Linear Mixed Effect Modelling for Analyzing Prosodic Parameters for Marathi Language Emotions

Trupti Harhare, Milind Shah
Dept. of Electronics and Telecommunications
Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India

***Abstract*—Along with linguistic messages, prosody is an essential paralinguistic component of emotional speech. Prosodic parameters such as intensity, fundamental frequency (F0), and duration were studied worldwide to understand the relationship between emotions and corresponding prosody features for various languages. For evaluating prosodic aspects of emotional Marathi speech, the Marathi language has received less attention. This study aims to see how different emotions affect suprasegmental properties such as pitch, duration, and intensity in Marathi's emotional speech. This study investigates the changes in prosodic features based on emotions, gender, speakers, utterances, and other aspects using a database with 440 utterances in happiness, fear, anger, and neutral emotions recorded by eleven Marathi professional artists in a recording studio. The acoustic analysis of the prosodic features was employed using PRAAT, a speech analysis framework. A statistical study using a two-way Analysis of Variance (two-way ANOVA) explores emotion, gender, and their interaction for mean pitch, mean intensity, and sentence utterance time. In addition, three distinct linear mixed-effect models (LMM), one for each prosody characteristic designed comprising emotion and gender factors as fixed effect variables, whereas speakers and sentences as random effect variables. The relevance of the fixed effect and random effect on each prosodic variable was verified using likelihood ratio tests that assess the goodness of fit. Based on Marathi's emotional speech, the R programming language examined linear mixed modeling for mean pitch, mean intensity, and sentence duration.**

Keywords—Prosodic parameters; a marathi language prosody model; a two-way analysis of variance; linear mixed-effect models; r programming language

I. Introduction

The COVID-19 pandemic has drastically altered people's lifestyles in many parts of the world. The lockdowns and social distancing norms eventually increased human-machine interaction applications. If computers can recognise emotions, they can communicate in a human-like manner. The prosodic features employed for emotion recognition play an essential role in the quality of the human-computer interaction that replicates human speech emotions. Supra-segmental features or the prosody features such as intensity, pitch, duration, etc., contribute additional information to speech known as paralinguistic information [1-4] and characterize the emotional speech. Developing a prosodic model for emotional utterances for less-studied languages is very challenging. It entails a lot of work, such as creating a database, processing it for analysis, investigating the fluctuation of prosodic elements about emotions using acoustic analysis, and establishing the relevance of these aspects using statistical analysis. In India's Maharashtra and Goa states, the Marathi language is spoken by over 73 million people. In comparison, in the Marathi language, there is less research on prosody aspects for emotional speech. Few of them includes, the syudy of the effect of focus shift in Subject-Object-Verb type Marathi sentences on prosodic features such as F0, duration, and intensity variations[5]. Authors analyzed that the speakers consistently provide acoustic cues with increased duration, higher mean F0, and higher intensity, differentiating focus location. The authors of [6] used broadcast radio transmission Marathi news that are available to the general public to investigate the significant prosodic aspects of the Marathi news-reading style. The authors observed prominence and boundary as the important prosody cues for Marathi's news reading style. Acoustically, the boundaries showed pre-boundary lengthening and pitch contour slope on the final syllable, and the prominence correlated with maximum F0 and maximum intensity and lesser duration. The authors analyzed MFCC features and energy ratios in [7] to investigate the Marathi emotion recognition for anger and happiness. The authors observed the anger emotion recognition rate higher than happiness and neutral emotions. However, the authors suggested generating more emotional speech databases from skilled Marathi speakers.

Considering comparatively less work towards prosody features of Marathi language and lack of emotional database from trained speakers, the paper focuses on emotion analysis for the Marathi language. We constructed a Marathi emotion database from professional speakers with a theatre background in a recording studio expressing anger, fear, happiness, and neutral emotions. The detailed study of the relationship between acoustic features such as mean intensity, mean pitch, sentence duration, and the emotions such as anger, happiness, fear, and neutral for Marathi's emotional speech showed various prosodic cues based on the emotions. Also, a comprehensive statistical analysis was conducted to construct a practical framework for assessing emotional speech data. A two-way ANOVA test for emotion, gender, and their interaction for mean pitch, mean intensity, and sentence time, as well as a linear mixed-effect analysis, were used in the statistical study. The LMM analysis is used to examine the relationship between the emotions and prosodic features data while considering the impacts of fixed and random effect variables and their connection. The two-way ANOVA analysis and a linear mixed model (LMM) analysis contributed while selecting the optimal prosodic features for constructing a prosody model.There is no comparable effort for the Marathi language that we are aware of.

This prosody model for the Marathi language using the acoustic and statistical investigation will help develop a human-machine interaction application such as emotion recognition from speech can help interpret students' answers and fit pupils with various learning abilities, Text-to-Speech systems (TTS) for Marathi storytelling, speaker recognition, speech recognition, online education etc. among other things.

The remainder of the paper is structured as follows: Section 2 contains a literature review, Section 3 explains the methodology and implementation for creating a Marathi database for various emotions and calculating prosodic features using the PRAAT speech analysis framework, Section 4 focuses on the results and discussion based on acoustic and statistical analysis to prepare a Marathi prosody model, and Section 5 summarises all of the discussions.

II. LITERATURE REVIEW

Speech is an important channel for the communication of emotion, yet studied little in the context of emotion. Speech conveys linguistic messages and includes a major paralinguistic part, prosody. The prosody of speech is defined in the linguistic literature as the suprasegmental properties of speech and include the pitch/F0, loudness/intensity, and rhythm/duration aspects (Brown 2005). Analyzing prosody features based on emotional speech is central to a few emotions. Although emotion classifications, in reality, are much larger, the majority of emotional speech statistics comprise four or eight emotions. Variations in prosodic elements concerning emotions, on the other hand, differ among languages and are dependent on culture and speaking style. As a result, it is vital to investigate the prosodic features for the emotional expressions specific to the language and culture. Fundamental frequency (F0), intensity, and duration are the essential acoustic characteristics influencing prosody. Fundamental frequency (F0) or pitch is the number of vibrations per second produced by the vocal cords, and the relative highness or lowness of a tone perceived by the ear determines pitch in speaking. The length of time a sentence, word, or syllable exists is called its duration. The intensity of a sound measures the energy contained in a given waveform signal. It is essential to analyze the prosodic features of emotion expression specific to the language and culture as emotions differ according to the cultural backgrounds, several international and national languages; the researchers are looking for acoustic correlates of prosody. Table I compares various prosody features studied in different languages based on distinct emotions and the corresponding dominant emotional signaling in the respective language.

Researchers often analyzed the database statistically after acoustic analysis to validate the acoustic analysis results and then select the best prosodic features to construct a prosody model [14-19]. Analysis of variance (ANOVA) findings investigate statistical discriminations of prosodic properties between various emotion classes. The success of ANOVA in identifying the best prosodic qualities to model the emotion recognition system has significantly reduced signal evaluation time. Hence, we have carried out a a two-way ANOVA analysis and linear mixed model (LMM) statistical analysis to design a prosody model for various emotions for Marathi. The LMM refers to using both fixed and random effects on the variables in the same analysis [20-23]. Due to the differences in prosodic variation patterns based on emotions, we examined three separate LMM models, one per prosody feature.

TABLE I. COMPARISON OF ACOUSTIC FEATURE VARIATIONS BASED ON VARIOUS EMOTIONS IN DIFFERENT LANGUAGES

Reference	Language Studied	Emotions	Prosody Features	Emotional Signaling
Bansal S., Agrawal, S., Kumar, A., 2019[8]	Hindi	neutral, fear, anger, surprise, sadness, and happiness	pitch, intensity and duration	The most intense emotion is anger, followed by neutral, happy, surprise, sadness, and fear. For all emotions, the pitch fluctuates in accordance with the intensity feature of speech.
J. Kaur, K. Juglan, V. Sharma, 2018.[9]	Punjabi	happiness, anger, fear. Sad, neutral	Mean Pitch, Intensity and formants	Mean pitch highest for happiness and lowest for sad, Intensity is highest for anger and lowest for fear.
Swain, M.; Routray, A., 2016.[10]	Odia	anger, fear, happiness, disgust, sadness, surprise.	pitch, energy, duration, and formant	In both males and females, the feeling "happy" has the greatest mean pitch value, followed by "surprise" in a close second. All other emotions have significantly lower energy levels than disgust and fear. Female respondents showed no discernible differences in the amount of energy levels for distinct emotions.
Hellbernd, N.; & Sammler, D., 2014.[11]	German	Criticism, naming, suggestion, doubt, warning, wish	Mean duration, mean intensity, mean F0, Pitch rise, harmonic-to-noise ratio	The loudest and most arching pitch contour were seen in warning stimuli. Naming stimuli having a low mean pitch, flat pitch contour, and low intensity.
Rao, K.; Koolagudi, S., 2013.[12]	Telugu	Anger, Disgust, Fear, Compassion, Neutral, Happiness, Sarcasm, Surprise	Mean duration, mean pitch, mean energy	Anger emotion with the highest energy Anger, happiness and neutral have high pitch values
Liu, P., Pell, M.D. 2012.[13]	Mandarin	anger, sadness, happiness, disgust, fear, pleasant surprise, neutrality	mean fundamental frequency, amplitude variations, speech rate (in syllables per second). mean harmonics-to-noise ratio, HNR variation	Anger and pleasant surprise had comparatively high mean f0 values and significant f0 and amplitude variations, but sadness, disgust, fear, and neutrality had relatively low mean f0 values and minor amplitude variations, while pleasure had a moderate mean f0 value and f0 variation.

III. Methodology and Implementation

Because prosody varies by language and speaking style, studying the relationship between emotions and the accompanying prosody variants is vital for all languages. This study aims to see how the prosodic aspects of Marathi's speech change with emotions, and four sub-questions are investigated concerning it as below.

1) Do Marathi speakers employ changes in prosody elements to help them communicate the emotion they want to convey in their speech?

2) If so, what precise variations in an utterance's prosody are used by speakers to differentiate one emotion from another?

3) Is it possible to create a predictive statistical model of prosody variation based on emotions in Marathi that can be utilized as a prosody model for a variety of applications such as emotion recognition, speaker recognition, speech recognition, text to speech synthesis systems, etc.?

4) Is it possible to consider neutral emotion as a baseline and analyze variations of prosodic features concerning neutral emotion and be used for emotion conversion applications?

The workflow for conducting out the research is depicted in the steps below.

1) Collection of sentences.

2) Selection of trained speakers.

3) Recording the sentences in anger, happiness, fear, and neutral read-out style emotions.

4) Collecting the database in .wav format.

5) Processing the .wav files with segmentation, annotation, and creating corresponding text grid files in the PRAAT speech processing toolbox.

6) Calculating the mean pitch, mean intensity, and sentence duration for all the .wav files.

7) Calculating and analyzing acoustic behavior of the above prosodic features based on the emotions, gender, speakers, and sentences.

8) Statistial analyzation of these prosodic features using two-way ANOVA and LMM analysis.

The corpus was constructed by identifying the sentences for the recordings, finding expert Marathi speakers, practicing, and recording their acted utterances in a recording studio. Each line was deliberately crafted to avoid provoking any emotion. There were three to nine words in each sentence. Eleven Marathi professional artists, four females and seven males with experience in drama and television, aged 18 to 40, participated in the experiment. The research objective was conveyed to the speakers, and two practice sessions were arranged to get acquainted with the sentences. The participants were paid incentives for this work. The selected ten sentences from different Marathi storybooks listed in Table II along with their English translations.

TABLE II. The English Translation of Ten Marathi Sentences used for Recording

	Marathi Sentences	**English Translation**
1.	आम्ही पण दहा बाय दहाच्या खोलीत राहतो .	We stay in 10 by 10 room.
2.	लेकीला सांगा तिचा बाबा आलाय.	Tell daughter that her father has come.
3.	मन मोठं असलं की सारं काही सामावून घेता येतं.	Big heart accommodates everything.
4.	अन्न वाया घालवू नये, त्याची किंमत कमवायला लागल्यावर कळेल.	You will value food when you start to earn.
5.	प्रत्येक दगड हा देव होतोच असं नाही.	Every stone does not become God.
6.	अंथरुण पाहून पाय पसरावे.	Spend as per your earning.
7.	गरीब माणसाची गम्मत करू नये.	Do not make fun of poor people.
8.	भाकरीची किंमत घाम गाळल्याशीवाय कळत नाही.	You never understand value till you won't work for it.
9.	डोकं शांत असेल तर निर्णय चुकत नाहीत.	A calm mind takes always the right decision.
10.	तुमचं बरोबर आहे.	You are right.

Each speaker repeated the given sentences with different emotions such as anger, happiness, fear, and neutral. The speakers recorded the utterances in a recording studio with a condenser microphone and a digital audio tape (DAT) recorder using a lossless 44kHz, 16bit audio format and saved at a sampling rate of 16kHz. Each speaker initially recorded ten sentences in a single emotion during the recording. Between the two sentences, the speakers left a reasonable pause. After recording all ten sentences in one emotion, the speakers took a short rest before recording all ten sentences in another emotion. The recording procedure took over three months to complete. Each speech file was an a.wav file with 2-4 seconds duration. The entire database of 440 sentences (eleven speakers, ten sentences, and four emotions) was available for further study. Each line was listened to by fifteen people (twelve Marathi native speakers and three non-Marathi speakers). They were able to identify emotions such as anger, happiness, fear, and neutrality in each recorded voice recording. The perceptually verified sentences were segmented in a PRAAT Text Grid. The .wav file of all the sentences is annotated manually in a sentence and word level for better accuracy. We observed variations in pitch contour, intensity contour, and duration for the same sentence comprising four emotions uttered by every speaker. It showed that there is some relationship that exists between the emotional utterances and corresponding prosodic features even for the Marathi language. The mean values of mean pitch, mean intensity and sentence duration of 422 sentences were calculated using the PRAAT speech analysis framework. The mean pitch was calculated by 'getting the Pitch' command and the mean intensity was calculated by 'getting Intensity (dB)' by selecting the sound interval in the PRAAT editor window. Sentence duration is calculated by selecting the portion of the utterance and reading the duration of the selection (in seconds) from the duration bar from the PRAAT editor window.

IV. Result and Discussion

A. Acoustic Analysis

The mean and standard deviations of prosodic parameters such as mean pitch, mean intensity, and sentence duration were determined for all 422 utterances to check for variation in mean pitch, mean intensity, and sentence duration for distinct emotions. Table III shows the overall descriptive statistics for the three prosodic variables for all four emotional utterances.

From Table III, we can see that the amount of variation or standard deviations (SD) are high for the mean pitch with 25.58% and the sentence duration with 27.62%, while mean intensity appears fairly consistent with the amount of variation (SD) of 6.6%. The standard deviation provides some insight into the patterns of variation occurring within the data. We analysed the variations of means and SDs of all three prosodic variables independently of anger, happiness, fear, and neutral emotions to acquire a clear picture of the prosodic variations based on emotional utterances, as shown in Table IV.

Table IV shows that the mean and standard variation values of all the three prosodic features for anger and happy emotions, and fear and neutral emotions, are nearly identical. To understand variations of prosody features for emotions, other factors such as gender, speakers and sentences are also important. Fig. 1 gives the analysis of variations of mean pitch, mean intensity and duration for gender, speakers and sentences.

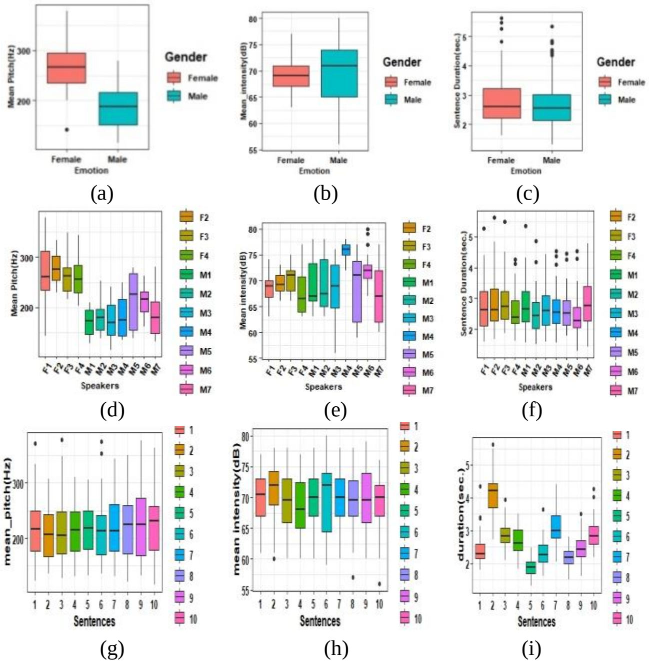
Fig. 1a demonstrated substantial differences in mean pitch values by gender, with males having lower values than females. Fig. 1b showed that males have variability in mean intensity than females, and Fig. 1c showed that both genders with similar observations for utterance duration. Fig. 1d, 1e, and 1f showed the variations of mean pitch, mean intensity, and sentence duration among the multiple speakers of the same gender. Fig. 1g, 1h, and 1i show the variations of the prosodic features concerning the ten different sentences. Fig. 1 shows that in the Marathi language, prosodic features vary for emotion change as well as variations in gender, speaker, and sentence. In Fig. 2, gives variation of prosodic features based on emotion and gender.

TABLE III. Descriptive Statistics for the Three Prosodic Variables

Prosodic Variables	Mean	Std. Deviation	Percentage
Mean pitch	217.70 Hz	55.69 Hz	25.58%
Mean intensity	69.74 dB	4.615 dB	6.6%
Sentence duration	2.69 sec.	0.743 sec.	27.62%

TABLE IV. Summary of Means and SDs of Each Prosodic Variable by Emotions

Emotion	Mean pitch (Hz)		Mean Intensity (dB)		Duration (sec.)	
	Mean	***S.D.***	***Mean***	***S.D.***	***Mean***	***S.D.***
Anger	254.7	46.6	73.69	2.58	2.27	0.57
Happiness	241.4	51.5	71.02	2.86	2.64	0.63
Fear	193.9	43.3	67.40	4.15	2.86	0.73
Neutral	179.5	42.2	66.77	4.71	3.01	0.82

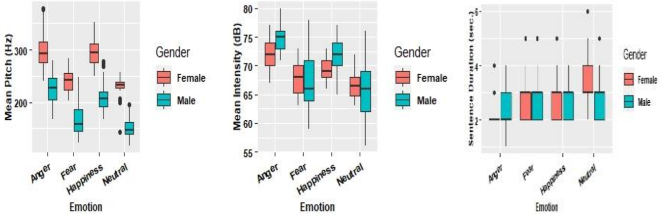


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 1. Box Plots Showing Variations in mean Pitch mean Intensity and Sentence Duration (Figure 1a, 1b, and 1c) for Gender, Speakers (Fig. 1d, 1e, 1f) and Sentence Duration Due to Multiple Speakers (Fig. 1g, 1h, 1i).



(a) (b) (c)

Fig. 2. Variations in (a). Mean Pitch, (b). Mean Intensity, (c). Sentence Duration for Emotions and Gender.

Fig. 2a and 2b showed variations in mean pitch and mean intensity for gender for all the emotions. There was little change in mean intensity levels for the fear emotion between male and female speakers. Each gender takes the same amount of time to speak the lines in fear and happiness observed in Fig. 2c. Female speakers took less time to express anger than male speakers, whereas female speakers' utterance duration was higher for neutral speaking style sentences. As a result, acoustic analysis of mean pitch, intensity, and duration revealed that prosodic features behave differently for emotions and gender.

For Hindi language, acoustic correlation of emotions were analysed for prosodic parameters such as pitch, intensity and duration in [24]. Authors have generated Hindi speech database with 10 speakers in all six emotions such as anger, Fear, Happy, Neutral, Sad and Surprise. Comparison between the prosodic parameters related to the emotions for Marathi and Hindi language given in Table V.

From Table V, comparing the prosodic variation patters for anger, happiness, fear and neutral emotions the pitch, intensity and duration features observed with the following order by taking neutral emotion as reference.

Pitch : Neutral > Anger > Happiness > Fear (for Hindi).

Pitch : Anger > Happiness > Fear > Neutral (for Marathi).

Intensity : Happiness >Anger > Neutral > Fear (for Hindi).

Intensity : Anger > Happiness > Fear > Neutral (for Marathi).

Duration : Fear > Happiness > Neutral >Anger (for Hindi).

Duration : Neutral > Fear > Happiness > Anger (for Marathi).

When the behavior of variations of prosodic elements dependent on emotions is compared for the two Devnagari languages of India, Hindi, and Marathi, the relevance of studying each language separately for prosodic patterns based on the emotions becomes clear.

B. *ANOVA Analysis*

A two-way ANOVA analysis explored the impact of emotions and gender and their interaction on the mean pitch, mean intensity, and duration. Both emotion (F = 205.7, p <0.0001) and gender (F=872.1, p < 0.0001) were significant for mean pitch, suggesting that gender is more responsible for mean pitch variations. The p-value for the interaction between emotion and gender was non-significant (F =1.468, p=0.223) indicated that the relationships between gender and mean pitch was independent of the emotion. For mean intensity, emotion (F = 104.2, p< 0.0001) and gender (F= 9.528, p <0.001) were statistically significant with emotion as the most significant factor variable. The p-value for the interaction between emotion and gender for mean intensity observed to be significant (F =7.274, p <0.0001) indicated that the relationships between gender and mean intensity depends on emotion. For sentence duration, emotion factor observed to be statistically significant(F = 16.142, p< 0.0001) but gender as non-significant (F= 0.923, p =0.337). The p-value for the interaction between emotion and gender non-significant (F =0.709, p = 0.547) indicated that the relationships between gender and sentence duration depend on emotion.

TABLE V. THE PROSDIC FEATURE VALUES FOR MARATHI AND HINDI LANGUAGE FOR ANGER, HAPPINESS, FEAR AND NEUTRAL EMOTIONS

Prosodic Parameters	**Emotions**	**Marathi**	**Hindi**
Mean pitch (Hz)	Anger	254.7	303
	Happiness	241.4	300
	Fear	193.9	295.5
	Neutral	179.5	304.4
Mean Intensity (dB)	Anger	73.69	84
	Happiness	71.02	84.5
	Fear	67.40	81
	Neutral	66.77	83
Duration (sec.)	Anger	2.27	1.39
	Happiness	2.64	1.67
	Fear	2.86	2.8
	Neutral	3.01	1.66

C. *LMM Analysis*

Differences in prosodic features in Marathi are attributable to emotion fluctuations and gender, speaker, and sentence variations. Also, even if the independent variables such as emotions and gender have a somewhat consistent impact on prosodic feature variations, it can vary amongst speakers of the same gender or even between the different sentences. Linear mixed models are a type of regression model that takes into account variation explained by the independent variables of interest, known as fixed effects, and variation not explained by the independent variables, known as random effects [22]. The model is mixed since it combines both fixed and random effects. Thus, to calculate variations in prosodic features, emotion and gender factors are of primary interest and added as fixed effect variables. The emotion factor with four factors: anger, happiness, fear, and neutral emotions and gender factors included males and females: the speaker and the sentence considered random effect variables. Each prosodic feature was then verified for the model fit considering these factors. The goodness of fit of prosodic features for fixed-effect and random-effect variables, as shown in equation (1).

$$\text{Prosodic feature} = \text{emotion} + \text{gender} + (1/\text{speaker}) + (1/\text{sentence}) \tag{1}$$

Equation 1 shows the variations in the prosodic feature for the variations in emotion and gender as fixed effect variables and speakers and sentences as random effect variables with 1/speaker and 1/sentence as random intercept different for each speaker and each sentence individually.

The likelihood ratio tests to assess the goodness of fit to verify the significance of the fixed effect and random effect for each prosodic variable. The goodness of fit test confirmed the relevance of the fixed and random effect variables for prosodic feature variations.

1) Modeling mean pitch: The impact of fixed-effect variables on the mean pitch calculated by comparing the null effect models with fixed effect = 1 and two models with emotion as a fixed effect factor and the other model with emotion and gender as two fixed-effect elements shown in equation (2), (3) and (4) respectively.

$$\text{Mean pitch} = 1 + (1/\text{speaker}) \tag{2}$$

$$\text{Mean pitch} = \text{Emotion} + (1/\text{speaker}) \tag{3}$$

$$\text{Mean pitch} = \text{Emotion} + \text{Gender} + (1/\text{speaker}) \tag{4}$$

Chi-square difference tests showed the significant p-value of emotion with $\chi2(1) = 434$, $p < 0.001$ and of gender with $\chi2(1) = 23$, $p < 0.001$. Both emotion and gender observed to be significant with p<0.001 and considered fixed effect variables for mean pitch modeling.

In addition, likelihood ratio tests examined the goodness of fit and confirmed the relevance of the random effect variables' influence on the mean pitch. We compared the two null effect models, with fixed effect = 1 and one without sentence intercept, and sentence intercept with the following equation (5) and (6) respectively.

$$\text{Mean pitch} = 1 + (1/\text{speaker}) \tag{5}$$

$$\text{Mean pitch} = 1 + (1/\text{speaker}) + (1/\text{sentence}) \tag{6}$$

Comparing the models with Chi-square difference tests, resulted $\chi2(1) = 0$, $p > 0.01$. It showed that the inclusion of a sentence is not significant for mean pitch calculation since it does not improve the model fit.

The final design of the model fit to calculate the mean pitch model for Marathi emotional speech calculated as shown in equation (7) as below,

$$\text{Mean pitch} = \text{Emotion} + \text{Gender} + (1/\text{speaker}) \tag{7}$$

As in equation (7), emotion and gender factors are fixed effect variables, and the speaker is a random effect variable for calculating the mean pitch model for Marathi's emotional speech. Table VI shows the impacts of fixed effect variables such as emotions (angry, happiness, fear, and neutral) and gender (male and female) on computing mean pitch values.

TABLE VI. FIXED EFFECT SUMMARY OF MEAN PITCH

Emotions	Estimate	Std. Error	t- value
Intercept (Neutral)	228.19	7.257	31.45***
Anger	75.51	3.19	23.64***
Fear	15.47	3.2	4.84***
Happiness	61.49	3.19	19.30***
Male	-77.40	8.756	-8.840***

***=p< 0.001

The neutral emotion is used as an emotion baseline, while the female gender is a gender baseline. The estimate of the intercept value of 228.19 indicates that the mean pitch value for neutral emotion and female gender is 228.19Hz. The mean pitch values for other emotions were calculated from Table V based on the neutral emotion mean pitch value estimates. The estimate for the mean pitch value of anger emotion is 228.19+ 75.51= 303.7Hz, which is significantly higher than for neutral emotion (t= 31.45, p<0.001). Similarly, the estimate for the mean pitch value of fears emotion is 228.19 + 15.47= 243.66Hz, which is significantly higher than for neutral emotion (t=4.98, p<0.001). Similarly, the estimate of the mean pitch value for happy emotion is 228.19 + 61.49= 289.68 Hz, and this is significantly higher than for neutral emotion (t= 19.30, p<0.001). Also, the estimated value of the mean pitch of males of -77.40 based on a baseline of female gender means the pitch of males is lower than that for females by 77.40Hz. With this we can calculate mean pitch values for male gender for each of the emotion as; anger = 303.7 – 77.4 = 226.3Hz, happiness = 289.68 – 77.4 = 212.28Hz and fear =243.66 – 77.4 = 166.26Hz.

The Fixed Effects table, similar to most methods such as ANOVA, MANOVA, multiple regression analyses only focuses on group differences in changes in mean pitch values for emotions and gender. Understanding the mean pitch change at both the group and individual levels will be helpful to capture a complete overview of developmental changes in mean pitch values. Table VII summarizes the random effect of individual speakers on the mean pitch model design below.

TABLE VII. RANDOM EFFECT ANALYSIS FOR MEAN PITCH

Groups Name	Variance	Std. Dev.
Speaker	181.3	13.46
Residual	534.5	23.12

The variance due to the speaker is 181.3 and hence the standard deviation of 13.46Hz. This means that there can be variations in the fixed effect values due to variability between the individual speakers. The residuals are the random deviations from the predicted values that are due to some factors outside of the purview of the experiment. The estimate of the residual variance, with a standard deviation equal to 23.12Hz, represents the variability in individual emotion pitch values due to unknown factors.

2) Modeling mean intensity: The relevance of fixed effects on mean intensity was established by comparing null effect models with fixed effect = 1 to two models, one with emotion as the fixed effect factor and the other with emotion and gender as fixed effect factors as shown in equation (8), (9) and (10).

$$\text{Mean intensity} = 1 + (1/\text{speaker}) \tag{8}$$

$$\text{Mean intensity} = \text{Emotion} + (1/\text{speaker}) \tag{9}$$

$$\text{Mean intensity}= \text{Emotion} + \text{Gender} + (1/\text{speaker}) \tag{10}$$

Chi-square difference tests showed the significant p-value of Emotion with $\chi2(1) = 274.63$, $p <0.001$ and gender with $\chi2(1) = 0.48$, $p > 0.1$. This means, emotion factor is significant for variations in mean intensity, but gender is non-significant.

A chi-square difference test calculated the inclusion of a random effect structure with random intercepts for speakers and sentences as shown in equations (11) and (12), respectively.

$$\text{Mean intensity} = 1 + (1/\text{speaker}) \tag{11}$$

$$\text{Mean intensity} = 1 + (1/\text{speaker}) + (1/\text{sentence}) \tag{12}$$

Comparing the models with Chi-square difference tests, we conclude that sentence inclusion is not significant for mean intensity calculation since it does not improve model fit, $\chi2(1) = 0$, $p > 0.01$.

The final design of the model fit to calculate the mean intensity model for Marathi emotional speech calculated as shown in equation (13) as below.

$$\text{Mean intensity} = \text{Emotion} + (1/\text{speaker}) \tag{13}$$

Table VIII gives the summary of fixed effect variables for calculating the mean intensity.

TABLE VIII. FIXED EFFECTS SUMMARY FOR ANALYSIS OF MEAN INTENSITY

	Estimate	Std. Error	t value
Intercept (Neutral)	66.76	0.73	91.39***
Anger	7.02	0.41	17.31 ***
Fear	0.68	0.41	0.096
Happiness	4.22	0.40	10.44 ***

***=p< 0.001

The estimate of the intercept value of 66.76 indicates that the mean pitch value for neutral emotion 66.76dB. The estimates of mean intensity values for other emotions are calculated based on the estimates of intercept, i.e., neutral emotion mean intensity. The estimate for mean intensity of anger emotion is 66.76 + 7.02= 73.78 dB and this is significantly higher than for neutral emotion (t= 17.31, p<0.001). The estimate for mean intensity for fear emotion is 66.76 + 0.68 = 67.44 dB and this is not showing any significance with neutral emotion (t=0.096, p > 0.1). Similarly, the estimate for mean intensity for happiness emotion is 66.76 + 4.22 = 70.98 dB and this is significantly higher than for neutral emotion (t= 10.44, p<0.001).

Table IX summarizes the random effect of individual speakers on the mean intensity model design below.

TABLE IX. Random Effect Analysis for Mean Intensity

Groups Name	Variance	Std. Dev.
Speaker	4.956	2.226
Residual	8.619	2.936

The variance due to speaker is 4.956, indicating the standard deviation in mean intensity is 2.23dB in the fixed effect values due to variability between the speakers. The residuals are the random deviations from the predicted values, with a standard deviation equal to 2.96dB representing the variability in intensity apart from speakers.

3) Modeling duration: The significance of fixed effects on the sentence duration was determined by comparing null effect models, where fixed effect = 1 and the two models one with fixed effect factor as emotion and the other with fixed-effect factors as emotion and gender as shown in equation (14), (15) and (16) respectively.

duration = 1 + (1/speaker) (14)

duration = Emotion + (1/speaker) (15)

duration = Emotion + Gender + (1/speaker) (16)

Chi-square difference tests for duration showed the significant p-value for emotion with $\chi2(1) = 640.8$, $p < 0.001$ but non-significant for gender with $\chi2(1) = 2$, $p = 0.15$. This suggests that the variation in sentence duration is due to emotion rather than gender.

Also, the two null effect models, one without speaker intercept and the other with a sentence and speaker intercept, compared to determine the duration model fit for random effect variables as shown in equation (17) and (18) respectively.

duration = 1 + (1/speaker) (17)

duration = 1 + (1/speaker) + (1/sentence) (18)

The Chi-square difference tests showed that inclusion of sentence as one of the random effects is significant for the mean duration and it improved model fit, $\chi2(1) = 410.73$, $p < 0.001$.

The final design of the model fit to calculate the sentence duration model for Marathi emotional speech calculated as shown in equation (19) as below.

duration = Emotion + (1/speaker) + (1/sentence) (19)

Table X gives the summary of fixed effect variables for calculating the duration given by the final design to calculate the duration for emotions equation.

TABLE X. Fixed Effects Summary for Analysis of Duration

	Estimate	Std. Error	t value
Intercept (Neutral)	2.99	0.19	15.52***
Anger	-0.73	0.04	-16.42 ***
Fear	-0.14	0.04	0.002**
Happiness	-0.36	0.04	-8.104 ***

***=p< 0.001, **=p< 0.01

The estimate of the intercept value of 2.99 indicates that the mean sentence duration for neutral emotion is 2.99sec. The estimates of mean sentence duration values for other emotions are calculated based on the estimates of intercept, i.e., the mean sentence duration for neutral emotion. The estimate for the sentence duration anger emotion is 2.99 - 0.73= 2.26 sec and this is significantly lower than for neutral emotion duration (t= -16.42, p<0.001). The estimate for the sentence duration for fear emotion is 2.99 -0.14 = 2.85 sec and this is not showing any significance with neutral emotion (t=0.002, p = 0.01). Similarly, the estimate for the mean sentence duration of happy emotion is 2.99 -0.36 = 2.63 sec and this is significantly higher than for neutral emotion (t = -8.104, p<0.001).

Table XI, giving the summary of the random effect of individual speaker and individual sentence on the sentence duration model design as below.

TABLE XI. Random Effect Analysis for Duration

Groups Name	Variance	Std. Dev.
Speaker	0.119	0.014
Sentence	0.59	0.35
Residual	0.32	0.10

The variance due to speaker is 0.014, and hence standard deviation of 0.119 sec. in the fixed effect values can be due to variability between the speakers. The variance due to the sentence is 0.35 and hence standard deviation of 0.59 sec. in the fixed effect values can be due to variability between the sentences. The residuals are the random deviations from the predicted values, with a standard deviation equal to 0.32 sec. represents the variation in duration values apart from speaker and sentence.

V. Conclusion

This work explains how to investigate acoustic clues for Marathi emotions, including anger, happiness, fear, and neutral. Eleven Marathi professional artists created a database of 440 words in anger, happiness, fear, and neutral emotions in a recording studio. According to an acoustic experiment, the features of mean intensity, mean pitch, and sentence duration vary depending on the emotions. A two-way ANOVA and a linear mixed-effect analysis provided a valuable framework for studying emotional speech data and, as a result, best practices for generating an emotional speech corpus.

The following is the prosodic model for emotions in the Marathi language, with emotion and gender as fixed effect variables and speaker and sentences as random effect variables.

Mean pitch = Emotion + Gender + (1/speaker).

Mean intensity = Emotion + (1/speaker).

duration = Emotion + (1/speaker) + (1/sentence).

A detailed analysis of Marathi's emotional speech will help develop a prosody model. This model will help select appropriate input features for machine learning algorithms used in emotion classification applications. In the future, it will be beneficial to examine some more prosodic aspects for the Marathi language emotions. Sadness, surprise, and sarcasm are examples of other basic emotions that may be investigated for Marathi speech. Children, young adults, and the elderly can all be studied separately in a similar way. When these meticulously generated prosodic elements are fed into a machine learning model, they can aid in emotion recognition, text-to-speech synthesis, and other human-machine interaction applications in the future.

REFERENCES

[1] G. Zhang, S. Qiu, Y. Qin and T. Lee, "Estimating Mutual Information in Prosody Representation for Emotional Prosody Transfer in Speech Synthesis," 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021, pp. 1-5.

[2] T. Wani, T. Gunwan, S. Qadri, M. Kartiwi, “A Comprehensive Review of Speech Emotion Recognition Systems”, IEEEAccess, vol. 9, pp. 47795–47814, April 2021.

[3] S. Bharadwaj and P. B. Acharjee, "Analysis of Prosodic features for the degree of emotions of an Assamese Emotional Speech," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1441-1452.

[4] V.Raju, H.Vydana, S.Gangashetty and A.Vuppala, "Importance of non-uniform prosody modification for speech recognition in emotion conditions," 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017, pp. 573-576.

[5] P. Rao, N. Sanghvi, H, Mixdorff, K Sabu., “Acoustic correlates of focus in Marathi: Production and perception”, Journal of Phonetics, 2017, vol. 65, pp. 110-125.

[6] S. Barhate, S. Kshirsagar, N. Sanghvi, K.Sabu, P. Rao, and N. Bondale, “Prosodic features of Marathi news-reading style”, IEEE Region 10 Conference (TENCON), 2016, pp.2215-2218.

[7] V. Degaonkar, Apte S., “Emotion modeling from speech signal based on wavelet packet transform”, Int J Speech Technol, 2013, vol. 16, pp. 1–5.

[8] S. Bansal, S. Agrawal, A. Kumar, “Acoustic analysis and perception of emotions in Hindi speech using words and sentences”, Int. j. inf. Technology, 2019, vol. 11, 807-812.

[9] K. Jasdeep, S. Vishal, “ Role of Acoustic Cues in Conveying Emotion in Speech”, Journal of Forensic Sci & Criminal Inves., 2018, vol. 11(1), pp. 555803.

[10] M.Swain, A. Routra, P. Kabisatpathy, J.Kundu,“Study of prosodic feature extraction for multidialectal Odia speech emotion recognition”, IEEE Region 10 Conference (TENCON), 2016, pp. 1644-1649.

[11] N. Hellbernd & D. Sammler, “Prosody conveys speaker’s intentions: Acoustic cues for speech act perception”, Cognitive Processing, vol. 15, 2014, S46-S46.

[12] K. Rao, S. Koolagudi, R.Vempada, “Emotion recognition from speech using global and local prosodic features”, Int J Speech Technol, vol. 16, 2013, pp.143–160.

[13] P. Liu, D. Pell, “Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal emotional stimuli”, Behav Res, vol. 44, 2012, pp.1042–1051.

[14] H. Chang, S. Young, K .Yuen, “Effects of the Acoustic Characteristics on the Emotional Tones of Voice of Mandarin Tones”, Proceedings of 20th International Congress on Acoustics, Sydney, Australia, 2010.

[15] A. Jacob, P.Mythili, “Upgrading the Performance of Speech Emotion Recognition at the Segmental Level”, IOSR Journal of Computer Engineering (IOSR-JCE) Volume 15, Issue 3, pp. 48-52, 2013.

[16] T. Iliou,C.Anagnostopoulos, “Classification on Speech Emotion Recognition - A Comparative Study”, International Journal on Advances in Life Sciences, vol 2 no 1 & 2, 2010.

[17] S. Ali, M. Andleeb, D. Rehman, “A Study of the Effect of Emotions and Software on Prosodic Features on Spoken Utterances in Urdu Language”, I.J. Image, Graphics and Signal Processing, vol. 4, pp.46-53,2016.

[18] M. Yusnita A, Paulraj, S. Yaacobb, N. Fadzilah, Shahriman A., “Acoustic Analysis of Formants across Genders and Ethnical Accents in Malaysian English using ANOVA”, International Conference On Design and Manufacturing, vol.64, pp. 385–394, 2013.

[19] A.Meftah, Y. Alotaibi, A. Selouani,“Evaluation of an Arabic Speech Corpus of Emotions: A Perceptual and Statistical Analysis”,. IEEE Access, PP. 1-1, 2018. M. Ayadi, M. Kamel and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, 2011, vol.44, pp. 572-587.

[20] H. Singmann, & D. Kellen, “An Introduction to Mixed Models for Experimental Psychology”’. In D. H. Spieler & E. Schumacher (Eds.), New Methods in Cognitive Psychology, 2019, pp. 4–31.

[21] Sherr-Ziarko, Ethan., PhD thesis, University of Oxford, 2017.

[22] B. Winter, “Linear models and linear mixed-effects models in R with linguistic applications”, Cognitive and Information Sciences, University of California, Merced, 2013.

[23] M. Rouch, Undergraduate Honors Theses, Williamsburg VA, 2019.

[24] S.Bansal, S. Agrawal, A. Kumar, “Acoustic analysis and perception of emotions in hindi speech using words and sentences, Int. j. inf. tecnol. 11, pp. 807–812, 2019.