1. 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]

Speech Emotions

The modelling of spectral and prosodic elements such as formants, pitch, loudness, timbre, speech pace, and pauses, which carry linguistic and semantic information, is fundamental to emotion identification systems [8]. However, the issue frequently involves one of the following basic emotion categories: Happy, Sad, Angry, Afraid, Surprised, and Neutral are all possible responses.

One of the key contributing elements for the low recognition accuracy gained during the development of speech-based systems is the emotion communicated by speech. When it comes to communication, human emotions influence the person's tone and speaking style. To solve the challenges of emotion identification from speech, more study in this field is required [9].

Speech emotion is a fundamental aspect of human expression. Since 1884, when William James attempted to define or answer it, a lot of definitions of emotions have been presented. Emotion is described as "an episode of interconnected, synchronised 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 the organism's major concerns." Natural emotions are those that arise spontaneously as a result of a set of events to which the brain responds. Artificial or performed emotions refer to feelings that are identical to those exhibited without the occurrence of situations to which the brain spontaneously responds.

There are numerous debates on the distinction between natural or real emotion and acted or fake emotion. The study required emotional speech, however because of the aforementioned issue, we were unable to capture the natural emotion, thus we created databases of performed or fake emotions that are copied. The artificial emotional Marathi speech database was created, and the experiment for emotion recognition was carried out on the created database [10, 11].

- **Artificial Emotion**

Artificial emotions are those that are not real but are imitated or acted by a person. Natural emotions are brain responses to internal or external stimuli. Any person will react to an occurrence in such a way that it becomes the natural emotion. Because it is impossible to capture spontaneous emotions, the best source for the study was acted or fake emotions that were reproduced. Everyone cannot accurately copy emotions; but, professional actors and actresses can mimic feelings that are most likely true emotions [12].

The researchers attempted to distinguish emotions in movies, TV shows, plays, and other media where people mimic emotions. Actors and actresses do the act and fulfil the emotional jobs in movies, TV shows, plays, and other sources, making it simple to carry out the emotion detection procedure on such data.

- **Natural Emotion**

Human speech has information and some meaning that relates the emotion, which comprises not only the linguistic content but also some emotions of the speaker, even if the emotion does not change the verbal content. There are numerous debates on whether natural or genuine emotion should be used. Natural emotions are difficult to record because they are responses to internal or external stimuli received by the brain. Because no one can foresee how a different person's brain would react to an experience, capturing natural emotions and categorising them becomes challenging. We needed emotions, however due to the aforementioned issue, we were unable to capture natural feelings.

- **Real Life Emotions**

In real life, people can show their emotions in a variety of ways, including facial expression, yelling, and touch. Speech is one of the most crucial outputs of a person's emotional state. The contribution of the vocal tract system activated by the excitation source signal results in a spoken signal.

- **Whispered Emotional Speech**

Whispered speech can also include emotional information such as prosodic traits such as short time energy and talk rate, voice quality parameters, formant, and spectrum to analyse emotional variances. When someone whispered, one might sense the emotions of others. Nowadays, with the widespread use of cell phones, individuals whisper to limit the amount of speech that is spelled out; whispered speech is frequently encountered for criminal analysis; and for laryngectomees, whisper is the only way of articulation. The whispered speech signal can be quantified using acoustic properties such as endpoint detection, abstraction of formant frequencies, and the related bandwidths. [13]

- **Mood Extraction**

One of the most challenging tasks is extracting emotions from speech. Sentiment analysis (SA) is critical in natural language processing. For domain-specific sentence level mood extraction, sentimental analysis can be performed using tasks to classify distinct moods such as Happy, Sad, Frustrated, Angry, Depressed, Temper, and so on [14].

Marathi Language:

Marathi is a member of the Indo-Aryan language family. It is primarily spoken by the Maharashtra people of Western India, and it has been the state's official language since 1966, as well as the co-official language of Goa state. During prehistoric times, Marathi was also known as Maharashtri, Marhatti, Mahratti, and other names. According to the 2011 census, India had 83 million Marathi speakers,

ranking it third after Hindi and Bengali and the fifteenth most spoken language in the world. It is the oldest type of Indo-Aryan regional literature. The language contains some of the earliest literature in all modern Indian languages, dating back to roughly 600 A.D. Marathi is thought to be about 1300 years old and evolved from Sanskrit, which was derived from Prakrit and Apabhramsha. Its grammar and syntax are claimed to be derived from Prakrit and Pali. The Marathi we hear now is the result of a long democratic cycle of reform and development. Marathi speakers can be found in both Israel and Mauritius. Marathi has around 42 dialects, with Tamil and Kannada loans heavily influencing the dialect used in Thanjavur and Tamil Nadu areas. Marathi is closely related to Konkani, Goanese, Deccani, Gowlan, Ihrani, and Varhadi-Nagpuri.[15,16].

2. Objective :

- To design and development of text corpora for Marathi Language.
- To collect speech samples from both Native and Non-Native speakers.
- To analysis of prosody features of both database.
- To design and develop the emotion recognition system for Marathi Language.

3. Motivation:

In the speech research generally research on Native speakers means who know the specific language from his/her childhood but still work on non-Native speakers are not done comparatively Native speakers. In the database collected data from Marathi speakers and non-Marathi speakers and analysis three sample of speech i.e Happy, Sad and Angry. Emotion can be defined as a positive or negative experience that is associate with a particular pattern of physiological activities.

4. Related work

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	Simulated

	persons	
German	51 School children (21M+30F)	Elicited
Spanish	8 Actors(4M+4F)	Simulated
Russian	61 Native speakers	Simulated

5. Methodology:

The following fig 1 shows basic workflow of work

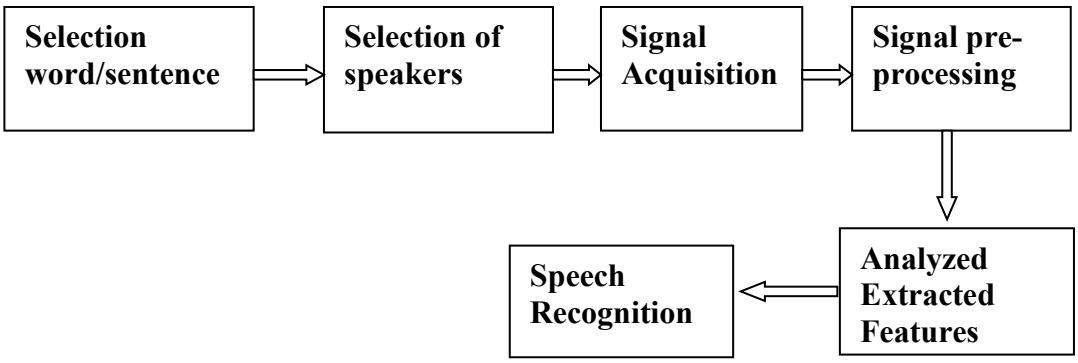


Fig 1 Methodology for Emotion recognition

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]

6. Design and Development:

i) 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.

ii) Developed Speech Database:

- 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	
		1008 utterances	

iii) Developed Speech Database:

- 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 .

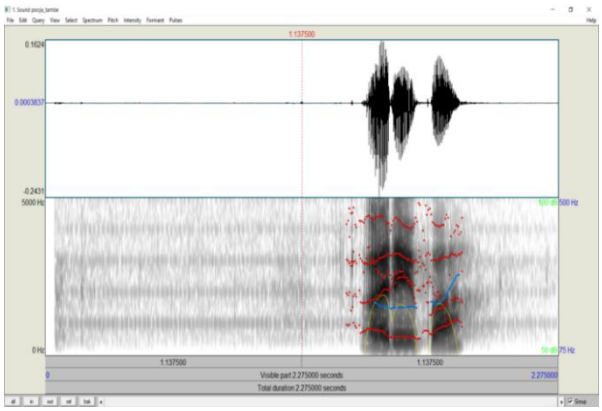


Fig.1 (a)Native female speakers voice sample With Noise

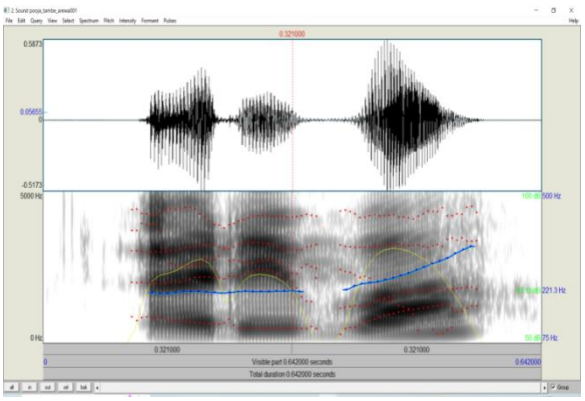
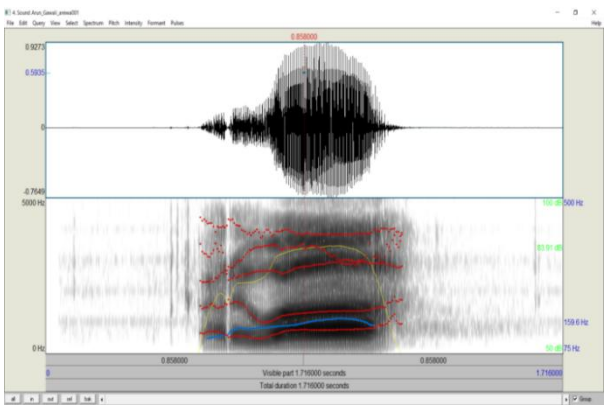


Fig.1 (b) Native female speakers voice Sample Without Noise



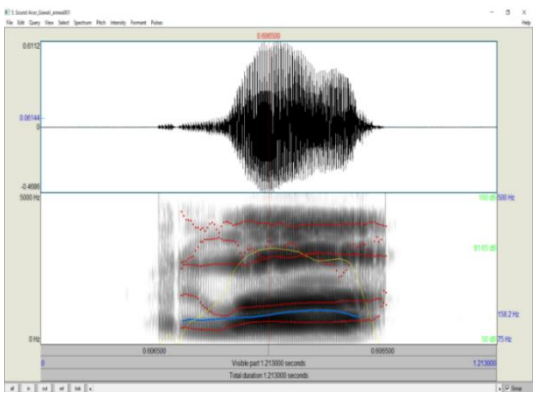


Fig.1 (c)Native male speakers voice sample With Noise

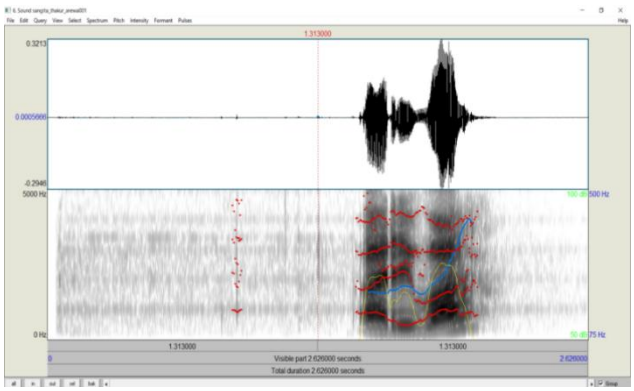


Fig.1 (d)Native male speakers voice sample Without Noise

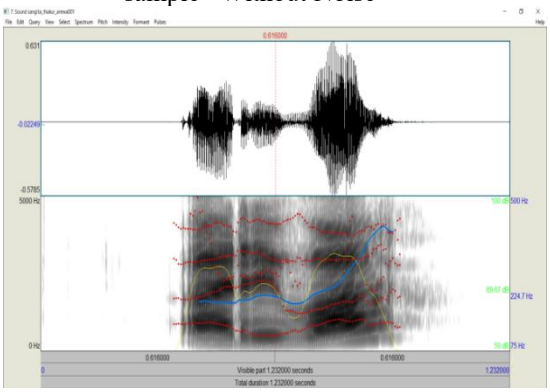


Fig.1 (e) Non-Native female speakers voice sample With Noise

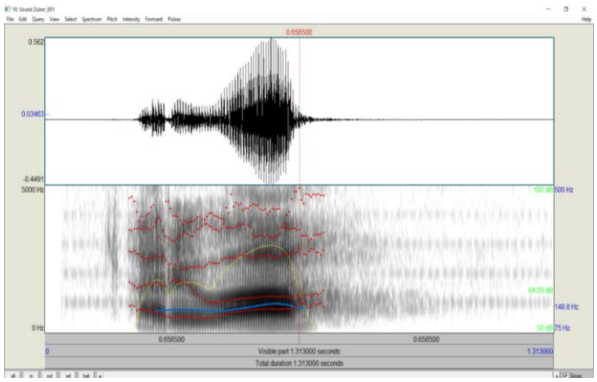


Fig.1 (e) Non-Native female speakers voice sample Without Noise

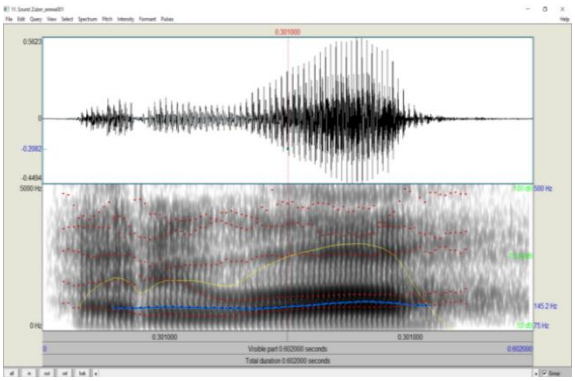


Fig.1 (g) Non-Native male speakers voice sample With Noise

Fig.1 (h) Non-Native male speakers voice sample Without Noise

7. Performance Analysis:

8.1 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.

8.1.1 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.

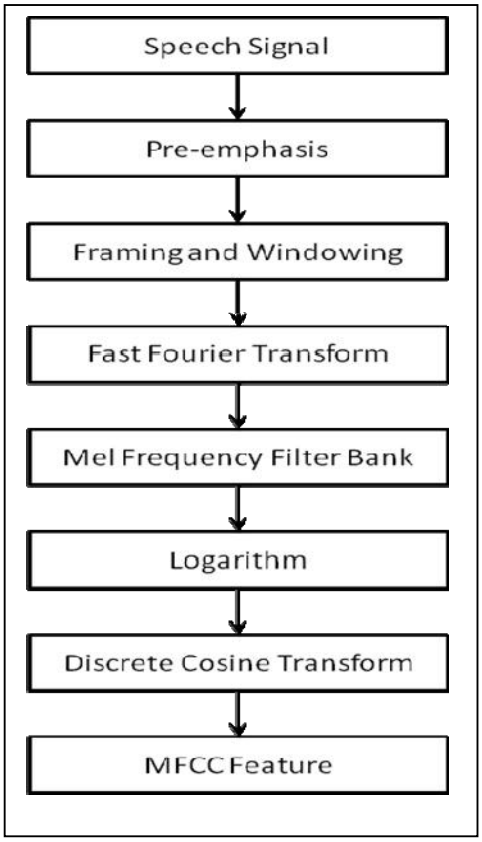
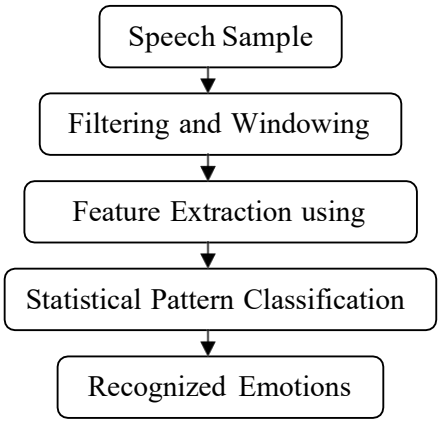


Figure 2: Block diagram of MFCC Feature Extraction method

8.1.2 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.



8.2 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:

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

Table Entries in confusion matrix

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

- **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

- **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

Result based on MFCC or LPC

	Native	Non-Native

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

8.3 Prosody-based emotion recognition system

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.

Confusion matrix for Prosody Features

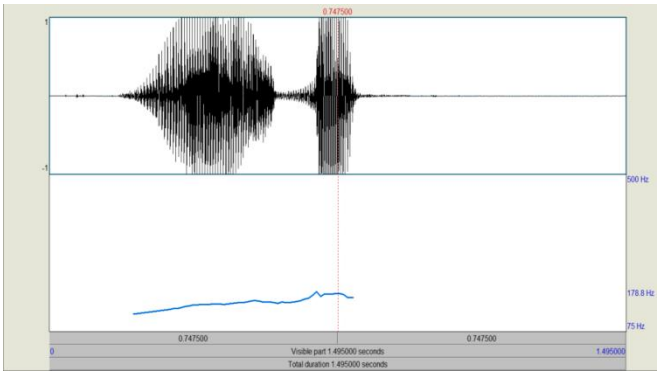
	Angry	Happy	Sad	Total number of samples
Angry	**27**	16	10	53
Happy	16	**21**	10	47
Sad	19	10	**21**	50
Total				150

Using speech features to extract emotions

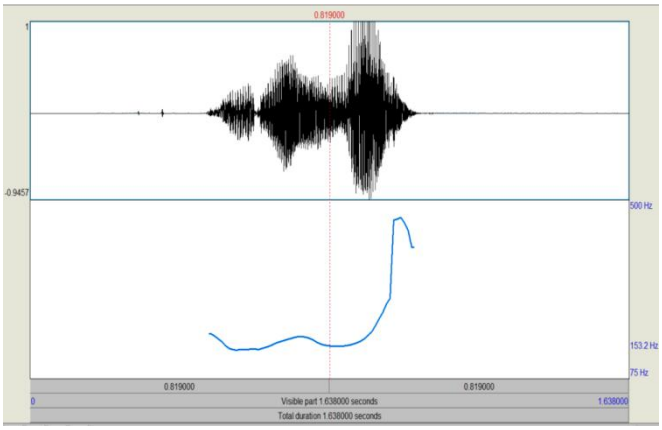
Prosodic Feature

- **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**.**

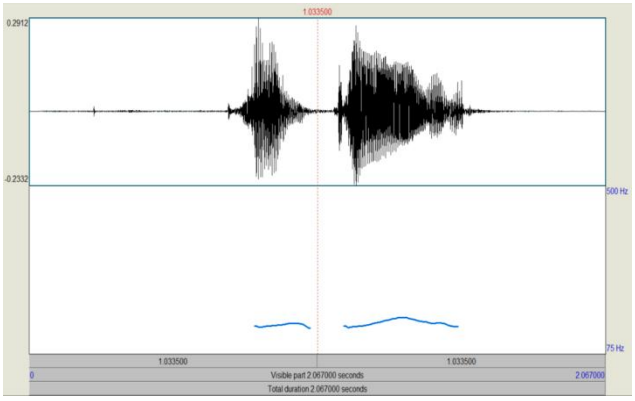
Below is a representation of the retrieved pitch contour from the voiced portions of the utterance.



A. Average pitch calculated speech sample for Angry

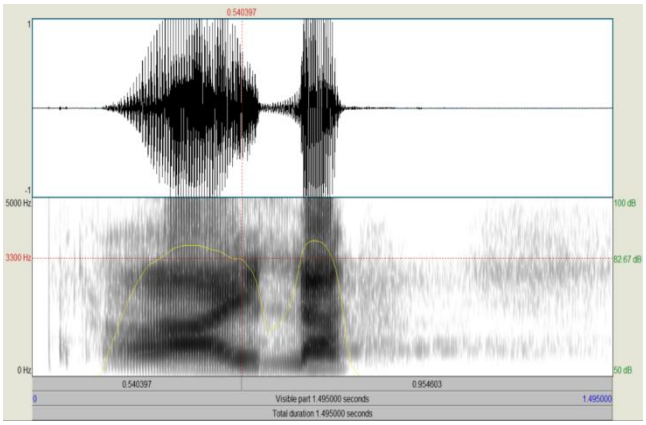


B. Average pitch calculated speech sample for Happy

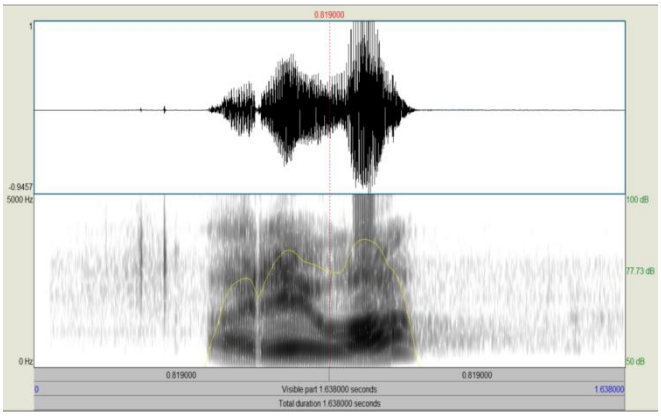


C. Average pitch calculated speech sample for Sad

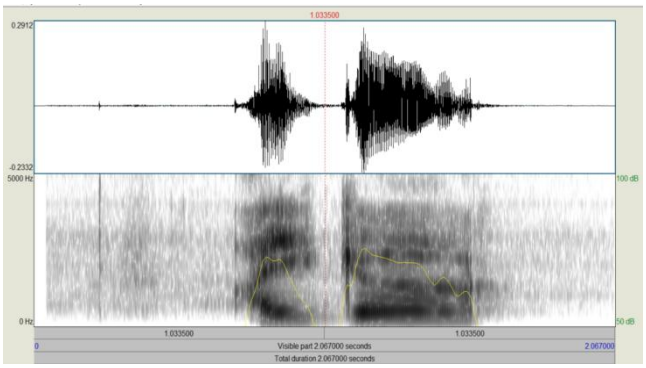
- **Enrgy:** We calculated the energy value using the first derivatives of the smoothed speech signal rather than the absolute signal amplitude in order to lessen the effect of loudness.It was possible to retrieve data on the energy statistic's mean, minimum, maximum, and standard deviation.



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.

Conclusion

Because there was currently no voice database for the Non-Native Marathi language, the primary goal of this research was to create one. It also created a library of isolated Marathi emotive terms. The Marathi language was receiving very little attention, and the current efforts are focused on gathering voice data via Non-Native and Native people. After completing a literature analysis, we discovered that language technologies can be extremely important in the creation of an effective governing system.

Mel frequency cepstral coefficient (MFCC), linear predictive coding, and the emotional speech database of native and non-native speakers were used in the investigation (LPC). We tested the Confusion matrix algorithm's ability to identify emotions.Comparing the recognition rate (%) for a chosen combination of feature sets for various classifier types was also investigated.

Reference

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.

13. 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.

14. 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.

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

16. 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.

17. 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.

18. 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).