BMISEC:Corpus Of Burmese Emotional Speech

Lwin Lwin Mar
Natural Language Processing
University of Computer Studies,Yangon.
Yangon,Myanmar
lwinlwinmar@ucsy.edu.mm

Win Pa Pa
Natural Language Processing
University of Computer Studies,Yangon.
Yangon,Myanmar
winpapa@ucsy.edu.mm

Tin Lay Nwe
Institute of Infocomm Research(I2R), Singapore.
Singapore.
tlnma@i2r.a-star.edu.sg

***Abstract*—Speech is the most popular method of communication and emotions play an important role in human to human communication. For naturalness in human to machine communication, machine needs to understand human emotions well. However, Speech Emotion Recognition (SER) is a challenging task for machine. In this paper, we propose a Burmese Movies Interviews Speech Emotion Corpus (BMISEC) for Burmese SER and present our analysis on collected emotion data. Emotion data are collected from Myanmar movies and interviews. There are seven emotions categories in speech corpus: Angry, Happy, Disgust, Fear, Sad, Surprise and Neutral. Four important Burmese tones are low tone, high tone, creaky tone and checked tone (stopped tone). Hot angry speech contains more high tone than other emotions. Fear speech has more low tone. Comparisons of pitch, intensity and formant of important Burmese tones are presented.**

***Keywords*— Speech Emotion Recognition, Burmese emotional speech corpus, Burmese tones**

I. INTRODUCTION

There are many emotional speech databases developed for different languages. Speech emotion has been studied since the 1950', but the investigation of emotional cues in speech is gaining growing attention. This is because of new developments in human machine interfaces that see applications of automatic recognition and simulation. The Berlin database is developed for German language, the Persian emotion detection database is developed for Persian language, the RAVDESS database is built for North America language.

This paper describes the Burmese dataset of acted emotions and nature of Burmese speech, and relationship between seven emotions and Burmese tones.

Emotion databases can be classified as two types: natural and acted. Natural databases are collected from the ordinary human conversions in daily life. In acted databases, emotional sentences are expressed by professional actors. Developing natural databases are very expensive and are commonly restricted. Burmese Movies Interviews Speech Emotion Corpus is mixed of acted and natural speech. From seven classes, some of disgust, neutral and happy speech samples are collected from interviews. To evaluate BMISEC, speech emotion classification is experimented with it and results are presented. As experimental results, it is useful for speech emotion recognition.

The next section are related speech emotion databases, nature of Burmese speech, data collection, choice of emotions and actor, labeling data, properties of four tones, experimenting data corpus, linguistics materials and conclusion.

II. RELATED WORK

The study in [1] describes an emotional speech database for German language. Emotions were simulated by ten actors (5 females and 5 males), producing 10 German utterances (5 short and 5 longer sentences) which could be used in everyday communication and are interpretable in all emotions. German actors performed the emotional utterances. The emotional database has four big emotions plus boredom and disgust. The material was evaluated in an automated listening test and each utterance was judged by 20 listeners with respect to recognizability and naturalness of the displayed emotion.

Another study in [2] introduces a large-scaled, validated database for Persian language called Sharif Emotional Speech Database (ShEMO). The ShEMO includes 3000 semi-natural utterances, equivalent to 3h and 25 min of speech data extracted from online radio plays. The database covers speech samples of 87 native-Persian speakers for five basic emotions including anger, fear, happiness, sadness, and surprise, as well as neutral state.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a validated multimodal database of emotional speech and song [3]. The dataset is collected for gender balance and utterances are from 24 professional actors, vocalizing lexically- matched statements in a neutral North American accent. This dataset includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song includes calm, happy, sad, angry, and fearful emotions.

III. NATURE OF BURMESE SPEECH

Burmese is a tonal, pitch-register, and syllable-time language. It is tonal language therefore phonemic contrasts can be made on the basis of the tone of the vowel. In Burmese, these contrasts involve not only pitch, but also phonation, intensity (loudness), duration, and vowel quality.

There are four contrastive tones in Burmese. They are low tone, high tone, creaky tone, and checked tone. The pitch contours of creaky and checked tones are very similar. Pitch is not the only contrastive feature in Burmese tones [9]. Creaky tone maintains a falling pitch contour generally while checked tone does not [9].

IV. DATA COLLECTION

Emotion speech corpus is collected from Myanmar movies and interviews. Myanmar movies are mainly collected from Mahar [1] .Interviews are collected from

[1] https://www.youtube.com/channel/UCJFGNmtLtwmX752P7OhcdvQ/videos

YouTube[2].The collected mp4 files are converted into wave files by using Format Factory[3] with sampling rate 16000 HZ. The speech samples are extracted from wave files by using Praat (doing phonetics by computer). And then they are labeled with naming rule.

V.CHOICE OF EMOTIONS AND ACTOR

There are seven emotions in Burmese dataset. They are Anger, Happy, Fear, Disgust, Surprise, Sad and Neutral. Anger can be divided into (normal or cold) anger and hot anger. Hot anger is produced with higher, more varied pitch, and even greater energy. In comparison to neutral speech, disgust is produced with a lower, downward directed pitch, (with energy lower fast formant) and first attack times similar to anger. Fear can be divided into panic and anxiety. Emotion utterances come from nine actors including five males and four females. Utterances are extracted from Myanmar movies and celebrity interviews. Male actors are classified as 01,71., etc. (denoted by odd numbers) and female actors are classified as 22,56., etc (denoted by even numbers).

VI.LABELING DATA

Speech utterances are labeled according to their emotion content. Each speech sample contains only one emotion type. The file name consists of four parts. The first part is emotion. Surprise is 1, anger is A, happy is H, fear is F, sad is S, disgust is D and neutral is N. The second part is actor.01,71,73,77,81 are male actors and 22,56,60,66 are female actors. The third part is number of sample and the last part is code for movie and interview. They are 08 to 19 code values. Example is 101surprise308.wav. The first digit '1' represents surprise emotion,01 is male actor,3 is 3rd sample in surprise dataset.08 is code of movie. Another example is A22ngry1202.The first character 'A' is angry emotion,22 is female actor,12 is 12th sample in angry dataset and 02 is code of movie. The remaining characters such as "ngry" have no definition. Another example is H73appy202.H is happy emotion,73 is male actor, and it is 2nd happy sample in happy dataset with movie code '02'. The lengths of the utterances range from 1 second to 3 seconds.

The labeling process involves two annotators:35-years - old student and 44-years-old man. Annotators labeled the speech samples according to procedure:(1) play the audio, (2) determine the emotion class which is most appropriate based on intonation and pronunciations, (3) label the speech sample according to emotion class.

The speech samples in .wav format are preprocessed for removing noise and background music. Audacity [4] is used to remove the noise and music from speech samples by using silence function.

[2] https://www.youtube.com/watch?

[3] Format Factory is free and multifunctional, multi-media processing tool. It can be used as audio and video converter, clipper, joiner, spliter, mixer, crop and delogo.

[4] Audacity is the free, open source, cross platform software for recording and editing sounds.

VII. PROPERTIES OF FOUR TONES

Table 1. shows the properties of four tones of Burmese language. Low tone has low pitch, low intensity, medium duration. The four tones are ကာ, ကား, က, ကတ်. High tone has high pitch, but it falls towards the end, the high intensity, and long duration.

TABLE I. PROPERTIES OF FOUR BURMESE TONES.

Tone	Pitch	Intensity	Duration
Low	Low	Low	Medium
High	High (falling towards the end)	High	Long
Creaky	High	High	Short
Checked	High	High	Very short

As presented in Fig1, hot angry speech involves 50% of words with high tone. And, (normal or cold) speech has more high tone.

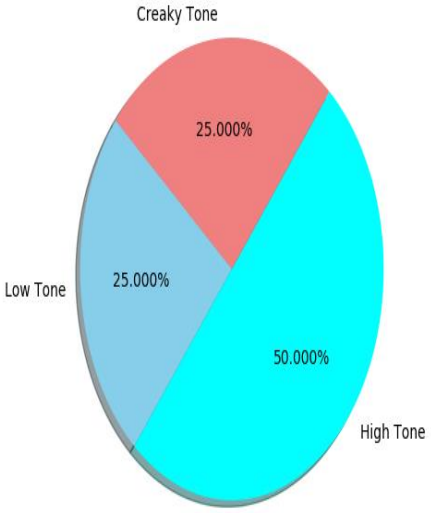

Fig.1. Percentage of four tones in angry speech.

Happy speech has more low tone than other tones.

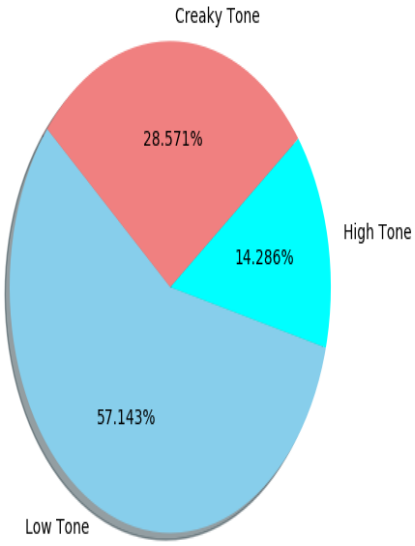

Fig.2. Percentage of four tones in happy speech.

A disgust speech has a little more creaky tones than other two tones and more creaky tones than checked tones.

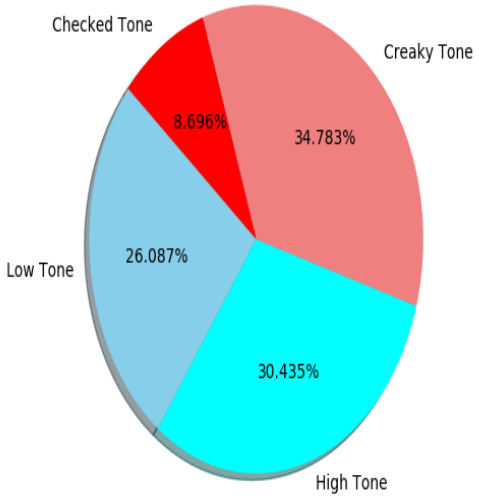

Fig.3.Percentage of four tones in disgust speech.

Fear speech contains many low tones than other classes of emotion.

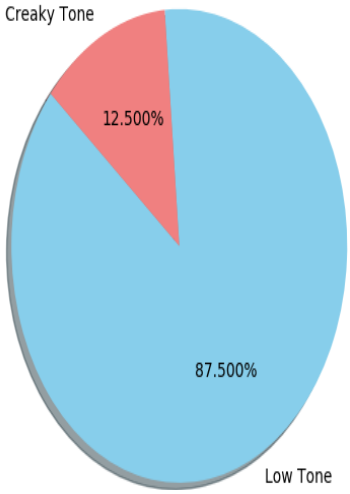

Fig 4. Percentage of four tones in fear speech.

As presented in figures, 5, 6 and 7 neutral speech, surprise and sad emotions have more words with low tone than other tones.

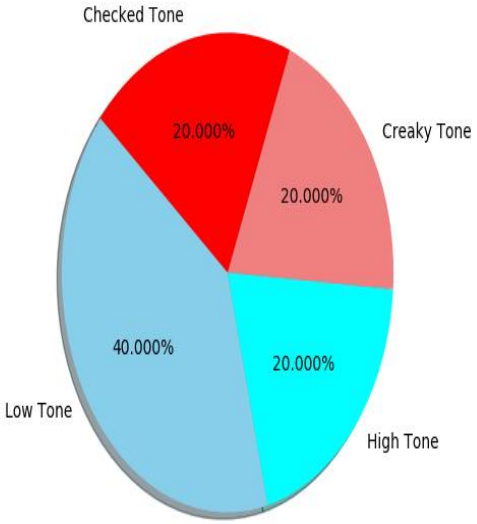

Fig.5. Percentage of four tones in neutral speech.

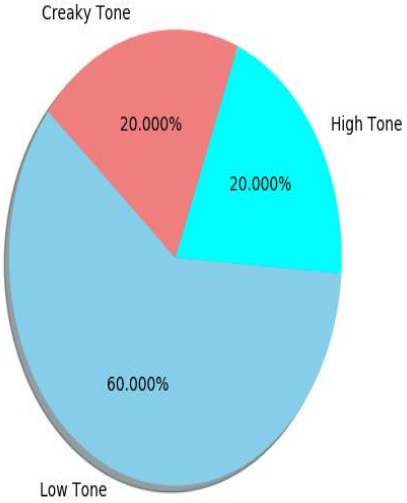

Fig 6. Percentage of four tones in surprise speech.

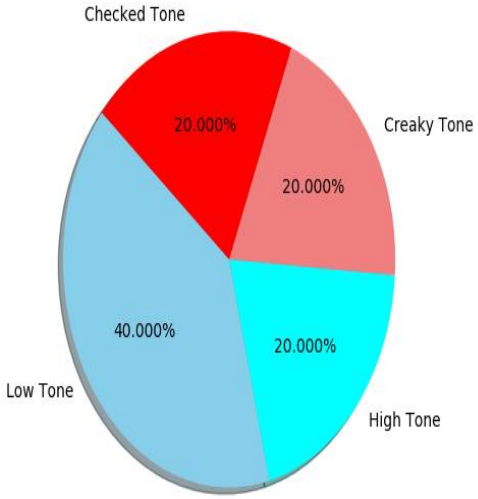

Fig.7. Percentage of four tones in sad speech.

In Fig.8, pitch and intensity of four tones of Burmese language are compared. Pitch of checked tone is the highest among all. Low tone has the highest intensity among all tones. Fig. 9 shows the comparison of formant of four tones of Burmese language. High tone gets highest formant with mean F5. The pitch, intensity and formant values are obtained from Praat[5].

[5] Praat is a free computer software package for speech analysis in phonetics.

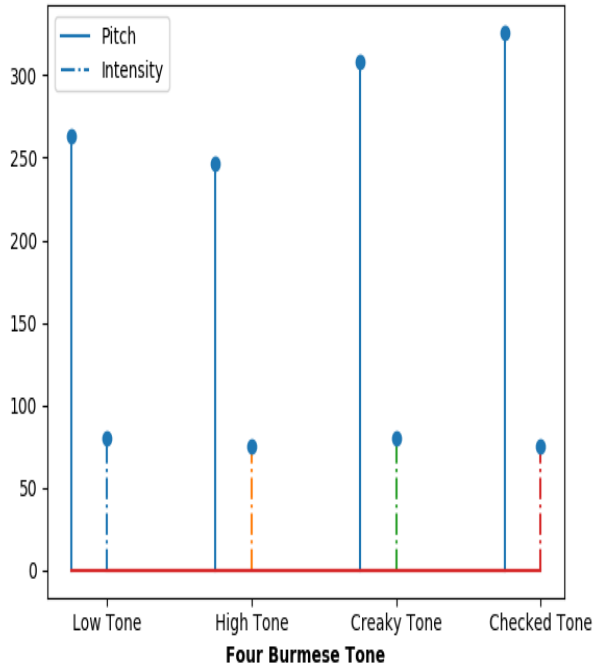

Fig.8. Plotting pitch and intensity of four tones of Burmese

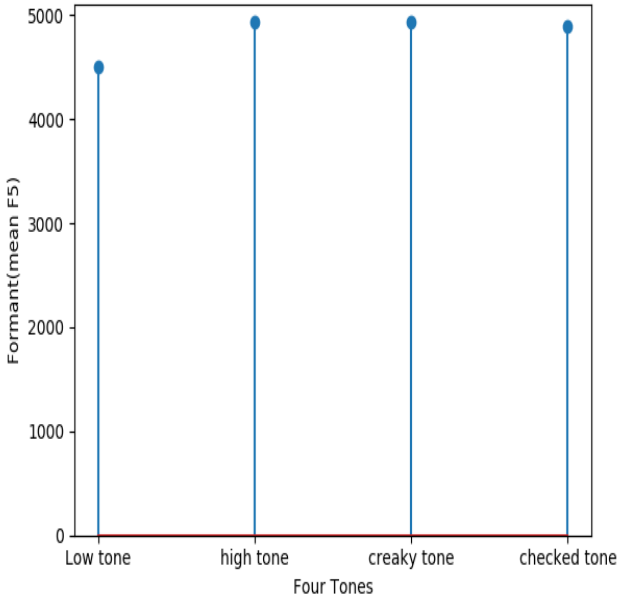

Fig.9.Formant of four tones of Burmese.

VIII. Emotion Classification Experiment

Burmese Movies Interviews Speech Emotion Corpus (BMISEC) has emotion utterances with total duration of 1 hour 56.645 seconds. We conduct emotion classification experiments on our proposed dataset using Level 3 Discrete Wavelet Transform and MFCC as acoustic features, and Multi-layer Perceptron classifier. In experiment, males and females are combined and classified. In this experiment, train dataset is 0.8 and test dataset is 0.2 of total dataset.

As feature extraction, level 3 DWT is used. It uses low-pass and high-pass filters to compute features.

Input speech signal is converted into 39 dimensional MFCC feature vectors.

Multi-layer Perceptron with 1 hidden layer with 10 hidden units is used as classifier.

In Fig.10, class distribution of Burmese dataset is shown. Angry, Happy, Disgust, Fear and Neutral classes have two hundred samples each. Class distribution is done equally. Figure.11 presents comparison of class accuracy across seven emotion classes. Except Angry, Disgust and Sad emotions we achieve classification accuracy of 100% for all other emotion classes.

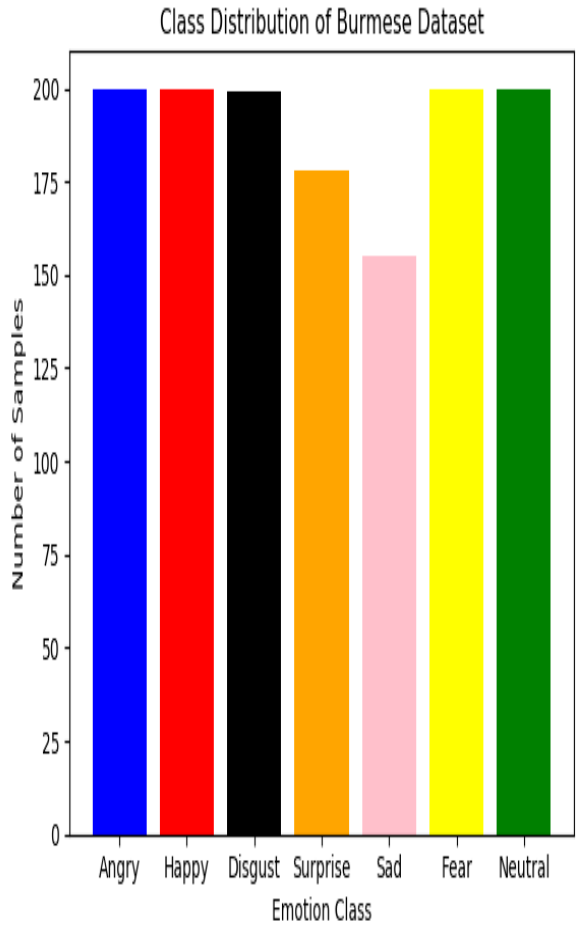

Fig.10. Comparison of class distribution in Burmese dataset

In Table.II, confusion matrix is presented. In Angry class, there are four samples predicted as Surprise. In Disgust, there are three samples wrongly predicted as Angry. In Sad, there is one sample predicted as Neutral. In other classes, there is no wrongly predicted class. In confusion matrix, alphabets such as A, D, H,1, S, N and F to define seven emotion classes.

TABLE II. CONFUSION MATRIX OF CLASSIFYING BURMESE DATASET

	A	**D**	**H**	**1**	**S**	**N**	**F**
A	36	0	0	4	0	0	0
D	3	37	0	0	0	0	0
H	0	0	40	0	0	0	0
1	0	0	0	36	0	0	0
S	0	0	0	0	30	1	0
N	0	0	0	0	0	40	0
F	0	0	0	0	0	0	40

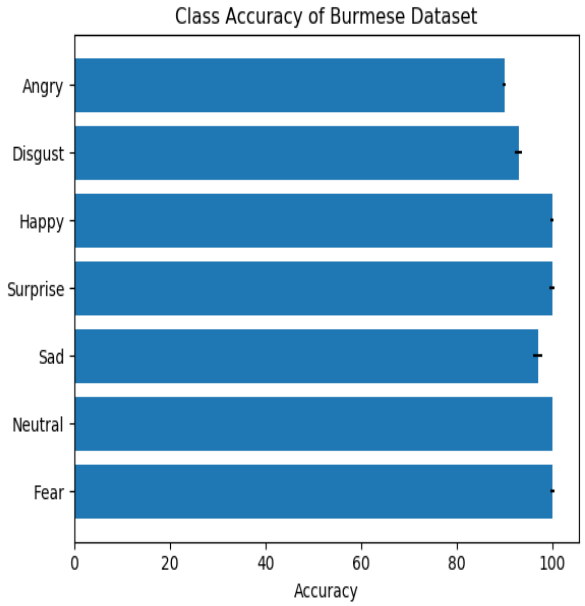

Fig.11. Class Accuracy of classifying Burmese dataset.

IX. LINGUSTIC MATERIAL

This is a list of phrases:

1. ထွက်ခဲ့စမ်း လို့ငါပြောနေတယ်နော် မင်းယောက်ျားဆို ဆင်းလာခဲ့။

   I'm telling you to get out, you guys came down.

2. ဟေ့ကောင် ကျော်ကြီး ထွက်ခဲ့စမ်းလို့ ငါပြောနေတယ်။

   Hey, I'm telling Kyaw Gyi to come out.

3. ပြဿနာရှာတာ ငါမဟုတ်ဘူး မင်း။

   I'm not the one who caused the problem.

4. ဟာ ညီမလေး လာပြီကွ။

   My sister is here.

5. ခုဒီခါပြန်လာတာက နင်အပြီးပြန်လာတာလားဟင်။

   Are you coming back now?

X. CONCLUSION

Speech emotion recognition is a challenging task for machines. And, emotion classification accuracies of machines are far below human level performance. Speech emotion database and corpus are built for different languages. In this study, speech emotion corpus (BMISEC) is built for Burmese language. Important four Burmese tones are studied and their properties are compared. Seven emotion classes are experimented for four Burmese tones. By using level3 Discrete Wavelet Transform and MFCC feature vectors and Multi-layer Perceptron classifier, dataset is classified. Our proposed dataset includes emotion utterances with total duration of 1 hour 56.645 seconds. Confusion matrix for classifying all 7 classes is presented. Accuracies of seven emotion classes are compared using graphical illustrations.

REFERENCES

[1] F.Burkhardt,A. Paeschke, M.Rolfes, W.Sendlmeier, B. Weiss," A Database of German Emotional Speech", Interspeech 2005, Lisbon, Portugal, September 4-8,2005

[2] Omid Mohamad Nezami,Paria Jamshid Lou,Mansoureh Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection", Lang Resources & Evaluation (2019),8 October 2018

[3] Steven R. Livingstone, 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 ",PLOS ONE|https://doi.org/10.1371/journal.pone.0196391May16,2018

[4] Makarova, V., "A database of Russian emotional utterances", in ICSLP 2002. (2002)

[5] Ya Li, Jianhua Tao, Linlin Chao, Wei Bao, Yazhu Liu," CHEAVD: a Chinese natural emotional audio–visual database", J Ambient Intell Human Comput DOI September 10,2016

[6] Giovanni Costantini , Iacopo Iadarola3, Andrea Paoloni, , Massimiliano Todisco, "EMOVOCorpus: an Italian Emotional Speech Database", Proceedings of the Ninth International Conference on Launguage Resources and Evaluation (LREC'14),May,2014,page 3501-3504

[7] https://en.wikipedia.org/wiki/Burmese_language#Tones

[8] Mimi Tiana, Albert Lee," Burmese Quotation Intonation",The 19th International Congress of Phonetic Sciences (ICPhS 2019) At Melbourne

[9] Justin Watkins," Tone and Intonation in Burmese",15thICPhS Barcelona

[10] Daniel Bowling1 , Bruno Gingras1 , Shui'er Han2 , Janani Sundararajan3 , Emma Opitz, "Tone of voice in emotional expression and its implications for the affective character of musical mode", Journal of Interdisciplinary Music Studies, May 19,2014

[11] Z. Esmaileyan , H. Marvi, "A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation", International Journal of Engineering, Vol 27,No.1. January,2014, pp 79-90

[12] Samuel Kim, Thomas Eriksson , Hong-Goo Kang, Dae Hee Youn, " A PITCH SYNCHRONOUS FEATURE EXTRACTION METHOD FOR SPEAKER RECOGNITION", ICASSP 2004

[13] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii," CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation", arXiv: 1802.05630v2[CS.SD] 11 Sept 2018

[14] Sudarsana Reddy Kadiri1, P. Gangamohan2, V.K. Mittal3and B. Yegnanarayana," Naturalistic Audio-Visual Emotion Database", 11th Intl Conference on Natural Language Processing, December 2014, pages 206-213

[15] P. Vijayalakshmi A. Anny Leema," Speech Emotion Recognition Using Support Vector Machine", International Journal of System and Software Engineering, Volume 2 Issue 1 June 2014

[16] Yi Luo, Zhuo Chen, Takuya Yoshika, "Dual-Path RNN: Efficient Long Sequence Modeling for Time-domain Single-Channel Speech Separation", arXiv :1910.06379v1[eess.AS] 14 Oct 2019

[17] Ryan M. Corey, Naoki Tsuda, and Andrew C. Singer, "Acoustic Impulse Responses for wearable Audio Devices", arXiv: 1903.02094v [eess.AS] 5 Mar 2019

[18] Leena Mary, Anish Babu K. K,Aju Joseph, Ginbin M.George, "Evaluation of Mimicked Speech using Prosodic Features", https://www.researchgate.net/publication/261345005

[19] Dorra Gargouri ,Med Ali Kammoun and Ahmed Ben Hamida, "A comparative study of formant frequencies estimation techniques", Proceedings of the 5th WSEAS International Conference on on Signal Processing,Istanbul ,Turkey ,May 27-29,2006 (pp15-19)