Investigating Marathi Prosody: Unraveling Features of an Understudied Language

Trupti Harhare[1], Milind Shah[2]

[1] Dept. of Electronics and Telecommunications, Agnel Charities' Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India.
[2] Dept. of Electronics and Telecommunications, Agnel Charities' Fr. C. Rodrigues Institute of Technology, Navi Mumbai, India.

ABSTRACT: Studying the detailed prosodic features of emotional speech helps us understand the rhythm, stress, and intonation patterns. Since prosodic features vary by language, researchers worldwide have been studying these features in different languages to create better models for human-machine communication. The Marathi language, mainly spoken in Maharashtra and Goa, has not been studied much in this area. The current freely available Marathi database from trained speakers is limited, making it difficult to analyze its prosodic features fully. To address this, we have created a large database with 440 Marathi sentences, showing four main emotions: anger, happiness, fear, and neutrality, spoken by trained Marathi speakers. Seven male and four female actors contributed to this research, capturing the differences in prosodic features for each emotion in Marathi speech. We used PRAAT software to study the recorded data, focusing on twelve sound features related to pitch, intensity, duration, and voice quality. Our study found that anger had the highest pitch and intensity, with average voice quality and the shortest duration. Happiness also had high pitch and intensity, short duration, and moderate voice quality. Fearful emotion showed low pitch and intensity, moderate voice quality, and long duration. These were compared to neutral emotion. We also observed differences between male and female speakers in pitch, intensity, duration, and voice quality across emotions. We examined the prosodic features of Hindi, English, Mandarin, German, Odia, and Punjabi languages and identified variations in the prosodic features specific to the Marathi language. Across nearly all the languages examined, variations in acoustic correlates of prosodic features were observed across emotions. Studying language-specific prosodic features is crucial as prosody depends on speaking style and language, providing valuable acoustic cues for designing speech-processing applications, particularly in the context of TTS systems. The Marathi emotional dataset from trained Marathi speakers and the emotion-based acoustic cues analyzed in this study can be used to design various human-machine interaction (HMI) applications, particularly for developing natural Text-to-Speech (TTS) systems in the Marathi language.

KEYWORDS: Acoustic correlates, Prosodic features, Marathi emotional expressions

1. Introduction

Effective human-machine interaction (HMI) relies heavily on speech, humans' primary mode of communication [1-10]. As humans primarily rely on speech to convey information, it becomes essential for machines to understand and produce speech accurately. The three major information categories used to express machine-produced speech are linguistic, paralinguistic, and non-linguistic [1]. Linguistic or segmental information encompasses discrete text characters, including vowels and consonants. It involves language's syntactic and semantic aspects, focusing on the words and their arrangement. While linguistic information is vital for conveying the content of the message, it may only capture part of the richness of human communication. The paralinguistic information or suprasegmental structure of speech goes beyond words and includes aspects such as the speaker's emotions, attitudes, and intentions. Paralinguistic cues, including prosody features, provide additional layers of meaning to speech and enhance the expression of emotions [11-16, 19]. Prosody, a key component of paralinguistic information, refers to the emphasis, rhythm, intonation, and other speech features contributing to effective communication. It includes acoustic aspects such as intensity, duration, fundamental frequency (F0) or pitch, voice quality, etc. By incorporating prosodic features into machine-generated speech, particularly for local languages with unique prosodic patterns, communication can become more natural and accessible. Speech researchers are developing HMI applications, such as text-to-speech (TTS) systems, that can work without language barriers by creating resources for less commonly studied languages and turning them into valuable products and services [1-10], [64-69].

In the present digital era, the proliferation of mobile phones, laptops, and other electronic devices has rapidly expanded beyond urban areas and reached rural regions. However, a prevailing challenge arises in HMI, as most applications and interfaces are predominantly available in English, other Western languages, and Indian languages [2-10], [15-16], [48], [50], [52]. Promoting and preserving regional languages such as Marathi is vital, and developing speech-processing programs specifically for Marathi is a step in that direction. Designing and implementing numerous HMI applications, such as speaker recognition, automatic speech recognition, emotion conversion, TTS systems, etc., requires careful consideration of prosody [2-10], [14-16], [30-31], [40], [44-45]. A detailed grasp of the language's phonetics, intonation patterns, and other prosodic elements contributing to emotional expression is necessary to develop a prosodic model for emotional utterances in Marathi. Building a unified database, segmenting and annotating it at different levels, examining the acoustic variations in prosodic features for emotional changes, and verifying the acoustic analysis using statistical methods are all crucial steps. However, as noted, Marathi has received less attention in prosodic studies, particularly for emotional speech, than in other languages, highlighting the need for more research in this area [17-22].

The availability of appropriate speech datasets is crucial in designing a prosody model for the TTS system. However, we

found a lack of publicly accessible, high-quality Marathi databases for such studies. Creating a suitable database of skilled Marathi speakers is essential for developing a prosodic model for emotional utterances in Marathi, emphasizing the importance of generating high-quality speech datasets with professional artists' assistance, as has been done for other languages [23-29]. Efforts to collect and create high-quality speech datasets for Marathi speakers can facilitate the development of a natural TTS system, leading to better user experiences and accessibility for Marathi speakers.

Statistical tests such as ANOVA, LMM analysis [18-20], [30-36], [53-55] need to be used for the Marathi language to determine the prosodic features for emotional expression based on neutral speech to convert neutral TTS system into natural or emotional output.

2. Literature Review

2.1. Importance of Prosodic Features

Prosody studies the elements of speech, such as stress, rhythm, and intonation, which reflect various aspects, including emotions, utterance form, and more. Languages influence speech production, resulting in differences in intonation patterns, rhythm, pacing, and temporal adjustments. These elements enable speakers to convey meaning that aligns with their linguistic backgrounds effectively. The speech signal exhibits notable variations across languages, underscoring the complex interplay between language and prosodic features. Prosodic analysis has proven useful in developing models for various languages, driving increased interest in prosody research across diverse linguistic contexts. Acoustic aspects of prosody, including intensity, duration, fundamental frequency (F0) or pitch, and voice quality, can be objectively measured and analyzed contributing to a deeper understanding of prosodic phenomena. Specifically, the fundamental frequency (F0)/pitch, duration, and intensity are crucial in influencing emotional expression in speech. By studying these factors, we can better understand prosody's role in emotional communication.

F0 / Pitch: The pitch period is the length of one glottal cycle, and the reciprocal of the pitch period is the corresponding pitch, also known as fundamental frequency. The fundamental frequency (F0), or F0 contour generated by the larynx during speech production, is a vital prosodic feature of speech, aligning with the pitch contour and representing the perceptual property of vocal tone. Pitch refers to the perceived high or low quality of a person's voice and plays a crucial role in shaping emotional expression. The speaker's modulation of pitch can significantly influence how emotions vary. For example, a higher pitch often indicates excitement, joy, anger, or surprise, while a lower pitch can convey seriousness or sadness. Pitch inflections and fluctuations also convey nuances of emotions, such as sarcasm or irony. Researchers commonly utilize distribution factors such as median, mean, minimum, maximum, and variance values to retrieve pitch or F0 features.

Duration: Duration measures time units that indicate the length of a sentence, word, or other linguistic unit in speech. While speaking, individuals may alter the length of a word or sentence to emphasize an important point. The duration pattern also varies according to speaking styles or languages and hints at the uniqueness of a spoken language sound. The duration of speech elements, such as syllables, words, or pauses, influences the emotional impact of an utterance. Prolonged sounds or elongated pauses can create suspense, anticipation, or emphasis, adding intensity to the emotional message. Shortened durations can convey urgency, excitement, or impatience.

Intensity: Intensity refers to the loudness or strength of a person's voice. Higher intensity levels often indicate anger, enthusiasm, or urgency, while lower intensity levels may convey calmness, sadness, or intimacy. Intensity variations can emphasize certain words or phrases, highlighting their emotional significance. Intensity, which represents the amount of energy in the signal, can be influenced by factors such as vocal stress and the resonance of the vocal tract, both of which may be affected by emotional and physiological effects.

The voice quality is determined by the structure of the vocal apparatus, which is influenced by the way the vocal folds vibrate. To assess voice quality in speech, various parameters like jitter, shimmer, and harmonic signal-to-noise ratio (HNR) are employed. Jitter refers to the frequency fluctuations of the F0 from one cycle to the next, while shimmer indicates variations in the amplitude of the sound wave. HNR provides an estimate of the ratio between periodic and non-periodic elements in voiced speech segments. It's important to note that even if two speakers produce the same pitch, differences in voice quality attributes may cause them to sound as though they are producing different pitches. By scrutinizing these prosodic cues, researchers and technology developers can enhance speech processing systems, including text-to-speech systems, emotion recognition algorithms, and virtual assistants. This progress enables a more profound comprehension and replication of human-like emotional communication.

2.2. Prosody Study for foreign languages

Researchers extensively researched prosody elements in various foreign languages, including French, English, Japanese, Mandarin, German, etc. The authors compared prosody elements in French, explicitly focusing on vowel F0, duration, and intensity, for various emotional utterances [36]. The authors observed anger emotion with more incredible F0, short pauses, high energy values at vowels, and disgust and grief emotions with low energy and short vowel length. Murray et al. [37] found that pitch-related attributes, including pitch contour shape, duration, level, and range, aid in distinguishing primary emotions like happiness, fear, anger, and sorrow. Conversely, voice quality was observed to differentiate secondary emotions such as surprise, grief, sarcasm, and affection. Mareüil et al. [38] investigated several acoustic cues for angry emotion in European languages such as English, French, and Spanish. The authors noticed shorter pauses and longer stressed syllables in English, a fundamental frequency (F0) rise in the final vowel in French, and a shorter and more abrupt F0 rise in Spanish.

Similarly, in their study of Mandarin emotional speech [39], authors observed that speech rate and F0 height were

more closely associated with anger and joy emotions than sorrow. Additionally, anger and sorrowful emotions lengthened than happy emotional utterances. To represent emotional utterances, the authors [40] generated an emotional dataset of Mandarin language recorded by native speakers. The authors performed an acoustic analysis of the perceptually valid database. They found that while fear, disgust, sadness, and neutrality had fewer amplitude variations and low mean f0 values, anger and pleasant surprise had high mean f0 values and more considerable amplitude variations, and pleasure had an intermediate mean f0 value and f0 variation. The authors of [41] created a simulated emotion database in English, Chinese, Tagalog, German, and Japanese to investigate the capacity of western listeners to comprehend speech emotion in western and non-western languages. They found that listeners detected emotional prosody more efficiently in their native language than in another.

2.3. Prosody Study for Indian Languages

The authors [16] examined how emotions conveyed through acoustic prosodic correlates in Hindi speech, including anger, happiness, sadness, fear, and surprise. They constructed a database of fifteen phrases and isolated words uttered by five males and five females, with each speaker repeating the sentences three times. They used the PRATT voice processing software application to evaluate the acoustic parameters that change in response to emotions. Anger had the highest intensity compared to other emotions. Another study [42] investigated twenty emotional voice recordings in German and Telugu languages and analyzed the signal energy and F0 for fear, sadness, anger, and neutral emotions. The authors observed that the average F0 was similar, with relative proportional changes based on emotions for both databases. Still, the mean energy values for Telugu were highest for the fear emotion, whereas the mean energy values for the German language were highest for the sad emotion. Authors D. Pravena and D. Govind created a database in Indian English, Malayalam, and Tamil by imitating emotions [43]. The authors recorded a database simultaneously, capturing speech and EGG signals using an electroglottograph. The excitation characteristics, such as instantaneous SoE and F0, were investigated.

2.4. Prosody Study in the Marathi Language

Studying prosody in the Marathi language has received comparatively little research attention. Some researchers have made efforts to create a small database by collecting data from either non-skilled individuals or a few speakers to analyze prosody features. The authors [13] assessed the acoustic features of the subject-object-verb type Marathi utterance spoken with various focus patterns, including intensity, length, and F0, and to determine the significant effects of both on-focus words and broad focus across the sentences, F0, intensity, and duration features at the subject, object, and verb positions served as the basis. The authors further demonstrated the significance of changes in the prosodic qualities for each word in neutral focus and a specific focus condition using one-way ANOVA analysis. The authors used publicly available radio broadcasts to examine variations in the Marathi news reading style prosodic features [19]. The authors observed the sentence boundaries with pauses, pre-boundary lengthening of the penultimate syllable, and prominence with maximum intensity and F0. The authors [21] discovered that neutral speech transformed into expressive speech by adjusting pitch frequency and time duration. The research focused on changing the basic frequency contour by utilizing a smaller database to obtain acceptable quality and naturalness for the Marathi TTS system. The authors [44] analyzed a database of emotions generated by male and female professional actors. To identify emotional states, the researchers investigated four acoustic features: energy, pitch, vocal tract frequency, and Mel-frequency cepstral coefficients (MFCC). The authors concluded that a specific range of pitch, MFCC, energy values, and vocal tract frequencies for each emotion in the Marathi language helped identify emotions based on the analysis results. In [45], authors focused on analyzing happiness and anger emotions in Marathi speech using three acoustic features: wavelet packet transforms, energy ratios, and MFCC. The researchers found more accurate recognition rates for angry emotions than happy and neutral ones.

2.5. Prosody study for text-to-speech systems

Text-to-speech (TTS) synthesis is a method that allows machines to generate human-like speech from text. Naturalness in synthetic speech enhances communication and user experience, facilitating effective and engaging interactions with technology. Consequently, improving the naturalness of synthetic speech has become a strong focus in speech synthesis research and development. However, to produce speech akin to a native speaker, machines must be capable of synthesizing segmental (word-level) and suprasegmental (prosodic) information. Current text-to-speech (TTS) systems perform well in handling segmental information, such as individual sounds and words. However, they often face challenges regarding suprasegmental aspects, particularly in regional languages. Incorporating a prosody model specific to regional languages poses a significant challenge in the design of TTS systems. Applying appropriate prosodic features can transform neutral speech into emotional speech in TTS systems. Figure 1 illustrates how a typical TTS system can generate emotional speech with the inclusion of suitable prosody variations.

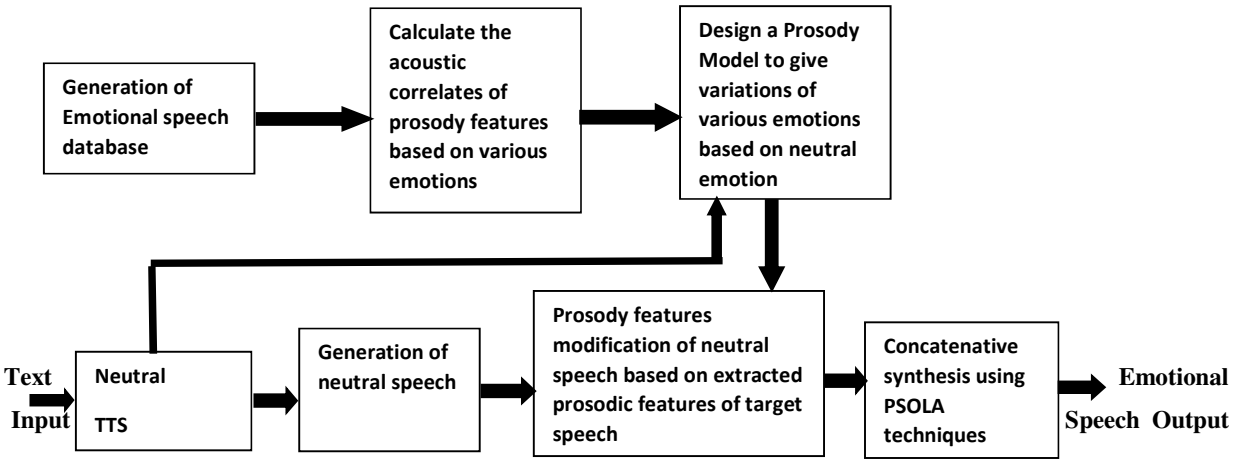


Figure 1: Neutral speech modification based on prosodic feature modification to generate natural text to speech output [3], [4], [5].

Transforming a neutral TTS system into an emotional or natural TTS system includes a series of steps, as illustrated in Figure 1. These steps are outlined below.

Emotional speech database creation: Develop a suitable emotional database that contains speech data representing different emotions.

Acoustic correlates extraction: Extract the acoustic correlates of prosody features in the emotional database for each emotion. The process involves analyzing the variations in pitch, duration, intensity, and other relevant prosodic parameters specific to each emotion.

Prosody model integration: Integrate the prosody model derived from the emotional database into the neutral-sounding TTS system. This integration involves modifying the prosodic characteristics of the neutral sound files based on the extracted prosody features of each emotion. The TTS system can generate speech output that reflects the desired emotions by applying the appropriate prosodic modifications. During the literature review, we observed that TTS systems incorporate similar steps discussed below.

The authors proposed a methodology that involved constructing an emotional speech corpus in Castilian Spanish by appointing eight actors to simulate the seven "basic" emotions [46] for a TTS system. They modelled emotions using acoustic correlates and developed guidelines for speech features related to emotional expressiveness. The authors calculated fundamental frequency (F0) features, including mean F0, F0 range, variability, sound pressure features, and timing parameters. The authors further identified the most crucial features for acoustical emotional modelling and integrated them into a speech synthesis system. In [47], the authors developed an emotional speech synthesis system that transformed neutral speech by modifying its prosodic qualities. They partitioned the emotional styles from the Berlin Emotional Speech (BES) database into word units and calculated prosody features, including energy, pitch, and duration. The researchers developed an emotional-specific prosody model that leveraged these features and employed it to modify the prosody characteristics of neutral speech. The TD-PSOLA approach was subsequently applied to generate the final synthesized emotional speech, ensuring smooth transitions and modified unit boundaries. The authors [48] examined the Assamese speech prosodic features to develop text-to-speech synthesis systems for various emotions. They observed that a high pitch and flattened speech characterized anger, while sadness was associated with increased fundamental frequency. Boredom was expressed through a fast speech tempo, while surprise was marked by low speech rates and increased fundamental frequency. The authors concluded that extracting and comprehending prosodic features is crucial for TTS, highlighting the necessity of a unified database for further research. In a study [49], analysis-synthesis techniques were employed to convert the prosodic features of neutral speech into question and exclamatory speech in Marathi. MATLAB and PRAAT calculated the mean, minimum, and maximum pitch values. By examining the prosodic characteristics differences between the neutral and original dynamic speech, the authors made appropriate modifications to the neutral speech. The authors [50] developed a rule set for various prosodic elements to transform neutral speech into storytelling speech. They compared storyteller utterances with neutral voice-generated utterances to establish diverse pitch contour, intensity, pause, duration, and tempo rule sets. The authors found that incorporating these prosody rule sets into the story-specific TTS system yielded benefits. They recommended studying and modifying the prosody algorithms to enhance the quality of synthetic storytelling speech. This research [52] presented a flexible technique for generating synthetic voices by modeling prosodic elements for different emotions. The authors calculated the prosody features associated with different emotions by estimating changes in emotional prosody features compared to neutral ones. The authors conducted experiments to modify prosody features and evaluated how listeners perceived emotions. The study highlighted the significance of a prosody model and an emotional speech corpus in developing synthesized systems.

2.6. The Details of the Existing Databases

A database in regional languages is vital for designing text-to-speech systems that accurately capture linguistic, emotional, and cultural aspects. It enhances naturalness, authenticity, and inclusivity in synthesized speech output, ultimately improving the overall quality and effectiveness of the systems. The speech processing community created several databases with the assistance of trained actors often used in research work. Regional languages possess unique phonetic, prosodic, and expressive characteristics, which researchers can accurately capture through a database for improved synthesis.

Developing a comprehensive database in regional languages contributes to the availability of linguistic resources for research and development purposes, benefiting various language processing applications. Customization and personalization of TTS systems become possible with a regional language database, allowing tailored synthesis for specific dialects, accents, and speaking styles. It facilitates overcoming language barriers and ensures equitable access to information and communication for everyone. A high-quality and natural database is critical in speech processing systems as it can impact the accuracy and reliability of the system's results. The researchers created emotional databases in different languages for research purposes, and many of these databases involve recordings of trained actors imitating speech. Table 1 provides an overview of various speech databases, including information on language, number of speakers, databases, and various emotions included in each database.

Table 1 The Details of The Foreign and Indian Languages Database

Language	**No. of Speakers**	**Database**	**Emotions**
German (2005) [23]	Five Males +five Females	800 sentences	Seven emotions
Spanish (2008) [24]	an actress and an actor	60 sentences	Six emotions
Japanese (2020) [25]	Three female professionals	100 sentences	neutral, angry, and happy
British (2014) [26]	Four male actors	480 British English utterances	the seven basic emotions
Mexican (2021), [27]	A female, a male, and a child	864 voice recordings	anger, fear, happiness, disgust, neutral, and sadness
Arabic (2021) [28]	twenty-three speakers	1596 audio files	neutral, happiness, sadness, surprise, and anger
Bangla (2021) [29]	ten males and ten females	7000 utterances,	Seven emotions
Hindi (2019) [16]	five males and five females	fifteen phrases and isolated words	anger, fear, happiness, sadness, surprise
Telugu (2013), [53]	five males and five females	Telugu emotion speech corpus	Happiness, Sarcasm, Anger, Compassion Neutral, Surprise, Disgust, Fear.
Malayalam (2016) [54]	ten females and two male actors	100 words and ten sentences	happy, angry, sad, and neutral

Punjabi (2021) [55]	58 men and 62 women	five Punjabi lines	happiness, sadness, anger, fear, and neutral
Gujarati (2020) [56]	Six males and three females	1,296 words	anger, disgust, sadness, surprise, fear, and happiness
Odia (2016) [57]	18 speakers	60 utterances	anger, fear, happiness, disgust, sadness, surprise
Sanskrit (2018) [58]	four males and four females	160 Sanskrit sentences	fear, anger, disgust, happiness, sadness, excited

Table 2 provides a detailed overview of a Marathi speech database, including the number of speakers who generated the database, the recording procedure, the number of databases generated, the prosodic features studied with corresponding results, and any gaps in the database.

Table 2 A Marathi Database's Details with Prosody Study and Related Results and Gap

Reference	**Speakers**	**Prosody Studied**	**Recording procedure**	**Database**	**Results**	**Gap**
IJISET, 2016, [49]	one male and one female	Pitch features	Audacity sound editor	500 sentences	Question mark sentence with high pitch, the words before punctuation mark with a high pitch. Emotional speech duration is less.	Less no. of speakers non-trained speakers Only the pitch feature analyzed
Journal of Phonetics, 2017, [13]	Six males and six females	intensity, duration, and F0	Praat recorder	168 utterances	Focused words observed by higher mean F0, increased duration, and higher intensity	Non-trained speakers Linguistic prosody analysis
TENCON 2016, [19]	Several female speakers	intensity, and F0	Prasar Bharati radio station	Ten sentences	The maximum intensity and maximum F0 observed for prominence	Analysis of news reading speech
I.J. Speech Tech. 2013, [45]	one male and one female	wavelet packet transform, energy ratios, and MFCC	Audacity sound editor	500 sentences	More recognition rates for angry emotions followed by happy and neutral emotions	There is a need to generate more databases for various emotions from trained Marathi speakers.

2.7. Statistical Analysis of prosodic features

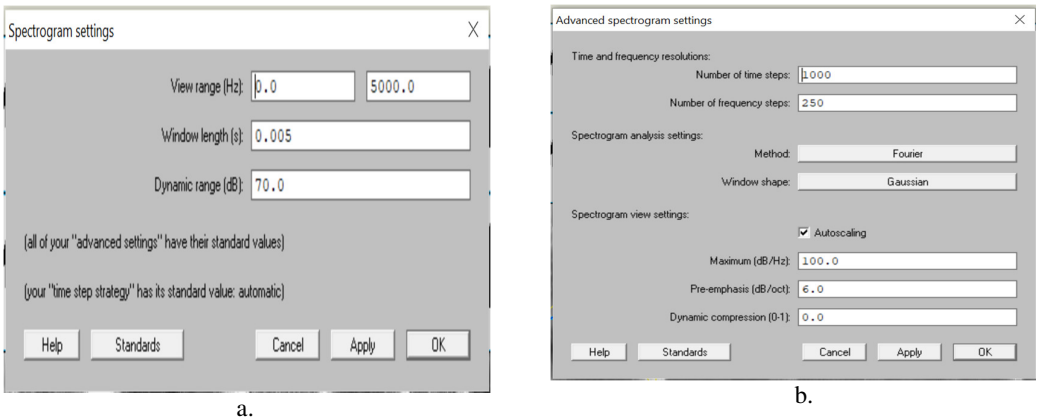
Statistical studies play a crucial role in analyzing variations in prosodic features in emotional speech and establishing standards for categorizing emotions based on prosody. One-way ANOVA analysis assesses the differences in prosodic features across various emotional categories. By comparing average values of prosodic features among different emotion groups, ANOVA analysis helps identify significant variations in prosody. This comprehensive analysis provides valuable insights into the influence of emotions on prosody. However, one-way ANOVA cannot incorporate additional factors, such as gender or individual speakers, into the analysis. A two-way ANOVA extends the analysis by considering two independent variables, emotions, and gender, to examine their individual and interactive effects on prosodic features. In addition, using linear mixed modeling analysis allows for investigating variations in prosodic features, accounting for fixed effects (emotions and gender) and random effects (individual speakers and sentences). This comprehensive approach captures systematic variations and individual-specific variations in the prosodic features. Table 3 provides an overview of the statistical analyses performed for various languages worldwide and the impact of such analyses on the calculation of prosodic features.

Table 3 Statistical Methods for Analyzing Statistical Results

References	**Language**	**Statistical Analysis Used**	**Impact**
2013[30]	English	Analysis of variance	Analysis of formant features for seven emotions generated by ten female speakers.
2010 [31]	German	Bivariate Correlation technique	From a database of seven emotions, applying the Bivariate Correlation technique, authors selected 35 prosodic components out of 133 for further analysis.
2018 [35]	Arabic	two-way ANOVA, chi-square, Tukey tests, etc.	Analyzed behavior of five emotions.
2015. [36]	English	LMM	The authors analyzed F0, HNR, pitch frequency, amplitude ratio, and first formant bandwidth of 80 acts and 80 radio interviews.
2017 [60]	Indonesian	linear mixed effect models	The authors analyzed how the emotions, subject, gender, and word affected amplitude, duration, and dwt coefficient.
2018 [61]	Japanese	mixed-effects regression model	prosodic indicators for the level of formality for F0, intensity, pause frequency, and articulation rate features.
2018 [62]	German	linear mixed-effect analyses	The authors calculated joy and sadness prosodic features considering language as a fixed effect variable and persons as a random effect variable.

3. Methodology and Implantation

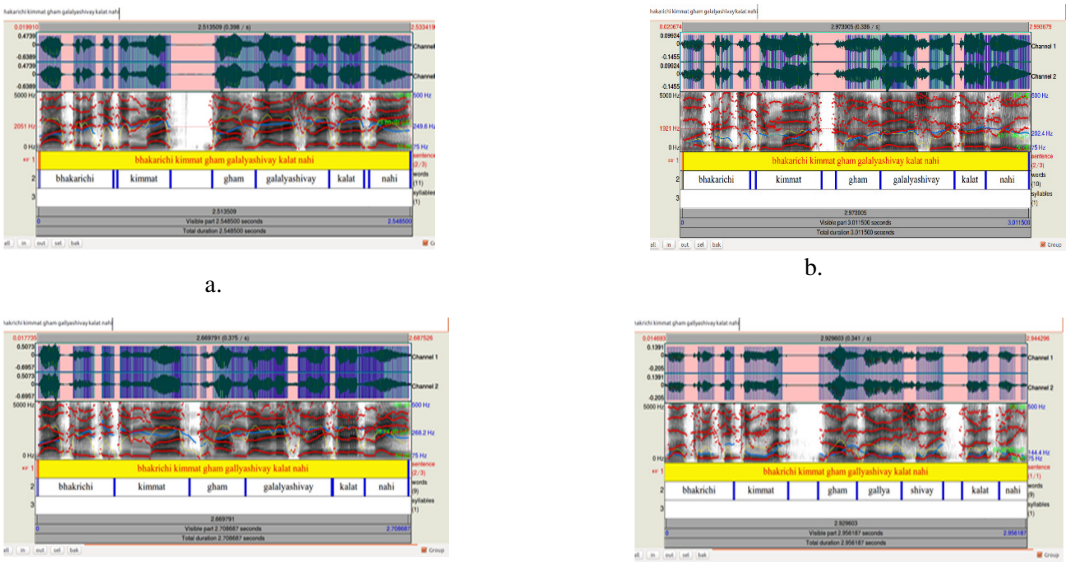
We collected a Marathi speech dataset featuring utterances expressing happiness, fear, anger, and neutral emotions from trained Marathi speakers. From seventeen candidates, we selected four female and seven male speakers with acting experience in drama or television. We chose the ten versatile and unbiased Marathi sentences, suitable for expressing various emotions. Over two to three meetings, we discussed research objectives with participants, presenting selected sentences and outlining expected outcomes. Based on participant suggestions, we modified two sentences. Each of the ten sentences was recorded with anger, fear, happiness, and neutral emotions. Sound experts reviewed recordings for errors or background noise. This process generated a database of 440 utterances across four emotion styles. Following recordings, emotional style verification was conducted through listening tests with thirty participants. Emotional utterances with a 75% recognition rate for the relevant emotions were selected for further analysis. Selected sound files were annotated into sentences and words using Praat [63], with spectrogram settings set to display frequency range to 0-5000 Hz and pitch frequency to 75-500 Hz. Spectrogram analysis utilized the Fourier method and Gaussian window shape, with standard autoscaling with maximum 100 dB/Hz and a pre-emphasis of 6 dB per octave. Figure 2 illustrates spectrogram details 2a and advanced spectrogram settings 2b.



a. b.

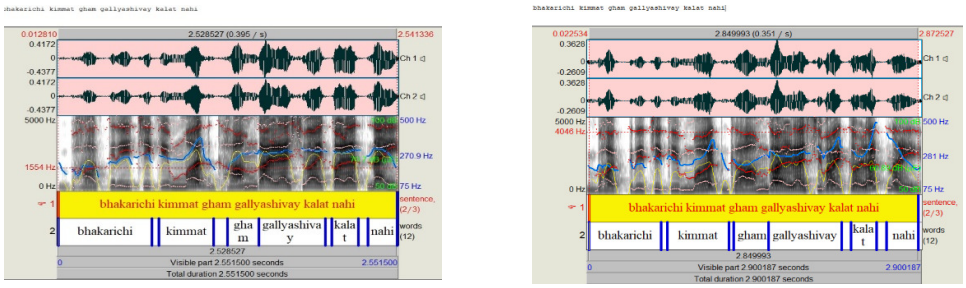
Figure 2 Diagram showing a. Spectrogram and b. Advanced Spectrogram Settings

To minimize segmentation errors arising from artificial segmentation, both segmentation and annotation were carried out manually. The utterances were annotated by examining the spectrogram, listening to the corresponding speech segment, and then determining the boundaries of a complete sentence or its constituent words. Figure 3 showcases the sentence and word-level annotation depicting a female speaker's expression across all four emotions for the sentence "bhakarichi kimmat gham gallyashivay kalat nahi" (meaning "you will realize the value of food whenever you earn"), while figure 4 presents a similar depiction for a male speaker.

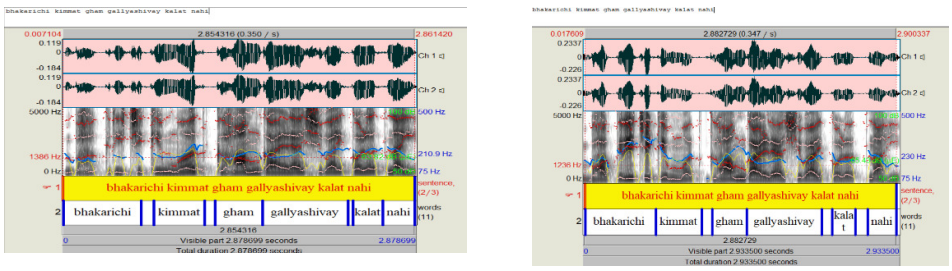


a. b. c. d.

Figure 3. Annotation of the femail speaker's statement "bhakarichi kimmat gham gallyashivay kalat nahi" in PRAAT in the following emotions: a. fear, b. happiness, c. anger, and d. neutral



a. b.



c. d.

Figure 4. Annotation of the male speaker's statement "bhakarichi kimmat gham gallyashivay kalat nahi" in PRAAT in the following emotions: a. fear, b. happiness, c. anger, and d. neutral

Figures 3 and 4 demonstrate how the spectrogram displays differences in intensity contour, pitch contour, and formants among all sentences and words, resulting from variations in both emotion and gender. To detect variances in prosodic characteristics linked to emotions, we utilized PRAAT to calculate twelve prosodic attributes related to pitch, intensity, duration, and voice quality.

3.1. Calculating Pitch-related Acoustic Features

We calculated all the pitch-related features by viewing the pitch contour by selecting a sentence from PRAAT. We calculated mean pitch values from the "Get pitch" command from the Pitch menu in the sound editor and text grid editor window. The minimum pitch values were calculated by selecting a sentence and "Get minimum pitch" from the Pitch menu in the sound editor and text grid editor window. The maximum pitch values were calculated by selecting a sentence and "Get maximum pitch" from the Pitch menu in the sound editor and text grid editor window.

3.2. Calculating Intensity-related Acoustic Features

Intensity-related features were computed in PRAAT by analyzing a selected sentence and examining the intensity contour. We calculated the mean intensity values (in dB) within a specific time range using the "Get intensity" command in the sound editor and text grid editor window from the intensity menu. The dB averaging technique was utilized to calculate the average intensity within a specific segment, yielding the mean of the intensity curve in dB. In this method, the mean intensity between times t_1 and t_2 is determined as

$$\text{Mean intensity} = \frac{1}{(t2-t1)}\int_{t1}^{t2} x(t)dt \qquad 1$$

Furthermore, the minimum intensity values within the given time range were determined using the parabolic interpolation method and expressed in dB by selecting a sentence "Get minimum intensity" from the intensity menu in the sound and text grid editor windows. The maximum intensity values within the specified time domain, expressed in dB, were calculated by selecting a sentence and "Get maximum intensity" from the intensity menu in the sound editor and text grid editor window.

3.3. Calculating Duration-related Acoustic Features

We calculated the duration-related acoustic features by selecting a sentence and the "Time range of SELECTION" command from the Pulses menu. We calculated the number of voice breaks and syllables by selecting the command "Number of voice breaks" and “Number of syllables per second” from the Pulses- voicing menu.

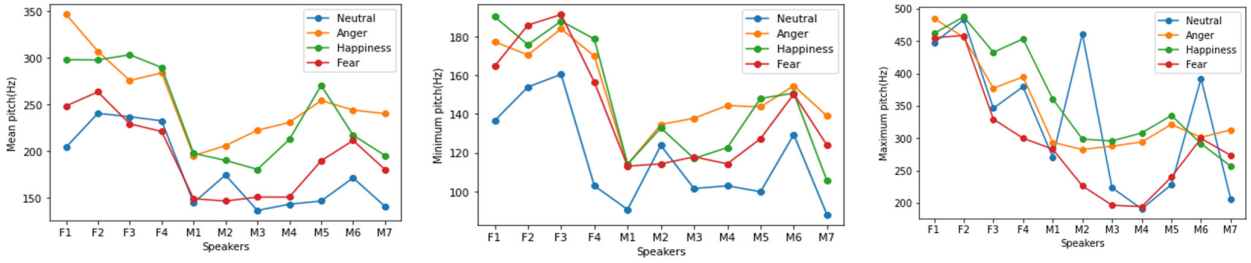
3.4. Calculating voice quality-related Features

We calculated the voice quality -related acoustic features such as jitter, shimmer, and H/N ratio from the Pulses- voicing menu.

4. Results and discussions

This section discusses the acoustic analysis of various prosodic features for different emotions in Marathi.

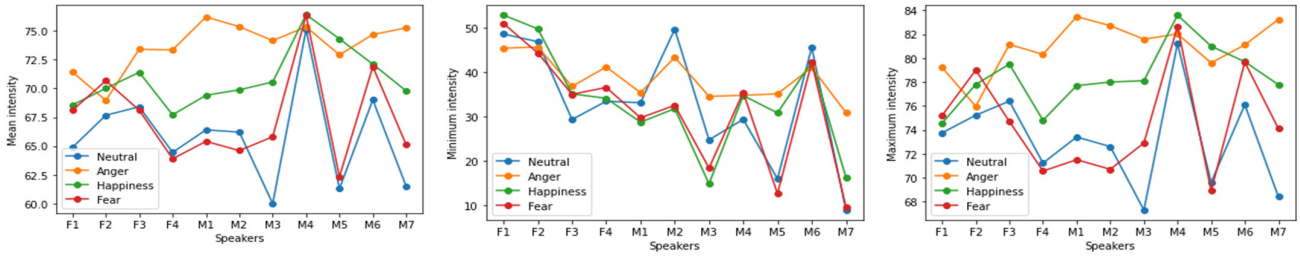
Figure 5 presents the line graph for the pitch-related features, such as mean pitch (measured in Hz), minimum pitch (Hz), and maximum pitch (Hz), across neutral, anger, happiness, and fear emotions for four female speakers and seven male speakers [database refered from [66]]. The figure 5 highlights significant variations in pitch features between male and female speakers, with males generally exhibiting lower pitch values across all emotions compared to females. Analysis of recordings from four female and seven male speakers revealed diverse responses, with considerable individual variances observed within both gender classes. Some individuals displayed relatively higher gender-specific values, while others exhibited lower values. These individual differences contribute to random effects on emotional shifts in prosodic characteristics.



a. b. c.

Figure 5. Line graphs depicting Mean Pitch, Minimum Pitch, and Maximum Pitch Variations among the Speakers for Neutral, Anger, Happiness, and Fear Emotions

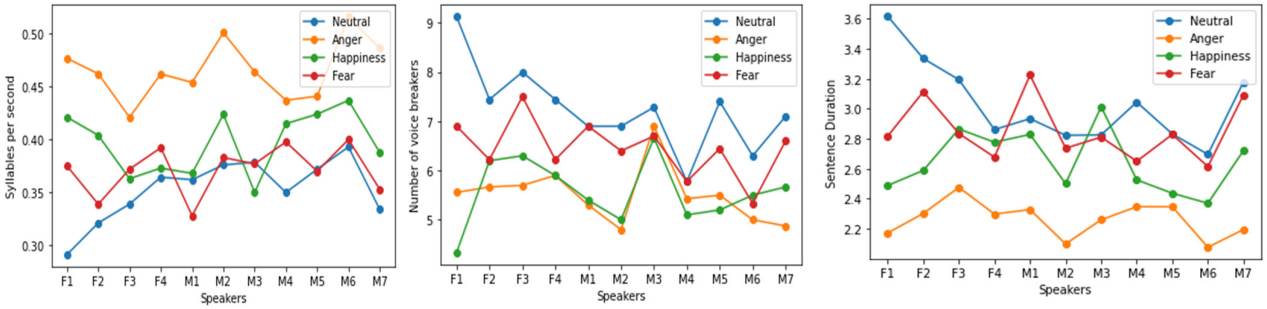
Figure 6 illustrates line graph of the diversity of intensity-related attributes, including mean intensity (measured in dB), minimum intensity (dB), and maximum intensity (dB), across neutral, anger, happiness, and fear emotions [database referred from [66]]. Analysis of intensity-related readings from both female and male speakers reveal that, in comparison to other emotions, anger consistently exhibits elevated intensity levels. The recordings of four female and seven male speakers indicate minimal variability in all intensity-related features.



a. b. c.

Figure 6. Mean Intensity, Minimum Intensity, and Maximum Intensity Variations between the Speakers for Anger, Happiness, Fear, and Neutral Emotions

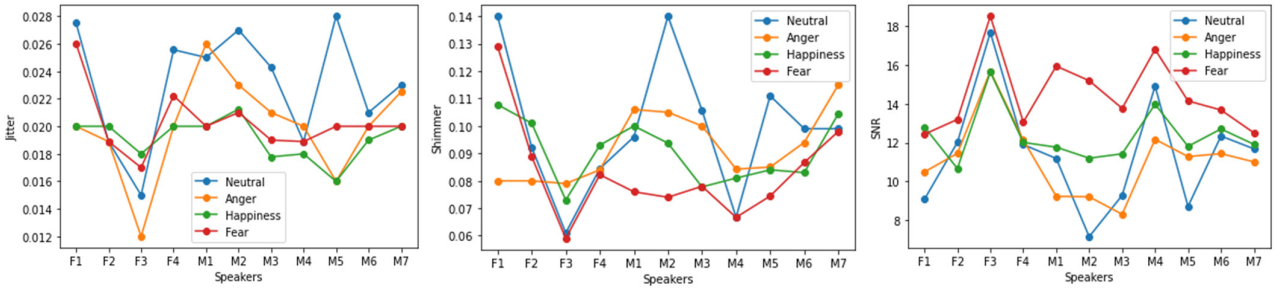
Figure 7 presents the distinctions in syllables per second, voice breaks, and sentence duration (in seconds) across all four emotions conveyed by the four female and seven male speakers [66]. Across the emotions, neutral expression exhibited the lowest syllables per second and the longest sentences. Happy and angry emotions displayed a comparable frequency of voice breaks, whereas the angry emotion demonstrated the shortest sentence duration and the highest syllables per second count.



a. b. c.

Figure 7. Syllables per Second, Number of Voice Breaks, Sentence Duration (Sec.) Variations between the Speakers for Anger, Happiness, Fear, and Neutral Emotions.

Figure 8 illustrates the voice quality metrics, including jitter, shimmer, and Harmonics-to-Noise Ratio (HNR) [66]. In the context of happy emotion, fewer instances of jitter were observed compared to other emotions, except for female speakers F2 and F3, and male speakers M4 and M6. Similarly, fear emotion exhibited fewer shimmer values compared to other emotions, except for female speaker F1 and male speaker M7. The HNR for fear emotion was relatively high, while for neutral emotion, it was low compared to anger and happiness emotions.



a. b. c.

Figure 8. The Voice Quality Measures in Terms of Jitter, Shimmer, and HNR for Anger, Happiness, Fear, and Neutral Emotions.

Gender-based differences in these acoustic features demonstrate significant variations in pitch values, notable variations in intensity values, and minimal differences in duration and voice quality features. Further, we have carried out separate female and male analysis for mean pitch, mean intensity, and sentence duration. Figure 9 represents separate female and male mean pitch analysis.

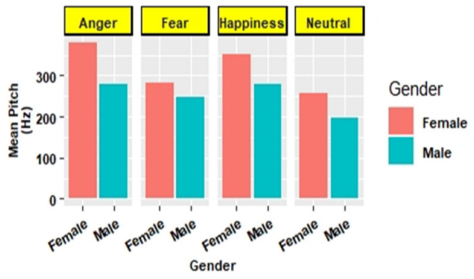


Figure 9: Separate mean pitch analysis for female and male speakers

Female mean pitch values observed high comparative to male mean pitch for all the four emotions. Males often have a lower pitch than females due to their longer and more muscular vocal cords (Male:17cm and female: 15cm long).

Figure 10 represents separate female and male mean intensity analysis.

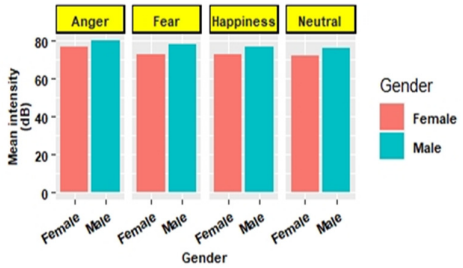


Figure 10: Separate mean intensity analysis for female and male speakers

The mean intensity of male speaker observed higher than female speakers for all the emotions. When we add stress to a speech sound, such as during anger or happiness emotions, we raise the pressure within the lungs, which creates a rise in subglottal pressure, and causes an increase in loudness or intensity for anger. During regular spoken the pressure produced immediately below the glottis, also known as subglottal pressure, is often consistent, as in neutral reading style.

Figure 11 represents separate female and male utterance duration analysis. The utterance duration for male and female speakers for anger and happiness emotions observed to be similar. The utterance duration for fear emotion is observed to be more for male speakers than female speakers whereas the utterance duration for neutral emotion is more for female speakers than male speakers.

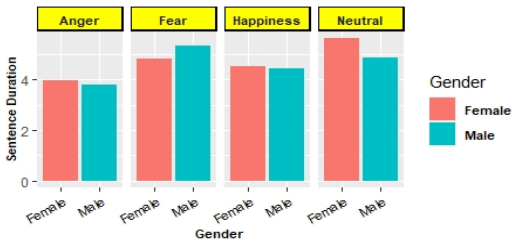


Figure 11: Separate utterance duration for female and male speakers

We compared the acoustic correlates of Hindi language [16] with our analysis of the acoustic correlates of the prosodic features of Marathi language. Despite both being Devanagari languages, Hindi and Marathi exhibit variations in prosodic features based on speaking style. In Hindi, fear has the lowest mean pitch, while neutral has the highest, following the sequence of fear, happiness, anger, and neutral. In Marathi, neutral has the lowest mean pitch, and anger has the highest, following the sequence of neutral, fear, happiness, and anger. The mean intensity in Hindi follows an ascending order: fear, neutral, anger, and happiness, with fear having the lowest intensity and happiness having the highest. In Marathi, the order for mean intensity is neutral, fear, happiness, and anger, with neutral having the lowest intensity and anger having the highest. In Hindi, utterance duration shows an ascending order: anger, neutral, happiness, fear, with the lowest duration for anger and the highest for fear. In contrast, in Marathi, the ascending order for duration is: anger, happiness, fear, neutral, with the lowest duration for anger and the highest for neutral. Across nearly all the languages examined, variations in acoustic correlates of prosodic features were observed across emotions. We examined the prosodic features of English, Mandarin, German, Odia, and Punjabi languages and identified variations in the prosodic features specific to the Marathi language in Table 4. The first column provides details of the research paper, including its title, journal or publication name, publication year, the language studied, and the emotions investigated. The second column presents the variations of these prosodic features based on the emotions studied in each language. Finally, the third column highlights the comparing these findings with the emotional signaling observed in Marathi.

Table 4: The variations in acoustic features related to different emotions between the Marathi language and English, Odia, German, and Mandarin Chinese languages

Publication Year and References No.	**Language Studied:**	**Emotions Studied:**	**Emotional Signalling**	**Comparing with Marathi Emotional Signalling**
2004, [64]	English	sadness, anger, happiness. neutral	Anger–higher pitch, longer utterance duration, shorter pauses, wider range energy values. Mean F0 is higher than neutral for sadness, anger, happiness emotions	Anger–higher pitch, shorter utterance duration, shorter pauses, high energy values. The mean F0 (pitch) of anger and happiness is higher than neutral emotions. Mean F0 of neutral and fear emotions have similar values.
2016, [65]	German	Criticism, doubt, naming, suggestion, warning, wish	warning stimuli- arched pitch contour Naming stimuli - low intensity, flat pitch contour	Anger- higher pitch, intensity, and short duration values than happiness, fear and neutral emotions
2012, [40]	Mandarin Chinese	anger, disgust, fear, sadness, happiness, pleasant surprise, and neutrality	Anger and pleasant surprise displayed higher mean fundamental frequency (f0) values with significant variations in both f0 and amplitude. Sadness, disgust, fear, and neutrality exhibited lower mean f0 values with minimal amplitude variations, Happiness showed a moderate mean f0 value and f0 variation.	Mean f0, mean intensity of anger and happiness observed to be relatively high than fear and neutral, The mean intensity of anger emotion was observed with fewer variations for all the male and female speakers than for other emotions.
2016, [57]	Odia	anger, fear, happiness, disgust, sadness, surprise.	Happy- the highest value of mean pitch in both males and females. Disgust and fear- higher energy values than all other emotions. Female participants displayed consistent energy levels across various emotions	Anger- the highest value of mean pitch in both male and females. Anger and happiness- higher energy values than fear and neutral.
2018, [48]	Punjabi	happiness, sadness, anger, fear, and neutrality in Punjabi sentences.	fear exhibited the lowest intensity feature, anger had the highest intensity, highest pitch characterized the happy emotion, while the sad emotion exhibited the lowest pitch.	Anger characterized with highest pitch, intensity, and shortest utterance duration Happy characterized with lower pitch than anger emotion Fear characterized with low pitch, low intensity compared to happy and anger emotions.
2013, [53]	Telugu	anger, disgust, fear, compassion, neutral, happiness, sarcasm, surprise	Anger emotion with the highest energy Anger, happiness and neutral have high pitch values	Anger characterized with highest pitch, intensity, and shortest utterance duration Happy characterized with lower pitch than anger emotion Neutral emotion has lowest pitch values.

5. Conclusion

The integration of prosodic features, including emphasis, rhythm, intonation, and other speech attributes, into machine-generated speech is pivotal for enhancing communication, particularly in local languages like Marathi. However, the scarcity of research and publicly accessible high-quality datasets for Marathi presents a significant challenge. To address this gap, we meticulously collected a Marathi speech dataset featuring expressions of various emotions from skilled Marathi speakers with acting experience.

In our investigation, we utilized PRAAT to compute twelve prosodic attributes linked to pitch, intensity, duration, and voice quality, with the objective of uncovering variations in prosodic traits associated with emotions. Anger exhibited the highest pitch and intensity, average voice quality, and the shortest duration among the emotions studied. Happy emotions were distinguished by elevated pitch and intensity, brief duration, and moderate voice quality, while fearful emotions were characterized by low pitch and intensity, moderate voice quality, and extended duration. These characteristics were compared to those of neutral emotions. The findings from this cues-based acoustic analysis could inform the development of a prosody model tailored for the Marathi language.

Our analysis revealed notable gender-based variations in pitch, intensity, duration, and voice quality metrics across different emotions. The findings underscore the importance of tailored research efforts and dataset creation to develop natural Text-to-Speech (TTS) systems that cater to Marathi speakers, ultimately enhancing user experiences and accessibility. Furthermore, our study highlights the need for continued exploration and investment in prosodic studies, particularly in underrepresented languages like Marathi, to promote linguistic diversity and inclusivity in technology-driven communication platforms. Studying emotional signalling through prosodic features across different Indian and Western languages like Hindi, English, Mandarin, German, Odia, and Punjabi, along with Marathi, uncovered notable differences in conveying emotions. These findings underscore the influence of linguistic backgrounds and speaking styles on prosody, emphasizing the necessity for a thorough examination of behavioural patterns in prosodic features to construct language-specific prosody models.

Future research can examine additional emotions including surprise, disgust, sorrow, etc. in a similar manner. To learn more specifically about the prosodic differences based on different emotions, word and syllable level analysis can also be done.

References

[1] H. Fujisaki, "Prosody, Models, and Spontaneous Speech," in Y. Sagisaka, N. Campbell, and N. Higuchi (eds.), *Computing Prosody, Springer*, New York, 1997.

[2] J. Prakash and H. Murthy, "Analysis of Inter-Pausal Units in Indian Languages and Its Application to Text-to-Speech Synthesis," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1616-1628, Oct. 2019, doi: 10.1109/TASLP.2019.2924534.

[3] S. Kayate, "Text-To-Speech Synthesis System for Marathi Language Using Concatenation Technique," Ph.D. Thesis, Dr. Babasaheb Ambedkar Marathwada University, 2018.

[4] P. Sarkar, A. Haque, A. K. Dutta, G. R. M., H. D. M., P. Dhara, R. Verma, N. P. Narendra, S. K. S. B., J. Yadav, and K. S. Rao, "Designing Prosody Rule-set for Converting Neutral TTS Speech to Storytelling Style Speech for Indian Languages Bengali, Hindi, and Telugu," in *Proceedings of the IEEE International Conference on Contemporary Computing (IC3 2014)*, 2014.

[5] L. He, H. Huang, and M. Lech, "Emotional Speech Synthesis Based on Prosodic Feature Modification", Engineering, 2013. doi: 10.4236/eng.2013.510B015.

[6] E. Zovato, "Towards emotional speech synthesis:a rule-based approach," in Interspeech, SSW5, 2004, pp. 219-220.

[7] F. Niu and W. Silamu, "Prosody-Enhanced Mandarin Text-to-Speech System," 3rd International Conference on Advances in Computer Technology, Information Science and Communication (CTISC), pp. 67-71, 2021.

[8] S. Raptis, "Towards Expressive Speech Synthesis: Analysis and Modeling of Expressive Speech," in 5th IEEE International Conference on Cognitive Info. Communications, Nov.5-7, 2014, Italy.

[9] D. Xin, S. Adavanne, F. Ang, A. Kulkarni, S. Takamichi and H. Saruwatari, "Improving Speech Prosody of Audiobook Text-To-Speech Synthesis with Acousticand Textual Contexts," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp.1-5, doi: 10.1109/ICASSP49357.2023.10096247.

[10] D. Jiang, W. Zhang, L. Shen, and L. Cai, "Prosody Analysis and Modeling for Emotional Speech Synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, 2005, pp. I/281-I/284.

[11] G. Pamisetty, S.Varun and K. Murty, "Lightweight Prosody-TTS for Multi-Lingual Multi-Speaker Scenario," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-2, doi: 10.1109/ICASSP49357.2023.10095839.

[12] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for storytelling applications," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1137-1144, July 2006, doi: 10.1109/TASL.2006.876129.

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

[14] T. Wani, T. Gunawan, S. Qadri, and M. Kartiwi, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47795–47814, April 2021.

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

[16] Bansal, Agrawal, and Kumar 2019, "Acoustic analysis and perception of emotions in Hindi speech using words and sentences," *Int. J. Inf. Technology*, vol. 11, pp. 807–812, 2019, DOI: 10.1007/s41870-017-0081-0.

[17] "Marathi people," Ethnologue, 22nd ed., 2019.

[18] https://www.meity.gov.in/content/technology-development-indian-languages-tdil

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

[20] S. Kshirsagar, "Determination of Acoustic Parameters of Marathi Prosody," M.E. thesis, Mumbai University, India, 2016.

[21] C. Manjare and S. Shirbahadurkar, "Pitch and duration modification for expressive speech synthesis in Marathi TTS system," *in International Conference on Pervasive Computing (ICPC)*, Pune, 2015, pp. 1-4.

[22] M. C. Madhavi, S. Sharma, and H. A. Patil, "Development of language resources for speech application in Gujarati and Marathi," in *International Conference on Asian Language Processing (IALP)*, Kuching, 2014, pp. 115-118[

23] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 1517-1520.

[24] R. Barra-Chicote, J. Montero, and J. Macias-Guarasa, "Spanish Expressive Voices:Corpus for emotion research in Spanish," in *Proc. of 6th international conference on Language Resources and Evaluation*, 2008.

[25] S. Takamichi, M. Komachi, N. Tanji, and H. Saruwatari," JSSS: free Japanese speech corpus for summarization and simplification," 2020.

[26] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (SAVEE) database,"University of Surrey: Guildford, UK, 2014.

[27] M. M. Duville, L. M. Alonso-Valerdi, and D. I. Ibarra-Zarate, "The MexicanEmotional Speech Database (MESD): elaboration and assessment based on machine learning," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 1644-1647.

[28] A. Meftah, M. Qamhan, Y. Seddiq, Y. Alotaibi, and S. Selouani, "King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison,"in IEEE Access, vol. 9, pp. 54201-54219, 2021, doi: 10.1109/ACCESS.2021.3070751.

[29] S. Sultana, M. Rahman, M. Selim, and M. Iqbal, "SUST Bangla Emotional Speech Corpus (SUBESCO) An audio-only emotional speech corpus for Bangla," *PLoS ONE*, vol. 16, no. 4, e0250173, 2021.

[30] A. Jacob and P. Mythili, "Upgrading the Performance of Speech Emotion Recognitionat the Segmental Level," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol.15, no. 3, pp. 48-52, 2013.

[31] T. Iliou and C. Anagnostopoulos, "Classification on Speech Emotion Recognition – A Comparative Study," *International Journal on Advances in Life Sciences*, vol. 2, no.1& 2, 2010.

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

[33] M. Yusnita A, M. P. Paulraj, S. Yaacobb, N. Fadzilah, and A. Shahriman,"Acoustic Analysis of Formants across Genders and Ethnical Accents in Malaysian English using ANOVA," in International 57Conference on Design and Manufacturing, vol. 64, pp. 385– 394, 2013.

[34] L. Hui, L. Ting, S. See, and P. Chan, "Use of electroglottograph (EGG) to find a relationship between pitch, emotion, and personality," in International Conference on Applied Human Factors and Ergonomics, 2015.

[35] A. Meftah, Y. Alotaibi, and A. Selouani, "Evaluation of an Arabic Speech Corpus of Emotions: A Perceptual and Statistical Analysis," in *IEEE Access*, vol. 1, pp. 1-1, 2018.

[36] J. Rebecca, A. Grass, M. Drolet, and J. Fischer, "Effect of Acting Experience on Emotion Expression and Recognition in Voice: Non-Actors Provide Better Stimuli than expected," *Journal of Nonverbal Behavior*, vol. 39, 2015.

[37] I. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097-1108, 1993.

[38] P. Boula de Mareüil et al., "Generation of Emotions by a Morphing Technique in English, French, and Spanish," *in Proc. in Speech Prosody*, 2002, pp. 187-190.

[39] Lin, HSin-Yi,and Janice Fon, "Prosodic and Acoustic Features of Emotional Speech in Taiwan Mandarin," *Proceedings of the International Conference on Speech Prosody*, 2012, pp.450-453.

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

[41] W. Forde Thompson and L. Balkwill, "Decoding speech prosody in five languages," *Semiotica*, vol. 158, pp. 407–424, 2006.

[42] Ramdinmawii and Mittal, "Emotional speech discrimination using sub-segmental acoustic features," TEL –NET, pp. 1-7, 2017. DOI: 10.1109/tel-net.2017.8343515.

[43] D. Pravena and Govind, "Development of simulated emotion speech database for excitation source analysis," *Int J Speech Technol*, vol. 20, pp. 327–338, 2017.

[44] Darekar and Dhande, "Toward Improved Performance of Emotion Detection: Multimodal Approach," *Proceedings of the International Conference on Data Engineering and Communication Technology, Advances in Intelligent Systems and Computing*, 2016, pp. 431-443.

[45] A. Degaonkar and R. Apte, "Emotion modeling from speech signal based on wavelet packet transform," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 1–5, 2013.

[46] S. Ignasi, G. Roger, et al., "Validation of an Acoustical Modelling of Emotional Expression in Spanish Using Speech Synthesis Techniques," in ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Newcastle, Northern Ireland, UK, 2000.

[47] Jianhua Tao, Yongguo Kang, and Aijun Li, "Prosody Conversion from Neutral Speech to Emotional Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp.1145-1154, 2006.

[48] K. Jasdeep, S. Vishal, "Role of Acoustic Cues in Conveying Emotion in Speech,"*Journal of Forensic Sci & Criminal Inves.*, vol. 11(1), pp.1-6, 2018.

[49] S. S. Patil and S. D. Apte, "Prosody Conversion from Neutral Speech to Emotional Speech," *IJISET – International Journal of Innovative Science, Engineering & Technology*, vol. 3, issue 2, February 2016.

[50] R. Jia, L. Gang, and P. Y.-P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Information Processing & Management*, vol. 45, no. 3, pp. 315-328, 2009.

[51] Bartkova, Jouvet and Roussarie, "Prosodic Parameters and Prosodic Structures of French Emotional Data," Speech Prosody, Boston, USA, 2016.

[52] Chul Min Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," in *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293-303, March 2005, doi: 10.1109/TSA.2004.838534.

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

[54] T.M. Rajisha, A.P. Sunija, and K.S. Riyas, "Performance Analysis of Malayalam Language Speech Emotion Recognition System Using ANN/SVM," *Procedia Technology*, vol. 24, pp. 1097-1104, 2016.

[55] K. Kaur and P. Singh, "Punjabi Emotional Speech Database: Design, Recording and Verification," *Int. J. Intell Syst Appl Eng*, vol. 9, no. 4, pp. 205–208, Dec. 2021.

[56] V. Tank, "Creation of speech corpus for emotion analysis in Gujarati language and its evaluation by various speech parameters," Int. Journal of Electrical and Computer Engineering, pp. 4752-4758, 2020.

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

[58] S. Kakodkar and S. Borkar, "Speech Emotion Recognition of Sanskrit Language using Machine Learning," *International Journal of Computer Applications*, vol. 179, pp. 23-28, 2018.

[59] B. Aarti, and S. Kopparapu, "Spoken Indian language identification: a review of features and Databases," *Sādhanā*, pp. 43-53, 2018.

[60] N. Wunarso and Y. Soelistio, "Towards Indonesian Speech-Emotion Automatic Recognition," 4th International Conference on New Media (CONMEDIA 2017), pp. 1-6, 2017.

[61] E. Sherr-Ziarko, "Prosodic properties of formality in conversational Japanese," Journal of the International Phonetic Association, pp. 1-22, 2018.

[62] M. Kraxenberger, W. Menninghaus, A. Roth, and M. Scharinger, "Prosody-Based Sound-Emotion Associations in Poetry," Front. Psychol., vol. 9, pp. 1284, 2018.

[63] P. Boersma, "PRAAT, a system for doing phonetics by computer," Glot International 5, pp. 341-345, 2001.

[64] Yildirim, Serdar & Bulut, Murtaza & Lee, Chul & Kazemzadeh, Abe & Lee, Sungbok & Narayanan, Shrikanth & Busso, Carlos, "An acoustic study of emotions expressed in speech," 10.21437/Interspeech, pp. 242, 2004.

[65] N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70-86, 2016.

[66] T. Harhare and M. Shah, "Analysis of Acoustic Correlates of Marathi Prosodic Features for Human-Machine Interaction," 2022 International Conference on Engineering and Emerging Technologies (ICEET), Kuala Lumpur, Malaysia, pp. 1-6, 2022, doi: 10.1109/ICEET56468.2022.10007268.

[67] T. Harhare and M. Shah, "Study of Acoustic Correlates Between Prosodic Features and Emotions in Marathi Language," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, pp. 1-5, 2019, doi: 10.1109/ICNTE44896.2019.8946095.

[68] T. Harhare and M. Shah, "An Acoustic and Statistic Study of Emotions Expressed in Marathi Speech," 2022 International Conference for Advancement in Technology (ICONAT), Goa, India, pp. 1-6, 2022, doi: 10.1109/ICONAT53423.2022.9725936.

[69] T. Harhare and M. Shah, "Linear Mixed Effect Modelling for Analyzing Prosodic Parameters for Marathi Language Emotions," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 12, 2021.