Engineering, **2013, 5, 73-77**
http://dx.doi.org/10.4236/eng.2013.510B015 Published Online October 2013 (http://www.scirp.org/journal/eng)

Emotional Speech Synthesis Based on Prosodic Feature Modification

Ling He[1], Hua Huang[1], Margaret Lech[2]
[1]School of Electrical Engineering and Information, Sichuan University, Chengdu, China
[2]School of Electrical and Computer Engineering, RMIT University, Melbourne, Australia
Email: ling.he@scu.edu.cn, margaret.lech@rmit.edu.au

Received November 2012

ABSTRACT

The synthesis of emotional speech has wide applications in the field of human-computer interaction, medicine, industry and so on. In this work, an emotional speech synthesis system is proposed based on prosodic features modification and Time Domain Pitch Synchronous OverLap Add (TD-PSOLA) waveform concatenative algorithm. The system produces synthesized speech with four types of emotion: angry, happy, sad and bored. The experiment results show that the proposed emotional speech synthesis system achieves a good performance. The produced utterances present clear emotional expression. The subjective test reaches high classification accuracy for different types of synthesized emotional speech utterances.

Keywords: Emotional Speech Synthesis; Prosodic Features; Time Domain Pitch Synchronous Overlap Add

1. Introduction

The modern speech synthesis system has a wide variety of applications. In the call-centers, the speech synthesizer could conduct dialogues with customers. The intelligent virtual agent devices could read loud to users using speech synthesis techniques, such as in the video games or children's toys. In the medicine field, the speech synthesizer could even be used to speak for sufferers who lost the use of their voice. The majority of modern speech synthesizers could produce voice (acoustic waveform) from text. However, few machines can "speak" in a totally natural way as a human being. One of the major drawbacks existing is that the machines could not speak with emotions. Emotion expression is a vital part in human communication, an effective human-to-human communication is virtually impossible without speakers could not express or understand affections. The emotional speech synthesis aims to add human emotions into synthesized speech to produce more natural affective speech.

Two major approaches to emotional speech synthesis dominate the literature: formant synthesis and concatenative synthesis [1]. Formant synthesis generates acoustic speech data entirely based on rules surrounding the acoustic correlates of the speech and does not utilize human speech recordings. Acoustic profiles for each emotion category are derived from the literature and manually adapted [2] to create a signal. In 1989, Jenet Cahn [3] implemented the synthesized emotional speech firstly using a formant synthesis system. After then, several researches have been done using the formant synthesizer [4]. Despite of the high degree of control over the acoustic parameters provided in this technique, formant synthesis is not widely applied, since the resulting speech, has an unnatural, mechanical sound. In contrast, concatenative synthesis [5] joins recordings of a human speaker to generate the synthetic speech. The generating utterances are more natural. However, in order to produce variety of emotions, the system requires a larger size of speech database to build a selecting units pool [6-9]. To solve this problem, several researchers incorporate prosodic strategies into unit selection [10,11]. In this way, smaller number of speech corpora is required. Different types of emotion could be added into the synthesized speech through modifying corresponding to acoustic parameters (like the fundamental frequency, or the duration of speech contour), and then applying the waveform concatenative approaches, such as the PSOLA (Pitch Synchronous OverLap Add) algorithm [12].

In this work, an emotional speech synthesis system is proposed based on prosodic feature modification and TS-PSOLA concatenative synthesis method.

2. Emotional Speech Synthesis System

In this work, a prosodic modification based emotional speech synthesis system is proposed. The block diagram

is illustrated in **Figure 1**.

In the experiment, the speech utterances, which are selected from emotional speech database with four different types of emotion: angry, happy, sad and bored, are divided into a set of units. For the concatenative speech synthesis, there are various possible choices for the type of unit, the most popularly used types include words, syllables, phonemes, demi-syllables and diphones. The speech unit is the basic synthesis building block. It is known that when the length of unit goes longer, the effect of context decreases and the quality of resulting synthesized speech increase. Considering the size of speech database, in this work, the chosen concatenative unit is word. For the segmented units, the prosodic analysis is applied to calculate the pitch, energy and duration rules. The length of silence is also calculated to decide the pause assignment. The prosodic feature templates for different types of emotion are then built up using the estimated parameters.

In the next step, for the neutral input speech, the utterances are segmented into units (word) firstly. Then the prosodic features are extracted for each unit, and modified according to the prosodic feature templates which are built in the previous step, with the corresponding emotion type. At last, the time-domain pitch synchronous overlap add (TS-PSOLA) concatenative synthesis method is used to smooth and modify unit boundary, and to produce the final synthesized emotional speech.

3. Speech Database

The publicly available Berlin emotional speech (BES) database [13] is used in this work, for the purpose of comparison with other experiments. In the construction of the database, the text materials are carefully chosen to achieve natural emotion arousal. 10 sentences (5 short and 5 long sentences) frequently used in everyday communication are selected. The speech utterances are produced by ten actors (5 female and 5 male). A total of 248 emotional recordings are selected in this work, with four different types of emotion: angry, happy, sad and bored, to build up the emotional prosodic feature templates. For the emotion "angry", it is a short clipped speech, with one word or syllable being more strongly stressed. For the emotion "happy", the voice tone is high pitched, speech is faster or louder than usual. Under "sad" emotional state, a person tends to speak slowly, and use a low voice tone. For the emotion "bored", the voice tone is cold and dull. For each emotion, the number of recordings is 62. The sampling frequency for the data is 16 kHz.

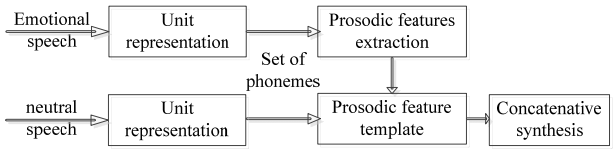

Figure 1. Block diagram of emotional speech synthesis.

In addition, 40 speech recordings under neutral state are used for testing. A neutral voice tone is even, relaxed, without marked stress on individual syllables. For each neutral speech, four synthesized utterances are produced with emotion type of angry, happy, sad and bored.

4. Prosodic Features Calculation

In this work, three prosodic features are extracted: fundamental frequency, energy and time duration.

4.1. Calculation of Fundamental Frequency

As illustrated in **Figure 2**, the fundamental frequency F0 of vocal folds vibration is estimated simultaneously in the time domain using the autocorrelation method [13], and in the frequency domain using the cepstral method [13]. The average value of these two measurements provides the final estimate of F0.

The frequency domain cepstrum method of the F0 estimation looks for a periodicity in the log spectrum of the signal; if the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum is expected to show a peak corresponding to the spacing between the harmonics: *i.e*. the fundamental frequency.

The time domain autocorrelation method, on the other hand, estimates the fundamental frequency directly from the waveform using the autocorrelation function which is expect to show peaks at delays corresponding to multiples of the glottal wave period (1/F0). The autocorrelation is calculated as:

$$R_m(k) = \sum_{j=0}^{N-k-1} s_m(j) s_m(j+k) \tag{1}$$

where s is the speech signal.

4.2. Calculation of Energy

Speech utterance is a non-stationary signal. However, it could be viewed as a stationary signal in a short-time roughly ranging between 16 ms and 32 ms.

In the experiment, the short-time energy is calculated

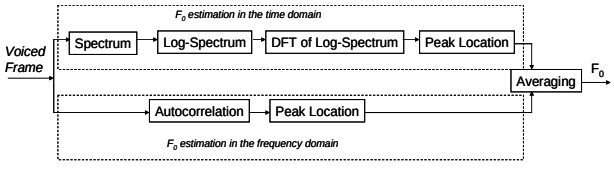

Figure 2. A flowchart of the fundamental frequency estimation method.

for the speech frame with the length of 16 ms and 50% overlap. The short-time energy for a speech signal $s[n]$ is calculated as:

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} (s[m]w[\hat{n}-m])^2 \quad (2)$$

where $s[m]$ is the speech signal, $w[\hat{n}-m]$ is the applied window. $\hat{n} = rR$, where R represents frame shift and r is the integer.

4.3. Calculation of Time Duration

The time duration of each unit under different emotional states is calculated to obtain the prosodic characteristics of speech signals.

Moreover, the duration of silence in each sentence is estimated, in order to get the pause assignment for each type of emotion.

The duration of silence is calculated through speech endpoint detection method. The endpoint detection algorithm aims to identify the speech signal from background noise. The short-time energy is calculated to detect voiced speech and short-time zero crossing rate is estimated to decide the voiceless speech. The length of silence is then calculated while removing the speech parts.

Table 1 shows the average value of prosodic features for the units under five different types of emotion.

5. TS-PSOLA Method

Time Domain Pitch Synchronous Overlap Add is a popularly used concatenative synthesis method. The basic contribution of TD-PSOLA technique is to modify the pitch directly on the speech waveform. There are three steps for TD-PSOLA: pitch synchronization analysis, pitch synchronization modification and pitch synchronization synthesis.

Pitch synchronization analysis is the core of TD-PSOLA method, it finishes two tasks: fundamental frequency detection and pitch mark. Let $x_m(n)$ denotes the windowed short time signal:

$$x_{\mathrm{m}}(n) = h_{\mathrm{m}}(t_{\mathrm{m}} - \mathrm{n})\mathrm{x}(\mathrm{n}) \quad (2)$$

where t_m is the mark point of pitch, h_m is the window sequence.

Pitch synchronization modification adapts the pitch mark by changing the duration (insert or delete the sequence with the length of pitch duration) and tone (increase or decrease the fundamental frequency).

The pitch synchronization synthesis adds the new sequence signal produced in the previous step. In this work, the Least-Square Overlap-Added Scheme method is used to get the synthesized signal:

$$\bar{\mathrm{x}}(n) = \frac{\sum_q a_q \bar{\mathrm{x}}_{\mathrm{q}}(n)\bar{h}_{\mathrm{q}}(\bar{t}_{\mathrm{q}} - \mathrm{n})}{\sum_q \bar{h}_q^2(\bar{t}_{\mathrm{q}} - \mathrm{n})} \quad (3)$$

Table 1. Average prosodic feature values of units under five emotional states.

Prosodic features	Emotion types				
	Neutral	Angry	Happy	Sad	Bored
Pitch (Hz)	149	251	203	126	169
Time duration (s)	0.24	0.16	0.18	0.26	0.25
Energy (db)	56.4	73.2	64.8	52.1	49.3

where $\bar{t}_{\mathrm{q}}$ is the new pitch mark, $\bar{h}_{\mathrm{q}}$ is the synthesized window sequence, a_q is the weight to compensate the energy loss when modifying the pitch value.

6. Experiments and Results

Figure 3 illustrates an example of the emotional speech synthesis applying the proposed method in this work. **Figure 3** shows the waveforms of the utterance "Das schwarze Stück Papier befindet sich da oben neben dem Holzstück" produced under neutral and four types of emotional states: angry, happy, sad and bored. **Figure 3** also shows the waveforms of the synthesized emotional speech under four different types of emotional states based on the prosodic feature modification algorithm.

In order to evaluate the performance of proposed emotional speech synthesis system, a subjective test is made. Six participators listened the synthesized emotional speech utterances, and selected which type of emotion they are. The subjective test results (confusion matrix) are listed in **Table 2**.

7. Conclusions and Discussion

In this work, a prosodic feature modification method combined with PSOLA algorithm is proposed in order to add the emotional color to a neutral speech. **Figure 3** shows the waveforms of the natural and synthesized speech signals under four different types of emotion. It is seen that the waveforms of the synthesized speech are distinguished among different types of emotion, and they are similar to the waveforms of natural speech pronounced by human beings. The subjective test as illustrated in **Table 2** indicates that the synthesized speech signals contain clear emotion colors, it is easy to classify emotion types from the synthesized utterances by human being. For the synthesized speech, the emotion "angry" is easiest to classify. This is because the natural emotion "angry" contains strong emotional arousal, resulting in distinguished prosodic characteristics. The emotion "bored" obtaines the lowest subjective classification accuracy, this is probably because the acoustic characteristics of emotion "bored" is not clear, this kind of emotion is mainly expressed through the linguistic information.

One of the shortcomings of this work is that the emotional speech data size is limited. In order to meet the

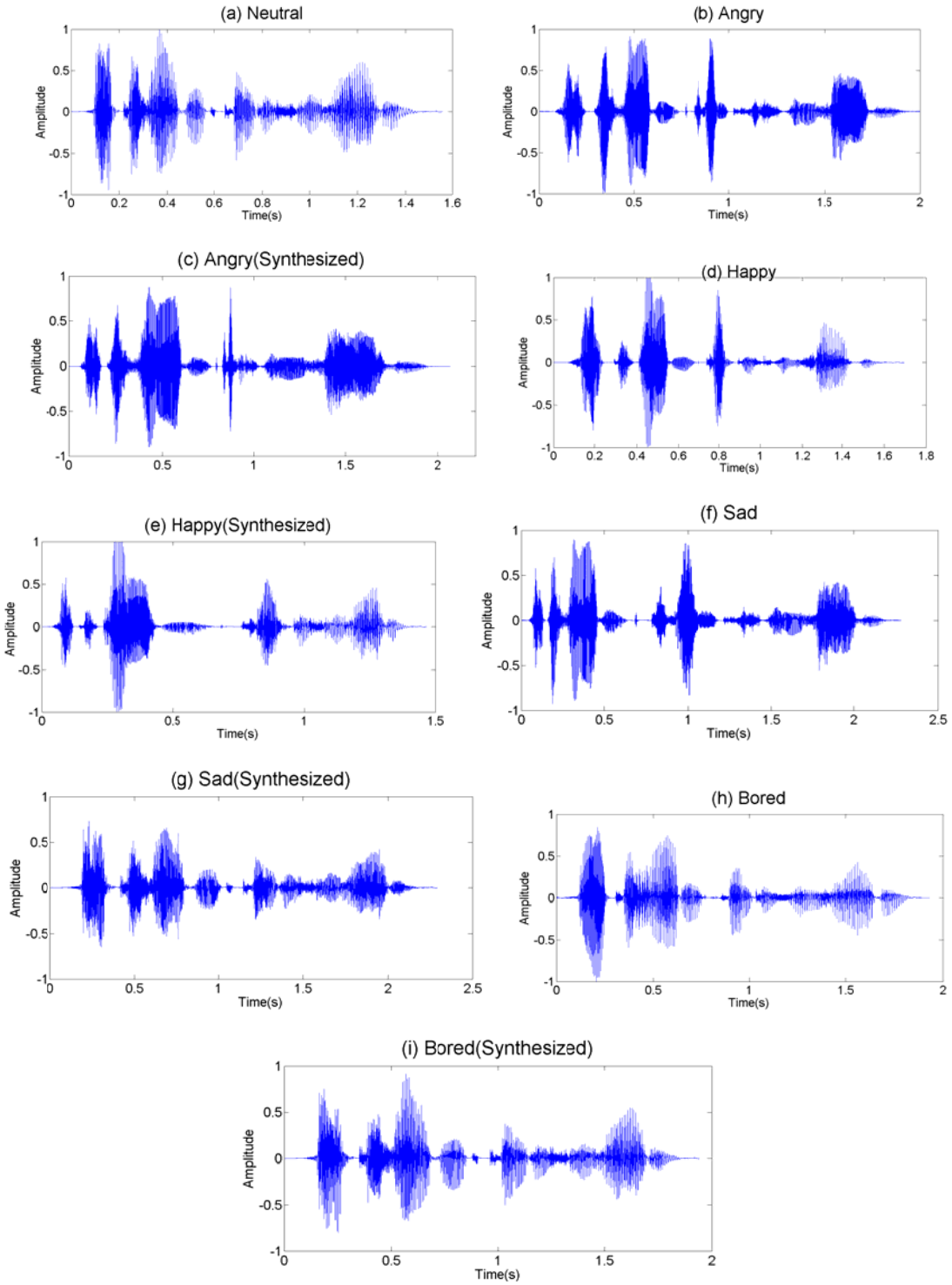

Figure 3. Waveforms of neutral speech (a), emotional speech (angry (b), happy (d), sad (f) and bored (h)) and synthesized emotional speech (angry (c), happy (e), sad (g) and bored (i)).

needs of natural conversation with rich emotion expression, a much larger size of emotional speech units database is required.

In order to produce more natural emotional speech, the concatenative unit selected in this work is "word", because the speech database provides corresponding words for each type of emotion. However, in the real-life application, there are essentially an infinite number of words. Therefore, there will be some words in the synthesized speech which is not in the "dictionary" database, only the

Table 2. Subjective test results.

Synthesized emotion types	Classified emotion types			
	Angry	Happy	Sad	bored
Angry	90.3%	8.1%	1.6%	0%
Happy	8.1%	87.1%	3.2%	1.6%
Sad	6.5%	6.5%	80.5%	6.5%
Bored	3.2%	8.1%	11.3%	77.4%

optimized word could be found through unit selection algorithm, resulting lower quality of the final speech. One of the solutions is to choose shorter length of unit type, like syllables, phonemes and so on. In this way, a smaller size of data set will be required.

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, "Emotion Recognition in Human-Computer Interaction," *Signal Processing Magazine, IEEE*, Vol. 18, No. 1, 2001, pp. 32-80. http://dx.doi.org/10.1109/79.911197

[2] M. Schröder, R. Cowie and E. Cowie, "Emotional Speech Synthesis: A Review," Eurospeech-2001, 2001.

[3] J. E. Cahn, "The Generation of Affect in Synthesized Speech," *Journal of the American Voice I/O Society*, Vol. 9, 1990, pp. 1-19.

[4] F. Burkhardt and F. Sendlmeier, "Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis," ISCA Workshop on Speech & Emotion, Northern Ireland, 2000, pp. 151-156.

[5] M. Bulut, S. Narayan and A. Syrdal, "Expressive Speech Synthesis Using a Concatenative Synthesizer," *Proceedings of ICSLP*, 2002, pp. 1265-1268.

[6] E. Eide, "Preservation, Identification, and Use of Emotion in a Textto-Speech System," *Proceedings of IEEE Workshop on Speech Synthesis*, 2002, pp. 127-130.

[7] A. W. Black and N. Cambpbell, "Optimising Selection of Units from Speech Database for Concatenative Synthesis," *Proceedings of EUROSPEECH*-95, 1995, pp. 581-584.

[8] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza and M. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English," *IEEE Transactions on Speech Audio Process*, Vol. 14, No. 4, 2006, pp. 1099-1108. http://dx.doi.org/10.1109/TASL.2006.876123

[9] W. Hamza, R. Bakis, E. Eide, M. Picheny and J. Pitrelli, "The IBM Expressive Speech Synthesis System," *Proceedings of ICSLP*, 2004.

[10] G. Hofer, K. Richmond and R. Clark, "Informed Blending of Databases for Emotional Speech Synthesis," *Proceedings of Interspeech*, 2005, pp. 501-504.

[11] M. Schroder, "Speech and Emotion Research: An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis," Ph.D. Thesis, Saarland University, Saarland, 2004.

[12] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals," Prentice-Hall, Inc., Englewood Cliffs, 1978.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, *et al.*, "A Database of German Emotional Speech," *Proceedings of Interspeech*, 2005.

