Neutral Speech to Target Speech Conversion by Prosodic Modification

Shreegowri A. J.
IV Sem, M.Tech,
Dept. of ECE, VVCE,Mysuru

D. J. Ravi
PhD, Professor & HOD
Dept. of ECE, VVCE,Mysuru

ABSTRACT
The dynamics of prosodic features are utilized for speech emotion conversion. In particular, emotion conversion of neutral speech to sad, fear, anger and happy speech is accomplished. The prosodic features considered for the study are pitch contour and duration. Subjective listening test results show that the effectiveness of perception of emotion is better in the case of pitch contour and duration for the whole utterance. The results show that the converted sad, fear, angry speech are perceived very close to natural sad, fear, anger and happy emotions.

Keywords
prosody (the study of rhythm, intonation, stress and related attributes in speech)

1. INTRODUCTION
The emotion in speech is the extra linguistic information which incorporates expressiveness to tell about mental state of a speaker. Attempts to add emotion effects to synthesized speech or neutral speech have existed for more than a decade and a limited work have been conducted on emotion conversion. Emotions play important role in expressive speech synthesis. Emotional state of a speaker is accompanied by physiological changes affecting respiration, phonation, and articulation. These changes are manifested mainly in prosodic patterns of pitch and duration. An approach to incorporate emotion into neutral speech is to modify emotion specific parameters to a neutral speech. Therefore, the objective is to analyze and modify these emotion specific parameters of the neutral speech to obtain speech of the target emotion. Change in prosodic features from neutral to emotional speech is analyzed and emotion conversion is accomplished by using the Gaussian distribution equation. With the help of Gaussian distribution equation we are going to calculate target pitch by using neutral pitch and both neutral and target mean and standard deviation.

$$Pt = \left(\frac{Pn-\mu n}{\sigma n}\right) * \sigma t + \mu t \qquad (1)$$

Where

Pt – target speech pitch value.

Pn – neutral speech pitch value.

μt – target speech mean pitch value.

μn – neutral speech mean pitch value.

σt – Standard deviation of target speech.

σn – Standard deviation of neutral speech.

It is observed that in the literature there has been lot of work on how the prosodic features vary for different emotions. But there are not much of papers discussing generating emotional speech from neutral utterance. Our interest is that, if we can suitably modify the prosodic parameters of neutral speech, we will be able to produce the emotive speech. The motivation has led to development of algorithm to convert neutral prosody to emotional prosody.

Emotion conversion from neutral to other emotions has interesting scope. For instance, when a computer reads out a story for a child, it will be effective if it is expressed emotionally with corresponding emotions. Besides this there are numerous applications of emotional speech such as announcements in railway stations, employing emotion specific robots in warfare and so on where in expressing emotionally according to circumstances will be much more effective.

2. SCOPE OF WORK
The aim is to convert the neutral speech into emotional (sad, fear, angry and happy) speech using prosody (pitch and duration) modification with the help of Gaussian distribution equation.

3. PROPOSED WORK

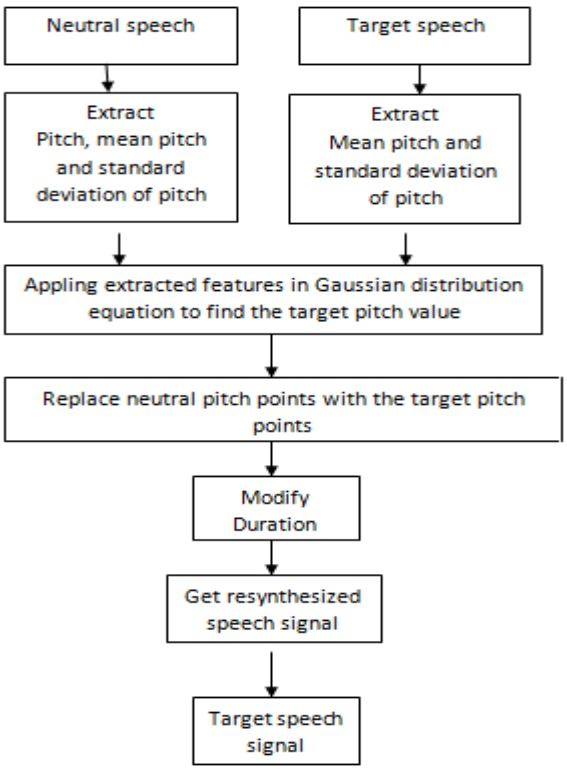

Figure 1: Flow of the proposed work.

Steps in proposed work:

Step 1: Open/load the neutral speech signal:

When we open the praat software we will obtain two windows, first one is the praat objects window and secondly praat picture window. Firstly loading/opening the saved neutral speech signal in to the praat objects window, by selecting "read from file option" option in the praat objects window. Now the object window contain neutral speech signal.

Step 2: Feature extraction

Initially we are loading the neutral speech signal to praat software, and then we are going to extract neutral speech signal pitch values(Pn), pitch mean(μn) and pitch standard deviation(σn). Select "view and edit" in the praat objects window and obtain the pitch counter of the neutral speech by using "pitch listing" option, it will obtain the list of pitch values of neutral speech in hertz with respect to time in seconds. Using praat we can obtain the neutral speech signal pitch values with time index. The mean pitch values and standard deviation values of neutral speech signal and targeted speech signal are substituted in the above equation(1). By substituting neutral pitch value, mean pitch and standard deviation of target and neutral speech we find out target speech pitch values with respect to time index.

Step 3: Replace the pitch values

In order to replace the pitch points of the neutral speech signal from the new/target pitch points we have to select both the target pitch tier and manipulated neutral speech signal then we obtain the 'replace pitch tier' option. By selecting replace pitch tier option it will replace the pitch points present in the manipulation window with the new/target pitch points. In that green color pitch tier represents the target pitch tier and gray color represents the neural speech pitch points.

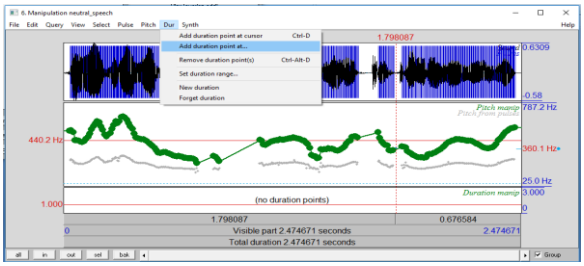

Figure 2. Adding duration point in modification window

Step 4: Modify duration

The duration speech signal can be modify by two methods, first one is duration modification by adding the point in the duration window (third sub-window) of the manipulation window. The duration points can be added by simply selecting the 'Add duration point at/Add duration point at cursor'. The option 'Add duration point at' will open the window for asking at which duration do you want to add the duration point so we have to set the timing at which we want duration point. The option 'Add duration point at cursor' will add the duration point at which the cursor is placed as shown in figure 15, the red line in the duration window indicates the cursor point. The second one is modifying the duration after getting resynthesis signal from manipulation by simply selecting the time range in the 'Scale times to' in 'Modify times' from 'Modify' option.

Step 5: Get resynthesis signal

The resynthesized signal is nothing but a target speech signal. After completing all the above steps we are obtaining target speech signal by selecting the 'Get resynthesis (overlap-add)' option in the Praat Object window.

4. PROSODY ANALYSIS OF EMOTIONAL AND NEUTRAL SPEECH

In this study, we considered neutral and emotional speech from 10 speakers (5 male and 5 female) for analyzing the prosody variation among various emotions, and also for the comparison of synthesized utterances with natural utterances. The prosodic features of interest are duration patterns, average pitch, pitch contour, and average energy of speech signal [4,2], and average prosody values are tabulated in Table 1. These characterize emotion specific information present in speech. From the analysis of the speech corpus and from the literature the several observations are made regarding emotion conversion. Anger emotions have smaller mean durations, whereas happy emotions have comparatively larger mean durations. The extreme emotions like anger, happy, fear contain emotion specific information in the first words. In comparison to neutral speech, anger speech is produced with higher and more varied pitch, higher intensity, and shorter duration, and faster attack times at the start of speech [4]. In comparison to neutral speech, sad and fear speech is produced with lower and less varied pitch, lower intensity, and longer duration. In comparison to neutral speech, happy speech is produced with medium and more varied pitch, average intensity, and longer duration.

Table 1. Average percentage of prosody values increases/decreases from neutral to emotional speech

speech (units)	**Duration (sec)**	**Mean Pitch (Hz)**	**Standard deviation (Hz)**
Female			
Angry	-16%	70%	80%
Fear	20%	20%	-28%
Sad	30%	20%	-28%
Happy	30%	40%	20%
Male			
Angry	-20%	40%	80%
Fear	20%	-10%	-10%
Sad	50%	-10%	-20%
Happy	40%	20%	50%

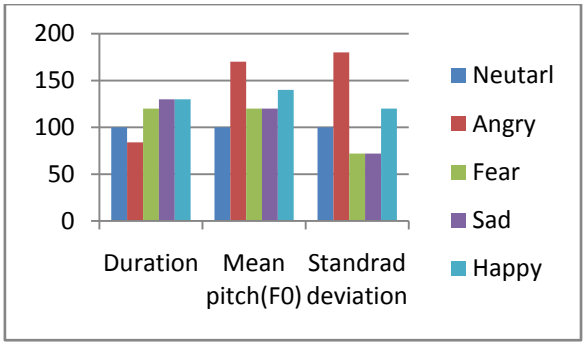

Figure1:Chart 1: Average prosody values increases/decreases from neutral to emotional speech for female

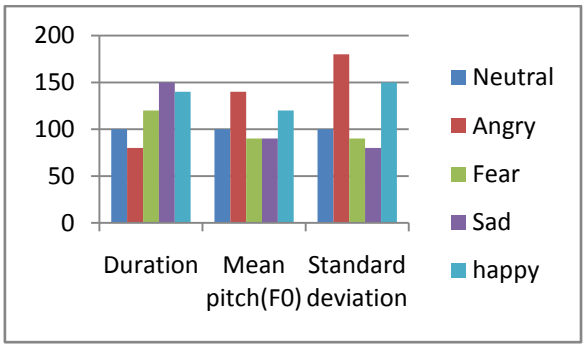

Figure2:Chart 2. Average prosody values increases/decreases from neutral to emotional speech for male

5. SUBJECTIVE TEST

The effectiveness of emotion conversion is evaluated by subjective listening test. This evaluation is carried out by 10 human subjects. The human subjects were previously made listened to the actual recordings by professional artists from the database. Subsequently we were given 10 sentences of 5 male and 5 female for recognition to find the efficiency of the converted emotion.

Table 2: Subjective test to find recognition rate of the converted signal

	SAD	**ANGRY**	**FEAR**	**HAPPY**
SAD	82%	-	18%	-
ANGRY	10%	88%	2%	-
FEAR	3%	1%	96%	-
HAPPY	60%	5%	15%	20%

6. CONCLUSION

It is an approach that would modify the neutral speech signal to produce different emotions like sad, angry, fear and happy version of it. Here prosodic features can be suitably modified to obtain sad, angry, fear and happy emotion's from neutral utterance. The pitch contours and duration are parameter which is considered for emotion conversion. From the results and recognition rate it is concluded that the propose work is more efficient. The emotion conversion method can be similarly accomplished for other emotions like surprise, disgust etc.

7. REFERENCES

[1] Amrita, Bageshree Pathak "Emotion Conversion Of Speech Signal Using Neural Network"; IJEEDC-2017

[2] "GMM-Based Emotional Voice Conversion Using Spectrum and Prosody Features"; American Journal of Signal Processing;p-ISSN: 2012.

[3] Vroomen, J., Collier, R., Mozziconacci, S.: Duration and intonation in emotional speech. Eurospeech 1, 577–580 (1993)

[4] Tao, J., Kang, Y., Li, A.: Prosody conversion from neutral speech to emo- tional speech. IEEE Transactions on Audio, Speech, and Language Processing 14, 1145–1154 (2006)

[5] Rao, K.S., Yegnanarayana, B.: Prosody modification using instants of significant ex- citation. IEEE Transactions on Audio, Speech and Language Processing 14, 972–980 (2006)

[6] Paeschke, A., Sendlmeier, W.F.: Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. In: Speech Emotion, pp. (2000)

[7] Koolagudi, S.G., Maity, S., Kumar, V.A., Chakrabarti, S., Sreenivasa Rao, K.: IITKGP-SESC: Speech database for emotion analysis. In: Ranka, S., et al. (eds.) IC3 2009. CCIS, vol. 40, pp. 485–492. Springer, Heidelberg (2009)

[8] Yegnanarayana, B., Murty, K.S.R.: Event-based instantaneous fundamental fre- quency estimation from speech signals. IEEE Transactions on Audio, Speech and Language Process 17(4), 614–625 (2009)