Modeling Prosody Patterns for Chinese Expressive Text-to-Speech Synthesis

Zhiyong WU, Lianhong CAI, Helen M. MENG
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems
Graduate School at Shenzhen, Tsinghua University, Shenzhen
zywu@sz.tsinghua.edu.cn, clh-dcs@tsinghua.edu.cn, hmmeng@se.cuhk.edu.hk

***Abstract*—This paper proposes an approach for modeling the prosody patterns of the acoustic features for Chinese expressive text-to-speech (TTS) synthesis. Based on the observation that the speaker usually tends to put more emphasis on one particular syllable within a multi-syllabic prosodic word, we identify such syllable as the core syllable that can be derived from the semantic stress and tone information of the text prompt. We then classify the syllables in speech into four classes, based on their relations with the core syllable in a prosodic word. We analyze the contrastive (neutral versus expressive) speech recordings for each of four classes, and develop a perturbation model that takes into account the prosody pattern to transform neutral speech to expressive speech. Perceptual experiments on both neutral speech recordings and neutral TTS outputs involving 13 subjects indicate that the proposed approach can significantly enhance expressivity in synthesizing expressive speech.**

Keywords-expressive text-to-speech (TTS); prosody pattern; non-linear perturbaton model

I. INTRODUCTION

There has been a rich repository of research work in the area of expressive speech synthesis [1-4]. Previous research has shown that the realization of expressivity can be achieved through speech prosody and their acoustic correlates, including intonation, amplitude, duration, timing and voice quality [5-8]. [9] attempted to synthesize four types of emotional speech (happiness, sadness, fear, anger) at three levels (strong, weak, medium) and found that the performance of linear modification model (LMM) is inferior to the approaches of Gaussian mixture model (GMM) and classification and regression trees (CART). The main reason is because the two latter approaches involve finer partitioning of the prosodic space based on stress and linguistic information. Similar findings were observed in [10] for expressive speech synthesis based on text semantics.

The variation of the acoustic features may reveal some *prosody patterns* while migrating from neutral speech to expressive speech. Several work on the analysis of expressive focus speech had proven this. Compared with neutral speech, the pitch and intensity of focus word generally increase, while the same features of words preceding the focus words tend to decrease in some language [11]. [12] analyzed the pitch and durations of vowels from focus speech, and found that the durations were shorter and the pitches were higher for high vowels than for low vowels. The durations were analyzed in [13] considering the distance between different phones and focus word. Results show that the closer the phone is to a focus word, the longer is the duration.

This work attempts to analyze the *prosody patterns* of Chinese syllables, based on their relative locations with respective to the *core syllables* in the prosodic words. Here the core syllables may be the stressed syllables, the syllables with falling tone, etc. We believe that such analysis will help us identify the patterns revealed by the variations of syllable acoustic features while migrating from neutral to expressive speech. The patterns can be utilized in improving our existing work on perturbation model to synthesize more natural and expressive speech for Chinese expressive TTS synthesis.

The rest of this paper is organized as follows: Section II presents the corpus with contrastive neutral and expressive recordings to support our investigation. Section III describes the analysis of acoustic features relating to prosody patterns. Section IV details the perturbation model to generate expressive speech from neutral speech by incorporating the prosody patterns. Section V describes our experimental design and perceptual evaluations of the proposed method. Finally, Section VI lays out conclusions and possible future directions.

II. CORPUS

This work is conducted in the context of a spoken dialog system in the tourist information domain, where TTS is used to generate expressive speech outputs to convey and emphasize the beauty and specialties of a scenic spot to the user.

A. Text Prompts

Text prompts are derived from text passages, corresponding to 20 scenic spots, which are sourced from the Discover Hong Kong website of the Hong Kong Tourism Board [14]. Each text passage begins with descriptive paragraph introducing the attractive features of a scenic spot, followed by an informative paragraph about opening hours and/or ticket prices, and finally a procedural paragraph about transportation and walking directions. The set of 20 text passages contains 60 paragraphs with a total of 357 utterances, 1,358 Chinese prosodic words and 3,340 Chinese syllables.

We have chosen the prosodic word as the basic unit for analysis and modeling since it is the smallest constituent at the lowest level of prosodic hierarchy, and it consists of a group of syllables uttered closely and continuously in an utterance [15].

The work is jointly supported by the National Natural Science Foundation of China (60805008, 60928005, 60910130), the National High Technology Research and Development Program of China (2009AA011905), and the Ph.D Programs Foundation of Ministry of Education of China (200800031015).

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

B. Text Annotation

The text prompts belong to three different paragraphs. The descriptive paragraph often contains commendatory words to describe scenic characteristics or specialties about a spot. The informative and procedural paragraphs are both used to provide useful facts. Speech synthesis for such text prompts will need to incorporate appropriate prosodic-word level prosody, with suitable emphasis to draw the attention of the tourist.

1) Semantic Expressivity Annoation

The PAD model [16] is adopted as the framework to annotate text prompts. A set of heuristics are designed [10,17] such that the *P*/*A* descriptors can be parameterized from the semantic expressivity of text prompts. The *P*/*A* values range from -1.0 to 1.0, with 0.0 representing neutral. Commendatory words would get high *P* values, while words with negative connotations would get low or even negative *P* values. Superlative words and the focus of the text message would get high *A* values.

2) Core Syllable Annotation

We choose the stress and tone as the cues for annotating the core syllable within a prosodic word. Stress is much related with the semantic meaning of the text prompts. Emphasis is usually placed on stressed syllables when uttering expressive speech. In our corpus, superlative words such as "最(most)", "极(super)", "很(very)", etc. and commendatory words such as "欢迎(popular)", "享誉(famous)", "漂亮(beautiful)", etc. are annotated as stressed syllables. High tone in Chinese Mandarin is another useful cue for the core syllable. This is because the uttering of a syllable with high tone (tone 1) usually requires more efforts and the speaker tends to put more emphasis on it.

C. Speech Recordings

A male native Mandarin speaker was invited to record in a sound-proof studio. The speaker has several years of research experience in expressive speech processing, and hence is professional in understanding the differences between neutral and expressive speech. For each text prompt, the speaker was asked to record contrastive versions of neutral and expressive speech. The expressive speech recordings should contain prosodic-word level expressivity that coveys and emphasizes the semantic of the prosodic words (e.g., the beauty of a scenic spot). The 60 text prompts in the 20 passages tend to be long and each may contain one to eight sentences, leading to 357 utterances in total. The sound files are saved in the wav format (16 bit mono, sampled at 16 kHz). This data is needed for data analysis and modeling. We set aside another disjoint set of 40 utterances to be used as the test set for experimentation.

III. Acoustic Analysis of Prosody Pattern

The objective of this work is to analyze how expressive elements are realized in the acoustic signal, and especially how syllable with a certain expressivity influences its neighbors to reveal a certain prosody pattern.

A. Acoustic Features

The acoustic features that are commonly associated with prosody include fundamental frequency (f_0), intensity and speaking rate. We choose to extract following features:

- f_0 mean, f_0 range, f_0 slope;
- root mean square (RMS) energy;
- duration of the syllable; and
- pause length before a prosodic word.

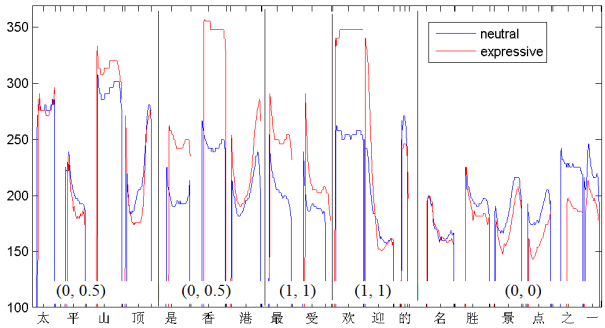

Figure 1. Comparison of pitch contours beween neutral speech and its expressive counterpart. The Chinese prosodic words from left to right may be translated as: "Victoria Peak", "is Hong Kong's", "most", "popular", "scenic spot".

B. Classification of Syllables

We categorized the syllables in the speech corpus into four classes, based on the location of the syllable in relation with the core syllable within a prosodic word:

- **Class 1**: core syllable in the prosodic word;
- **Class 2**: syllable before the core syllable;
- **Class 3**: syllable after the core syllable;
- **Class 4**: all other remaining syllables.

C. Analysis of Acoustic Features for Prosody Pattern

The variations of the acoustic features are not exactly the same for all the syllables within a prosodic word while migrating from neutral to expressive speech.

Fig. 1 illustrates the difference of pitch contours between neutral speech and its expressive counterpart. It can clearly be seen that, for the two consecutive prosodic words "最受(the most)" and "欢迎的(popular)", higher emphasis (i.e. higher pitch contour) is placed on the superlative "最" and the commendatory "欢迎" than on the functional syllables "受" and "的". For the prosodic word "是香港(is Hong Kong's)", higher emphasis is placed on the syllable "香(pronounced as 'xiang1' in Pinyin)" with high tone than other two syllables. All these superlative, commendatory or high tone syllables are the core syllables as defined in Section II.B.

To analyze the prosody patterns of the acoustic features, the recorded speech utterances are first automatically segmented into syllables with a home-grown segmentation tool and then the syllable boundaries are checked manually. Acoustic measurements are then taken from the contrastive recordings (neutral versus expressive) of each utterance. We compute the ratio of each acoustic feature (F^{exp}/F^{neu}, in %) between the expressive (F^{exp}) and neutral counterparts (F^{neu}), where F denotes any of the acoustic features described above. The mean ratio is then computed for each of the class, by averaging

TABLE I. STATISTICS OF MEAN RATIO BETWEEN EXPRESSIVE AND NEUTRAL ACOUSTIC FEATURES FOR DIFFERENT (P, A) VALUES IN CLASS 1

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
f0 mean	1.03	1.10	1.09	1.21
f0 range	1.08	1.35	1.35	1.36
f0 slope	1.80	2.06	1.74	0.92
energy	1.15	1.22	1.20	1.36
duration	1.03	1.04	1.06	1.20
pause	1.12	1.21	1.12	1.50

TABLE II. STATISTICS OF MEAN RATIO BETWEEN EXPRESSIVE AND NEUTRAL ACOUSTIC FEATURES FOR DIFFERENT (P, A) VALUES IN CLASS 2

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
f0 mean	1.01	1.12	1.06	1.18
f0 range	1.03	1.28	1.30	1.31
f0 slope	1.79	2.03	1.58	1.05
energy	1.12	1.18	1.17	1.28
duration	1.02	1.03	1.05	1.18

TABLE III. STATISTICS OF MEAN RATIO BETWEEN EXPRESSIVE AND NEUTRAL ACOUSTIC FEATURES FOR DIFFERENT (P, A) VALUES IN CLASS 3

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
f0 mean	1.02	1.08	1.05	1.19
f0 range	1.06	1.30	1.32	1.34
f0 slope	1.78	1.99	1.73	1.01
energy	1.12	1.20	1.20	1.33
duration	1.03	1.03	1.06	1.21

TABLE IV. STATISTICS OF MEAN RATIO BETWEEN EXPRESSIVE AND NEUTRAL ACOUSTIC FEATURES FOR DIFFERENT (P, A) VALUES IN CLASS 4

(P, A)	(0, 0)	(0, 0.5)	(1, 0.5)	(1, 1)
f0 mean	0.98	1.03	1.02	1.13
f0 range	1.02	1.25	1.28	1.29
f0 slope	1.75	1.88	1.36	1.01
energy	1.08	1.10	1.12	1.21
duration	0.98	1.01	1.03	1.09

the ratio values of all the syllables that are belong to the same class. Results are shown in Table I to Table IV.

We can observe that the average ratios of the acoustic features between expressive and neutral speeches are mostly the biggest for Class 1 and the smallest for Class 4. The ratios for Class 2 and Class 3 lie between Class 1 and Class 4, while the ratios for Class 3 are mostly bigger than those of Class 2. These observations agree with the common perception that the speaker has a tendency to put more emphasis on the core syllables (i.e. Class 1), which will have more influence on the succeeding syllables (i.e. Class 3) than the preceding syllables (i.e. Class 2). This influence tends to decrease when the position of the syllables moves farther from the core syllable.

IV. PERTURBATION MODEL FOR PROSODY PATTERN SYNTHESIS

We have proposed a nonlinear perturbation model [10,17] that incorporates the (P, A) values of a text prompt to transform neutral speech into expressive speech, as shown below:

$$\frac{F^{exp}}{F^{neu}} = C_1 P \exp(-C_2 A) + C_3 A \exp(-C_4 P) + C_5 \qquad (1)$$

where F^{exp} can be any of the acoustic features (as shown above) from expressive speech, F^{neu} is the corresponding feature from neutral speech, F^{exp}/F^{neu} is the ratio of acoustic feature between expressive and neutral speech, and $C_1,\ldots,C_5$ are coefficients. $C_1,\ldots,C_5$ are the parameters of the perturbation model, and can be estimated from the contrastive speech recordings using the non-linear least-squares regression method [10,17].

A. Perturbation without Considering Prosody Pattern

In our previous work [10,17], all the speech recordings are grouped together to estimated the coefficients, notwithstanding the fact that the acoustic features of the core or non-core syllables in a prosodic word may vary greatly. Only one set of model coefficients are estimated, and then used in transforming the neutral speech into expressive speech.

B. Perturbation Considering Prosody Pattern

This work categorized the Chinese syllables into four classes, according to the location of the syllable in relation with the core syllable in a prosodic word. The speech recordings are then grouped into four categories according to the classes of the syllables. Four sets of model coefficients are estimated, with one set of coefficients for each class. Following steps are then utilized to synthesize expressive speech by considering prosody pattern:

- Decide the core syllable of a prosodic word by virtue of the principles as shown in Section II.B;
- Determine the classes of syllables based on their location relations with the core syllable;
- Choose model coefficients according to syllable class;
- Obtain F^{exp}/F^{neu} value based on the perturbation model (1) using the chosen model coefficients and P, A values of the prosodic word;
- Synthesize the expressive speech by perturbing the acoustic features of neutral speech using the following realization method.

C. Realization of Perturbation with STRAIGHT

STRAIGHT algorithm [18] is utilized to implement perturbation. Neutral speech input is first analyzed by STRAIGHT to obtain spectrum, energy, pitch and durations. These acoustic features are then modulated with the proposed perturbation model to generate target acoustic features for expressive speech. The perturbed features and spectrum are then fed into STRAIGHT to resynthesize expressive speech.

Let $S_i(t)$ be the i-th syllable waveform of the neutral speech with boundaries $[b_i, e_i]$, i.e. begin/end time step; $S'_i(n)$ be the corresponding i-th syllable waveform of the target expressive speech. The perturbation is then realized through five steps:

1) Modifiying pitch contour

Let $P_i(t)$ be the pitch contour of the syllable waveform $S_i(t)$. Let $P_{mean,i}$, $P_{range,i}$ and $P_{slope,i}$ be the f_0 mean, f_0 range and f_0 slope for syllable i; R_{mean}, R_{range} and R_{slope} be the perturbation

ratios for f_0 mean, f_0 range and f_0 slope respectively. Then the new pitch contour $P'_i(t)$ for expressive speech is calculated as:

$$\begin{aligned}
&P'_{mean,i} = P_{mean,i} \times R_{mean} \\
&P'_{range,i} = P_{range,i} \times R_{range} \\
&P'_{slope,i} = P_{slope,i} \times R_{slope} \\
&\hat{P}_i(t) = P_i(t) + \left(P'_{slope,i} - P_{slope,i}\right)\left(t - \bar{t}\right) - P_{mean,i} \\
&\hat{P}_{range,i} = \max\left[\hat{P}_i(t)\right] - \min\left[\hat{P}_i(t)\right] \\
&P'_i(t) = \hat{P}_i(t)\frac{P'_{range,i}}{\hat{P}_{range,i}} + P'_{mean,i} \\
&t \in [b_i, e_i], \qquad \bar{t} = (b_i + e_i)/2
\end{aligned} \tag{2}$$

2) Modifying syllable duration and pause length

Let D_i be the duration of syllable i and Z_i be the pause length before syllable i; $R_{duration}$ and $Z_{duration}$ be the perturbation ratios of syllable and pause duration. The boundaries $[b'_i, e'_i]$ for expressive speech $S'_i(t)$ of syllable i are computed as:

$$\begin{aligned}
&D'_i = D_i \times R_{duration} = (e_i - b_i + 1) \times R_{duration} \\
&Z'_i = Z_i \times Z_{duration} = (b_i - e_{i-1} - 1) \times Z_{duration} \\
&b'_i = \sum_{j=1}^{i-1} D'_j + \sum_{j=1}^{i} Z'_j \\
&e'_i = b'_i + D'_i
\end{aligned} \tag{3}$$

3) Synthesizing expressive speech with STRAIGHT

STRAIGHT algorithm is used to synthesize intermediate expressive speech $S''_i(t)$ (without energy modification) for syllable i based on the new pitch contours $P'_i(t)$, time-axis mapping information $T_i(t)$ and the original spectrum $M_i(t)$:

$$\begin{aligned}
&T_i(t) = \frac{e_i - b_i}{\tilde{e}_i - \tilde{b}_i}\left(t - \tilde{b}_i\right) + b_i \\
&S''_i(t) = f\left(M_i(t), \tilde{P}_i(t), T_i(t)\right), \; t \in \left[\tilde{b}_i, \tilde{e}_i\right]
\end{aligned} \tag{4}$$

where $f(\cdot)$ represents the synthesis process of STRAIGHT algorithm, details of which can be found in [18].

4) Modifying energy

Let R_{energy} be the perturbation ratio for energy. The energy of the intermediate speech $S''_i(t)$ is scaled by R_{energy}, and further smoothed by a Hamming window $W_i(t)$ to generate the final expressive speech $S'_i(t)$ for syllable i:

$$\begin{aligned}
&S'_i(t) = S''_i(t) R_{energy} W_i(t), \; t \in \left[b_i, e_i\right] \\
&W_i(t) = 0.53836 - 0.46164 \cos\left(\frac{2\pi(t - b_i)}{e_i - b_i}\right)
\end{aligned} \tag{5}$$

5) Generating final expressive speech

Finally, the entire expressive speech is generated by concatenating all the N syllable waveforms:

$$S'(t) = \{S'_1(t), ..., S'_i(t), ..., S'_N(t)\} \tag{6}$$

V. Experimental Results

We devised a set of perceptual experiments to evaluate and compare the performance of the perturbation model with or without synthesizing prosody pattern.

A. Experiment on Neutral Speech Recordings

The first experiment was conducted on the neutral speech recordings to validate the efficacy of the proposed perturbation model. We selected 20 text prompts within our tourist domain. Each text prompt was tokenized into prosodic words with a home-grown tool and annotated with (P, A) values for each prosodic word according to the heuristics. We also verified the annotated prosodic words have a good coverage of (P, A) value combinations. We then ran the listening test where each text prompt is presented to the subject in four speech files:

(a) the neutral speech recording from the original speaker who recorded the speech corpus;
(b) the expressive speech recording from the same speaker;
(c) a perturbed speech signal without prosody pattern from the neutral speech recording (a); and
(d) a perturbed speech signal with prosody pattern from (a).

13 native Mandarin speakers (nine male, four female) were recruited to be subjects for the listening test. The speech files were played to the subjects either in the order (a)-(b)-(x) or (b)-(a)-(x), where (x) refers to perturbed speech and may be (c) or (d). Each subject was presented with the text prompt while listening, and was asked to judge whether (x) sounded more similar to its counterpart (a) or (b). Results shown in Table V indicate that the perturbation model with prosody pattern can generate appropriate expressivity for over 82% of the speech files, which is 9% higher than the perturbation model without prosody pattern.

B. Experiment on Neutral TTS Outputs

The purpose of this work is to generate expressive speech from neutral speech outputs of an existing concatinative TTS synthesizer, which utilizes voice libraries from *different female* speaker. To access the extensibility of the proposed method, a new evaluation experiment was conducted to compare between perturbation of neutral speech recordings and neutral TTS outputs. Another 20 text prompts were selected, tokenized into prosodic words, and annotated with (P, A) values. Each text prompt was presented in the form of seven speech files:

- the EXPressive and NEUtral speech RECordings from the original speaker (denoted as EXP_REC and NEU_REC respectively);
- the speech generated from perturbation of NEU_REC with No Prosody Pattern (denoted as NPP_REC);
- the speech generated from perturbation of NEU_REC With Prosody Pattern (denoted as WPP_REC);
- NEUtral TTS synthetic speech (denoted as NEU_TTS);
- the speech generated from perturbation of NEU_TTS with No Prosody Pattern (denoted as NPP_TTS); and
- the speech generated from perturbation of NEU_TTS With Prosody Pattern (denoted as WPP_TTS).

The same 13 subjects were recruited in the experiment, and

each subject was asked to score each speech file based on a five-point Likert scale:

5 Expressive – natural and expressive like human speech;

4 Natural – appropriate for the semantics of the message;

3 Acceptable – flat intonation with some expressivity;

2 Unnatural – robotic with little expressivity;

1 Erratic – low intelligibility and weird.

The average mean opinion score (MOS) for each speech file over all subjects was then computed. Results are shown in Fig. 2. The perturbation without prosody pattern (NPP_REC and NPP_TTS) applied to NEU_REC and NEU_TTS increases the MOS by 0.3 and 0.6 respectively. While the perturbation with prosody pattern (WPP_REC and WPP_TTS) further increases the MOS by 0.4 and 0.5 respectively. The *t*-test showed these increments are statistically significant with α=0.01. The results demonstrate the efficacy and extensibility of the proposed model to synthesize prosody pattern for both neutral speech recordings and neutral synthetic speech of TTS outputs. The results also indicate that the method can be successfully extended to new speakers.

TABLE V. PERCEPTUAL EVALUATION OF PERTUBATION WITH OR WITHOUT MODELING PROSODY PATTERN, MEASURED BY THE NUMBER (#) AND PERCENTAGE (%) OF SPEECH FILES JUDGED TO BE CLOSER TO THE EXPRESSIVE VERSUS NEUTRAL RECORDING

	Without Prosody Pattern	**With Prosody Pattern**
# of speech files	190	214
% of speech files	73.1	82.3

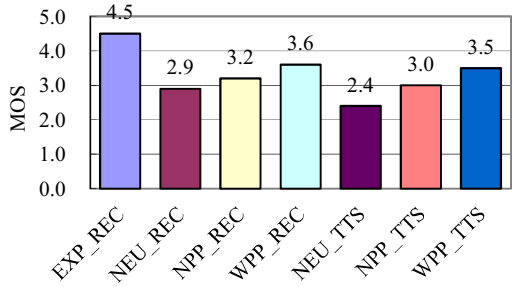

Figure 2. Comparison between neutral speech recordings, neutral synthetic speech and their perturbed renditions with or without prosody pattern based on MOS

VI. CONCLUSIONS AND FUTURE WORK

This work aims to enhance expressivity of text-to-speech (TTS) outputs. We analyze the acoustic features of the contrastive (neutral versus expressive) speech recordings, and find that the speaker usually tends to put more emphasis on one particular syllable within a multi-syllabic prosodic word while uttering the expressive speech. We call this particular syllable the *core syllable* in the prosodic word, which can be identified from the stress or tone information. Based on this observation, we categorize the syllables in speech into four classes, based on their locations in relation with the core syllable within a prosodic word. Acoustic feature analyses are then performed with respect to each class. The results indicate that the ratios between expressive and neutral speeches acoustic features for the four classes have the trend of the *prosody pattern* that core syllables > syllables succeeding the core syllable > syllables preceding the core syllable > all other remaining syllables. This prosody pattern is then utilized in the perturbation model to transform neutral speech to expressive speech with different model parameters for different syllable classes. Two perceptual experiments are then conducted by comparing the performance of the perturbation model with or without considering prosody pattern. The experimental results indicate that the perturbation model with prosody pattern can generate appropriate expressive 9% higher than the perturbation model without prosody pattern, and the model with prosody pattern can achieve MOS score 0.5 higher than the model without prosody pattern. These results demonstrate the efficacy and extensibility of the perturbation model for prosody pattern.

REFERENCES

[1] N. Campbell, "Towards synthesizing expressive speech: Designing and collecting expressive speech data," in Proc. Eurospeech, 2003.

[2] W. Hamza, R. Bakis, E. M. Eide, M. A. Picheny, and J. F. Pitrelli, "The IBM expressive speech synthesis system," in Proc. ICSLP, 2004.

[3] M. Bulut, S. Narayanan, and J. Johnson, "Synthesizing expressive speech: overview, challenges, and open questions," in S. Narayanan and A. Alwan. Text-to-Speech Synthesis: New Paradigms and Advances, pp. 175-201, Prentice Hall.

[4] E. Eide, R. Bakis, W. Hamza, and J. F. Pitrelli, "Toward expressive synthetic speech,", in Narayanan, S. and Alwan, A., Text-to-Speech Synthesis: New Paradigms and Advances, pp. 219-248, Prentice Hall.

[5] M. Schröder, "Expressing degree of activation in synthetic speech," IEEE Trans. A, 14(4): 1128-1136, 2006.

[6] N. Campbell, "Accounting for voice-quality variation," in Proc. Speech Prosody, Nara, pp. 217-220, 2004.

[7] N. Campbell, and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in Proc. Congr. Phonetic Sciences, pp. 2417-2420, 2003.

[8] T. Banziger, and K.R. Scherer, "The role of intonation in emotional expressions," Speech Communication, 46:252-267, 2005.

[9] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," IEEE Trans. Audio, Speech and Language Processing, 14(4): 1145-1154, 2006.

[10] Z. Wu, H. M. Meng, H. Yang, and L. Cai, "Modeling the expressivity of input text semantics for text-to-speech synthesis in a Chinese speaking avatar," IEEE Trans. Audio, Speech and Language Processing, 17(8): 1567-1577, 2009.

[11] S. W. Chen, B. Wang, and Y. Xu, "Closely related languages, Different ways of realizing focus," in Proc. Interspeech, 2009.

[12] F. Costa, "Intrinsic prosodic properties of stressed vowels in European Portuguese," in Proc. Speech Prosody, 53-56, 2004.

[13] A. Barbosa, P. Arantes, and L. S. Silveira, "Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule," in Proc. Speech Prosody, 49-52, 2004.

[14] http://www.discoverhongkong.com.

[15] C. Tseng, S. Pin, and Y. Lee, "Speech prosody: issues, approaches and implications," in From Traditional Phonology To Modern Speech Processing, G. Fant, H. Fujisaki, J. Cao, and Y. Xu, Eds. Beijing: Foreign Language Teaching and Research Press, pp. 417-438, 2004.

[16] A. Mehrabian, "Framework for a domprehensive description and measurement of emotional states," Genet Soc Gen Psycol Monogr, 121(3): 339-361, 1995.

[17] H. Yang, H. M. Meng, and L. Cai, "Modeling the acoustic correlates of expressive elements in text genres for expressive text-to-speech synthesis," in Proc. Interpseech, 2006.

[18] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," MAVEBA 2001, Firentze Italy, 2001.