Modeling Prosody Pattern of Chinese Expressive Speech and Its Application in Personalized Speech Conversion

Zhang Zhang[1,2], *Zhiyong Wu*[2], *Jia Jia*[1], *Lianhong Cai*[1,2]

[1]Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

zhangzhang09@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn,
{jjia, clh-dcs}@tsinghua.edu.cn

Abstract

This paper proposes an approach for modeling prosody patterns of acoustic features of Chinese expressive speech. In a Chinese multi-syllabic prosodic word, a syllable is identified as the core syllable based on the observation that speaker usually puts more emphasis on such syllable. The variations of the acoustic features migrating from neutral to expressive speech are then analyzed for both the core and non-core syllables. It is found that the acoustic variations of the core syllable are the most significant; the variations of the non-core syllables are influenced by the core syllable; such influence decreases while the non-core syllable moves farther from the core syllable. A double-layer perturbation model is then proposed to model such prosody patterns, which is further applied to generate personalized prosody patterns for personalized speech conversion. Experimental results indicate that our model can catch and regenerate the personality of prosodic features in expressive speech.

Index Terms: prosody pattern, expressive speech, prosodic features, personalized speech conversion

1. Introduction

There has been much research in the area of expressive speech processing [1][2][3]. Previous research has shown that expressivity can be realized through speech prosody and related acoustic features, including intonation, amplitude, duration, time, etc. [3]. The variation of the acoustic features may reveal some prosody patterns while migrating from neutral speech to expressive speech. Compared with neutral speech, the pitch and intensity of the focus word generally increase, while the same features of words preceding the focus word tend to decrease in some language [4]. Meng [5] analyzed acoustic features relating to focus in English and found that the variations of features are correlated to the phone position in relation with stressed syllables in focus words. Li [6] presented a prominence model for lexical stress and pitch accent detection, in which the differences of acoustic features between neighboring syllables are found to be important.

This work attempts to analyze the prosody patterns of Chinese syllables based on their positions relative to the *core syllable* in the prosodic word. In Chinese, the prosodic word is the smallest constituent at the lowest level of prosodic hierarchy [7], which consists of a set of syllables uttered closely and continuously in a sentence. And it is found that the variations of acoustic features are not exactly the same for all the syllables within a prosodic word while migrating from neutral speech to expressive counterpart. In a prosodic word, the core syllable is identified as the syllable whose acoustic variations are the most significant; and the acoustic variations of the other syllables are influenced by the features of the core syllable. A double-layer perturbation model is then proposed to describe and model the above prosody patterns. The differences of prosody patterns from different speakers are further investigated based on the parameters of the double-layer perturbation model; and these parameters are the representative of different speakers for personalized speech conversion.

2. Corpus

2.1. Text prompts and expressivity annotation

Text prompts are sourced from the Discover Hong Kong website of Hong Kong Tourism Board [8]. Each text prompt introduces the attractive features of a scenic spot. The final set of text prompts contains 25 utterances, 120 phrases, 416 prosodic words and 1231 syllables in total.

The PAD model [9] has been adopted to describe the expressivity of prosodic words in our previous work [10]. A set of heuristics have also been designed to parameterize the semantic expressivity of text prompts [10], where the A (arousal-nonarousal) descriptor measures the expressive degree (e.g. superlative, comparatives, etc.) of the words. It is found that prosodic words with $A > 0$ introduce the most attractive features of a scenic spot and speakers tend to put more emphasis on such words during speaking. Finer partitioning of the A values of such words will derive more levels of expressivity in speech. Hence in this work, five degrees of A values ($A = 0.2, 0.4, 0.6, 0.8, 1.0$) are adopted to describe the degree of expressivity of the prosodic words. Larger A value indicates stronger expressive degree. There are 272 prosodic words with $A > 0$ in our corpus.

2.2. Speech recordings

Four native Mandarin speakers (two males and two females) were invited to record our corpus. The text prompts with the same expressivity annotation of A values were presented to the speakers in advance so that they could get familiar with the text prompts and get a thorough understanding of the expressivity at prosodic word level based on the semantic meaning of the words (e.g. emphasize the beauty of a scenic spot). Then each speaker was asked to record the text prompts twice - first with neutral intonation throughout the utterances and second with expressive intonation according to the annotation labels of expressivity. We have 50 files of speech recordings for each speaker. The sound files are saved in wav format (16 bit mono, sampled at 16 kHz).

3. Acoustic analysis of prosody pattern

The objective is to analyze how acoustic features are related to expressive elements of the speech to reveal a certain prosody pattern, and how such prosody pattern differs between the speech recordings of different speakers.

3.1. Acoustic measurements

Acoustic features that are commonly associated with prosody include fundamental frequency ($F0$), intensity and speaking rate. Following acoustic features are extracted:

- Mean $F0$ (in Hz)
- $F0$ range (in Hz)
- Duration per syllable (in ms)
- Mean of root mean square (RMS) energy (in dB)

From the contrastive recordings (neutral versus expressive) of each prompt, the ratios of the acoustic feature values between corresponding expressive and neutral syllables are computed as in Equation 1 and adopted as the acoustic measurements for acoustic analysis of prosody patterns.

$$R = \sum_{i=1}^{n} \frac{F_i^{exp}}{F_i^{neu}} \tag{1}$$

Where n is the total number of syllables having the same expressivity (i.e. the same A value), F denotes any of the acoustic features described above, F_i^{exp} is the value of F for the i-th syllable in expressive speech, F_i^{neu} is its counterpart in neutral speech, and R is the average ratio of the feature F.

3.2. Classification of core and non-core syllables

Previous studies show that the variations of acoustic features are not exactly the same for all the syllables in a prosodic word [5][6]. The acoustic variations are the most significant for a certain syllable, which is identified as the core syllable in this work. The other syllables are identified as non-core syllables. The acoustic variations of non-core syllables are influenced by the core syllable and such influence is related to the distance between core and non-core syllables. The variations of acoustic features for the core and non-core syllables are expected to reveal a certain prosody pattern.

3.3. Acoustic analysis of prosody pattern

This section provides the analysis of the variations of acoustic features for the core and non-core syllables in a prosodic word while migrating from neutral to expressive speech, and investigates how the acoustic features of the core syllable influence the features of neighboring non-core syllables by taking into account the distance between them.

To analyze the prosody patterns of the acoustic features, the speech recordings are first automatically segmented into syllables with a home-grown segmentation tool and then the syllable boundaries are checked manually. The acoustic measurements of the average ratio R for different acoustic features F are calculated using the Equation 1.

Table 1 shows the average ratios between expressive and neutral speech of different acoustic features for the core syllables with different A values (272 core syllables in total). As can be seen from the table, the average ratios for all acoustic features have the positive correlation with the value of expressivity descriptor A. These observations agree with the fact that the speaker has a tendency to put more emphasis on the syllables with higher expressive degrees.

Table 1: *Average ratios of different acoustic features between expressive and neutral speech for the core syllables with different A values.*

R	A				
	0.2	0.4	0.6	0.8	1.0
Mean $F0$	1.09	1.11	1.14	1.16	1.18
$F0$ range	1.12	1.16	1.19	1.25	1.31
Duration	1.06	1.09	1.10	1.11	1.13
RMS energy	1.20	1.38	1.54	1.71	1.94

For the non-core syllables, (615 non-core syllables in total), the acoustic variations are influenced by the core syllables and such influences are related to the distance between core and non-core syllables. Such expectation can be observed from Table 2 3 4 5 which shows the average ratios of the parameters between expressive and neutral speech for non-core syllables with different A values and different distances to the core syllable.

Table 2: *Average ratios of mean $F0$ between expressive and neutral speech for the non-core syllables with different A values and different distances to the core syllable.*

Distance to the core syllable	A				
	0.2	0.4	0.6	0.8	1.0
1	1.07	1.08	1.08	1.09	1.09
2	1.04	1.05	1.05	1.06	1.06
3	1.02	1.02	1.03	1.03	1.04

Table 3: *Average ratios of $F0$ range between expressive and neutral speech for the non-core syllables with different A values and different distances to the core syllable.*

Distance to the core syllable	A				
	0.2	0.4	0.6	0.8	1.0
1	1.02	1.06	1.09	1.05	1.07
2	1.01	1.02	1.03	1.04	1.05
3	0.97	0.98	1.01	1.02	1.06

The first row of Table 1 shows the average ratios of Mean $F0$ for the core syllables. Comparing them with Table 2, it can be observed that the average ratios of mean $F0$ between expressive and neutral speech are bigger for the core syllables than for the non-core syllables. The average ratios of mean $F0$ for the non-core syllables are negatively correlated with the distance to the core syllables. Besides mean $F0$, the average ratios of other three features ($F0$ range, duration and RMS energy) follow the similar disciplines which can be seen from Table 3 4 5 compared with the other three rows of Table 1. These observations

Table 4: *Average ratios of duration between expressive and neutral speech for the non-core syllables with different A values and different distances to the core syllable.*

Distance to the core syllable	A				
	0.2	0.4	0.6	0.8	1.0
1	1.05	1.06	1.06	1.07	1.07
2	1.01	1.01	0.99	0.99	0.99
3	0.93	0.92	0.90	0.89	0.88

Table 5: *Average ratios of RMS energy between expressive and neutral speech for the non-core syllables with different A values and different distances to the core syllable.*

Distance to the core syllable	A				
	0.2	0.4	0.6	0.8	1.0
1	1.11	1.30	1.48	1.68	1.88
2	1.03	1.17	1.31	1.42	1.58
3	0.97	1.07	1.16	1.27	1.55

agree with the common perception that the speaker tends to put more emphasis on the core syllables, which will influence the neighboring syllables. This influence tends to decrease when the non-core syllables move farther from the core syllable.

And we can see from the tables that the average ratios of RMS energy are much bigger than the other three features for both core and non-core syllables. This indicates that the speaker tends to emphasize the expressive words by increasing their volume rather than other ways, such as changing the duration, pitch, or the range of the words.

3.4. Analysis of different speakers' prosody patterns

The prosody pattern of a specific speaker describes the personalized prosodic characters for that speaker while uttering expressive speech. For example, to put emphasis on the same syllable, different speaker may use different acoustic characteristics either by raising $F0$ contour or by lengthening duration and the levels of acoustic variations are even different.

The variations of the acoustic features for the core and non-core syllables are further investigated using the data from four different speakers as described in section 2.2.

For each speaker's speech, we classified core and non-core syllables as introduced in section 3.2 and found that different speaker may choose different syllable as the core syllable for the same prosodic word. To avoid the influence of such differences, 268 prosodic words (out of 272 all prosodic words with $A > 0$) whose core syllables are the same across different speakers were finally used in the analysis.

Figure 1 and 2 depict the variations of the parameters for the four different speakers. Figure 1 shows the results of the core syllables with different expressivity of A values (268 in total). While Figure 2 shows the results of the non-core syllables (604 in total) in relation to the distance between core and non-core syllables.

As can be seen from the figures, the acoustic variations of the core syllables (Figure 1) have obvious differences between different speakers. This indicates that different speakers may tend to use different acoustic characteristics to express the same intension of expressivity. For example, comparing the Duration and RMS energy curves of speaker 1 and speaker 2 in Figure 1, we can know that speaker 1 tends to lengthen the duration while speaker 2 tends to increase the energy while emphasizing (A increases from 0 to 1) the same prosodic words to utter the expressive speech. And Figure 2 shows the variations of the non-core syllables. It can be seen that the differences are not significant.

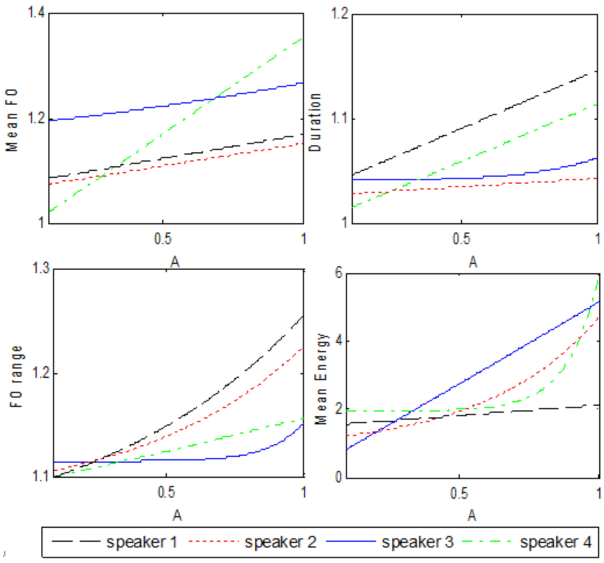

Figure 1: *Parameters of different speakers for the core syllables related to A values.*

These indicate that the personalized prosody patterns of different speakers are mainly revealed by the core syllables, while the influences of the core syllables to the neighboring non-core syllables are similar between different speakers.

4. Modeling prosody pattern

4.1. Double-layer perturbation model of prosody pattern

A double-layer perturbation model is proposed in this work to model the prosody patterns.

The first layer of the model is to describe the acoustic variations of the core syllables. Its position in the prosodic word was confirmed in the training part as we described in section 3.2. A non-linear exponential function is utilized for this layer of the model by following our previous work [10], as in Equation 2, where R_c is the ratio of acoustic variation of the core syllable, A is the expressivity annotation of the prosodic word, and C_1,C_2,C_3 are the constant parameters of the model.

The second layer of the model describes the influences of the acoustic features of core syllable on the features of non-core syllables. In this layer, the ratio of the acoustic feature variation for the core syllable is represented by R_c that is derived from the first layer; while the influence of the acoustic feature between core and non-core syllable is quantified by another exponential function of distance, as in Equation 3, where dis is the distance between the core and non-core syllable, C_4,C_5,C_6 are the constant parameters of the model, R_c and R_{nc} are the ratio of acoustic feature variation for the core and non-core syllable respectively.

$$R_c = C_1 + C_2 exp(C_3 A) \qquad (2)$$

$$R_{nc} = C_4 R_c exp(-C_5 dis) + C_6 \qquad (3)$$

The nonlinear least-squares regression is utilized to esti-

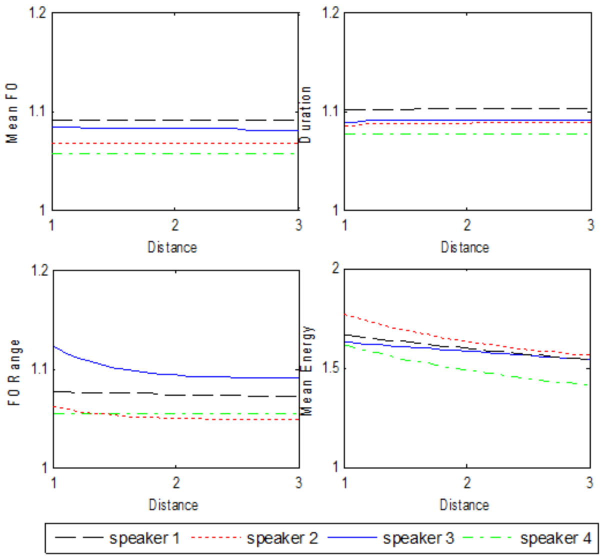

Figure 2: *Parameters of different speakers for the non-core syllables related to the distances to the core syllables ($A = 1$).*

mate the constant parameters in Equation 2 and 3. To produce an accurate finite-difference gradient, the initial values of these constants are set to 1, and the maximum number of iterations is 100.

4.2. Application in personalized speech conversion

To generate personalized speech, generally two aspects are involved: 1) to generate speech with target user's voice quality, and 2) to generate speech with target user's prosody characteristics. Prosody information plays a very important role in converting personalized speech. As has been discussed in section 3.4, the variations of acoustic features related to prosody patterns are speaker dependent, which lead to the possibility of converting personalized speech patterns for personalized speech conversion.

This work focus on converting personalized prosody patterns (prosody characteristics). The purpose is to generate the speech that just sounds like the source speaker is speaking in the way of the target speaker. To realize this, we calculate the acoustic feature of the target expressive speech ${F_t}^{exp}$ from the source neutral speech ${F_s}^{neu}$ by Equation 4.

$$ {F_t}^{exp} = ({F_s}^{neu} + \Delta_t - s^{neu}) \times R_t \qquad (4) $$

Where $\Delta_t - s^{neu}$ is the acoustic feature difference of the neutral speech between target and source speeches, and Rt can be R_c or R_{nc} that is calculated by the double-layer perturbation model as in Equation 2 or 3.

5. Experiment

5.1. Experiment on prosody pattern synthesis with double-layer perturbation model

The first experiment was conducted to evaluate the performance of the double-layer perturbation model in generating the prosody patterns while transforming the neutral speeches to expressive ones of the same person. 10 text phrases, each containing about 10 syllables, were randomly selected from our corpus for each speaker and used in the experiment. Three speech files were designed for each phrase:

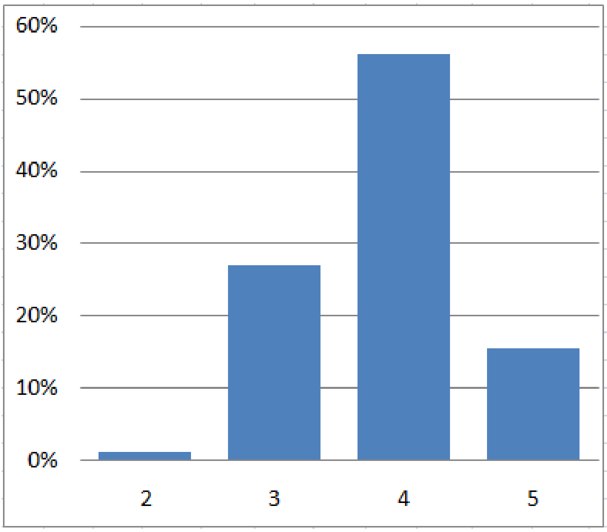

Figure 3: *Perceptual experiment results for evaluating the model used into transforming neutral to expressive speech.*

- (a) the neutral speech recording of a speaker;
- (b) the expressive speech recording of the same speaker;
- (c) the transformed speech from (a) generated by the double-layer perturbation model.

Total 60 speech files were generated. We divided them into 20 groups each of which corresponds to a phrase. For each group, the speech files were played in the order of (a)-(b)-(c)-(a)-(b)-(c). Each subject was asked to judge whether (c) sounded similar to its counterpart (b) and give a score from 1 to 5 indicating the level of the similarity between (c) and (b). 15 native Mandarin speakers were recruited to be subjects for the listening test. The bigger the score is, the more similar (c) (b) is, and the better the model is.

The result of the subjective test is shown in Figure 3. The average score is 3.94. About 90% of the data (c) is considered as more similar to (b) than to (a). The result shows that this model is better than our previous work [10] where the best result is 84.5%.

5.2. Experiment on personalized prosody pattern generation

A further experiment was devised for evaluating the performance of the proposed method in converting prosody pattern from one speaker to another. Another 10 phrases (43 prosodic words in total) were selected from the corpus. Four speech files were designed for each prompt:

- (d) the expressive speech recording of speaker 1;
- (e) the expressive speech recording of speaker 2;
- (f) the transformed speech from the neutral speech recording of speaker 3 using speaker 1's model;
- (g) the transformed speech from the neutral speech recording of speaker 3 using speaker 2's model.

Speech (f) and (g) are generated from the source speaker to imitate the prosody characters of expressive speech of the target speaker using our double-layer perturbation model.

Total 40 speech files were generated. We divided them into 20 groups. Each prompt corresponds to 2 groups, (d)-(e)-(f) and (d)-(e)-(g). The 20 groups of recordings are presented to the subjects in a random order. In each group, the files are played in the conformed order of (d)-(e)-(x), where (x) may be (f) or (g). Each subject was asked to determine whether (x) is imitating (d) or (e). The same 15 native Mandarin speakers were recruited to be subjects for the listening test. We defined the accuracy as in Equation 5 to describe the results.

$$Accuracy = \frac{\#Correct\_Speech(i)}{\#Trans\_Speech(i)} \times 100\% \qquad (5)$$

Where i denotes the speaker serial number may be 1 or 2 here. $\#Trans\_Speech(i)$ denotes the number of the speeches transformed from the neutral speech of speaker 3 using the model of speaker i. $\#Correct\_Speech(i)$ represents the number of the speeches chosen by a subject as it sounded that speaker 3 is imitating speaker i, while it is actually transformed using the model of speaker i. The bigger the accuracy is, the better our model reflected personality.

The results are shown in Table 6. It shows that our model for the conversion of personalized speech is able to achieve good results, indicating that our model could reflect the personalized features of different speakers.

Table 6: *Accuracy of the experiment for speaker 1 and 2.*

speaker i	1	2
Accuracy	74.6%	72.7%

6. Conclusions and future work

This paper proposes an approach for modeling prosody pattern of Chinese expressive speech. According to the data analysis, the syllables in a prosodic word are categorized into two classes: core syllable and non-core syllable. The acoustic features, including mean $F0$, $F0$ range, duration and RMS energy, for the two classes of syllables are further analyzed. Results indicate that, when migrating from neutral to expressive speech, the acoustic feature variations of syllables are positive correlated to the semantic expressivity and the influence on the core syllable is more significant than the influence on non-core syllables. The acoustic variations of non-core syllables are influence by the core syllable, and such influence is related to the relative distance between core and non-core syllables. The farther the non-core syllable is from the core syllable, the smaller the influence is. A double-layer perturbation model is then proposed for generating prosody patterns, which is further applied to generate personalized prosody patterns for personalized speech generation. Results of perceptual experiments indicate that our proposed method could describe the prosody patterns of the acoustic feature variations while migrating from neutral to expressive speech as well as model the prosody characteristics of different speakers for personalized speech generation.

7. Acknowledgement

The work is jointly supported by the National Natural Science Foundation of China (60805008, 60928005, 61003094) and the Ph.D. Programs Foundation of the Ministry of Education of China (200800031015).

8. References

[1] Campbell, N., “Towards Synthesizing Expressive Speech: Designing and Collecting Expressive Speech Data”, Proc. Eurospeech, 2003.

[2] Hamza, W., Bakis, R., Eide, E.M., Picheny, M.A. and Pitrelli, J.F., “The IBM Expressive Speech Synthesis System”, Proc. ICSLP, 2004.

[3] Banziger, T. and Scherer, K.R., “The Role of Intonation in Emotional Expressions”, Speech Communication, vol. 46, pp.252-267, 2005.

[4] Chen, S. W., Wang, B. and Xu, Y., “Closely Related Languages, Different Ways of Realizing Focus”, Proc. Interspeech, 2009.

[5] Meng, F., Meng, H., Wu, Z. and Cai, L., “Synthesizing Expressive Speech to Convey Focus using a Perturbation Model for Computer-Aided Pronunciation Training”, Proc. Interspeech, 2010.

[6] Li, K., Zhang, S., Li, M., Lo, W. and Meng, H., “Prominence Model for Prosodic Features in Automatic Lexical Stress and Pitch Accent Detection”, Proc. Interspeech, 2011.

[7] Tseng, C., Pin, S. and Lee, Y., “Speech prosody: issues, approaches and implications” Traditional Phonology To Modern Speech Processing, pp.417-438, 2004.

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

[9] Mehrabian, A., “Framework for a Domprehensive Description and Measurement of Emotional States” Genet Soc Gen Psycol Monogr, vol. 121, no. 3, pp.339-361, 1995.

[10] Yang, H., Meng, H. and L. Cai, “Modeling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis” Proc. ICSLP, 2006.