International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Conversion of Neutral Speech into Emotional Speech in Hindi Language based on Supra Segmental Features

Archana Agarwal[1], Prashant Johri[2]

Submitted:14/02/2023 **Revised:**15/04/2023 **Accepted:**09/05/2023

Abstract: Speech form is becoming a prevalent and imperative practice as interface between a user and computers systems and smart machines. Mode of input and output in computers and other similar devices in the coming era will be dominated by speech form. To add the spontaneity and intelligibility, emotions play expressive role to ordinary speech. Therefore, it is latest emerging area of research. The practice of converting neutral speech into the intended emotional speech involves examining, identifying, and then changing the characteristics of the speech utterance. The functions of prosodic characteristics, in particular the fundamental frequency F0, pitch, and intensity, are studied for various emotions and employed for emotion conversion. This paper discusses a methodology for conversion of emotions from neutral to Happy, Sad and others for Hindi language speech. The algorithm is developed based using Linear Modification Model where direct modifications are done on F0 contours, pitch and duration on the segmented part of the speech. The continuous speech sentence is segmented into different parts such as words and analysis is done for above said parameters between different emotions of the same neutral speech. Based on the analysis these parameters are modified in the target emotion.

Keywords: *Emotion Conversion, F0 Contour, Pitch, Prosodic features, Speech processing.*

1. Introduction

When we talk of speech, then emotions can’t be ignored. Even in ancient times when there were no specific language to communicate with each other, with the help of gestures and emotions people used to interact with each other. Gestures and emotions are two basic requisites for human to human communication along with language. Now a days Speech form is becoming a popular and imperative method to communicate with computers and other smart devices. Many applications are developed using speech technology which are used in our day to day life. For example speech controlled remote system for TVs, Bluetooth devices, browsing internet through speech, speech is taken as input to recognize the person, and speech as output can be very important for visually impaired persons etc. etc..

One human can interact and express his emotions with other humans with gestures or face expressions learned from the environment and regular interaction experience. Expressing the same emotions with same gestures is a great challenge, to interact with the machines.

There is endless list of applications of speech processing or natural language processing. Emotional conversion is highly valued in many applications. Direct applications include call centres, computer games for enjoyment, and human-machine interaction for any layman, whether literate or illiterate, but especially for those who are blind or visually handicapped. Speech recognition systems, artificial intelligence systems, etc., can work more effectively if emotional traits are converted into speech. for oblique use. It could be used in children's storytelling settings where different expressions need to be made in different tale situations in order to be effective and catch the audience's attention. It can be used to enhance the effectiveness and naturalness of human-computer interaction as a part of a dialogue system. The managers of the contact centres can assess the emotional maturity of the operators by using expressive speech analysis to ascertain their emotional states when chatting with customers. E-mail, for example, may be misunderstood because writing lacks emotional prosody. Orthographic conventions that mark or replace prosody include the use of commas (,), exclamation points (!), question marks (?), scare quotes ("), and typographic emphasis (italic, bold, and underlined text).

In most of these applications use of English language is very common. The same type of research is going on literally in every other languages and in different dialects. Therefore, it is prevalent to do research in many Indian languages especially Hindi language. An analysis and modification of the parameters of the input speech emotion to the target emotion is done during an emotion conversion (or resynthesis) process, which is followed by resynthesizing the final output using the updated parameters.

Our research is based on the Linear Modification Model. This method will modify values of F0 contours, pitch, durations, and intensities on the segmented portion of speech after detailed analysis. The spoken utterance is

[1] Himalayan University, Ita Nagar, Arunachal Pradesh, INDIA
ORCID ID : 0000-0003-3876-798X
[2] Galgotias University, Greater Noida, G.B. Nagar, U.P., INDIA
ORCID ID : 0000-0001-8771-5700
johri.prashant@gmail.com
** Corresponding Author Email: arch.agarwal03@gmail.com*

segmented into words and analysis is done for above said parameters between different emotions of the same neutral speech.

Based on the analysis these parameters are modified in the target emotion. Emotion conversion in this method is mainly reliant on division of the words of a continuous speech, and stress, intensity, pitch and duration of different words.

2. Related work

Lot of research has been done in area of speech processing for English, Mandarin and other languages, now is the time to focus on Indian languages especially Hindi language, because Hindi is the 3rd most spoken language in the world, by around more than 300 million individuals mainly in central and northern India. In the past several researches are being conducted in the area of Hindi language prosody and intonation building.

The F0 contour in a regular declaratory sentence has an inclination to decrease all through the utterance from the starting word onwards till the final word of the sentence. For example in the sentence 'बाहर बादल छाए हुए है ।' the Fig.1 can be checked that pitch pattern starts with low L tone at the first syllable of the starting word 'bahar' and increases as H tone towards the last syllable.

There are many researchers who are focusing on emotion conversion for English spoken language using many different techniques like some focused on emotion conversion for English spoken language with using Gaussian Mixture Model, some have used Regression Tree technique, and some have used Codebook approach, but little focus on work for spoken Hindi language.

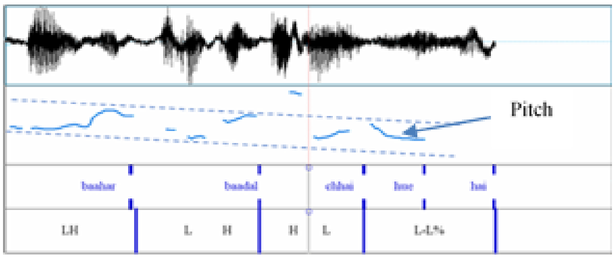


Fig. 1. Pitch pattern (blue colour curve) and F0 Contour of declarative sentence 'बाहर बादल छाए हुए है ।' **(L H** represents low and High pitch tones respectively). LH is low to high. L-L%: low phrase accent, low boundary tone.

Some have worked on Mandarin language and used different techniques, like linear Modification Model (LMM) in which they found that direct modification can result in loss of many prosody features which may exist inside the utterance before modification so they might not reflect the same emotion. Gaussian mixture model (GMM) is numerical method and can provide smooth conversion but may not combine linguistic information. Who worked on CART model concluded that it can include linguistic parameters and analyzed that the GMM scheme is more appropriate when small training set is used, whereas the CART model is better approach and can give more appropriate emotional speech output when large trained set is used.

Preceding research have proven that anger emotion and happy or joyful emotion are generally shown with greater mean pitch, wider pitch variation. Sad emotion is generally shown as lower in mean pitch, barely narrow pitch range, and slower rate of speaking.

Utilizing excitation features that are independent of linguistic and lexical content is suggested by recent research [1]. [1] suggested that emotions should be viewed as deviations from the neutral in order to achieve a robust emotion recognition. The goal is to analyze and utilize these variances utilizing features connected to the speech manufacturing system's excitation element [1].

Reference [2] investigated that prosodic features are not dependent of speaker. Reference [3] stated however, in conversion of emotional voice, emotion is fundamentally supra-segmental and multifaceted that comprises both spectrum features and prosody.

Reference [4] proposed a emotional voice conversion, which is not dependent on any speaker and that can convert speaker's emotion without the need for equivalent data. Reference [4] proposed a VAW-GAN based encoding and decoding structure to acquire the spectrum and prosody mapping. Reference [4] tried to accomplished prosody conversion by using their own ESD database which is open for all, so that there is no need to create own internal database due to unavailability of common database to all.

For the purpose of converting emotional speech, a new sequence-to-sequence-based "Emovox" framework was created by Kun Zhou, Berrak Sisman, and others in 2022 [5]. The suggested Emovox architecture provides effective, fine-grained control over emotion strength. Several points were made. They created a method for simulating emotional arousal and proposed a system for controlling emotional arousal based on comparative features. They showed that their proposed mechanism outperformed other competing controlling strategies in terms of managing speech quality and emotion intensity [5]. They provided a complete analysis and illustrated several prosodic elements of human capacity to control emotion intensity by demonstrating how they interacted with speech duration, pitch envelope, and speech energy in order to comprehend the relationship between emotion intensity and these prosodic properties.

Sequence-to-sequence In the study of [6], a neural network for acoustic modelling in voice conversion is introduced as ConvErsion NeTwork (SCENT). The training stage of a SCENT model involves aligning the feature sequences of the source and target speakers by implicitly utilizing the attention mechanism [6]. The unified acoustic model is utilized at the conversion stage to simultaneously modify the source utterance lengths and acoustic properties [6]. The performance of the system is much improved by using the text transcriptions of the parallel training data [7]. Experimental results show that multi-task learning with linguistic labels reduces seq2seq voice conversion errors [7].

A useful neutral-to-emotional VC model (GAN and VAE) was proposed using the training model Variational Autoencoder (VAE) with a Generative Adversarial Network (VA-GAN) [8], which consists of two effective generating models. They used Cross Wavelet Transform (CWT) to methodically extract and analyze the VA-GAN model-appropriate F0 features at different time scales (CWT-F0 features) [8].

Reference [9] developed and tested a unified model for spectral and prosodic mapping in vocal emotion conversion in a multilingual scenario. The i-vector technique employed non-parallel, mixed-lingual utterances for spectral training [9]. Testing and evaluations revealed that the framework could be utilized to effectively transform the emotions of speakers of the three languages—German, Telugu, and English [9]. A brand-new method for speech emotion conversion that doesn't require parallel training data was provided by [12]. By only weakly depending on a cycle-GAN schema, their approach reduces the reconstruction error brought on by switching back and forth between emotion pairs. The GAN generators were divided into trainable and static parts. Using the momenta latent embedding that the trainable component produced, the fixed component distorted the source F0 contour [12]. In both an objective and subjective sense, their model performed better than the most cutting-edge methods at the moment [12].

The quadratic multivariate polynomial (QMP) has been studied by [13] for the aim of converting neutral speech to emotional target speech. With the exception of neutral/sad conversion, the comparative mean opinion scores CMOS analysis for the Toronto emotional database demonstrates that the altered speech can be partially read as the intended emotion [13]. The CMOS value of the proposed method performs better than the gross and initial-middle-final methods but is inferior to the syllable approach for the German database [13]. The suggested method (which includes English and Chinese speech) was tested using the Emotional Speech Dataset and a Japanese emotional speech dataset [14]. The transformer improved the quality of the transformed speech by expanding the model's temporal dependency over a wider range [14]. The Choi, H., & Hahn, M. (2021) research describes a method for increasing emotional voice conversion (EVC) that has controllable emotional intensity and length. The recommended solution, which uses a sequence-to-sequence network to generate speech with different rhythms and variable utterance times, is based on the desired emotion [15].

Speech can by analyzed with most commonly used and dedicated software tool called PRAAT, evolved by Boersma and Weenink, which is freely available on the internet. The primary degree of the evaluation is to differentiate the speech signals into unvoiced and voiced parts, differentiate between words, and syllables to allow a quantifiable elements of particularly equivalent and consequently similar elements of every utterance. Then you can still abstract various voice signals relevant to speech emotion such as fundamental frequency, rate of speech, silences, intensity, voice beginning time, pitch disturbances, the frequency and bandwidth of energy hikes within the spectrum due to natural tones of the vocal tract referred to as formants. Numerous measures can be received for every sort of sign.

How to design speech corpus with emotions, the decision can deeply have an effect on the eminence of emotion recognition and synthesis research. The characteristics of emotional speech corpora are naturalness, scope, and context. The naturalness refers to whether or not the emotions have been formed spontaneously or artificially.

3. Speech Database

To carry out experiments and verification of results, a Hindi speech corpus was organized. A corpus is designed, which include 60 speech sentences which are containing almost all phonemes of the Hindi language in all effective phonetic situations, and may be termed as phonetically stable and prosodically wealthy randomly decided on from the Hindi text books, and common newspapers of Hindi and from other sources. For building the database of speech, 11 individuals were selected who can record Hindi speech, and were given 20 sentences each to record utterances in different emotions like ‘Happy’, ‘Sad’, ‘Anger’, ‘Surprise’ and in ‘Neutral’. All these individuals were from the age range of nineteen to twenty-two years, commonly graduate scholars and post-graduate scholars. They recorded the speech using ‘PRAAT’ software for speech synthesis in ordinary lab situations. The average duration of the recorded speech sentences is from 1 to 3 seconds, relying on the emotions. The sentences had been recorded via every speaker at the sampling rate of 44.1 KHz and sixteen-bit precision with Mono channel. These speech files are saved in format of wav files.

4. Methodology

In this paper an algorithm is proposed based on Linear Modification Model using supra segmental features. The supra-segmental features play very important role the speech of any language. Vowels and consonants can be considered to be the segments of which the utterance of a speech is composed. Together they form syllables, and words which in turn make up utterances. For the conversion of ‘Neutral’ speech into other emotions, these segmentation is proved to play very important role. The features taken in this study are mainly F0, Pitch, Intensity and duration.

Several studies have divided a spoken utterance into its beginning, middle, and end as part of a segmentation process. Some researchers are doing segmentation based on syllables. In this paper the conversion of neutral to emotional speech is suggested through splitting the speech sentence into words using supra-segmental features. There are several learnings in which only F0 and pitch is considered for segmentation of words. In some research pauses are taken as end of the words and beginning of word, but only pauses cannot be considered as end of the word and start of next word because normally there are few pauses in continuous speech. After careful analysis of a speech utterance following observation was done.

- In the neighbourhood of word boundary, either its pitch is undefined means very near to zero or its intensity starts falling down or it may be both see Fig. 2.
- Intensity keeps on raising and falling but every place where it decreased may not word boundary. The lowest intensity point within some duration of time may be word boundary.
- We may have several regions where intensity is falling, but no two word boundaries can exist very close to each other within the time span of 0.10 seconds.
- There would be one point with lowest intensity value, that may be the word boundary point, refer to Fig.2.

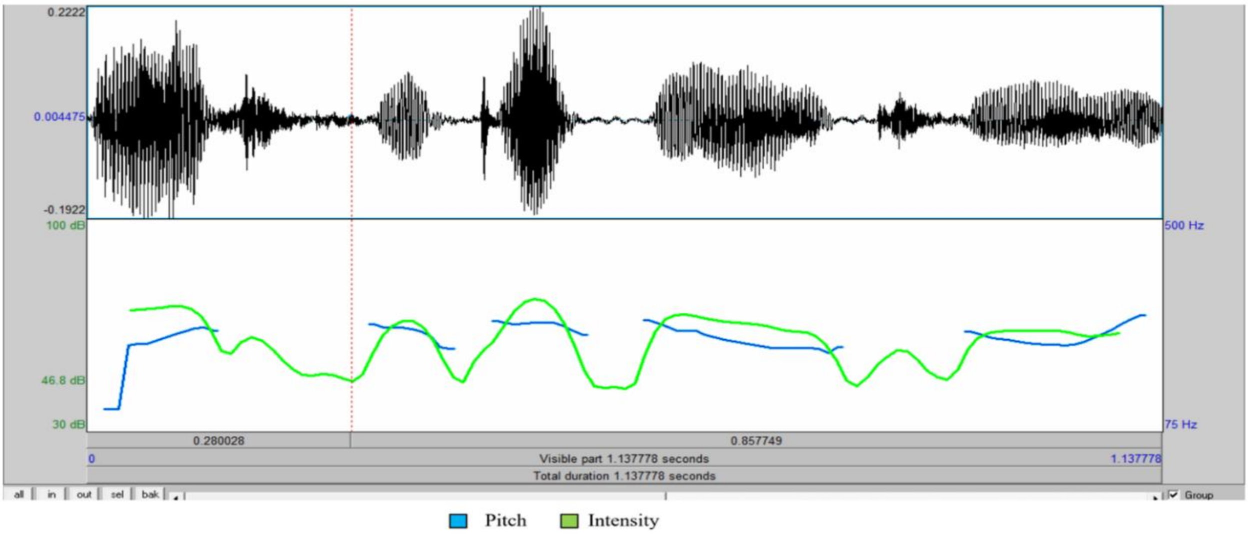


Fig. 2. Pitch (blue line) and Intensity Pattern (Green line) of the Sound File Uttered in Neutral Emotion आज क्रिकेट मैच है।.

How to detect word borders i.e. the position where on word is completing and new word is starting, from continuous speech of Hindi language is already discussed in our previous paper. Once we segmented our sentence into words the next target is to apply this technique to change the emotion of that sentence. The technique this paper refers for the conversion of emotion, is based primarily on modification of two prosodic parameters ‘pitch’ and ‘duration’ of every word in a sentence of spoken utterance. To analyze the changes in form or behavior of pitch values, every word of a sentence pronounced in a “Happy” emotion is matched for the difference to the respective word of same sentence pronounced in a “Neutral” emotion. Entire analysis is recorded for all spoken utterances of the created speech corpus.

Word boundary information was collected after processing of all speech files. Check word boundary points detected by that algorithm in Fig. 3. We can see intensity points information of each speech file in Fig. 3 after processing for finding the word distinctions. Starting and ending time of the speech file is recorded for all speech file to calculate duration of the sentence. Value of time at start of the word and value of the time at end of every word in a sentence is taken, final number of

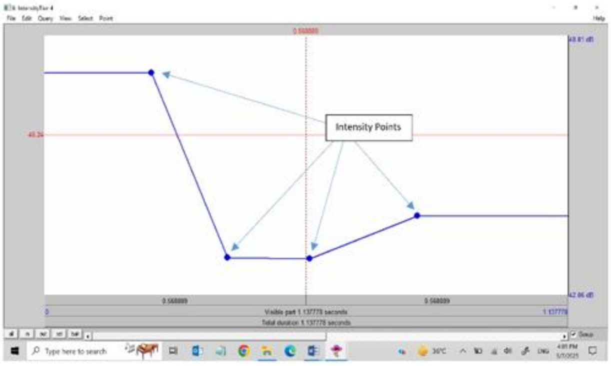

File Edit Query View Select Point
Help
Intensity Points
Visible part 1.137778 seconds
Total duration 1.137778 seconds
Type here to search

Fig. 3. Intensity Points of Detected Word Boundaries of Sentence आज क्रिकेट मैच है।.

word boundaries is recorded, along with their time values and their intensity values in the speech file. All these data were recorded using the PRAAT software referred above.

The data is collected in two different files as ‘Neutral’ and ‘Happy’ respectively. Duration of each word is stored in a datasheet as Starting time and ending time. Using a script made in PRAAT, pitch tier information is extracted with their time value and pitch values for every 0.01 seconds interval. For neutral speech and for happy emotional speech, different files were created, refer to Table 1 and Table 2.

Table 1. Time and Pitch values every 0.01 seconds for Neutral Emotion Sentence आज क्रिकेट मैच है।

c:\archana\tohappy\pitch points\neutral\04_Points.txt			
Time	**Pitch values**	**Word 1 Pitch values**	**Word 2 Pitch values**
0.023888889	120.7947126	120.7947	
0.033888889	121.9218091	121.9218	
0.043888889	247.4858653	247.4859	
0.053888889	250.6541364	250.6541	
0.063888889	252.0159937	252.016	
0.073888889	258.314576	258.3146	
0.083888889	264.5860795	264.5861	
0.093888889	270.2414121	270.2414	
0.103888889	275.5147683	275.5148	
0.113888889	282.8125283	282.8125	
0.123888889	284.2116492	284.2116	
0.133888889	277.3236929	277.3237	
0.303888889	289.985554		289.9856
0.313888889	284.564732		284.5647

Table 2. Time and Pitch values every 0.01 seconds for Happy Emotion Sentence आज क्रिकेट मैच है।

c:\archana\tohappy\pitch points\happy\03_Points.txt			
Time	**Pitch values**	**Word 1 Pitch values**	**Word 2 Pitch values**
0.024716553	326.7942623	326.7943	
0.034716553	328.5057052	328.5057	
0.044716553	341.2796055	341.2796	
0.054716553	364.0957758	364.0958	
0.064716553	398.7394759	398.7395	
0.074716553	417.696904	417.6969	
0.084716553	430.6319312	430.6319	
0.094716553	432.1752891	432.1753	
0.104716553	415.6913952	415.6914	
0.254716553	340.2036264		340.2036
0.264716553	336.3021237		336.3021

For each word in a sentence, pitch values at the starting time and at ending time of the word are recorded in the different data files. Each word's highest pitch and lowest pitch values are also stored. Repeat this practise for each sentence in the neutral speech database and the happy speech database for the corresponding sentences in each emotion. Refer to Table 3.

Difference between pitch values of neutral and happy emotion for every word is calculated for all the sentences. For each word of each speech phrase, the difference between the greatest pitch value and the lowest pitch value, as well as the difference between the respective word's beginning and ending times, are determined, refer to Table 4.

Difference of pitch values between neutral and happy and sad for Each Word for sentence आज क्रिकेट मैच है। is shown in Table 5.

For all words of each sentence 14 groups were made for recording differences in pitch values of start, end, maximum pitch in the word and minimum pitch of the word , for example from 0 to 60Hz, 60 to 120 etc. and 0 to -60, and -60 to -120 etc. refer to Fig. 4.

Table 3. Time Value and Pitch at beginning and at end of the all words

S.No.			Word 1				Word 2			
	Sentence	Emotion	Start	Pitch	End	Pitch	Start	Pitch	End	Pitch
1	Aaj College Jaana Hai	Happy	0.02	263.662	0.245	275.661	0.245	304.030	0.440	282.982
2		Neutral	0.02	122.630	0.209	258.550	0.209	255.879	0.540	214.349

3	Aaj Cricket Match Hai	Happy	0.02	326.794	0.218	415.691	0.218	340.204	0.375	251.790
4		Neutral	0.02	120.795	0.285	277.324	0.285	289.986	0.398	241.351
5	Mujhe Aaj Kaam Karna Hai	Happy	0.02	273.770	0.058	302.619	0.058	310.247	0.426	298.648
6		Neutral	0.02	271.390	0.175	264.173	0.175	278.671	0.543	298.161
7	Aaj Office Nahin Jaana Hai	Happy	0.02	298.399	0.142	317.852	0.142	322.327	0.335	315.324
8		Neutral	0.02	128.450	0.153		0.153	128.450	0.325	324.762

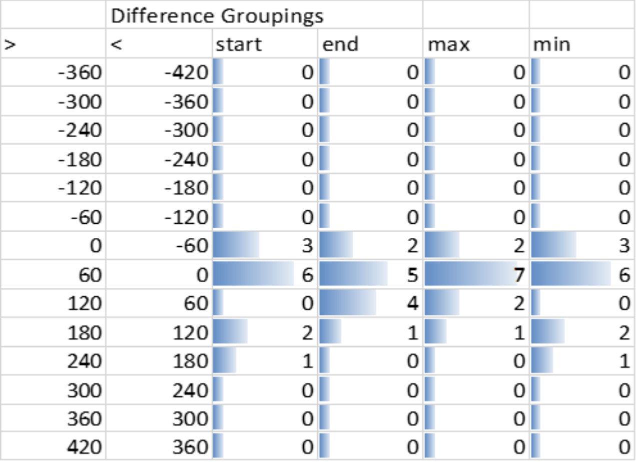
Table 4. Difference of Pitch Values between Neutral and Happy for Word1

		Difference (Happy - Neutral) Word1			
Sno	**Sentence**	**Start**	**End**	**max**	**min**
1	Aaj College Jaana Hai	141.03	17.11	36.62	142.94
2	Aaj Cricket Match Hai	206.00	138.37	147.96	206.00
3	Mujhe Aaj Kaam Karna Hai	2.38	41.77	58.58	15.60
4	Aaj Office Nahin Jaana Hai	169.92	14.77	15.48	169.92
5	Dukaan Band Ho Gyi Hai	-29.29	31.68	15.31	-3.07
6	Kal Subah Ghar Par Milna Hai	16.92	67.45	45.99	47.40
7	Weh Tumhare Ghar Mein Rehta Hai	37.14	25.61	40.47	36.36
8	Mujhe Aaj Kaam Karna Hai	45.44	69.02	59.46	49.56
9	Dukaan Band Ho Gyi Hai	14.18	-2.27	-7.71	6.39
10	Hum Ghar Pahunchne Waale Hain	55.86	94.03	80.20	57.26
11	Yahan Phool Todna Mana Hai	-25.39	-6.77	-19.07	-13.60
12	Tum Paas Ho Gaye Ho	-33.60	91.05	87.60	-30.80

Table 5. Difference of Pitch Values between Neutral and Happy and Sad for Each Word for sentence आज क्रिकेट मैच है।

Difference of pitch values (Happy - Neutral)				
Words	Start	End	max	min
word1	206.00	138.37	147.96	206.00
word2	50.22	10.44	50.22	10.44
word3	59.84	47.28	68.11	47.28
word4	67.65	-28.28	67.65	-17.58
word5	0.21	-67.76	-31.21	-9.26
Difference of pitch values (Happy - Sad)				
word1	163.72	8.34	9.08	162.04
word2	17.39	1.24	17.39	-4.92
word3	0.35	44.57	43.52	-4.03
word4	6.47	-10.29	6.47	-2.71
word5	-35.11	-100.69	-66.54	-41.64

It was checked that 1st word of every sentence lies in which group for all pitch differences. The group with maximum number of entries was selected for the pitch differences of 1st word. Similarly, each word was checked for a particular group with maximum entries, refer to Fig. 4.



	Difference Groupings				
>	<	start	end	max	min
-360	-420	0	0	0	0
-300	-360	0	0	0	0
-240	-300	0	0	0	0
-180	-240	0	0	0	0
-120	-180	0	0	0	0
-60	-120	0	0	0	0
0	-60	3	2	2	3
60	0	6	5	7	6
120	60	0	4	2	0
180	120	2	1	1	2
240	180	1	0	0	1
300	240	0	0	0	0
360	300	0	0	0	0
420	360	0	0	0	0

Fig. 4. Different Groups for a Word with Maximum Entries

An algorithm was developed, with the whole analysis done above. This algorithm will take neutral speech file, and segment the sentence into words. Calculate differences in pitch and differences in length of every word, and then based on these differences neutral speech is converted into emotional speech of happy state. The algorithm can be used for different number of words perceived from a spoken sentence. (In the study sentences are taken having four to six words). Algorithm is explained as follows:

- Take input and read neutral speech file .
- Run the PRAAT script to detect and count number of word boundary points and total number of words from a neutral emotion speech file.
- Record the starting value of pitch, ending value of pitch of every word.
- Record the Min, and Max pitch of every word of speech sentence.
- Stylize the pitch at 5 Hz to reduce number of pitch points.
- Modify pitch tier of sound file. Increase or decrease values of all pitch points by the higher range of the group to which it belongs to, of frequency (f0) value.
- Every word of speech file's pitch tier is changed, including the starting pitch, ending pitch value, maximum pitch, and minimum pitch. Swap previous pitch tier with the modified one.
- Abstract the duration tier and adjust duration of words of the speech file, and exchange with the modified duration.
- Result is the transformed emotional speech in happy state.

Fig. 6. depicts the sentence uttered in neutral emotion converted in happy emotion using above algorithm. Same sentence uttered in happy emotion is shown in Fig. 7.

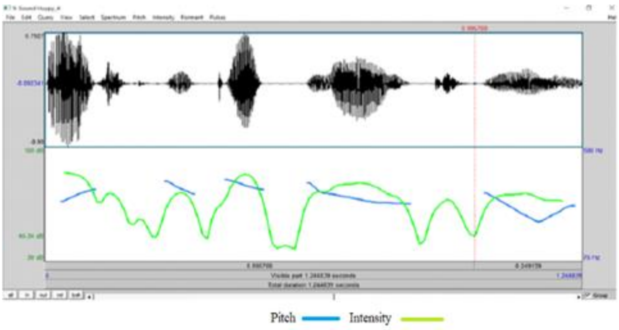


Fig. 5. Pitch and Intensity of a Sentence in Happy Emotion Converted from Neutral Emotion 'आज क्रिकेट मैच है।

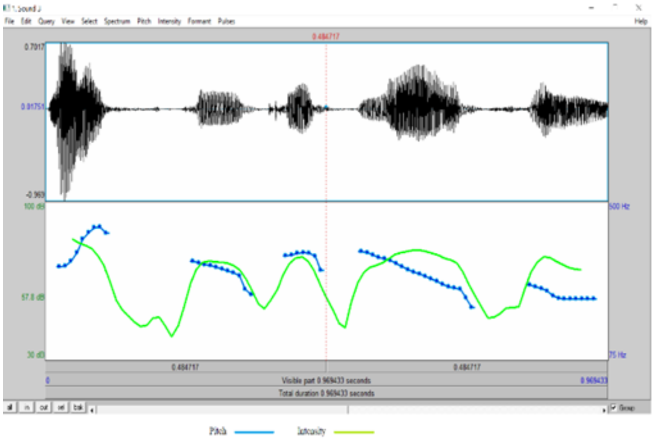


Fig. 6. Pitch and Intensity of a Sentence uttered in Happy Emotion 'आज क्रिकेट मैच है।'

The above algorithm was implemented for every sentence uttered in neutral emotion and converted into different emotions like happy, anger, and sad. The translation of neutral speech into happy, sad, angry, and surprised emotional speech was done using four separate scripts, each representing one of the four emotions. Following are the pitch differences of the speech file converted from neutral to happy emotion, see fig. 7.

The above algorithm was implemented for every sentence uttered in neutral emotion and converted into different emotions like happy, anger, and sad. The translation of neutral speech into happy, sad, angry, and surprised emotional speech was done using four separate scripts, each representing one of the four emotions. Following are the pitch differences of the speech file converted from neutral to happy emotion, see fig. 7. Pitch difference from neutral to sad are depicted in Fig. 8 and neutral to anger are depicted in Fig. 9.

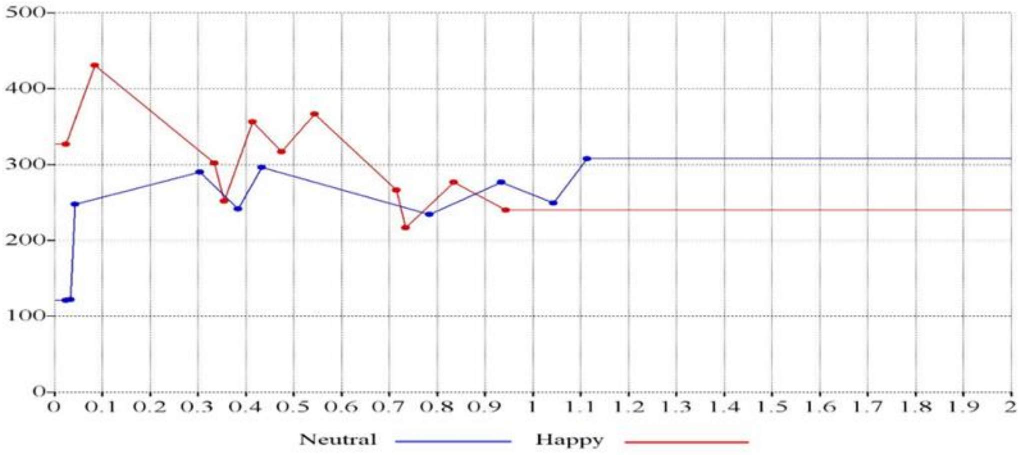


Fig. 7. Pitch Differences Of The Sound File From Neutral To Happy 'आज क्रिकेट मैच है।'

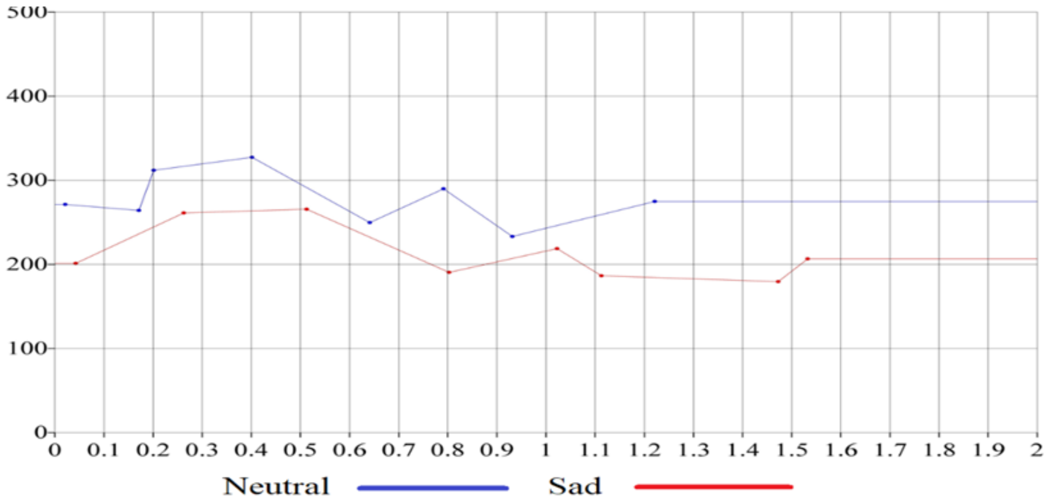


Fig. 8. Pitch Differences Of The Sound File From Neutral To Sad 'आज क्रिकेट मैच है।'

5. Result and Analysis of Emotion Conversion

For assessment of the result of the conversion system, a detailed system was formed. Emotions were manually identified by different individuals, which were requested to recognize the emotion by listening in the given utterance. Sometimes listeners were not able to correctly recognize the particular emotion, we asked them to put that utterance in ''Not decided" category. This option was given them as one of the available choices. For better comparison, the original recorded utterance in a particular emotion was given to listeners and then they were asked to listen the same sentence with converted emotion.

For each emotion, ten different statements were provided to ten different listeners. Participants had to listen to one of the converted emotional speech utterances for the identical sentence before listening to the neutral emotion speech utterance. Some listeners were able to clearly recognize the converted emotion of speech and some were confused between the different emotions i.e. confused between sad with neutral or between happy with neutral etc.. A listener's perception matrix was made for that and is depicted in Table 7.

With these result we can see that 'happy' emotion is sometimes confused with neutral. The 'anger' emotion is confused with 'surprise'. 'Surprise' and 'Sadness' were recognized very evidently by the listeners. As these results are dependent on the output given by the listeners, they are also dependent on speaker's emotional state that which kind of emotions the speaker is able to communicate.

Table 6. Listener's Perception Matrix for Four Emotions

Emotional State	Neutral	Happy	Sad	Anger	Surprise	Not decided
Happiness	20%	55%	0	0	0	25%
Sadness	10%	0	75	0	0	10%

			%			
Angry	10%	0	0	70%	20%	0
Surprise	20%	0	0	0	70%	10%

6. Conclusion

After careful analysis of Hindi language intonation pattern and the role of supra-segmental features like F0 contour, pitch, intensity, and duration, we got evident clues which helped us to split speech utterance into words, and further helped us to do transform the neutral state of speech into target emotional state by using Linear Modification Model. The results are inspiring. Sometimes word boundary couldn't be clearly identified because some words were merged together or some were rejected due to the deletion of intensity points, while speaking in continuous speech. Similarly the some points are syllable boundary points which are discovered as word boundary points. If word boundary detection is improved, this will help in conversion of emotion also.

There is great scope of improvement in the research of conversion of emotions from neutral to others. Database can be further improved by carefully choosing the speakers which can correctly pronounce and fluently speak in Hindi.

7. Acknowledgment

This research was supported Himalayan University, Arunachal Pradesh. I would like to express my special thanks of gratitude to my guide, Prof. (Dr.) Prashant Johri, Professor in School of Computing Science and Engineering, Galgotias University, Greater Noida, G.B. Nagar, for the continuous support to do my research on such a realistic application, with her supervision. We thank our colleagues from Inderprastha Engineering College, Ghaziabad who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. I gratefully acknowledge Dr. Ajay Kr., Director, and Dr. Vijay Kr. Singh, Professor and head of computer science department, Inderprastha Engineering College, Ghaziabad for helping me with their valuable suggestions and comments that greatly improved the manuscript

References

[1] Kadiri, S. R., Gangamohan, P., Gangashetty, S. v., Alku, P., & Yegnanarayana, B. (2020). Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference. Circuits, Systems, and Signal Processing, 39(9), 4459–4481. https://doi.org/10.1007/s00034-020-01377-y

[2] Liu, S., Cao, Y., Kang, S., Hu, N., Liu, X., Su, D., Yu, D., & Meng, H. (2020). Transferring Source Style in Non-Parallel Voice Conversion. http://arxiv.org/abs/2005.09178.

[3] Schuller, D. M., & Schuller, B. W. (2021). A Review on Five Recent and Near-Future Developments in Computational Processing of Emotion in the Human Voice. Emotion Review, 13(1), 44–50. https://doi.org/10.1177/1754073919898526

[4] Zhou, K., Sisman, B., Liu, R., & Li, H. (2022). Emotional voice conversion: Theory, databases and ESD. Speech Communication, 137, 1–18. https://doi.org/10.1016/j.specom.2021.11.006A. Agarwal and A. Dev, "Emotion recognition and conversion based on segmentation of speech in Hindi language," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2015, pp. 1843-1847.

[5] Kun Zhou, Berrak Sisman, Rajib Rana, Bjorn W. Schuller, Haizhou Li, 2022, "Emotion Intensity and its Control for Emotional Voice Conversion", arXiv:2201.03967v3 [cs.SD] 18th July 2022.

[6] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai, "Sequence-to-sequence acoustic modeling for voice conversion," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 27, no. 3, pp. 631–644, 2019.

[7] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improving sequence-to-sequence voice conversion by adding text-supervision," in ICASSP 2019. IEEE, 2019, pp. 6785–6789.

[8] Zhaojie Luo, Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki, 'Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks', APSIPA Transactions on Signal and Information Processing , Volume 8 , 2019 , e10 DOI: https://doi.org/10.1017/ATSIP.2019.3

[9] Susmitha Vekkota, Deepa Gupta, 'Fusion of spectral and prosody modelling for multilingual speech emotion conversion' , Elsevier, Knowledge-Based Systems Volume 242, 22 April 2022, 108360.

[10] Sandeep Kumar, , MohdAnul Haq, , Arpit Jain, , C. Andy Jason, , Nageswara Rao Moparthi, Nitin Mittal5, Zamil S. Alzamil, 'Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance', Computers, Materials & Continua Tech Science Press, DOI: 10.32604/cmc.2023.028631.

[11] Kreuk, F., Polyak, A., Copet, J., Kharitonov, E., Nguyen, T., Rivière, M., Hsu, W., Mohamed, A., Dupoux, E., & Adi, Y. (2021). Textless Speech

Emotion Conversion using Decomposed and Discrete Representations. ArXiv, abs/2111.07402.

[12] Ravi Shankar, Jacob Sager, Archana Venkataraman, Non-parallel Emotion Conversion using a Deep-Generative Hybrid Network and an Adversarial Pair Discriminator, INTERSPEECH 2020, October 25–29, 2020, Shanghai, China.

[13] Singh, J.B., Lehana, P. STRAIGHT-Based Emotion Conversion Using Quadratic Multivariate Polynomial. Circuits Syst Signal Process 37, 2179–2193 (2018). https://doi.org/10.1007/s00034-017-0660-0.

[14] Singh, J.B., Lehana, P. STRAIGHT-Based Emotion Conversion Using Quadratic Multivariate Polynomial. Circuits Syst Signal Process 37, 2179–2193 (2018). https://doi.org/10.1007/s00034-017-0660-0.

[15] Choi, H., & Hahn, M. (2021). Sequence-to-Sequence Emotional Voice Conversion With Strength Control. IEEE Access, 9, 42674-42687.

[16] Vekkot, S., & Gupta, D. (2019). Emotion Conversion in Telugu using Constrained Variance GMM and Continuous Wavelet Transform-F_{0}. TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 991-996.

[17] Haque, A., & Rao, K.S. (2015). Modification and incorporation of excitation source features for emotion conversion. 2015 International Conference on Computer, Communication and Control (IC4), 1-5.

[18] GurunathReddy, M., & Rao, K.S. (2015). Neutral to happy emotion conversion by blending prosody and laughter. 2015 Eighth International Conference on Contemporary Computing (IC3), 342-347.

[19] Saheer, L., Na, X., & Cernak, M. (2015). Syllabic Pitch Tuning for Neutral-to-Emotional Voice Conversion.

[20] Haque, A., & Rao, K.S. (2015). Analysis and modification of spectral energy for neutral to sad emotion conversion. 2015 Eighth International Conference on Contemporary Computing (IC3), 263-268.

[21] Reddy, M.G., & Rao, K.S. (2017). Neutral to Joyous Happy Emotion Conversion. 2017 14th IEEE India Council International Conference (INDICON), 1-6.James, J., Tian, L., & Watson, C. I. (2018). An open source emotional speech corpus for human robot interaction applications. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September, 2768–2772. https://doi.org/10.21437/Interspeech.2018-1349

[22] Vekkot, S., & Gupta, D. (2019). Prosodic transformation in vocal emotion conversion for multi-lingual scenarios: a pilot study. International Journal of Speech Technology, 22, 533 - 549.

[23] Haque, A., & Rao, K.S. (2016). Modification of energy spectra, epoch parameters and prosody for emotion conversion in speech. International Journal of Speech Technology, 20, 15 - 25.

[24] James, J., Tian, L., & Watson, C. I. (2018). An open source emotional speech corpus for human robot interaction applications. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018-September, 2768–2772. https://doi.org/10.21437/Interspeech.2018-1349

[25] Alaria, S. K., A. . Raj, V. Sharma, and V. Kumar. “Simulation and Analysis of Hand Gesture Recognition for Indian Sign Language Using CNN”. *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 4, Apr. 2022, pp. 10-14, doi:10.17762/ijritcc.v10i4.5556.

[26] Vyas, S., Mukhija, M.K., Alaria, S.K. (2023). An Efficient Approach for Plant Leaf Species Identification Based on SVM and SMO and Performance Improvement. In: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. (eds) Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol 959. Springer, Singapore. https://doi.org/10.1007/978-981-19-6581-4_1