Neutral to Target Emotion Conversion Using Source and Suprasegmental Information

D. Govind [1], *S. R. M. Prasanna*[1] *and B. Yegnanarayana* [2]

[1]Department of Electronics and Electrical Engineering, IIT Guwahati, Assam, India
[2]International Institute of Information Technology Hyderabad, A.P, India
{dgovind, prasanna}@iitg.ernet.in, yegna@iiit.ac.in

Abstract

This work uses instantaneous pitch and strength of excitation along with duration of syllable-like units as the parameters for emotion conversion. Instantaneous pitch and duration of the syllable-like units of the neutral speech are modified by the prosody modification of its linear prediction (LP) residual using the instants of significant excitation. The strength of excitation is modified by scaling the Hilbert envelope (HE) of the LP residual. The target emotion speech is then synthesized using the prosody and strength modified LP residual. The pitch, duration and strength modification factors for emotion conversion are derived using the syllable-like units of initial, middle and final regions from an emotion speech database having different speakers, texts and emotions. The effectiveness of the region wise modification of source and supra segmental features over the gross level modification is confirmed by the waveforms, spectrograms and subjective evaluations.

Index Terms: Emotions, ZFF, strength of excitation, instantaneous pitch, duration

1. Introduction

The emotions in speech carry extra linguistic information about the context, add expressiveness and characterize the mental state of speaker. The problem of incorporating the emotions in the synthesized speech is a difficult and challenging task. An approach for the emotional speech generation is by incorporating the emotion specific parameters in a neutrally recorded speech [1] [2]. The emotion specific parameters are obtained by analyzing the source and vocal tract parameters of different emotions. The recent work in [3] proposed a set of source features based on the zero frequency filtering (ZFF) of the electroglottograph (EGG) and also speech signals. These include features based on the instantaneous pitch and strength of excitation. The parameters are found to be showing a systematic variation across different emotions and hence may be useful for emotion conversion. The objective is therefore to analyze and modify these emotion specific source parameters of the neutral speech to obtain speech in the target emotion.

The scope is converting neutral speech to the target emotion speech by modifying the source and suprasegmental (duration) parameters. This is achieved by the prosody modification based on the instants of significant excitation [4]. The instants of significant excitation correspond to instants of glottal closure in case of voiced speech and onset of events like burst and frication in case of unvoiced speech [5]. The recent prosody modification method based on ZFF is demonstrated to be providing significantly improved performance, both in terms of accuracy of detecting instants of significant excitation, and also computational time [5], [6]. Finally, this resulted in an improved quality of the prosody modified speech.

The speech signals of neutral and target emotions are processed by the ZFF method to estimate the source parameters based on the instantaneous pitch and strength of excitation as described in [3] [7]. For estimating various emotion specific source parameters, the whole speech signal is divided into three regions (initial, middle and final) and parameters from each of these regions are averaged separately and used for neutral to target emotion conversion. Similarly, the durations of syllable-like units are also obtained. The instantaneous pitch based parameters and duration information are incorporated into the linear prediction (LP) residual of the neutral speech by the prosody modification process. The Hilbert envelope (HE) of the prosody modified LP residual is computed using the analytic signal definition. The HE of the LP residual is used as the representation of strength contour and modified according to the strength of target emotion. The LP residual is reconstructed from the modified HE of LP residual using the cosine of LP residual phase. The vocal tract parameters of the neutral speech are updated according to the duration modification factor. The target emotion speech is then synthesized using the prosody and strength modified LP residual and neutral speech LP coefficients as vocal tract information.

The organization of the paper is as follows. Section 2 explains methods to incorporate the source and suprasegmental parameters. The analysis and estimation of source parameters of various emotions are given in Section 3. Section 4 describes the neutral to target emotion conversion. Finally, Section 5 concludes with a mention on the future scope of the present work.

2. Methods to incorporate source and suprasegmental parameters

2.1. Prosody modification using instants of significant excitation

In the present work prosody modification refers to both pitch and duration modification. The prosody modification using the instants of significant excitation involves the following steps [4]:

- Finding the locations of the instants of significant excitation
- Deriving the modified instant locations according to the desired pitch and duration modification factors
- Deriving the prosody modified LP residual
- Synthesizing the speech using the prosody modified LP residual

2.1.1. Finding the instants of significant excitation

Instants of significant excitation can be obtained from the ZFF of speech as follows [5]:

- Difference input speech signal $x(n) = s(n) - s(n-1)$
- Compute the output of cascade of two ideal digital resonators at 0 Hz $y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n)$, where a_1 = 4, a_2 = -6, a_3 = 4, a_4 = -1
- Remove the trend i.e., $\hat{y}(n) = y(n) - \bar{y}(n)$, where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{n=-N}^{N} y(n)$, where $2N + 1$ corresponds to the average pitch period computed over a longer segment of speech
- The trend removed signal $y(n)$ is termed as zero frequency filtered (ZFF) signal.
- The positive zero crossings of the ZFF signal will give the instants location.

2.1.2. Deriving the modified instants location

For obtaining the modified instants location, the epoch intervals are derived by finding the interval between successive instants location. These epoch intervals are then scaled (in case of pitch modification) or re-sampled (in case of duration modification) according to the desired pitch and duration modification factors. The new epoch locations are obtained by starting from the first epoch and using the linearly interpolated modified epoch intervals.

2.1.3. Constructing the prosody modified LP residual

The prosody modified LP residual is obtained using the modified instant locations and the LP residual of neutral speech. Perceptually relevant 30% number of LP residual samples of the neutral speech around the original epochs are copied to the corresponding regions around the modified instants location to obtain the prosody modified LP residual.

2.2. Incorporating the strength of excitation

The strength modification is incorporated by scaling the HE of the prosody modified LP residual. If $e(n)$ is the LP residual, then the HE of LP residual is defined as $h_e(n) = \sqrt{e^2(n) + e_h^2(n)}$, where $e_h(n)$ is the Hilbert transform of $e(n)$. The strength and prosody modified LP residual is reconstructed by multiplying the scaled HE of LP residual with the cosine of the phase of the prosody modified LP residual [3].

2.3. Synthesis of target emotion speech

The target emotion speech is synthesized by exciting the time varying vocal tract filter with the prosody and strength modified LP residual. The parameters of the vocal tract filter are still from the neutral speech, but copied according to the duration of the target emotion, to keep their length same as that of modified LP residual.

3. Estimation of source and suprasegmental parameters of different emotions

The source parameters like instantaneous pitch and strength of excitation are obtained by analyzing the speech signal across different emotions. Along with these source parameters, the duration of syllable-like units provided in the database is also considered as an emotion specific parameter. These parameters estimated from the speech of five different emotions (Neutral, Angry, Happy, Boredom and Fear) across 9 texts of 8 speakers (5 females and 3 males) from German emotional speech database are given in Table 1 [8]. The angry and fear emotions have the largest average pitch and the lowest average excitation strength, compared to all other emotions. Alternatively, the neutral and boredom emotions have the largest strength of excitation and lowest average pitch. This is because, at higher pitch frequency, the vocal folds start vibrating at a higher rate with reduced suction pressure resulting in lowering of excitation strength. Similarly, at lower pitch frequency, the vocal folds open and close slowly with higher suction pressure which increases the strength of excitation [3]. Also the table indicates that the boredom has the largest duration and, fear and neutral are having shortest duration. Table 2 presents the pitch, duration and excitation strength modification factors for converting neutral speech to target emotion speech. The modification factors given in the table are obtained by scaling the source and duration parameters given in Table 1 by the corresponding parameters of the neutral speech. For instance, the pitch modification factor of 1.67 to convert neutral to target angry emotion is obtained by taking the ratio of the average pitch of the angry to that of the neutral speech.

Table 1: Average of source parameters and duration of different emotions estimated from speech of the whole text .

Emotion	Mean Pitch	Mean Dur. (ms)	Mean Strength
Neutral	180.86	2260	0.51
Angry	301.59	2590	0.41
Happy	287.17	2430	0.38
Boredom	175.60	2700	0.52
Fear	249.10	2240	0.42

Table 2: Pitch, duration and strength modification factors obtained by taking ratio of target emotion parameters with respect to the neutral emotion .

Target Emotion	Pitch Mod.	Dur. Mod.	Strength Mod.
Angry	1.67	1.15	0.80
Happy	1.59	1.08	0.76
Boredom	0.97	1.20	1.02
Fear	1.38	0.99	0.83

4. Neutral to Target Emotion Conversion

The neutral to target emotion conversion can be achieved by modifying the pitch, duration and strength of the neutral speech according to the modification factors given in Table 2 using the methods described in Section 2. The best that can be achieved is for the case where we have reference speech of target emotion for the same text. Then the source and duration parameters can be analyzed across each syllable-like unit and use it for modification. Figure 1((e)-(f)) shows the waveform, pitch contour, excitation strength and spectrogram of the synthesized target emotion in this case. The instantaneous pitch and strength contours of the synthesized target emotion match closely with that of the reference target emotion. The spectrograms of the synthesized

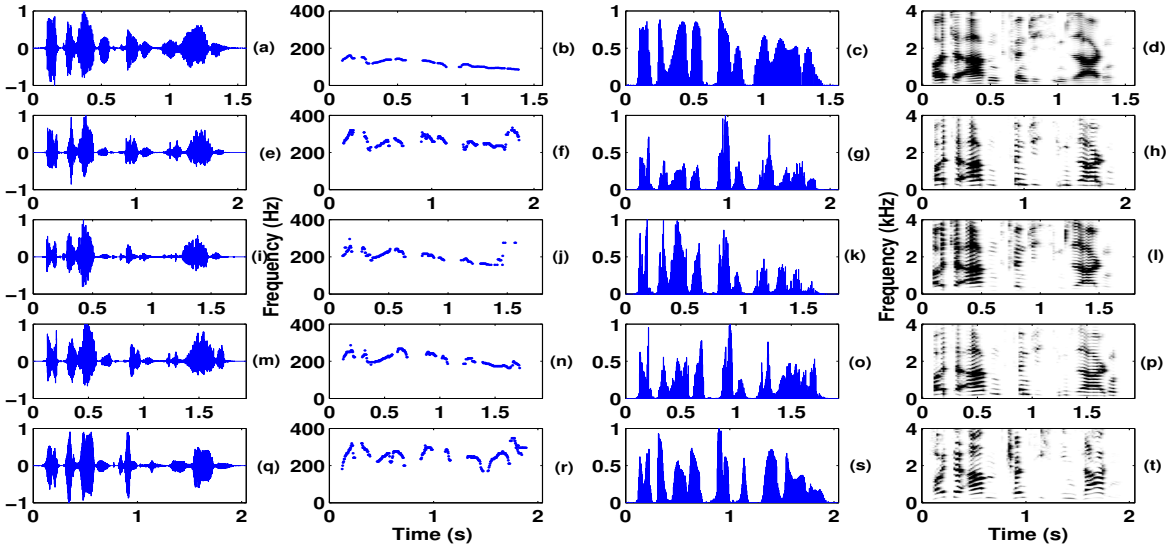

Figure 1: *Neutral to target emotion conversion. Speech waveform, pitch contour, excitation strength and spectrogram of the neutral ((a)-(d)), synthesized target emotions by deriving the scale factors from the target emotion syllables ((e)-(h)), by the gross level modification ((i)-(l)) and initial, middle and final region wise modification ((m)-(p)) and original target emotion ((q)-(t)).*

emotions indicate that the there are no spectral and temporal distortions. This shows the effectiveness of chosen source parameters and the duration information for emotion conversion. The reference target emotion speech is seldom available in practice. In such cases, we need to have *a priori* knowledge about the modification factors. This can be obtained by analyzing the different emotion speech signals taken from different speakers and texts. We can have one modification factor for the entire sentence and use it for modification, termed as gross level modification in this work. Figure 1((i)-(l)) shows the relevant plots of gross level modification. Even though the synthesized speech due to the gross level modification sounds like the target emotion, the dynamics of source parameters can be captured better by a finer analysis. This is because, the emotion specific information will be varying within a given text as we move across syllable-like units, demonstrated in Figure 1((e)-(h)). However, to have reliable estimation for each syllable-like unit, a large database is needed. A general observation is that the initial and final region syllables are relatively more affected compared to the middle region. A via media is therefore to average the effect of emotions across the initial, middle and final regions of the text. Two syllable-like units at the beginning and end of the text form the initial and final regions of the text, respectively. The syllable-like units between the initial and final regions constitute the middle region of the text. The average pitch, duration and strength of excitation modification factors are derived for each region of the neutral emotion speech by comparing the corresponding regions of the target emotion. The pitch, duration and strength modification factors computed for 5 different emotions in 9 texts of 8 speakers are given in the Table 3. Variability of pitch, duration and strength modification factors across the initial, middle and final regions show the variable effect of the emotions across different regions in a given text. The neutral speech syllable-like units are then converted to syllable-like units in the target emotion using these modification factors derived for each region. These neutral to target emotion converted

Table 3: Pitch, duration and excitation strength modification factors of initial, middle and final regions of sentences. I, M and F represents the initial, middle and final regions of the sentence, respectively.

Emotion	Pitch Mod.			Dur. Mod.			Strength Mod.		
	I	M	F	I	M	F	I	M	F
Angry	1.63	1.86	2.11	1.27	1.28	1.04	0.77	0.84	1.30
Happy	1.62	1.80	1.95	1.15	1.06	1.03	0.75	0.75	0.93
Boredom	1.19	0.94	0.98	1.24	1.25	1.19	0.99	1.04	1.13
Fear	1.38	1.53	1.83	1.11	0.98	0.90	0.67	0.84	1.33

syllable-like units are concatenated to obtain the speech of the target emotion. Figure 1((m)-(p)) shows the relevant plots for the region wise modification. From the Figure 1((n)) it is to be noted that the shape of the modified pitch contour matches more closely with the target emotion pitch contour than that of the neutral emotion. Same trend can be observed in the strength of excitation plot also. Alternatively, in Figure 1(j), even though the range of the pitch values match with that of the target emotion, the gross trend of pitch contour matches closely to that of the neutral emotion speech. Similar observation can be made with respect of excitation strength of gross level modification. Comparing the Figures 1((m)-(p)) and 1((i)-(l)), it may be noted that the shape of the pitch contour and strength of excitation of the target emotion are preserved better in the speech synthesized using region wise modification than the gross level modification. Also by comparing the Figures 1((e)-(h)) and 1((m)-(p)), it is to be noted that the shape of the pitch contour and excitation strength of synthesized angry emotion of region wise modification match closely to the syllable level modification.

In all the cases the speech is synthesized by exciting the LP

Table 4: Ranking used in perceptual test to judge the similarity of the synthesized emotion with the target emotion.

Rating	Speech Quality	Description for evaluating synthesized emotions
1	Very poor	sounds exactly like neutral
2	Poor	sounds slightly different from neutral
3	Good	sounds different from neutral
4	Very Good	sounds more different from neutral
5	Excellent	sounds exactly like target

Table 5: Comparison mean opinion scores for the emotional speech synthesized by modifying parameters of each syllable, gross level and initial,middle and final regions of speech in neutral emotion.

Method	CMOS			
	Angry	Happy	Boredom	Fear
$Syllable$	3.60	3.53	3.83	3.50
$Gross$	2.78	2.42	3.44	2.53
$Initial-Middle-Final$	3.38	3.17	3.72	3.22

coefficients obtained from the LP analysis of the neutral speech, with the modified residual in which the source and supra segmental modifications are incorporated. The spectral features of the neutral are modulated according the source and supra segmental modification in the neutral residual that make the spectrum of the synthesized speech looks like from the target. This can be confirmed by comparing the Figures 1((h)) and 1((t)). From the figures, the spectrogram of the speech synthesized by modifying the neutral speech syllables according to target angry syllables are more similar to that of the original target angry emotion spectrogram. This shows the significance of the selected source and supra segmental features in carrying the emotional information. The spectrogram of the speech synthesized by modifying initial, middle and final region syllables of the neutral speech shows the spectral characteristics similar to that of the target emotion. The smooth transitions of spectral characteristics indicate that there are no spectral and temporal distortions present in all the three modification methods.

A subjective evaluation is performed among 15 research scholars in the department to compare the quality of the synthesized emotion. All the subjects are presented with original neutral and target emotions and synthesized emotions using the syllable, gross level and region wise methods. The subjects were asked to compare the synthesized emotions in the coded file with the original neutral and target files and rate them according to the descriptions given in Table 4. A total of 40 (5X4X2) speech files synthesized from a male and a female speakers of the German emotional speech corpus are used. The comparison opinion scores obtained for the each of the emotion are averaged to get the comparison mean opinion scores (CMOS). The CMOS obtained for each emotion for all three methods are presented in Table 5.

It has to be observed that the the emotions synthesized using region wise modification of the source and suprasegmental features shows higher CMOS than the gross level modification of the parameters from neutral. The fear and happy emotions shows the lowest CMOS indicating the most confusable emotions with neutral. No significant difference in CMOS values are observed in case of boredom emotion. Some of the synthesized emotion speech samples are provided in the following link $http : //www.iitg.ac.in/ece/emstlab/emotionconversion.htm$.

5. summary and future work

The present work demonstrated an approach for neutral to target emotion conversion using the source and suprasegmental information. The emotion specific source and duration modification factors are estimated by the analysis of an emotional database. The prosody and strength of the LP residual of neutral speech is modified according to the estimated modification factors. The target emotion speech is then synthesized using the modified LP residual and LP coefficients of neutral speech. Finally, the initial, middle and final region wise modification is demonstrated to be more effective for emotion conversion.The comparison subjective study indicates that the synthesized speech of the region wise modification gives comparatively more effective neutral to emotion conversion than the gross level modification.

The future work should include the target emotion vocal tract modification to further increase the effectiveness. The proposed approach needs to be tested across different languages, speakers and other emotions.

6. Acknowledgement

This is a part of ongoing (2007-2011) UKIERI project between IIT Guwahati, India, IIIT Hyderabad, India and CSTR, University of Edinburgh, UK

7. References

[1] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1145–1154, Jul. 2006.

[2] M. Theune, K. Meijs, D. Heylen, and R. Ordelman, "Generating expressive speech for story telling applications," *IEEE Trans Audio, Speech and Language Proc.*, vol. 14(4), pp. 1099–1108, July 2006.

[3] S. R. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech," in *Proc. INTERSPEECH*, Sep. 2010, pp. 781–784.

[4] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 972–980, May 2006.

[5] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 8, pp. 1602–1614, Nov. 2008.

[6] S. R. M. Prasanna, D. Govind, K. S. Rao, and B. Yenanarayana, "Fast prosody modification using instants of significant excitation," in *Proc Speech Prosody*, May 2010.

[7] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech and Language Process.*, vol. 17, no. 4, pp. 614–625, May 2009.

[8] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlemeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.