Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/081L25

Tone Enhancing Model for Disyllable Words in Chinese Mandarin Speech

Jianbo Jiang[3,4,5], *Jia Jia*[1,2,3,*], *Ye Tian*[3,5], *Yongxin Wang*[1,2,3] *and Lianhong Cai*[1,2,3]

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

Received: 1 Aug. 2012, Revised: 26 Aug. 2012, Accepted: 27 Aug. 2012
Published online: 1 Apr. 2014

Abstract: Tone recognition is the core function in Chinese speech perception. The tone perception ability of people with sensorineural hearing loss (SNHL) is often weaker than normal people. Automatically tone enhancement would be useful in helping them understand Chinese speech better. In this paper, we focus on the tone enhancing model for Chinese disyllable words. We first analyze the acoustic features related to tone perception. By agglomerative hierarchical clustering method, the first and second syllables of disyllable words are clustered into 6 clusters respectively. Discriminative features of these clusters are experimentally determined from a set of possible features related to tone perception, such as the pitch value, pitch range and position of minimum pitch, etc. We further propose a practicable tone enhancing model with these discriminative features: 1) an input pitch contour is classified by calculating the distance between it and the centroid of each cluster, and 2) selecting the smallest distance, then the unclassified pitch contour belongs to this cluster, 3) the pitch contour is modified for tone enhancement with model parameters corresponding to this cluster using TD-PSOLA. Both statistical and subjective experiments show that higher hit rate of tone recognition can be obtained after tone enhancement with the proposed model. Especially, the proposed enhancing model can also avoid traditional tone recognition, which is more convictive and less laborious.

Keywords: Chinese Mandarin, disyllable word, tone cluster, tone enhancing model, pitch contour

1 Introduction

Chinese Mandarin is a tonal language. Most of Chinese syllables are pronounced with one of the four lexical tones: the high-level tone (tone 1), the mid-rising tone (tone 2), the low-falling-rising tone (tone 3) and the high-falling tone (tone 4). The four tones of Mandarin Chinese are used to distinguish the meaning of words. The main difference between different tones is their pitch contours: the pitch contour of tone 1 is high and flat, the pitch contour of tone 2 rises after a possible short falling, the pitch contour of tone 3 rises after a longer falling, and the pitch contour of tone 4 falls sharply [1].

For describing pitch contours of different tones, Chao proposed the five-scale method [1]. In the five-scale method, the four tones in isolated syllable could be represented as following: 555 for tone 1, 345 for tone 2, 214 for tone 3 and 531 for tone 4. In a continuous speech, the pitch contour of a syllable will be affected by surrounding syllables. Tao proposed Syllable Pitch Stylized Parameters (SPiS) to describe the pitch contour, and used the parameters to realize stress and intonation [1]. Various tone patterns were brought forward, which was tend to be perfect; however, researches on parameters related to tone perception are still not sufficient.

For the tone perception researches, most existing related researches focus more on contrastive tones. Wang and Liu analyzed the effect of some pitch features on the perception of tone 1 and tone 2 [2]. They found that onset, offset, and turning of the pitch contour (defined as the duration from the onset of the tone to the point of change in F0 direction) are relevant to tone perception. Moore and Jongman, who worked on tone 2 and tone 3, reached the conclusion that both the turning point and

* Corresponding author e-mail: jjia@tsinghua.edu.cn

ΔF0 (the decrease in F0 from the onset of the tone to the turning point) would affect tone perception [3]. Afterwards, researchers paid more attention on tones 2 and 3, and reached their own observations [4,5,6,7]. They described some key features influencing tone perception, yet they didn't come up with an enhancing model or a modifying method with these features.

On the research of improving perceiving tones, Lu et al. used a tone modification factor "contrast weight" to achieve tone enhancement [8]. They put forward a digital processing method for modifying tone contrast that was defined as the greatest difference in frequencies between peaks and valleys of pitch curves. But their method could only apply to monosyllables. In fact, in Mandarin Chinese, there were often not only monosyllables but also disyllables, and the effect of surrounding syllables on tone perception couldn't be ignored. So, how to modify for disyllables is a major task.

The ability of perceiving and distinguishing the tone information, is an important component of the people's Mandarin speech perception ability. The tone perception ability of people with sensorineural hearing loss (SNHL) is often weaker than normal people. Therefore, it is significant to research on the specific role of different acoustic features in tone perception, and propose an algorithm for automatically tone enhancement.

In this paper, we focus on the tone enhancing model for disyllable words in Chinese Mandarin speech. We analyzed the discriminative acoustic features related to tone perception of disyllable words, based on a disyllable word corpus with different level of tone emphasis, which will be described in Section 2. And then in Section 3, we propose a tone enhancing model for disyllable words using those discriminative features. Both statistical and subjective experiments are carried out in Section 4 to validate the correctness and reliability of the model. Finally, Section 5 gives conclusions and future work.

2 Analysis on the Discriminative Features Related to Tone Perception

To form a tone enhancing model, we analyze features related to tone perception first. For monosyllable, J. Lu et al. recommended a tone enhancing algorithm with factor "contrast weight" on pitch value [8]. But without changing timing features which describe the shape of pitch contours, such as positions of minimum and maximum pitch points, tone 2 and tone 3 are still not easy to be identified by listeners. In addition, the condition for disyllable words is more complex because of coarticulation. In this context, we analyze a speech corpus with 2 different levels of tone emphasis to find tone perception related features.

There are totally 230 disyllable words in our corpus, each word is recorded two times by a female speaker with normal and emphasized tone. The recordings with normal tone would be referred as the regular corpus in the rest of the paper, while the recordings with emphasized tone would be called emphasized corpus. The data are digitized using a sampling rate of 16kHz with 16-bit resolution, and saved in single channel wav files.

A. Tone Perception Test

A perception test is carried out to validate that there is an obvious perceptional difference between the regular corpus and the emphasized one.

Fifteen subjects aged between 20 and 30 take part in the test. The subjects are all native speakers of Mandarin Chinese from mainland China. None reports any hearing or speech disorders.

The disyllable corpus is manually segmented into single syllables. Each syllable is normalized to the same root-mean-square (RMS) level. In either of the regular and emphasized corpus, we built 16 tables with 25 syllables in each by random selection.

To simulate the hearing condition for people with hearing disabilities, voices are played at the hearing threshold of each subject. Each subject listens to 4 tables, two from regular corpus and two from emphasized corpus, and is asked to reply which tone the syllable belongs to. The hit rate of the perception test is shown in Figure 1.

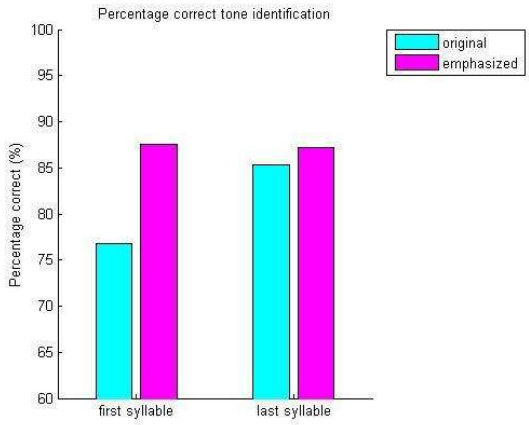


Fig. 1: Hit rate of tone perception test

From the hit rate we can find that the emphasized corpus has a higher hit rate than the regular one. In other words, there is an apparent perceptual difference between the two corpora. What's more, for regular corpus, 30.5% of the first syllables with tone 2 are mistakenly identified as tone 3, while 29.7% of the second syllables with tone 3 are as tone 2. For example, /wu2 shu4/ ("countless") identified as /wu3 shu4/ ("martial art"), and /ke2 yi3/ ("may") identified as /ke3 yi2/ ("suspicious"). This indicates the necessity to use timing features (features describing the shape of pitch contours) besides pitch value as the discriminative features for tone 2 and tone 3.

By analyzing and comparing the two corpora, we can find the features related to tone perception.

B. Clustering of Pitch Contours

Related studies showed that the perceived tone is mainly determined by the shape of pitch contour of a Chinese syllable [12], so we extracted the pitch contours of all finals in the corpus. For different type of pitch contour, the features affecting tone perception may not be the same, which brings about the significance of reclassification. Although there are only four lexical tones in Mandarin Chinese, the pitch contour varies during coarticulation in continuous speech. This means there may be more than four types of pitch contours to be found in continuous speech for these four lexical tones. Without any prior knowledge, clustering, a method of unsupervised learning, is used to classify the pitch contours.

To avoid the influence of the lengths of pitch contours, each pitch contour is resampled to 20 equidistant points which are adequate to describe outline of the whole pitch contour. Each point of the resampled pitch is calculated based on the two adjacent fundamental frequencies in original contour using logarithmic interpolation. After resampling, the data are standardized to z-score values to help avoid dependence on the choice of measurement units taking the following steps: calculating the mean value; calculating the mean absolute deviation; calculating the standardized measurement, or z-score of resampled fundamental frequency.

Euclidean distance is selected to measure the distance between normalized pitch contours, and during clustering, inner square distance (Ward's distance) is selected measure the distances between clusters to guarantee that each cluster are well separated from others.

Agglomerative hierarchical clustering is selected to process clustering [9]. In this algorithm, initially each sample is assigned to a separated cluster, and on each time the two clusters that are the nearest among all the clusters are combined into a new cluster. The termination criterion of clustering in chosen to be the inner squared distance between any of two categories is greater than 11.5. This threshold can guarantee that the result shows observable difference between clusters.

The first syllables and second syllables in the regular corpus are clustered separated. They both got six clusters after clustering. The pitch contours in each cluster are shown in Figure 2 and Figure 3, representing the first and the second syllables respectively.

The clustering result is consistent with lexical tones, which means that samples in the same cluster mainly have the same tone. Therefore, each cluster is named as the tone it originally belongs to. As there are more than four clusters in the result, some lexical tones are separated into two clusters. For the first syllable in the corpus, the 2^{nd} tone and the 4^{th} tone are separated in two clusters each. So the six clusters for the first syllable in the corpus are called tone I-1, tone I-2-1, tone I-2-2, tone I-3, tone I-4-1 and tone I-4-2. The clusters for the second syllable in the corpus are called tone II-1, tone II-2, tone II-3-1, tone II-3-2, tone II-4-1 and tone II-4-2, as in this time, the 3^{rd} tone and the 4^{th} tone are separated in two clusters. Further analysis of the result can also reveal what effect the context has on pitch contour. For example, the tone I-2-1 cluster for the first syllable mainly comes from words whose second syllables are the 2^{nd} or 3^{rd} tone, while the tone I-2-2 cluster mainly comes from words whose second syllables are the 1^{st} and 4^{th} tone.

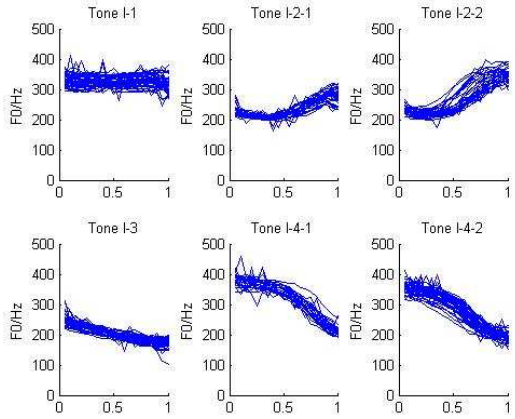


Fig. 2: The pitch contours of each cluster of the first syllables

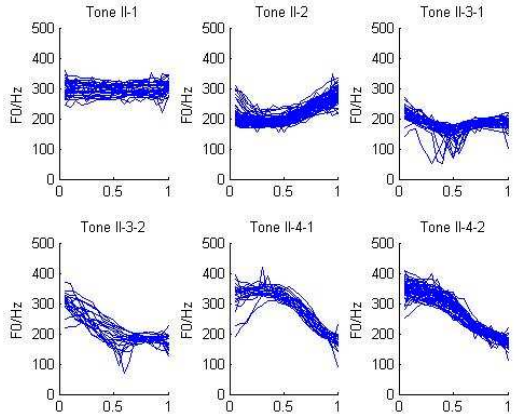


Fig. 3: The pitch contours of each cluster of the second syllables

C. Centroid of the pitch contours

The 20-point pitch contour representation is used in clustering to avoid dimensionality problems, but it is not enough to represent detailed shape of pitch contours in the study. And as the original pitch contour is not smooth enough when we look at the details, we then smoothed the pitch contours of the final of each syllable in the corpus. High order polynomial curve fitting (the order is 4 in our study because the curve has no more than 3 turning points) is experimental determined to smooth the pitch contours.

To analyze the discriminative features related to tone perception, the centroid of the pitch contours in each cluster is calculated, defined as the average of the 100-point fitted pitch contour:

$$F0_{i,center}(k) = \frac{\sum_{j \in set(i)} F0_{j,smooth}(k)}{\sum_{j \in set(i)} 1} \quad (1)$$

where $F0_{i,center}(k)$ represents the k-th frame of centroid of set i, $F0_{j,smooth}(k)$ represents the k-th frame of pitch contour No.j.

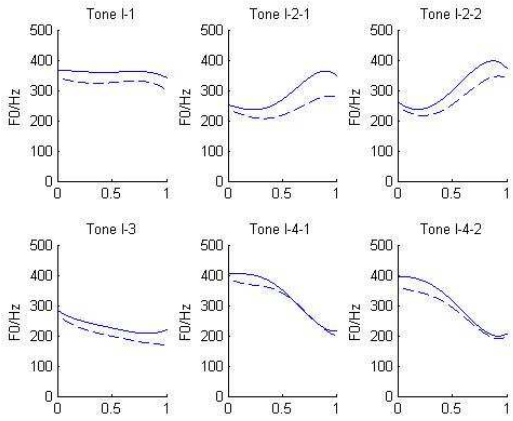


Fig. 4: The pitch contours of each cluster of the first syllables

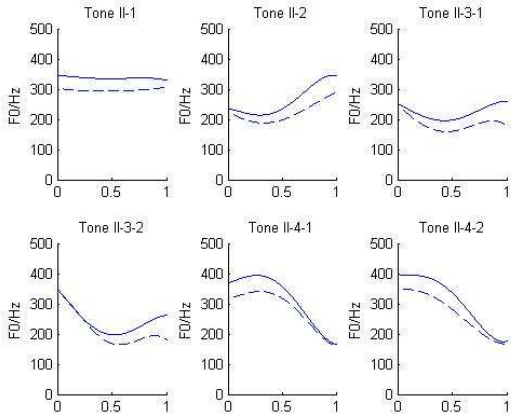


Fig. 5: The pitch contours of each cluster of the second syllables

The result is shown in Figure 4 and Figure 5. The dash line represents the centroid of the regular corpus, while the solid line represents that of the emphasized corpus. As can be seen from the figure, the difference between the pitch contour from the regular corpus and the emphasized corpus mainly are in pitch value and timing features. Specifically, pitch value features contains mean pitch, pitch range, minimum pitch, maximum pitch, and the onset and offset of the pitch contours, while timing features means positions of minimum and maximum pitch point.

D. The Discriminative Features Influencing Tone Perception

Inspired by SPiS (Syllable Pitch Stylized Parameters) proposed by Tao [1], we computed the following features for pitch contours in the 100-point representation:

- –Pitch value feature: mean F0 (mean, in Hz), range of F0 (range, in Hz), minimum F0 relative to mean F0 (relative min, in Hz), maximum F0 relative to mean F0 (relative max, in Hz), F0 at pitch onset (onset, in Hz), F0 at pitch offset (offset, in Hz);
- –Timing feature: position of minimum F0 divided by the length of F0 (pos of min), position of maximum F0 divided by the length of F0 (pos of max).

These features are calculated for each syllable, and the features that are significantly different between the regular corpus and the emphasized corpus in a cluster are selected and shown in Table 1 and Table 2. The value in the table means the ratio of the feature corresponding emphasized corpus and the one corresponding regular corpus.

Table 1: Values of discriminative features for the first syllables of disyllable words in tone emphasis

Feature	Tone I-1	Tone I-2-1	Tone I-2-2	Tone I-3	Tone I-4-1	Tone I-4-2
mean (M)	1.10	1.22	1.12	1.14	1.04	1.08
range (K_{mean})	1.10		1.20		1.01	1.01
rel_min (K_{min})						
rel_max (K_{max})		1.57				
min_pos (R_{min})		0.69	0.75	0.91		
max_pos (R_{max})						
onset (K_{begin})				1.01		
offset (K_{end})				1.07		

Table 2: Values of discriminative features for the second syllables in disyllable words in tone emphasis

Feature	Tone II-1	Tone II-2	Tone II-3-1	Tone II-3-2	Tone II-4-1	Tone II-4-2
mean (M)	1.13	1.18	1.21	1.13	1.12	1.15
range (K_{mean})	1.01					
rel_min (K_{min})					1.21	1.09
rel_max (K_{max})		1.47				
min_pos (R_{min})		0.81	0.91	0.82		
max_pos (R_{max})						
onset (K_{begin})			1.01	1.01		
offset (K_{end})			1.20	1.28		

With the further analysis of the perception test results in Section 2-A, it can be confirmed that these parameters are related to tone perception. For example, in the perception test, quite a number of samples from the regular corpus belonging to tone I-2-1 for the first syllable

are perceived as tone 3, but nearly none of those belonging to tone I-2-2. The centroid values of the features in these clusters are listed in Table 3.

Table 3: The average values of discriminative features in clusters tone I-2-1, tone I-2-2, tone I-3 for the first syllable.

Feature	Corpus	Tone I-2-1	Tone I-2-2	Tone I-3
mean (in Hz)	regular	236	271	203
	emphasized	289	311	231
range (in Hz)	regular	80	138	99
	emphasized	133	165	77
rel_min (in Hz)	regular	-31	-57	-37
	emphasized	-56	-75	-26
rel_max (in Hz)	regular	49	81	63
	emphasized	77	90	51
min_pos	regular	0.32	0.24	0.93
	emphasized	0.22	0.18	0.85

From Table 3 we can see that the features of tone I-2-2 differ much more from tone I-3 than those of tone I-2-1 for regular corpus. The features from clusters of the 2^{nd} tone differ more from the 3^{rd} tone in the emphasized corpus then in the regular corpus, while higher perception correctness rate is gotten in emphasized corpus in perception test.

In this section, based on the analysis on both the clustering result from the regular and emphasized corpus and perception test, we experimentally determine the discriminative features that are related to tone perception.

3 Tone Enhancing Model for Disyllables

With the discriminative features above as parameters, we proposed the disyllabic tone enhancing model.

A. Disyllabic Tone Enhancing Model

We use $F0(t)$ to represent the fundamental frequency at sample time t, t ranges from 0 to the total length T. The tone enhancing model is based on several basic transformations of the pitch contour, which are introduced as following:

–Stretch transform:

$$Stretch(F0(t),K,M,type) = K_{type}\cdot(F0(t)-F0_{type})+M\cdot F0_{type} \quad (2)$$

where K is the range parameter with normal value 1, M is the mean parameter with normal value 1, $type$ stands for the type of the transform, $type = mean$ for the transform based on mean value, corresponding to when mean belongs to discriminative parameters; and it is similar with the condition $type = min$ and $type = max$.

–Position transform:

$$Position(F0(t),R,index) = \begin{cases} F0(t/R) & for \quad 0 \le t \le R\cdot index \\ F0(T-(T-R\cdot index)(T-t)/(T-index)) & for \quad R\cdot index < t \le T \end{cases} \quad (3)$$

where $index$ stands for position parameter, or simply the sample time when minimum F0 occurs as the max position isn't used in the model. R is the ratio parameter with normal value 1, or simply R_{min}, ranging from 0 to $T/index$. Other variables defined as before.

–Endpoint transform:

$$EndPoint(F0(t),index,K_{begin},K_{end},M) = \begin{cases} K_{begin}\cdot(F0(t)-F0_{index})+M\cdot F0_{index} & for \quad 0 \le t \le index \\ K_{end}\cdot(F0(t)-F0_{index})+M\cdot F0_{index} & for \quad index < t \le T \end{cases} \quad (4)$$

where K_{begin} and K_{end} are the range parameter with normal value 1. Other variables defined as before.

With these transformations, then the disyllable tone enhancing models are constructed. Different models are used for different tones, as we can see from Tables 1 and 2 that the discriminative features for different tones are different.

–***Model 1***: Mean-range enhancing model

$$F0_{new}(t) = Stretch(F0(t),K,M,type) \quad (5)$$

This model applies to clusters I-1, I-4-1, I-4-2, II-1, II-4-1 and II-4-2, with the values of variables shown in Tables 1 and 2. This model reflects the change of pitch range and mean pitch value of the contour.

–***Model 2***: Mean-range & min_pos enhancing model

$$F0_{new}(t) = Stretch(Position(F0(t),R,index),K,M,type) \quad (6)$$

This model applies to clusters I-2-1, I-2-2 and II-2, with the values of variables shown in Tables 1 and 2. This model reflects the change of position of minimum pitch, pitch range, and mean pitch value of the contour.

–***Model 3***: Endpoint & min_pos enhancing model

$$F0_{new}(t) = EndPoint(Position(F0(t),R,index),index,K_{begin},K_{end},M) \quad (7)$$

This model applies to clusters I-3, II-3-1 and II-3-2, with the values of variables shown in Tables 1 and 2. This model reflects the change of position of minimum pitch, mean pitch value, and endpoint of the contour.

B. Tone Enhancing Application

In order to demonstrate the effectiveness of our proposed model, we apply the disyllable tone enhancing model to our continuous Chinese Mandarin speech tone enhancing application. In our previous tone enhancing application, the input is continuous speech, and after syllable segmentation, we use only tone enhancing model for monosyllable to modify pitch contours [10].But under the restriction of syllable segmentation accuracy, we often get not only monosyllables but also disyllables. So we further apply the proposed disyllable tone enhancing model to our application. That means, after syllable segmentation and F0 extraction, 1) an unclassified pitch contour will be classified by calculating the distance between its 20-point representation and the centroid of each cluster of both monosyllable and disyllable models, and 2) selecting the smallest distance, then the unclassified pitch contour belongs to this cluster, 3) the pitch contour is modified for tone enhancement with model corresponding to the cluster using TD-PSOLA, as TD-PSOLA method is a high quality speech synthesis method which is also be used on Chinese [11]. Furthermore, this method can also avoid traditional tone recognition, which is more convictive and less laborious.

4 Experiment and Discussion

To evaluate the proposed disyllable tone enhancing model, we first set up an experimental dataset “enhanced corpus” with 230 disyllable words by modifying the regular corpus using the proposed tone enhancing model. In order to examine the validity of the proposed model, we conduct two experiments with both statistical and subjective methods.

Statistical Experiment: A clustering experiment like the one described in Section 2-B and 2-C is carried out. The purpose of this experiment is to verify that after tone enhancement, the distances between tone clusters are larger than regular corpus.

Subjective Experiment: Two contrastive perception experiments are carried out on regular corpus and enhanced corpus. The purpose of this experiment is: 1) to examine how naturalness and perceived quality will behave along the synthesis, and 2) to verify that after tone enhancement, there is a clear improvement in tone perception by subjects.

A. Clustering

Clustering described in Sections 2-B and 2-C is reconducted on the enhanced corpus. The average cluster distance (Ward’s distance) increases from 38.5 to 40.6 for the first syllable, and from 36.7 to 43.8 for the second. Particularly, the cluster distance increases from 23.9 to 27.9 between cluster I-2-1 and I-3, from 42.0 to 44.5 between I-2-2 and I-3, from 29.3 to 31.2 between II-2 and II-3-1, from 24.2 to 28.5 between II-2 and II-3-2. That is, with the increase of cluster distance, the perceptual distances of the tone between clusters will also gain, which further means different tones will be distinguished more easily using enhanced corpus, especially for the clusters the major confusion in regular corpus belongs to. These results indicate the proposed tone enhancing method is feasible.

B. Naturalness along the Synthesis

This experiment is carried out to examine the perceived quality along the synthesis. Disyllables from regular corpus and enhanced corpus are mixed, and then 25 of them composed a table, constructing 18 tables in all. Each subject is asked to indicate the naturalness of each testing item, based on 5-point Likert scale, i.e.: ‘1’ (Bad), ‘2’(Poor), ‘3’(Fair), ‘4’(Good), ‘5’(Excellent).

Twelve subjects aged between 20 and 30 from mainland China participate in the listening test. Voices are played at the comfortable level. The mean opinion score (MOS) is calculated by averaging the five-point Likert scale over all the subjects.

The average MOS corresponding regular corpus is 4.75, while the MOS corresponding enhanced corpus is 4.49, a bit smaller. Analysis shows that almost all the disyllables are considered as good or excellent, so the synthesizing process ensures the perceived quality.

C. Tone Perception

The perception experiment is then to verify that after the modification of the discriminative features by the tone enhancing model, it will be easier for listener to perceive the intended tone type.

The listening materials are chosen randomly from the regular and enhanced corpus. Eighteen tables are constructed with 25 disyllables in each table. The syllables are normalized to the same RMS level.

Eighteen subjects aged between 20 and 30 are invited to this perception experiment. The subjects were all native speakers of Mandarin Chinese from mainland China. The test process is the same as in Section 2-A.

The hit rates for each cluster obtained by Section 2 for the first and the second syllable are shown in Figures 6 and 7. From the results, we can figure out that the enhanced corpus has higher hit rate than the regular one, ranging from about 3% to an incredible more than 30%, which validates the effectiveness and robustness of the proposed model.

In Section 2-A, we mentioned that for regular corpus, 30.5% of the first syllables with tone 2 and 29.7% of the second syllables with tone 3 are mistakenly identified. From Figures 6 and 7, we can find that both tone 2 of the first syllables (I-2-1, I-2-2) and tone 3 of the second syllables (II-3-1, II-3-2) have a remarkable increase of hit rate, especially tone 2 of the first syllables. These results confirm that the tone enhancing method can help people identify the correct tone.

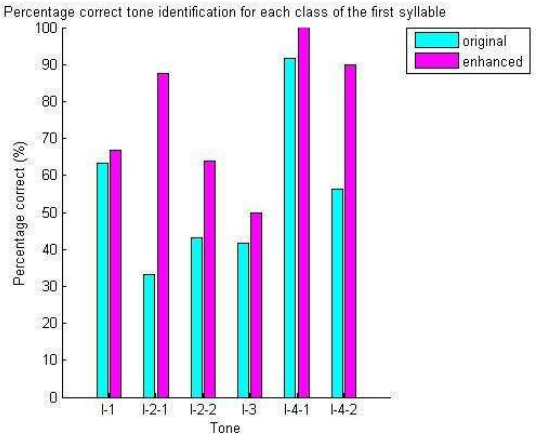


Fig. 6: The hit rates for all clusters of the first syllables

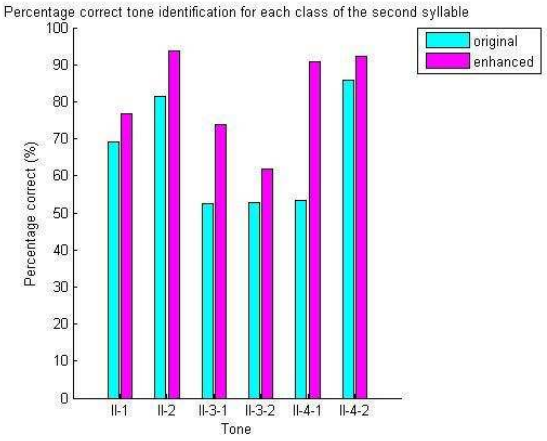


Fig. 7: The hit rates for all clusters of the second syllables

The modifications around the discriminative features shown in Section 2-D these features are also similar. When these features are modified more drifting the normal value within limits, the hit rate is always increasing sharper, e.g. tone 1-4-1 and tone 1-4-2. And if we consider the three vertical range features (range of F0, minimum F0 relative to mean F0 and maximum F0 relative to mean F0) as similar features, it also holds true. These results also show that the combination of these features can help people perceive the tone information in evidence.

5 Conclusion and Future Work

Based on analysis of a disyllable speech corpus with different levels of tone emphasis, we experimentally determine the discriminative features related to disyllable tone perception. A tone enhancing model for Chinese disyllable words are then proposed based on the discriminative features. Experiments indicate the relation of those discriminative features and tone perception, and show that the proposed tone enhancing model can lead to higher hit rate in tone perception. Especially, tone enhancement can be achieved using the model without automatic tone recognition, which is practicable and effective.

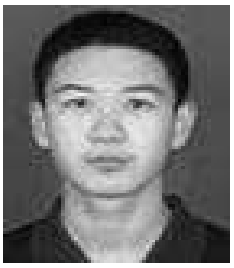
A separated evaluation on how useful pitch value features and timing features are would be interesting, so we will concentrate how the acoustic features contribute to tone perception independently.

Acknowledgement

This work is supported by the National Basic Research Program (973 Program) of China (2013CB329304),the National Natural, and Science Foundation of China (6137002361003094),and the National High Technology Research and Development Program(863Program) of China (2012AA011602). We would also like to thank Microsoft Research Asia-Tsinghua Univertity Joint Laboratory for its funds.

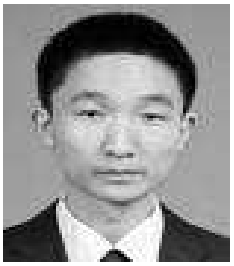
References

[1] L. Cai, D. Huang and R. Cai, Applications and Fundaments of Modern Speech Technology, Beijing: Tsinghua University Press, (2003).
[2] R. Yang, P. Liu and J. Kong, The Influence of F0 on the Perception of Tone 1 and Tone 2 of Mandarin in the First Syllable of Disyllables, Proc. PCC/ISPF, (2008).
[3] C. B. Moore and A. Jongman, Speaker Normalization in the Perception of Mandarin Chinese, Journal of Acoustical Society of America, **102**, 1864-1877, (1997).
[4] R. Cao and P. Sarmah, A Perception Study on the Third Tone in Mandarin Chinese, UTA Working Papers in Linguistics, **2**, 51-66 (2007).
[5] Y. Chen and J. Yuan, A Corpus Study of the 3rd Tone Sandhi in Standard Chinese, Interspeech, **2007**, 2749-2752 (2007).
[6] C. Leea, L. Taob and Z. Bond, Identification of Acoustically Modified Mandarin Tones by Native Listeners, Journal of Phonetics, **36**, 537-563 (2008).
[7] C. Wen and Z. Jinsong, Tone-3 Accent Realization in Short Chinese Sentences. Tsinghua Science and Technology, **13**, 533-539 (2008).
[8] J. Lu, N. Uemi, G. Li and T. Ifukube, Tone Enhancement in Mandarin Speech for Listeners with Hearing Impairment, IEICE Trans Inf Syst, **E84-D**, 651-661 (2001).
[9] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., San Francisco: Morgan Kaufmann Publisher, (2001).
[10] T. Ye, J. Jia, J. Jiang, L.Cai, Tone Enhancement for Chinese Mandarin Speech, NCMMSC, (2011).
[11] H. Cai, Z. Yu and L. Zhang, Prosodic Synthesis of Chinese Speech Based on TD-PSOLA Algorithm, Bulletin of Science and Technology, **18**, 6-8 (2002).
[12] F.Meng, Z.Wu, J.Jia, H.Meng, L.Cai, Synthesizing English Emphatic Speech for Multimodal Corrective Feedback in Computer-Aided Pronunciation Training, Multimedia Tools and Applications Journal, (2013).



Jianbo Jiang Jilin Province, China. Birthdate: May, 1989, is Computer Science B.Eng., graduated from Department of Computer Science and Technology, Tsinghua University. He is a graduate student of Department of Computer Science and Technology, Tsinghua University. And research interests on speech signal processing.

Jia Jia Beijing, China. Birthday: May, 1981, is Ph.D., graduated from Department of Computer Science and Technology, Tsinghua University, China. Her research interests are affective computing, and computational speech perception. She is now an associate professor of the Department of Computer Science and Technology, Tsinghua University.



Ye Tian He Nan Province, China. Birthday: November, 1982, is B.S., graduated from Department of Computer Science and Technology, Tsinghua University, China. He is a graduate student of Department of Computer Science and Technology, Tsinghua University. And research interests on speech signal processing.

Yongxin Wang Beijing, China. Birthday: October, 1982, is B.S., graduated from Department of Computer Science and Technology, Tsinghua University, China. He is a Ph.D. candidate of Department of Computer Science and Technology, Tsinghua University. And research interests on semantic analysis.

Lianhong Cai Beijing, China. Birthday: August, 1945, received the B.E. degree from Tsinghua University, Beijing, China. She is now a Professor with the Department of Computer Science and Technology, Tsinghua University. She directs the Human-Computer Speech Interaction Laboratory. She has been awarded Scientific Progress Prizes and the Invention Prizes from the Ministry of Mechanism and Electronics, and the Ministry of Education, P.R. China.