ANALYSING VOCAL PATTERNS TO DETERMINE EMOTIONS USING LSTM

[1]Dr. Suprava Patnaik, [2]Ritika Kaushal, [3]Shivani Kadam, [4]Dipti Kathayat, [5]Vaibhav Nardekar

[1] Head of Department, [2]Student, [3]Student, [4]Student, [5]Student
[1] Department of Electronics and Telecommunication,
[1] Xavier Institute of Engineering, Mumbai, India

Abstract: A speech signal contains multitude of emotions. A prodigious amount of research has been done in emotion recognition using speech in the recent years. This paper aims at providing a comparative study of two different database, emphasizing on the importance of a larger database. Ryerson Database and Berlin Database are the two acted speech corpses taken into consideration with 24 actors and 10 actors, respectively. MFCC have been extracted and LSTM is used for classification. 'Happy', 'Sad', 'Angry', 'Fearful' and 'Calm' are the emotions that have been taken into consideration. The results reveal the impact of more number of utterances i.e. a larger database aids in acquiring higher accuracy.

IndexTerms **– Berlin, LSTM, MFCC, Ryerson**

I. INTRODUCTION

Speech is a complex quasi-stationary signal that contains ample amount of information about the message, emotion, language, state of mind and so on. One of the fastest and the most effective methods in natural communication between human beings is the speech signal. A major challenge faced by the researchers in the domain of Human-computer interaction and machine learning is to find ways to make a machine or a network emotionally intelligent; many works have been done in this respect. Emotion recognition through speech has a multitude of applications some of which are, a system developed by Nakatsu et al. 2000[12] that showcased a concept of Interactive move, similar to Charles et al. 2009[13] with storytelling as well as an E-tutoring system developed by Ververidis and Kotropoulos 2006[14]. In recent years, with the development of machine learning techniques and deep learning, researchers have been successfully able to do some meaningful analysis of speech for emotion recognition. The human voice consists of sounds produced by a human being using the vocal folds for carrying out acoustic activities such as talking, singing, etc.; the human voice frequency is specifically a part of the human sound production mechanism in which the vocal cords or vocal folds are the primary source of generating sounds [6].

In this paper, along with the literature review, various extraction techniques available and used, various classifiers that are available and have been used are explained. The result along with the future scope has been explained in the last section.

II. LITERATURE REVIEW

Table 1 Literature Review

Paper	Types of Features	Classifier	Database	Emotions	Recognition rate
1. Daniel Neiberg, Kjell Elenius and Kornel Laskowski	MFFC, MFCC-low	GMM	Swedish voice controlled telephone services, English meetings	Neutral, positive, negative	Maximum 85%
2. Akshay S. Utane Dr. S.L.Nalbalwar	Pitch, energy, MFCC	GMM, SVM	Own database	Angry, happy, sad, surprise and neutral.	76% for neutral, 72.49% for angry
3. Ghai,M. Lal, S., Duggal, S., and Manik, S.	MFCC, energy	Random Decision Forest, SVM, Gradient Boosting	Berlin Database	Anger, boredom, disgust, fear, happiness, sad &neutral	Maximum 81.05 % for Random Decision Forest classifier
4. Luengo I, Navas E, Herna´ez I, Sanchez J	LFPC, Prosodic features	GMM, SVM	Basque database	Anger, fear, surprise, disgust, joy and sadness	98.4% with GMM, 92.3%. with SVM
5. Altun H, Polat G	MFCC, LPC, Prosodic and Energy	Multi class classifiers R2W2, LSBOUND, MUTINF	Berlin Emotional Speech Database-EmoDB.	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Average accuracy with R2W2 is 81%, MUTINF is 78%
6. J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen	Jitter, shimmer, LPCC, MFCC, LFPC	k-NN classifier	Private corpus (Mandarin speech)	Angry, Happy, sad, neutral, bored	Best 86%

III. DATABASE

Performance of vocal emotion recognition is totally dependent on the quality of the database generally research deals with database of acted by actor, induced or completely spontaneous emotions the complexity of the system increases with the naturalness, each database consist of corpus of human speech pronunciation under different emotional conditions with respect to authenticity there seems to be three types of databases [10]. Figure 1 depicts the types of database and their complexities.

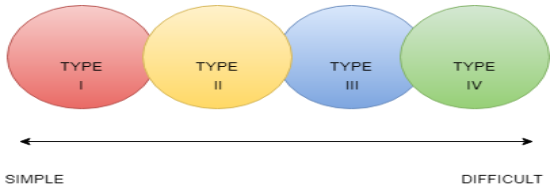

Fig.1 Database types and their complexities

Type I- Acted
Type II- Real Life Systems
Type III- Elicited
Type IV- Real life database

The first used database is 'The Ryerson Audio-Visual Database' (RAVDESS). It is an acted, validated, multimodal database of emotional speech. The database is comprised of 24 actors with 12 female and 12 male actors. The language is English with a North American accent. The database takes into consideration 8 emotions (neutral, calm, happy, sad, angry, fearful, surprise, disgust). Each emotional sample is recorded at two levels of intensity (normal and strong) other than 'neutral' with only one level of intensity. The bifurcation of the naming of each voice signal in this database has been represented below.

D[0]:Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
D[1]:Vocal channel (01 = speech, 02 = song).
D[2]: Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
D[3]: Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
D[4]:Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
D[5]: Repetition (01 = 1st repetition, 02 = 2nd repetition).
D[6]: Actor 01 to 24. G= Odd numbered actors are male, even numbered actors are female.

Each actor has 4 neutral utterances, with a total of 24 actors, the total utterances for the Neutral emotion is 96. For the other specified emotions, each actor has 9 utterances, hence with a total of 24 actors, the total utterances for each emotion is 192. The total number of utterances for the Ryerson Emotion Speech Corpus is 1440.

The second database used is 'The Berlin Database of Emotional Speech'. It is an acted emotion speech corpus. The language of this dataset is German. It consists of recordings from 10 actors with 5 male and 5 female actors. The emotions taken into consideration are anger, boredom, disgust, fear, happiness, sadness and neutral. It consists of 800 sentences. This database is selected because of the quality of its recording and its availability.

Every utterance is named according to the same scheme:
Positions 1-2: number of speaker
Positions 3-5: code for text
Position 6: emotion (sorry, letter stands for German emotion word)
Position 7: if there are more than two versions these are numbered a, b, c and so on.
Example: 03a01Fa.wav is the audio file from Speaker 03 speaking text a01 with the emotion "Freude" (Happiness).

Table 2 Code of emotions:

letter	emotion (english)	letter	emotion (german)
A	anger	W	Ärger (Wut)
B	boredom	L	Langeweile
D	disgust	E	Ekel
F	anxiety/fear	A	Angst
H	happiness	F	Freude
S	sadness	T	Trauer
N = neutral version			

Berlin Database, on the other hand has a total of 10 actors and only 535 utterances taking into consideration all of its specified emotions.

III. FEATURE EXTRACTION TECHNIQUE

4.1 MFCC

Mel Frequency Cepstrum Coefficient (MFCC) is a method of feature extraction of voice signals. MFCC is a representation of the short-term power spectrum of a sound which is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. [8]

In a nutshell, pre-emphasis filter is applied on the signal; then the signal is sliced into (overlapping) frames and a windowing function is applied to each frame. Further, Fourier transform is performed on each frame, calculate the power spectrum; and compute

the filter banks. Discrete Cosine Transform (DCT) is performed and it is applied on the filter banks, to obtain MFCCs, where a number of the resulting coefficients are kept while discarding the rest coefficients. Finally mean normalization is performed. Normalization is a pre-processing technique which retains the emotional content and eliminates speaker and recording variability.

4.1.1 SEQUENCE TO CALCULATE MFCC

- Pre-Emphasis:

The first step is pre-emphasis of the signal. The signal is passed through a pre-emphasis filter for amplification of the high frequencies. In speech processing, pre-emphasis filter is required after the sampling process. The purpose is to obtain a smoother spectral form of speech signal frequency. So, this filtering process is performed to reduce noise during sound capture.

- Frame blocking:

After pre-emphasis, signal is split into short-time frames. In this process, the sound signal is divided into multiple overlapped frames, so that not even single part of signal is deleted. Frame size is kept as 25 ms and a 10 ms stride (15 ms overlap) [9]

- Windowing:

Windowing technique is used to reduce the amplitude of the discontinuities at the end of each sample. After framing, window function such as the Hamming window is applied to each frame. Hamming window is used as it reduces the amplitude of both sides i.e. side lobes, the FFT will be more focused on the middle part of time domain i.e. frequencies in the middle of the signal will have more preference.

- Fourier-Transform and Power Spectrum:

A function with limited period can be expressed in Fourier series. The frame that has undergone the windowing process is now converted into a frequency spectrum. FFT is a fast algorithm of Discrete Fourier Transform (DFT) which is useful for converting every frame to N samples from time domain into frequency domain, where N is typically 256 or 512 and then compute the power spectrum [11]

- Mel-Frequency wrapping/ Filter banks:

The perception of the human ear against the sound frequency is not linear. The scale of Mel-Frequency is a low frequency which is linear under 1000 Hz and a logarithmic high frequency above 1000 Hz.

We can convert between Hertz (f) and Mel (m) using the following equations:

$m = 2595 \log_{10}(1 + f700)$

$f = 700\,(10m/2595 - 1)$

The final step to computing filter banks is applying triangular filters

- Cepstrum (MFCC):

Humans listen to voice information based on time domain signals. At this stage Mel-spectrum will be converted into time domain by using Discrete Cosine Transform (DCT). The result is called Mel-frequency cepstrum coefficient (MFCC). Selecting an appropriate set of features is of key importance to obtain high accuracy in a recognition system.

- Mean Normalization:

To balance the spectrum and improve the Signal-to-Noise (SNR), mean of each coefficient from all frames is simply subtracted. [11]

V. MACHINE LEARNING APPROACH FOR EMOTION RECOGNITION FROM SPEECH

5.1 WHAT IS RNN?

A vanilla feed forward network is a simple input-output network where input can be any image or vector that is fed and passed through a set of hidden layers where the number of hidden layers is based on the type of application and a single output is obtained after. Although, with the development in Machine learning, it was realized that more flexibility was required in context to the input and output that the network models can process. The basic structure of an RNN is where input is fed to an RNN core cell, its singular hidden state gets updated every-time and the output is hence obtained using the hidden state of the previous cell. Figure. 2 represents the Vanilla RNN structure

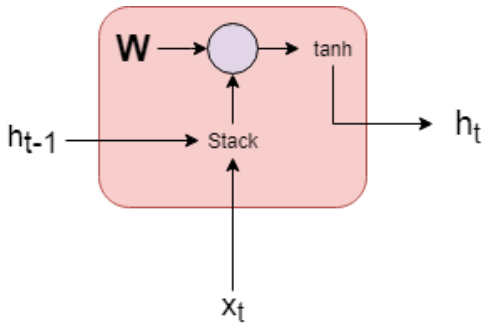

Fig.2 Vanilla RNN Structure

The hidden state of the cell is a function of the previous state and the input vector.

$ht = fw(ht\text{-}1, xt)$

Where, ht = new state

fw = some function with parameter 'w'

ht-1= previous state

xt = input vector

The training process of the represented RNN cells is simply stacking the current input vector on top of the previous state, this is multiplied by the specified weight matrix and squashed through a tan-hyperbolic function.

5.1 LONG SHORT TERM MEMORY

The preferred solution to the Vanishing Gradient problem of RNN is LSTM. The LSTM units categorize the information into long and short term memory cells. Doing so enables RNNs to figure out what data is important and should be remembered and looped back into the network, and what data can be forgotten [7]. LSTM is designed to have a better gradient flow. It has two hidden states as opposed to one from the Vanilla RNN; the first is identical to the Vanilla RNN and the second state could be termed as a Cell state. Figure. 3 represents the LSTM architecture

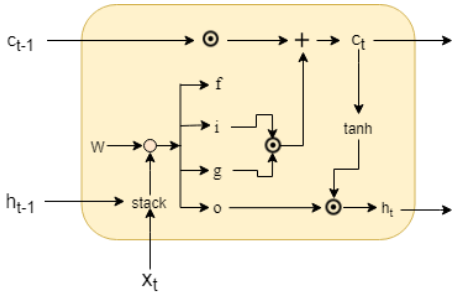

Fig.3 LSTM Architecture

VI. PROPOSED METHOD

Our proposed system consists of 3 parts: Speech acquisition, Feature Extraction and Emotion classification.

In this method MFCC technique for feature extraction is used. Initially, the input is given as .wav file to the network. Pre-processing like dropping the unvoiced part is performed on the input signal. After this, framing of the speech sample is done and windowing technique using Hamming window is applied. Then the frame is converted to frequency domain by applying Fast Fourier Transform and estimation of power spectrum is done. The next step is to calculate the sub-band energies with the help of Mel- filter banks, which is a non-linear filter bank. Log of the filter bank energies is acquired. Finally, with the help of Discrete Cosine Transform (DCT), MFCC is computed. After this computation, an excel file is formed which has coefficients corresponding to respective emotions. This file is further used for the classification purpose.

For classification, LSTM model is implemented. The coefficients are shuffled and are given to the system for training and testing purposes. In training part, the system learns the emotion characteristics of the speakers. The data (file containing features) is fed sequentially through the LSTM network, which has multiple dense layers. Different optimizers such as binary cross entropy for two-class classification, categorical cross entropy for multi-class classification have been used. Also different combinations of output Activation function like sigmoid, softmax, tanh, relu and leaky-relu have been used. The network is further compiled and fit to achieve higher recognition rate. Figure.4 shows the flowchart for the proposed method.

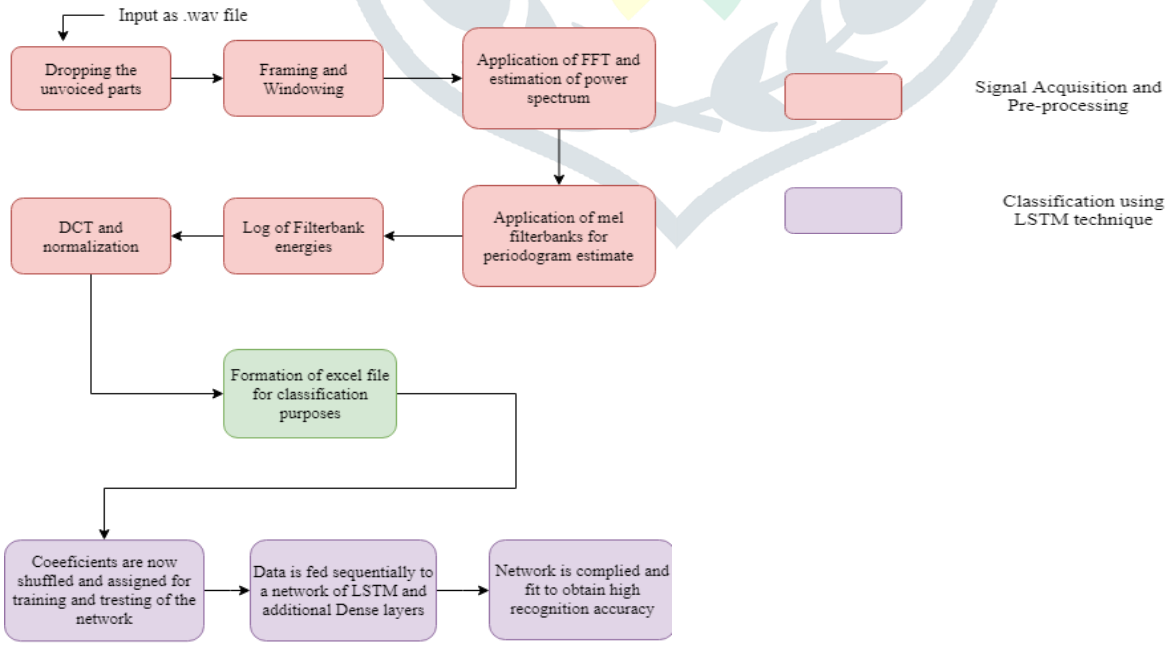

Fig.4 Flowchart for proposed method

VII. RESULT ANALYSIS

Table 3 Comparison of accuracies for Ryerson & Berlin database

Emotion-Combination	Ryerson Database	Berlin Database
Happy-Sad	94.60%	88.76%
Happy-Angry	89.97%	81.45%
Happy-Fearful	97.87%	91.67%
Sad-Angry	93.33%	85.56%
Sad-Fearful	94.45%	89.97%
Fearful-Angry	90.00%	93.22%
Disgust-Neutral	93.76%	84.56%
Happy-Disgust	93.33%	89.97%
Sad-Disgust	96.54%	92.89%
Fearful-Disgust	91.77%	86.92%
Angry-Disgust	91.00%	94.32%

The Table.2 projects the numerical accuracies obtained in selection of two different databases. The Berlin Database is a 10 actor, acted speech corpus whereas the Ryerson Database is a 24 actor, acted database; the number of sentences per actor, with different intensities of the said emotion is higher in case of the Ryerson Database. It is observed the bigger database aids in providing a higher accuracy.

Figure.5 represents the graph for the accuracies obtained for various combinations of emotions for Ryerson as well as Berlin database. It is observed that the Ryerson database obtains a higher accuracy as compared to Berlin database for most of the emotion combinations.

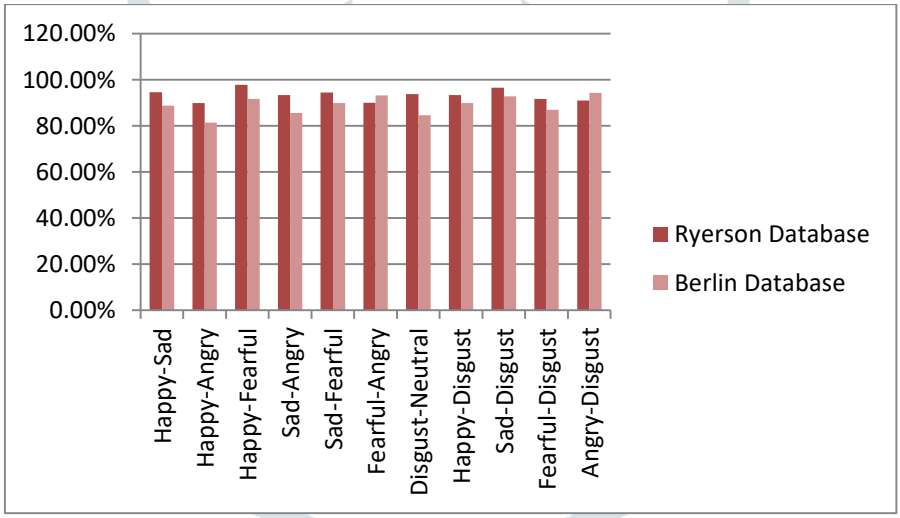

Fig.5 Comparison of accuracy for two database

VIII. CONCLUSION

The effect of a larger database on the acquired accuracy has been represented in the results explained in the above section. A larger database i.e. hours of speech helps in appropriate training and validation of the created network. Multiple combinations of various emotions have been used for rigorous analysis of the data-sets and hence its impact on the final accuracy or the recognition rate.

REFERENCES

[1] Daniel Neiberg, Kjell Elenius and Kornel Laskowski," Emotion Recognition in Spontaneous Speech Using GMMs"

[2] Akshay S. Utane Dr. S.L.Nalbalwar, " Emotion Recognition Through Speech Using Gaussian Mixture Model And Hidden Markov Model" in International Journal of Advanced Research in Computer Science and Software Engineering, April 2013

[3] https://sci-hub.tw/10.1109/ICBDACI.2017.8070805

[4] Luengo I, Navas E, Herna´ez I, Sanchez J (2005) Automatic emotion recognition using prosodic parameters. In: The proceedings of Inter speech, pp 493–496

[5] Altun H, Polat G (2009) Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst Appl 36:8197–8203

[6] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, and Y.-T. Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011.

[7] https://searchenterpriseai.techtarget.com/definition/recurrent-neural-networks

[8] github.com/Sangramsingkayte/MFCC_MATLAB

[9] Mel Frequency Cepstral Coefficient (MFCC) tutorial- http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[10] Burkhardt F., Ajmera J., Englert R., Stegmann J., Burleson W. Detecting anger in automated voice portal dialogs. Proc. INTERSPEECH'2006, Pittsburgh, 2006.

[11] Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between- https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[12] Nakatsu, R., Nicholson, J., & Tosa, N. (2000). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. Knowledge-Based Systems, 13, 497–504.

[13] Charles, F., Pizzi, D., Cavazza, M., Vogt, T., & Andr, E. (2009). Emoemma: Emotional speech input for interactive story telling. In Decker, Sichman, Sierra, & Castelfranchi (Eds.), 8th int. conf. on autonomous agents and multiagent systems (AAMAS 2009), Budapest, Hungary, May 2009 (pp. 1381–1382).

[14] Ververidis, D., & Kotropoulos, C. (2006). A state of the art review on emotional speech databases. In Eleventh Australasian international conference on speech science and technology, Auckland, New Zealand, Dec. 2006.