sensitivity model to happy class the most, from 77.4% to 81.2%.

The final experiment with the fusion of all transformations yielded the best performance so far. The composition of pitch, speed, and trimming transformations increased the performance of the model by 6.3% from 66.6% to 72.9%. It has increased the accuracy of all classes, affecting the angry class the most increasing recall on this class from 72.1% to 89.0%. However, this data augmentation method affected the accuracy of surprised and sad classes the least with an overall increase of less than 1%. The transformations methods that increase the accuracy of these classes are yet to be determined. Similarly, the neutral class still retains the property of the least accurate and most confused (to calm) class. The surprised class, who had not been affected by augmentation much, also remains one of the most confused (to happy) classes only second to neutral. Comparison of our models with the literature presented in Figure 5-2.

5.2 EMO-DB

EMO-DB is the less balanced dataset with fewer samples and different languages. Nevertheless, the model has shown higher accuracy in comparison to RAVDESS database on the baseline experiment. According to the confusion matrix, the model

Classes	Baseline (Ex1)	Pitch (Ex2)	Speed (Ex3)	Full (Ex4)
neutral	80.60	73.61	75.34	76.39
boredom	66.67	81.08	71.43	69.01
happiness	57.81	71.43	67.80	74.19
sad	89.66	92.73	93.10	96.36
anger	88.18	84.82	89.83	89.17
fear	78.79	82.54	80.65	85.25
disgust	85.00	73.17	82.50	84.62

Figure 5-4: EMO-DB: True Positive Rates (TPR) of classes (in %)

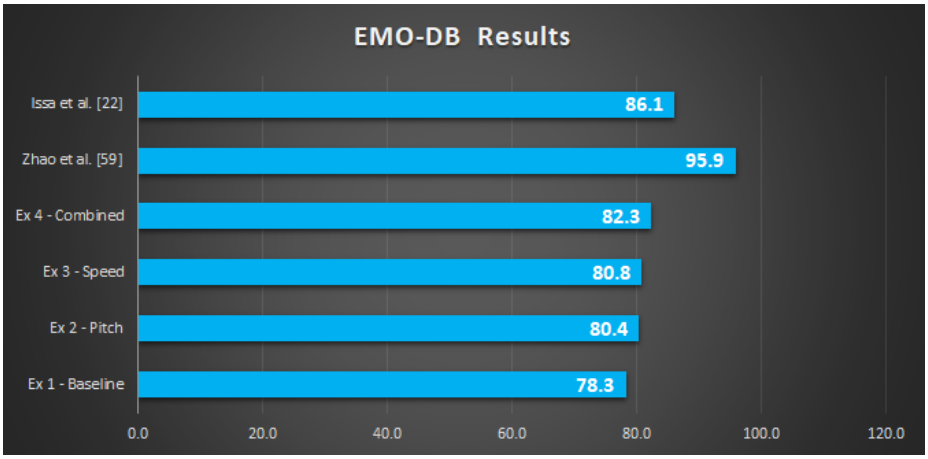

Figure 5-5: EMO-DB: Comparison of our results with literature [61, 22]

learned to recognize sad and anger classes most accurately with the sensitivity values 89.6% and 88.2% respectively. As it has been observed from previous dataset results, the model still has the highest confusion rate on classes between boredom and neutral, and between happiness and anger. It seems recognizing less pronounced emotions is universally complex despite the differences between speech sets.

As for the results of the next experiment, the application of pitch transformations affected positively the overall accuracy of the model increasing it to 80.4%. The sadness class still remains as the highest recognizable class with an increased TRP of 92.7%. Pitch transformation affected the boredom and happiness class the most positively increasing their TPR from 66.7% to 81.1% and from 57.8% to 71.4% respectively. However, this transformation had an adverse effect on the disgust class, greatly decreasing the sensitivity of the class from 85.0% to 73.2% compared to baseline results.

The next experiment, with speed alterations, also demonstrated higher overall ac-

curacy in comparison to baseline, around the same level of performance as the model learned on pitch transformations. Similarly, the model still struggles to differentiate boredom and neutral classes, and sadness remains as a class with the highest sensitivity with recall as high as 93.1%. Despite resulting in overall near accuracy, pitch transformation had a greater effect on few chosen classes while speed transformation's effect spread evenly on all classes decreasing its intensity. This behavior will be discussed further in Section 5.4.

The final experiment, the composite augmentation including both pitch and speed transformations, demonstrated the highest overall accuracy of 82.3%. The most affected class was happiness which is correctly recognized in 74.2% of cases, a 16.4% increase from the baseline model. However, it is the only class that has been greatly influenced by the combined augmentations: the impact was minor to other classes, quite similar behavior to speed the only augmentation. Nevertheless, the combination of pitch, speed, and trimmings proved to be the most efficient data augmentation so far for the EMO-DB dataset with a 4.0% increase in overall accuracy, which is higher than SRHA method [39] with 2.28% for the 4-class subset (angry, happy, neutral, and sad) of EMO-DB. Figure 5-5 shows our results in comparison with the results of other researchers.

5.3 IEMOCAP

The first experiment determined baseline accuracy and confusion matrix for further comparative analysis on the largest and least balanced dataset. The model that was trained on the original dataset without any augmentation demonstrated an accuracy of 54.5%. Confusion matrix indicates high confusion of happiness, a least populated class, with neutral class. Latter is often confused with sadness, yet these two classes are the ones the trained model is most sensitive towards, with TPRs of 58.7% and 66.2% respectively. It is one of the disadvantages of the imbalanced dataset. Despite the low accuracy level, it is beneficial in observing data augmentation behavior in imbalanced conditions.

Classes	Baseline (Ex1)	Pitch (Ex2)	Speed (Ex3)	Full (Ex4)
neutral	58.73	60.83	65.90	62.50
sad	66.16	65.44	66.61	61.44
anger	49.47	56.45	48.94	52.82
happy	16.55	16.79	16.91	18.05

Figure 5-6: IEMOCAP: True Positive Rates (TPR) of classes (in %)

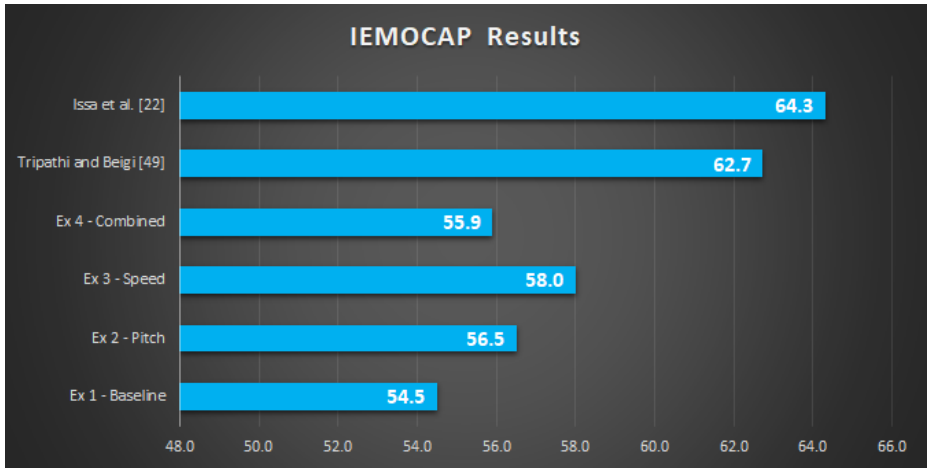

Figure 5-7: IEMOCAP: Comparison of our results with literature [50, 22]

Augmenting set with pitch-shifted versions of speech illustrated a positive effect on the overall accuracy of the model. In terms of impact on sensitivity across classes, the anger class benefited from the transformation the most, with a 7.0% increase in TPR. Despite the overall positive response, the sadness and happiness classes experienced minimal effect. As far as the sample size is concerned, this particular augmentation triples the samples of each class further increasing the gap between imbalanced classes by order of magnitude. One of the interesting observations is that this particular augmentation turned out to be unbiased towards the most represented class affecting both overpopulated and underpopulated classes.

Altering the speed of speech and augmenting it to the training set further raised the overall accuracy of the model to 58.0%. As experienced by previous experiments, high confusion rate between neutral and sadness, classes with the largest sample sizes, stay invariant to augmentations. Neutral class is one and only class that model has improved its sensitivity towards, with TRP increasing by 7.2%, while all other classes experienced less than 1% change in TPR. This outcome indicates a strong

Classes	Pitch			Speed			Combined		
	RAVDESS	EMO-DB	IEMOCAP	RAVDESS	EMO-DB	IEMOCAP	RAVDESS	EMO-DB	IEMOCAP
neutral	-1.72	-6.99	2.11	-9.68	-5.25	7.17	6.10	-4.21	3.77
sad	-4.67	3.07	-0.72	5.17	3.45	0.45	2.79	6.71	-4.73
anger	-3.43	-3.36	6.97	0.33	1.65	-0.53	12.71	0.98	3.34
happy	1.71	13.62	0.24	6.88	9.98	0.35	12.31	16.38	1.50

Figure 5-8: Change in TPR among common classes (in %)

bias towards the most represented class. Nevertheless, despite the effect on a single class, due to imbalanced weight, this model exhibits the highest improvement of 3.5% in overall accuracy for the IEMOCAP dataset. In comparison, VTLP augmentation [13] reported to achieve the highest of 1.1% increase for the same IEMOCAP dataset.

Full augmentation that increased sample size ten times and incorporated both speed and pitch variations displayed higher accuracy over the baseline model. However, while pitch and speed alteration showed improvements in performance in isolation, a combination of these augmentations resulted in lower accuracy. Analogously, one of the justifications for such behavior can be a gap problem which is strongly pronounced as the sample sizes as well as the gap between most and least represented classes dramatically increased. In terms of class sensitivity, sadness is the only class that experienced the detrimental effect, while all other classes benefited from this augmentation. Magnitudes of all changes were around the same level, both positive and negative, indicating augmentation does not favor a particular class. In the end, speed transformation demonstrated the highest improvement in performance despite the tendency to favor the most represented class. Comparison of our outcomes with the literature are demonstrated in Figure 5-7.

5.4 Common Observations

In this section, we analyze common classes and dynamics in their TRP as a result of data augmentation. Figure 5-8 shows the difference in true positive rates of common classes compared to baseline across augmentation types and datasets. Any values and details mentioned in this section can be discovered in this figure. There is a total of four classes that are present in all databases: neutral, sad, anger, and happy.

Neutral. Neutral is one of the most confused classes usually confused with boredom and calm. It has been concluded that combined transformation of pitch and speed has the most beneficial effect with the positive sum and mean of changes across datasets as 5.7% and 1.9% respectively. On the other hand, both speed and pitch in isolation showed a detrimental effect on the sensitivity of the neutral class with the sum and mean of changes as negative. However, it should be noted that the neutral class is over-represented in IEMOCAP which resulted in consistent positive changes across all augmentation experiments. Speed augmentation is the least effective and most detrimental of augmentations that showed sum and mean as -7.8% and -2.6% respectively.

Sad. Sadness experienced most benefits from speed transformations with the highest sum and mean as 9.1% and 3.0% respectively. Pitch ended up as the only detrimental augmentation type with negative values for sum and mean. Although combined augmentation had also a positive effect on the sensitivity of sadness in models in terms of sum and mean, speed augmentation in isolation is the only one that had a beneficial impact across all datasets.

Anger. Combined transformation tends to affect the anger class the most with sum and mean as high as 17.0% and 5.7% respectively. Although all the augmentation methods had positive values in terms of sum and mean, the difference of values between combined and isolated augmentations is quite large. In fact, pitch transformation had barely any effect on TPR with sum and mean of changes around 0.2% and 0.1% respectively. Moreover, combined augmentation is the only one that brought positive change across all datasets.

Happiness. Happiness is the single emotion class that had experienced universal positive effect by all augmentation methods across all datasets. It is also true for the IEMOCAP database where the happiness class is the least represented class in a set. Combined augmentation prevails with the highest impact on the sensitivity of happiness with sum and means as 30.2% and 10.1% respectively, the highest values throughout all experiments. As a result, observations illustrate that happiness is the most affected class in terms of value and magnitude of the changes by augmentation.

One of the speculations is that the intensity of the emotion is pronounced in the audio clearly (i.e. amplitude of the voice, frequent change in pitch). However, an anger class that has the same properties does not exhibit similar behavior in terms of sensitivity.

Chapter 6

Conclusion

In conclusion, we have tested the effect of three different data augmentation methods that incorporate two important audio properties, pitch, and speed, on three different datasets. As a result of augmentation experiments with these properties in isolation as well as in combination different patterns have emerged. First, in terms of the performance of the dataset, we found that augmentation with a combination of pitch and speed transformations had the highest effect in balanced datasets. In particular, RAVDESS and EMO-DB benefited most from this augmentation method. Secondly, we discovered that augmentation with speed alterations is a decent choice for an imbalanced dataset from the results of IEMOCAP database. Finally, we conducted a comparative analysis of the effects of augmentations methods on a subset of common classes. As a result, we learned that combined data augmentation has the highest positive effect in the majority of common classes followed by speed transformation in isolation.

6.1 Limitations

One of the limitations of the work is the resolution of the transformations. Pitch and speed augmentations in isolation have been altered in too few variations. For example, the pitch had only been shifted 4.5 steps up and down, ignoring steps in-between, It was a necessary sacrifice towards acquiring the strongest effect with

minimal computation time (i.e. more samples longer the training and testing time).

The next limitation is the usage of few metrics. TPR is a decent metric to access the class-specific performance, there are other model-specific or class-specific metrics that can also be utilized. Although some of them highly correlate with TPR they still can expose certain behavior of classes.

6.2 Future Work

In the future, we plan on applying accumulated knowledge about the properties of transformations and their effect on each class to further sophisticate the augmentation process. Specifically, we are interested in the selective augmentation of classes. In other words, instead of applying one type of augmentation to the whole dataset, select a custom augmentation method that favors a particular class. It has two main benefits: firstly, we may expect improvement in the performance of the whole model as we are certain each transformation is advantageous for each class; secondly, it will be possible to balance the imbalanced dataset by applying different degrees of augmentation for overpopulated and underpopulated classes.

Bibliography

[1] Ossama Abdel-Hamid, Li Deng, and Dong Yu. Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech*, volume 11, pages 73–5, 2013.

[2] Corina Albu, Eugen Lupu, and Radu Arsinte. Emotion recognition from speech signal in multilingual experiments. In *6th International Conference on Advancements of Medicine and Health Care through Technology; 17–20 October 2018, Cluj-Napoca, Romania*, pages 157–161. Springer, 2019.

[3] Surekha Reddy Bandela and T Kishore Kumar. Emotion recognition of stressed speech using teager energy and linear prediction features. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, pages 422–425. IEEE, 2018.

[4] Fang Bao, Michael Neumann, and Ngoc Thang Vu. Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *INTERSPEECH*, pages 2828–2832, 2019.

[5] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[7] YMD Chathuranga and KL Jayaratne. Automatic music genre classification of audio signals with machine learning approaches. *GSTF Journal on Computing (JoC)*, 3(2):13, 2013.

[8] Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, and Shrikanth Narayanan. Data augmentation using gans for speech emotion recognition. In *Interspeech*, pages 171–175, 2019.

[9] Dan Chazan, Ron Hoory, Gilad Cohen, and Meir Zibulski. Speech reconstruction from mel frequency cepstral coefficients and pitch frequency. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1299–1302. IEEE, 2000.

[10] Wan Ding, Mingyu Xu, Dongyan Huang, Weisi Lin, Minghui Dong, Xinguo Yu, and Haizhou Li. Audio and face video emotion recognition in the wild using deep neural networks and small datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 506–513. ACM, 2016.

[11] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[12] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

[13] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint arXiv:1802.05630*, 2018.

[14] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017.

[15] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[16] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5005–5009. IEEE, 2017.

[17] Christopher Harte, Mark Sandler, and Martin Gasser. Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 21–26, 2006.

[18] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. Speaker identification using mel frequency cepstral coefficients. *variations*, 1(4), 2004.

[19] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[20] Hai Huang, Hao Zhou, Xu Yang, Lu Zhang, Lu Qi, and Ai-Yun Zang. Faster r-cnn for marine organisms detection and recognition using data augmentation. *Neurocomputing*, 337:372–384, 2019.

[21] Yongming Huang, Ao Wu, Guobao Zhang, and Yue Li. Speech emotion recognition based on deep belief networks and wavelet packet cepstral coefficients. *International Journal of Simulation: Systems, Science and Technology*, 17(28):28–1, 2016.

[22] Dias Issa, M Fatih Demirci, and Adnan Yazici. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894, 2020.

[23] Navdeep Jaitly and Geoffrey E Hinton. Vocal tract length perturbation (vtlp) improves speech recognition. In *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, volume 117, 2013.

[24] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116. IEEE, 2002.

[25] Pengcheng Li, Yan Song, Peisen Wang, and Lirong Dai. A multi-feature multi-classifier system for speech emotion recognition. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, pages 1–6. IEEE, 2018.

[26] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, and Yazhu Liu. Cheavd: a chinese natural emotional audio–visual database. *Journal of Ambient Intelligence and Humanized Computing*, 8(6):913–924, 2017.

[27] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1–4. IEEE, 2016.

[28] Gang Liu, Wei He, and Bicheng Jin. Feature fusion of speech emotion recognition based on deep learning. In *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pages 193–197. IEEE, 2018.

[29] Steven R Livingstone and Frank 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*, 13(5):e0196391, 2018.

[30] Mohamed Loey, Gunasekaran Manogaran, and Nour Eldeen M Khalifa. A deep transfer learning model with classical data augmentation and cgan to detect covid-19 from chest ct radiography digital images. *Neural Computing and Applications*, pages 1–13, 2020.

[31] Reza Lotfidereshgi and Philippe Gournay. Biologically inspired speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5135–5139. IEEE, 2017.

[32] Jiang-Jing Lv, Xiao-Hu Shao, Jia-Shui Huang, Xiang-Dong Zhou, and Xi Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184–196, 2017.

[33] Kasiprasad Mannepalli, Panyam Narahari Sastry, and Maloji Suman. A novel adaptive fractional deep belief networks for speaker emotion recognition. *Alexandria Engineering Journal*, 56(4):485–497, 2017.

[34] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

[35] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231. IEEE, 2017.

[36] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. A study of language and classifier-independent feature analysis for vocal emotion recognition. *arXiv preprint arXiv:1811.08935*, 2018.

[37] Sarala Padi, Dinesh Manocha, and Ram D Sriram. Multi-window data augmentation approach for speech emotion recognition. *arXiv preprint arXiv:2010.09895*, 2020.

[38] William Hughes Perkins and Raymond D Kent. *Textbook of functional anatomy of speech, language, and hearing.* Taylor & Francis, 1986.

[39] Bagas Adi Prayitno and Suyanto Suyanto. Segment repetition based on high amplitude to enhance a speech emotion recognition. *Procedia Computer Science*, 157:420–426, 2019.

[40] Sonali T Saste and SM Jagdale. Emotion recognition from speech using mfcc and dwt for security system. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, volume 1, pages 701–704. IEEE, 2017.

[41] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–577. IEEE, 2004.

[42] Christine Senac, Thomas Pellegrini, Florian Mouret, and Julien Pinquier. Music feature maps with convolutional neural networks for music genre classification. In *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, pages 1–5, 2017.

[43] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems*, 26(5):1019–1034, 2014.

[44] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223, 2011.

[45] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[46] Bo Sun, Qihua Xu, Jun He, Lejun Yu, Liandong Li, and Qinglan Wei. Audio-video based multimodal emotion recognition using svms and deep learning. In *Chinese Conference on Pattern Recognition*, pages 621–631. Springer, 2016.

[47] Linhui Sun, Sheng Fu, and Fu Wang. Decision tree svm model with fisher feature selection for speech emotion recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019(1):2, 2019.

[48] Fei Tao and Gang Liu. Advanced lstm: A study about better time dependency modeling in emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910. IEEE, 2018.

[49] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[50] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.

[51] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309, 2017.

[52] Panagiotis Tzirakis, Jiehao Zhang, and Bjorn W Schuller. End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5089–5093. IEEE, 2018.

[53] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[54] Curtis Wigington, Seth Stewart, Brian Davis, Bill Barrett, Brian Price, and Scott Cohen. Data augmentation for recognition of handwritten words and lines using a cnn-lstm network. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 639–645. IEEE, 2017.

[55] Ao Wu, Yongming Huang, and Guobao Zhang. Feature fusion methods for robust speech emotion recognition based on deep belief networks. In *Proceedings of the Fifth International Conference on Network, Communication and Computing*, pages 6–10. ACM, 2016.

[56] Lu Yi and Man-Wai Mak. Adversarial data augmentation network for speech emotion recognition. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 529–534. IEEE, 2019.

[57] Lu Yi and Man-Wai Mak. Improving speech emotion recognition with adversarial data augmentation network. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[58] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):3030–3043, 2017.

[59] Shiqing Zhang, Xiaoming Zhao, Yuelong Chuang, Wenping Guo, and Ying Chen. Feature learning via deep belief network for chinese speech emotion recognition. In *Chinese Conference on Pattern Recognition*, pages 645–651. Springer, 2016.

[60] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE, 2018.

[61] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical Signal Processing and Control*, 47:312–323, 2019.

[62] Lianzhang Zhu, Leiming Chen, Dehai Zhao, Jiehan Zhou, and Weishan Zhang. Emotion recognition from chinese speech for smart affective services using a combination of svm and dbn. *Sensors*, 17(7):1694, 2017.