NATIVE AND NON-NATIVE EMOTIONAL SPEECH RECOGNITION FOR MARATHI LANGUAGE

Abstract

One of the most important subdomains of emotional computing is emotion recognition from speech. When a person delivers a sentence while experiencing a strong emotion, the tone of speech can radically shift the meaning of the sentence. Speech signals include emotions as well as words and meaning. The low recognition accuracy achieved during the development of speech-based systems is mostly due to the emotion communicated by speech. Emotions influence a person's tone and speaking style when it comes to human speech. To solve these challenges of emotion identification from speech, more study is needed in this field. However, the problem usually revolves around two main emotion categories: happy, sad, And Angry. Afraid, Surprised, and Uncertain.

Exploring basic spectral properties such as MFCCs and LPC derived from complete voice utterances for speech emotion identification using a block processing approach. Extracting static and dynamic aspects of prosodic contours, such as for emotion classification, consider energy and pitch. We have been discovered as promising models for constructing emotion detection systems that identify emotions based on prosodic properties.

Keywords: Speech Emotion Recognition, Feature extraction, MFCC, LPC, Native, Non-Native

Authors

Bharati D. Borade (Student)
Department of Computer Science & IT
Dr. B.A.M.University
Aurangabad, India.
Bharatiborade3@gmail.com

Ratnadeep R. Deshmukh
Department of Computer Science & IT
Dr. B.A.M.University
Aurangabad, India.
rrdeshmukh.csit@bamu.ac.in

I. INTRODUCTION

Speech is vocalized form of communication used by humans, which is based upon the syntactic combination of items drawn from the lexicon. The vocal abilities enables humans to produce speech. Speech consists various features like emotion, loudness, tempo, and rhythm, which provides us lot of meaningful information about speakers [1]. In the research recognize the different emotion of Native and non-Native for Marathi language [2][3]. The basic requirement of data is text which would recorded from various Marathi and non-Marathi speakers. The data will one sentence or one word that show such type of emotion.[4] The database consist the speakers speech collection in the three type of emotion i.e Happy, Sad and Angry. In the research we have analyzed the emotion of native and non native speakers how they react their emotion.[5]

- **Native:** A Native speakers is someone who learned to speak a language as part of his or her childhood development. A Native speakers language is usually their parent or county.[6]

- **Non-Native :** Non-Native speakers of a language on the other hand are people who have learned this particular language as second or third language.

- **Emotion:** Emotion is often intertwined with mood, temperament, personality, disposition, and motivation. In the research we analyzed the three type of emotion Happy, Sad, and Angry with the Native and Non-native speakers of Marathi language.[7]

1. **Speech Emotions:** Emotion recognition systems focuses on modelling of spectral as well as the prosodic features such as formants, pitch, loudness, timbre, speech rate and pauses which contains the linguistic and semantic information [8]. However the problem usually deals with the following basic emotion categories: Happy, Sad, Angry, Afraid, Surprise, Neutral.

   The emotion expressed by speech is one of the major influencing factors for the low recognition accuracy achieved during the development of speech based systems. When it comes to speech human emotions affects the tone and the speaking style of the person. The research in this area is needed to overcome these problems of emotion recognition from speech [9].

   Speech emotion is one of the important things in human expression. A number of definitions of emotions have been proposed from 1884 when William James first tried to define or give the answer of it. The emotion has been defined as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism". The natural emotion means the emotion which are expressed spontaneously when a series of event occur to which the brain responses accordingly. The artificial or acted emotions means to mimic the natural emotions which are similar to those emotions expressed without the occurrence of events to which the brain responses spontaneously.

There are many arguments over the selection of natural or real emotion and acted or artificial emotion. The research required emotional speech but due to the above mentioned problem we were unable to capture the natural emotion so we developed databases of acted or artificial emotions which are mimicked. The artificial emotional Marathi speech database was developed and performed the experiment for emotion recognition on the developed database [10, 11].

2. **Artificial Emotion:** The artificial emotions are the emotions which are not real but they mimicked or acted by an individual. Natural emotions are responses to the internal or external stimulus to the brain. Any person will react to an event so it becomes the natural emotion. The problem is that it's not at all possible to capture the natural emotions so for the study the best source was the acted or artificial emotions which are mimicked. Everyone cannot mimic the emotion accurately however, the professional actors and actress can mimic the emotions to most probable real emotion [12].

   The researchers have tried to recognise the emotions from movies, TV shows, plays and other sources where a person mimics the emotions. In movies, TV Shows, Plays and other sources the actors and actress performs the act and the complete the emotional tasks so it gets easy to carry out the emotion recognition process on such type of data.

3. **Natural Emotion:** The human speech contains the information and some meaning which relates the emotion which contains not only the linguistic content but also contains some emotions of the speaker even though the emotion does not alter the linguistic content. There are many arguments over the selection of natural or real emotion. To capture the natural emotions it is very difficult as these emotions are responses to the internal or external stimulus received by the brain. No one can predict how the brain of different person will react to an event so it becomes difficult to capture the natural emotions and their classification. We did require emotions but due to the above mentioned problem we were unable to capture the natural emotions.

4. **Real Life Emotions:** In real life human can express their emotion by many ways such as by facial expression, by yelling, by touch. Speech is one of the important outcomes of the emotional state of human beings. A speech signal is produced from the contribution of the vocal tract system excited by excitation source signal.

5. **Whispered Emotional Speech:** The whispered speech can also carry emotional information like prosodic features, including short time energy and speak rate, voice quality parameters, formant, and spectrum to analyze the differences between emotions. One could perceive others feelings at that moment when someone whispered. Nowadays, with the widespread of the cellular phone, people whisper so as to reduce the amount of speech being spell out; to public safety, whispered speech are often encountered for criminal analysis; and to laryngectomees, whisper is the only means of articulation. With the help of acoustic features like endpoint detecting, abstraction of formant frequencies and the corresponding bandwidths the whispered speech signal can be measured. [13]

6. **Mood Extraction:** Mood extraction from speech is one of the difficult task. The Sentiment analysis (SA) plays a vital role in natural language processing. The sentimental analysis can be done with the help of tasks to classify the different moods such as Happy,

Sad, Frustrated, Angry, Depressed, Temper etc for domain-specific sentence level mood extraction [14].

II. MARATHI LANGUAGE

Marathi is the southern branch of the Indo-Aryan language group. It is mainly spoken by the Maharashtra people of Western India, and since 1966 is the state's official language and co-official language of Goa state. Marathi was also called Maharashtri, Marhatti, Mahratti, etc. during prehistoric times. According to the 2011 census, India had 83 million Marathi speakers, making it the third most commonly spoken mother tongue after Hindi and Bengali, similar to the fifteenth most spoken in the world. It is the oldest of the Indo-Aryan regional kinds of literature. The language consists of some of the oldest literature from around 600 A.D. in all modern Indian languages. Marathi has developed out of Sanskrit, which ultimately originated from Prakrit and Apabhramsha, and is expected to be over 1300 years old. It is said that its grammar and syntax came from Prakrit and Pali. The Marathi we hear today is the product of the democratic cycle of reform and development over the years. In Israel as well as Mauritius, we can find Marathi speakers. The language of Marathi consists of approximately 42 dialects, of which Tamil and Kannada loans greatly influenced the dialect used in Thanjavur and Tamil Nadu districts. Marathi is closely connected to languages like Konkani, Goanese, Deccan, Gowlan, Ihrani, andVarhadi-Nagpuri.[15,16].

III. RELATED WORK

Table 1: Related work for Non Native speakers

Language	Speakers	Utterances	Duration	Specials
English	96	15000	-	Proficiency rating
English, French, German,Italian,Crech,Dutch	161	72000	133 h	City Names
NATO M-ATC	36	622	9833	17 h
Marathi	100	3000	-	Numeric
English	200	68000	-	Proficiency rating

While the emotional work for non-native speakers is still unfinished, most of the linguistic work has already been completed for native speakers, as illustrated below,

Language	Speakers	Types of database
English	22 patient and 19 healthy persons	Simulated
German	51 School children (21M+30F)	Elicited
Spanish	8 Actors (4M+4F)	Simulated
Russian	61 Native speakers	Simulated

IV. METHODOLOGY

The following fig 1 shows basic workflow of work

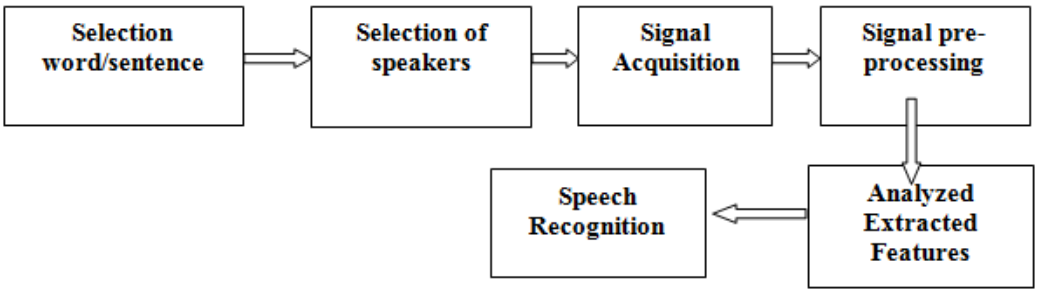


Figure 1: Methodology for Emotion Recognition

The process work is divided into several steps; we begin with the selection of text i.e word sample we have collect from the Native and Non-native speakers.[17] Then we select the emotion speak of Native and Non-native speakers. Once the speech samples are collected we will be performing pre-processing. The final step is to analyze the extracted features of speech of Native and Non-native speakers of Marathi. The final step is to recognize the speech of native and non-native speakers for Marathi language. [18]

V. DESIGN AND DEVELOPMENT

1. Acquisition Environment, Speaker and Instrument Setup

- The speech data has been collected from individuals' belonging to two districts of Marathwada region i.e. Aurangabad.
- Microphones was randomly selected i.e. Quantum High-Tech to experience how it will be work.
- The Microphone was approximately 5 cm from the mouth of the speaker. Each speaker was requested to speak the word from
- developed text corpus. Three utterances of each word. The speech samples recorded and the recorded speech file was stored in.wav format with PRAAT software. Annotation of speech sample is done with PRAAT. Linguistic Data Consortium for
- Indian Languages (LDC-IL) Recording standards are followed during the speech sample collection.

2. Developed Speech Database

Table 2: Isolated Word Speech Database for Native Speakers:

Word Count	Frequency Standard	Native Speakers	
		Gender	Age
24 words	16000 Hz, 16-bit mono	8 M, 6 F	20-30 year
Total		14	
		1008utterances	

3. Developed Speech Database

Table 3: Isolated Word Speech Database for Non-Native Speakers

Word Count	Frequency Standard	Non-Native Speakers	
		Gender	Age
24 words	16000 Hz, 16-bit mono	6 M, 4 F	20-30 year
Total		10	
		720 utterances	

Following figure1 (a,c,e,g) shows the spectrogram with noisy speech sample and also fig 1. (b,d,f,h) shows the spectrogram with noise free speech sample of Native and Non-Native Male female .

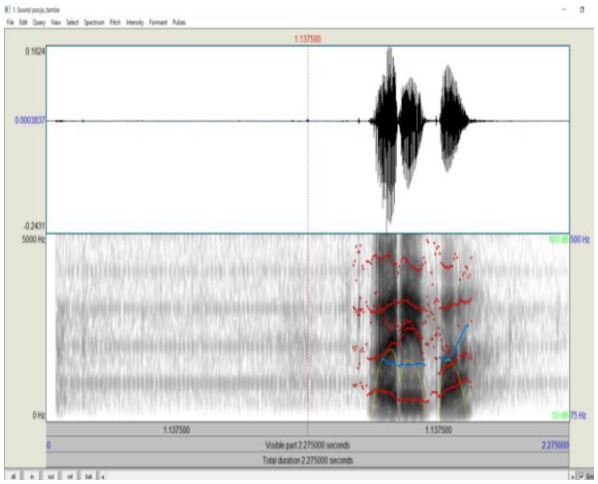


Figure 1 (a): (a)Native female speakers voice sample With Noise

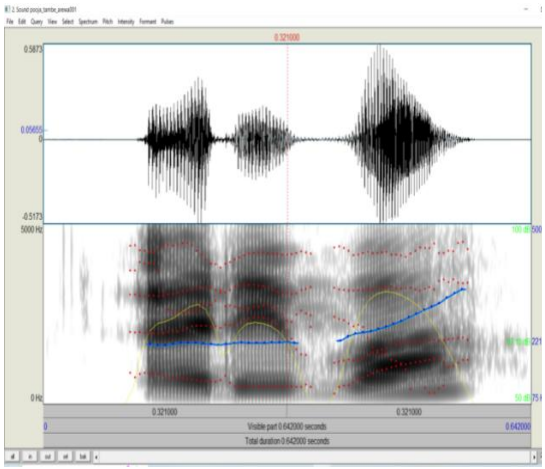


Figure 1 (b) : Native female speakers voice Sample Without Noise

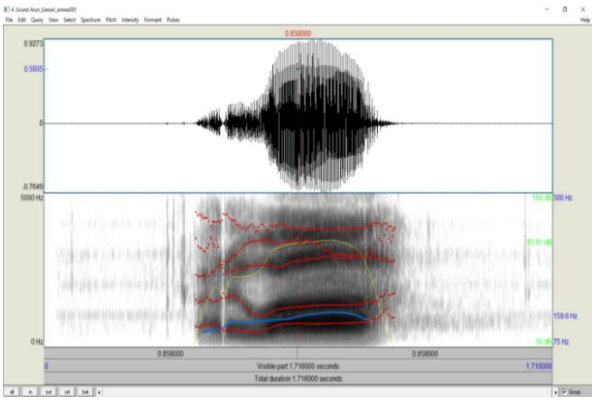


Figure 1 (c): Native male speakers voice sample

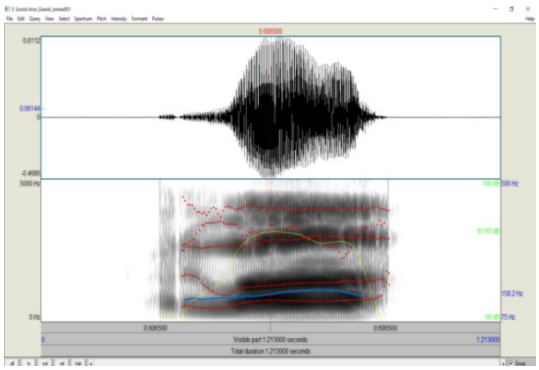


Figure 1 (d): Native male speakers voice Without Noise

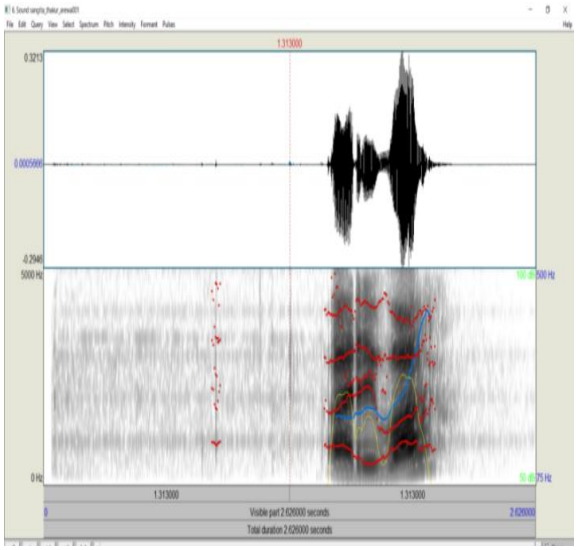


Figure 1 (e): Non-Native female speakers voice sample With Noise

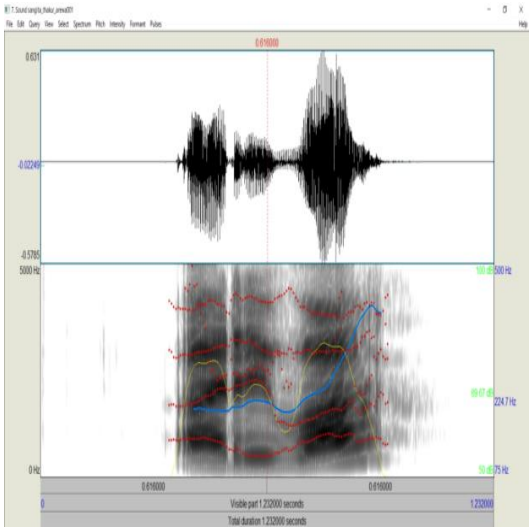


Figure 1 (e) : Non-Native female speakers voice sample Without Noise

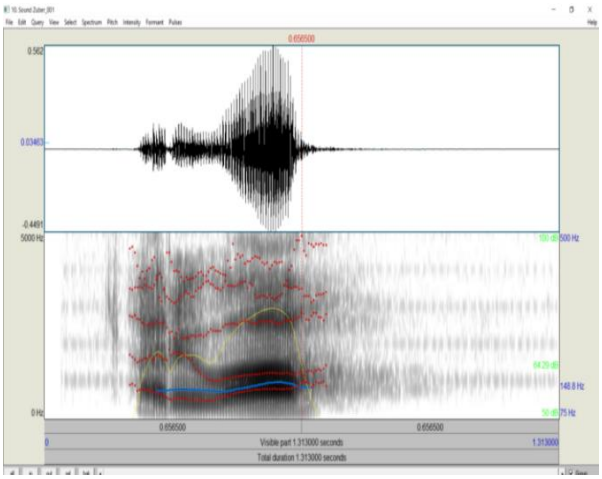


Figure 1 (g) : Non-Native male speakers voice sample With Noise

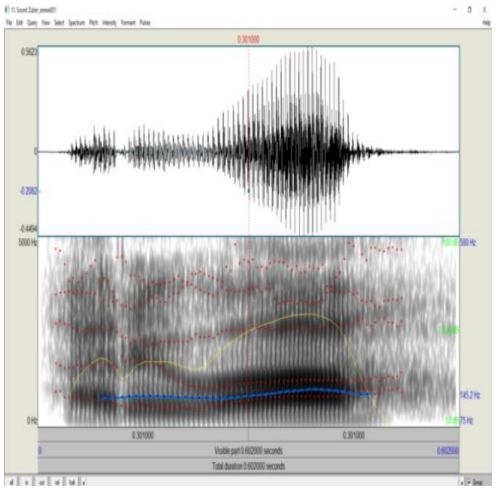


Figure 1 (h): Non-Native male speakers voice sample Without Noise

VI. PERFORMANCE ANALYSIS

1. **Feature Extraction:** In feature extraction, the signal's essential information is kept while the unnecessary and redundant information is removed. It may also entail translating the signal into a format suitable for models used in classification. The basic objective is to identify a set of utterance characteristics that have acoustic correlates in the speech signal, in order to compute or estimate the parameters through waveform processing. In this experiment, the features were extracted using MFCC and LPC.

- **Mel Frequency Cepstral Coefficient (MFCC):** The number of filters included in the Filter bank for the MFCC by the associated author is defined by the FB in the implementations. These implementations take various sample rates into account. Pre-emphasizing, Framing and Windowing, Fast Fourier Transform, Mel-Frequency Filter Bank, Logarithm, and Discrete Cosine Transform are the stages that are taken to compute the features using MFCC.

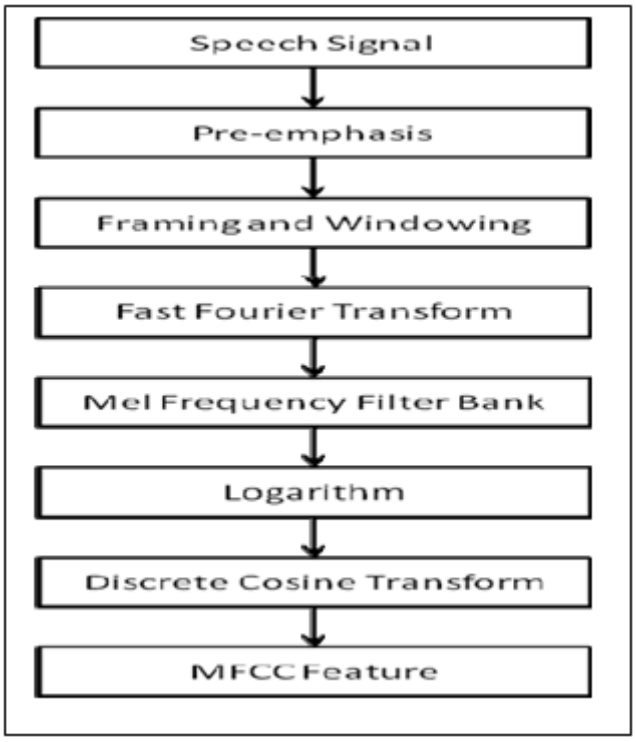


Figure 2: Block diagram of MFCC Feature Extraction method

- **LPC-based Speech Emotion Recognition:** After the creation of emotional speech databases in the Marathi language, the experiment was conducted. Linear Predictive Coding (LPC) was employed in this work for the feature extraction process. Features of Linear Predictive Coding (LPC) transmit specific emotional information.

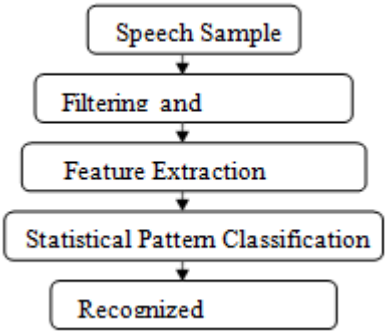


Figure 3: Block Diagram of Emotion Recognition System Using LPC

2. **Confusion Matrix:** Information about the actual and anticipated categorization performed by the classification system is contained in a confusion matrix. The data in the matrix is frequently used to evaluate the performance of such systems. The confusion matrix for the class classifier is displayed in the following table.

According to four studies, the entries in the confusion matrix mean the following:

- If an TN is the number of correctly predicted events, then a case is unfavourable.
- FP is the number of times a positive case was predicted incorrectly.
- FN is the number of false predictions that a negative instance and
- TP is the number of correctly predicted instances that are positive.

Table 4: Table Entries in confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Table 5: Confusion matrix of MFCC

	F_ Angry	**F_ Happy**	**F_Sad**	**M_ Angry**	**M_ Happy**	**M_ Sad**	**Total no of sample**	**Recognition Rate**
F_Angry	91	21	2	2	0	19	135	67.40
F_Happy	11	92	0	7	3	17	130	70.76
F_Sad	10	12	87	9	12	13	140	62.14
M_Angry	21	2	11	95	4	12	145	65.51
M_Happy	13	4	2	1	98	12	130	75.38
M_Sad	4	22	1	1	12	91	131	69.46

Table 6: Confusion matrix of LPC

	F_Angry	F_Happy	F_Sad	M_Angry	M_Happy	M_Sad	Total no of sample	Recognition Rate
F_Angry	89	21	3	13	7	0	133	66.91
F_Happy	9	87	2	12	29	0	139	62.58
F_Sad	8	12	92	21	22	0	152	60.52
M_Angry	11	2	11	87	25	0	136	63.97
M_Happy	6	4	12	10	92	0	124	92.80
M_Sad	12	2	0	2	12	88	116	75.86

Table 7: Result based on MFCC or LPC

Feature Extraction	Native		Non-Native	
	Female	Male	Female	Male
MFCC	70.3	75.5	62.7	62.7
LPC	75.7	67.7	61.7	73.7

Speech features can be divided into two categories: prosodic features and phonetic features.The prosodic qualities pertain to the musical aspects of speech, such as rising or falling tones, accents, or stresses, while the phonetic features are primarily concerned with the sorts of sounds included in speech, such as vowels and consonants, and their pronunciation. The key components for speech emotion prosody are the fundamental frequency, duration, and energy qualities. Information concerning intonation, accent, and rhythm is conveyed via prosodic qualities. The intensity contours and a linear approximation of F0 serve as the foundation for the prosodic features.

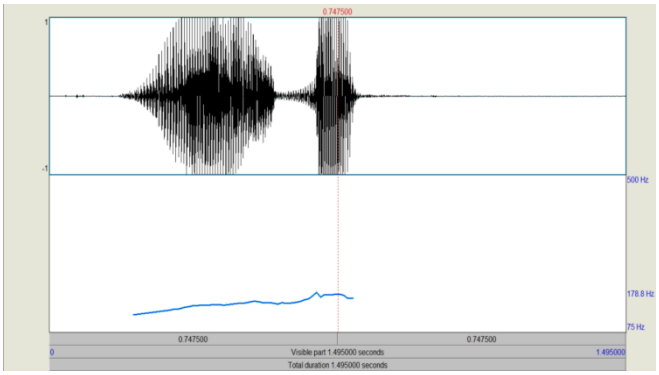
Table 8: Confusion matrix for Prosody Features

	Angry	**Happy**	**Sad**	**Total number of samples**
Angry	27	16	10	53
Happy	16	21	10	47
Sad	19	10	21	50
Total				150

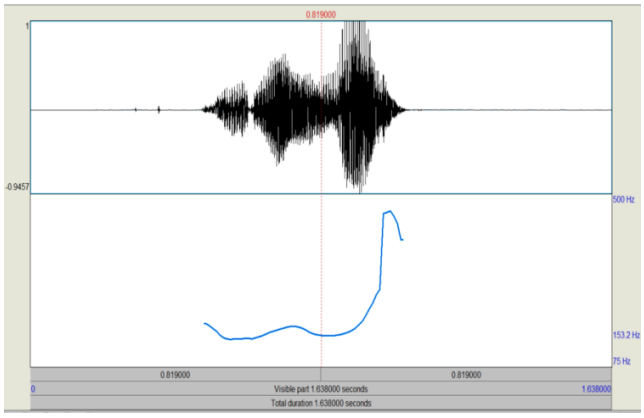
Using speech features to extract emotions

3. Prosodic Feature

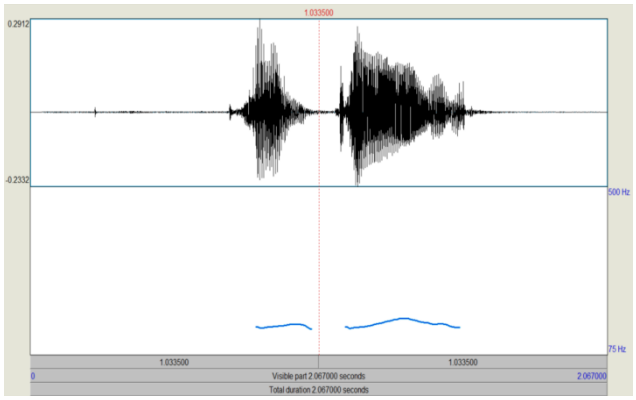
Pitch: In this study, we showed how pitch analysis can be useful for speech recognition tasks, such as identifying emotions in voice signals. The most prevalent prosodic property is the fundamental frequency. Pitch and intonation have been linked to a variety of speech functions. The experiment's objective was to develop a trustworthy automated speaker relative pitch estimate system for speech emotion recognition.
Below is a representation of the retrieved pitch contour from the voiced portions of the utterance.



A. Average pitch calculated speech sample for Angry

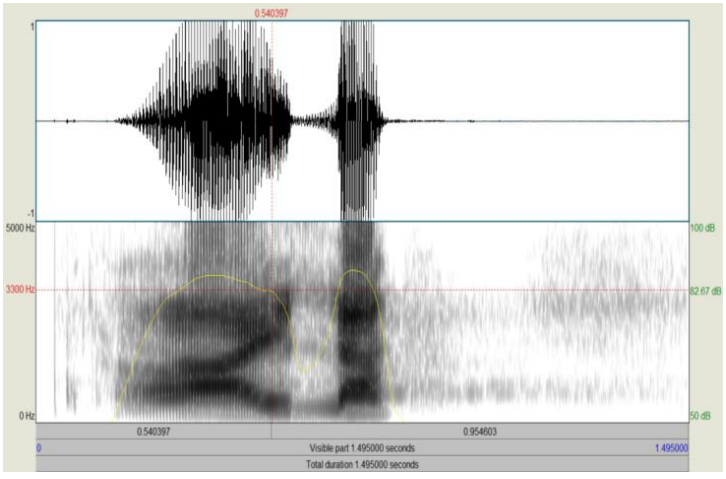


B. Average pitch calculated speech sample for Happy

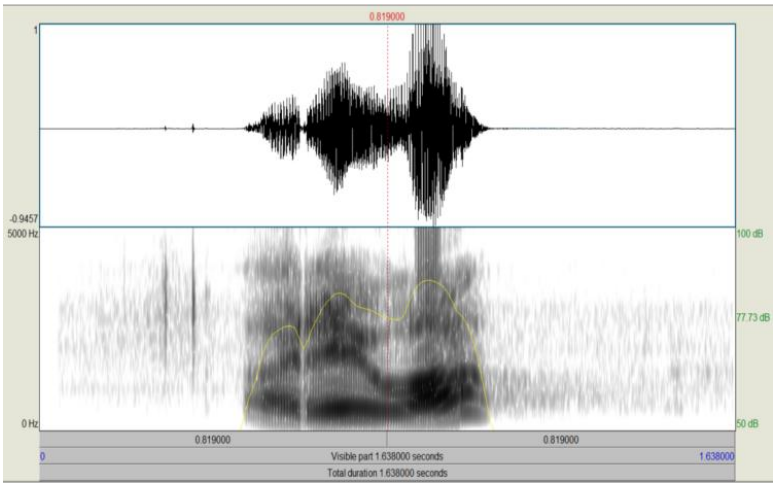


C. Average pitch calculated speech sample for Sad

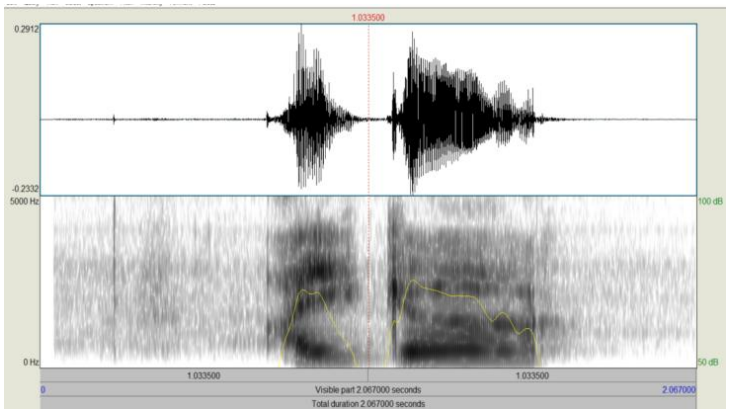
Energy: We calculated the energy value using the first derivatives of the smoothed speech signal rather than the absolute signal amplitude in order to lessen the effect of loudness.It was possible to retrieve data on the energy statistic's mean, minimum, maximum, and standard deviation.



D. Energy calculated speech sample for Angry



E. Energy calculated speech sample for Happy



F. Energy calculated speech sample for Sad

According to the categories of happiness, sadness, and rage, the energy level of the speech sample is represented in the image above. Depending on how much energy is expended in pronouncing the single Marathi emotional word, a yellow line that represents the energy level changes.

REFERENCES

[1] New, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. Speech communication, 41(4), 603-623.

[2] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3), 572-587.

[3] Kadam, M. A., Orena, A. J., Theodore, R. M., & Polka, L. (2016). Reading ability influences native and non-native voice recognition, even for unimpaired readers. The Journal of the Acoustical Society of America, 139(1), EL6-EL12.

[4] Arora, V., Lahiri, A., & Reetz, H. (2018). Phonological feature-based speech recognition system for pronunciation training in non-native language learning. The Journal of the Acoustical Society of America, 143(1), 98-108.

[5] Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018, April). Non-native children speech recognition through transfer learning. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6229-6233). IEEE.

[6] Livescu, K. (1999). Analysis and modeling of non-native speech for automatic speech recognition (Doctoral dissertation, Massachusetts Institute of Technology).

[7] Livescu, K., & Glass, J. (2000, June). Lexical modeling of non-native speech for automatic speech recognition. In 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100) (Vol. 3, pp. 1683-1686). IEEE.

[8] Pukhraj Shrishrimal, R. R. Deshmukh, Vishal Waghmare, (2012, July) "Indian Language Speech Database: A Review". International Journal of Computer Application (UCA) Vol 47, No.5 pp. 17-21

[9] Yu Zhou, Yanging Sun, Lin Yang, Yonghong Yan, "Applying articulatory features to speech emotion recognition". 2009 International Conference on Research Challenges in Computer Science, 978-0-7695-3927-009, IEEE 2009, pp. 73-76.

[10] Vishal B Waghmare, Ratnadeep R Deshmukh, Pukhraj P Shrishrimal (2012, July) "A Comparative Study of the Various Emotional Speech Databases". International Journal on Computer Science and Engineering, Vol 4, issue 6, pp. 1236-1240

[11] Klaus R. Scherer, "What are emotions? And how can they be measured?" (2005) Trends and developments: research on emotions, Social Science Information Vol 44-no 4, pp. 695-729.

[12] Vishal B Waghmare, Ratnadeep R. Deshmukh (2014, February) "Development of Artificial Marathi Emotional Speech Database" in proceeding of 101st Indian Science Congress, Jammu, India, 2014. Gong Chenghui, Zhao Heming, Zou Wei, Wang Yanlei, Wang Min, "Preliminary Study on Emotions of Chinese Whispered Speech" International Forum on Computer Science-Technology and Applications, 978-0-7695-3930-0/09, IEEE 2009 pp. 429-433.

[13] Neethu Mohandas, Janardhanan P. S. Nair, Govindaru V., "Domain Specific Sentence Level Mood Extraction from Malayalam Text" 2012 International Conference on Advances in Computing and Communications IEEE 2012 pp 78 81.

[14] https://www.outsourcingtranslation.com/resources/history/marathi-language.php Accessed on 13/06/2020.

[15] Szczurowska I, Jozkowiak WK, Smolka E. The application of Kohonen and Multilayer Perceptron Networks in the speech non fluency analysis. Archives of Acoustics. 2014;31(4):205–10.

[16] Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology, 3(12), 18006-18016.

[17] Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A review on different approaches for speech recognition system. International Journal of Computer Applications, 115(22).

