A REAL-TIME TONE ENHANCEMENT METHOD FOR CONTINUOUS MANDARIN SPEECHES

Ye Tian, Jia Jia, Yongxin Wang, Lianhong Cai
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

ABSTRACT

Chinese Mandarin is a tonal language. Tone perception ability of people with sensorineural hearing loss (SNHL) is often weaker than normal people. To help the SNHL people better perceive and distinguish tone information in Chinese speech, we focus on real-time tone enhancement method for mandarin continuous speeches. In this paper, based on the experimental investigation on the acoustic features most related to tone perception, we propose a practical tone enhancing model which employs the unified features independent of Chinese tonal patterns. Using this model, we further implement a real-time tone enhancement method which can avoid syllable segmentation and tonal pattern recognition. By the tone identification test for the normal and SNHL people under both quiet and noisy backgrounds, it is found that the enhanced speeches with the proposed method gains an average 5% higher correct rate compared to original speeches. And the time delay of the enhancement method can be controlled within 800ms, which can be further used in hearing aids to benefit the SNHL people in their daily life.

Index Terms — Tone enhancement, sensorineural hearing loss, real-time system, continuous Mandarin speech

1. Introduction

Chinese Mandarin is a tonal language. Tone perception ability of people with sensorineural hearing loss (SNHL) is often weaker than normal people. To help the SNHL people better perceive and distinguish Chinese speech information, we focus on real-time tone enhancement method for mandarin continuous speeches.

Many recent researches make progress in improving intelligibility of tonal speeches. For example, by conducting a comparison study in a group of hearing impaired adults, Grant [1, 2] found that fundamental frequency information, especially frequency change, yield superior reception of stress and intonation. J. Lu et al. [3] proposed that the direction and the slope of F0 change in the pitch contour is the main cue for tone perception. With these considerable efforts, J. Lu et al. [4] put forward a monosyllable tone enhancement model of modifying the slope of F0. The tone identification test results showed that modified monosyllables with enhanced F0 slope gains higher correct rate of tone identification when compared to unmodified speech tones. However, this model has no effect on tone 1 (high tone). By analyzing perceptional differences between regular disyllables and tone emphasized disyllables, J. Jiang et al. [5] found that the emphasized ones exaggerate not only in the slope of F0 but also in F0 mean. Based on the above analysis, our previous work [6] proposed a tone enhancement method for Chinese Mandarin speech. Speech is first segmented into syllables, and the slope and F0 mean of each syllable are modified according to the tonal pattern of the syllable. The method was experimentally proved to be effective for both normal people and SNHL people. A unified parameter set is also proposed in [6] to avoid tonal pattern recognition, but syllable segmentation is still required, which it is not suitable for real-time tone enhancement of continuous speeches due to the low precision and time performance of the automatic syllable segmentation algorithm.

In this paper, we propose an effective real-time tone enhancement method for mandarin continuous speeches. Based on the experimental investigation on the acoustic features most related to tone perception, we establish a tone enhancement model to expand the perceptional differences between different tonal patterns. Compared with previous studies, the proposed enhancement model employs the unified feature values independent of Chinese tonal patterns. We further implement a real-time tone enhancement method which can avoid syllable segmentation and tonal pattern recognition. By the tone identification test for normal and SNHL people under both quiet and noisy backgrounds, it is found that the enhanced speeches gains an average 5% higher correct rate compared to original speeches. The time delay of the enhancement method can be controlled within 800ms, which means it can be further applied in hearing aids algorithm to benefit the SNHL people in their daily life.

2. MODEL

Since the slope of F0 as well as the F0 mean has been proved to be related to tone perception [5,6], we consider establishing a tone enhancement model for continuous speech based on the monosyllable tone enhancement model proposed in [6].

2.1. Model Description

The monosyllable tone enhancement model proposed in [6] can be described by the following formula:

$$f0_{new}(n) = K \times [f0_{origin}(n) - F0_{mean}] + M \times F0_{mean} \quad (1)$$

Where n stands for the frame index, $f0_{new}(n)$ is the enhanced fundamental frequency, $f0_{origin}(n)$ is the original fundamental frequency. $F0_{mean}$ is the average fundamental frequency of the syllable. K and M are tone modification factors.

When K=1.0, the slope of target pitch contour is the same as the original one; while K greater than 1.0 produces pitch contour with enhanced slope. When M=1.0, the F0 mean of the target pitch contour is the same as the original one; M greater than 1.0 produces higher $F0$ mean.

As an example, Figure 1 illustrates the enhanced pitch for syllables with four different Chinese Mandarin tonal patterns, with K>1.0 and M>1.0. The solid line is the enhanced pitch contour, while the dashed line is the original one. The duration of the syl-

lable is normalized to 1. The horizontal index stands for the normalized time, and the vertical index stands for *F*0.

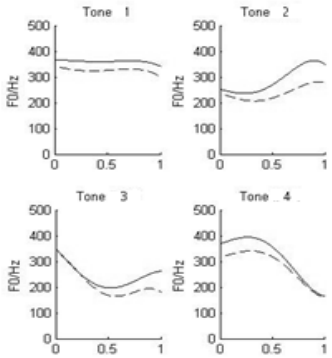


Figure 1: tone enhancement examples for four Chinese Mandarin tonal patterns when *K>1.0* and *M>1.0.*

In [6], the model was experimentally proved to be effective for both normal people and SNHL people, and a unified *K/M* pair for all tonal patterns is also found to avoid tonal pattern recognition. However, syllable segmentation is still required as the $F0_{mean}$ used in Eq (1) is the mean pitch of the target syllable.

2.2. Enhancement Model with Unified *K*/*M* Pair and Word Level Mean Pitch

In [6], the experiments only show that the unified *K*/*M* pair has comparable performance with the best *K*/*M* pair for each tonal pattern in quiet environment. In real situations, the environment is usually noisy. In this section, we would show with experiment that the unified *K*/*M* pair also has comparable performance with the best *K*/*M* pair in noisy environment.

In [6], $F0_{mean}$ used in Eq (1) is the mean pitch of the target syllable, which cannot be acquired without syllable segmentation. A word level mean pitch is used here to test the performance of the model when syllable segmentation is not available. The value of the unified *K/M* pair is taken from [6] to be: *K*=1.3, *M*=1.13.

The experiment is a tone identification test. Listeners are seated in soundproof rooms. After hearing one utterance, the listeners are required to repeat the utterance. No feedback is given as to whether the repeated utterance is correct. To avoid the listener guessing the content from context, only monosyllable or disyllable utterance is used. Listeners are allowed up to seconds to respond before the initiation of the next trial. Stimulus presentation level is set to each listener's most comfortable level (MCL). The procedure is recorded for amendment.

The speech signals for the tone identification test were recorded by one male and one female who speak natural mandarin at an average conversation-speaking rate. The signals were digitized at 16 kHz sampling rate. The multi-talker speech babble is added as noised at 0dB SNR level.

Six conditions are tested, which are the original speech, speech enhanced with the unified parameters, speech enhanced with the best parameters under quiet and noisy environment. Word level F0 mean is used in all conditions.

Eighteen subjects were invited to participate in this test. They are hearing-impaired subjects with pure-tone averages (PTAs) between 45 and 75 dB HL.

The test result is shown in Table 1.

Table 1. Identification correct rate of each tone modified with unified and optimum *K/M* value under quiet and noisy environments

		Tone 1	Tone 2	Tone 3	Tone 4
Quiet	No Enhancement	72.2	66.7	50.4	87.2
	Unified *K/M*	82.4	80.1	63.3	91.8
	Best *K/M*	83.8	78.8	66.7	97.3
Noisy	No Enhancement	36.0	45.3	44.3	46.5
	Unified *K/M*	40.5	54.5	52.4	53.6
	Best *K/M*	38.6	56.4	54.2	57.7

As can be seen in Table 1:

- Both the two of *K/M* pairs produce an average of more than 8% increase in correct rate under both quiet and noisy environment;
- The difference of increase between the two pairs of *K/M* is small;

The result shows that the unified *K*/*M* pair has comparable performance with the best *K*/*M* pair in both quiet and noisy environment, and it is feasible to employ the unified *K/M* pair independent of tonal patterns.

Also, using word-level pitch mean does not degrade the performance of the model much, which allows the model to be used in real-time systems without syllable segmentation.

3. METHODOLOGY AND SYSTEM

Based on the experiment result in Section 2, we build a real-time tone enhancement system as shown in Figure 2. With unified *K/M* pair and word-level mean pitch, syllable segmentation and tonal pattern recognition would not be necessary.

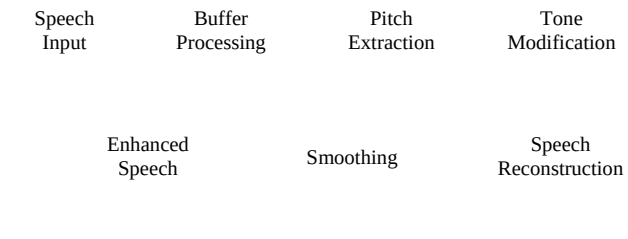


Figure 2: Real-time Tone Enhancement System

The buffer processing is essential for real-time processing of the input speech. As the speech signals are processed piece by piece, discontinuities may occur at piece boundaries. The smoothing of the boundary is also a big problem of real-time process. These problems would be addressed in the following sections.

3.1. Buffer Processing

In real-time processing, speech signals must be buffered before processing. The system uses a circular linked buffer list consisting of fixed length buffers, and would start processing when a buffer is full.

However, using fixed length buffers may cause one syllable separated in two adjacent buffers. As the mean pitch in Eq (1)

must be determined according to the speech signals being processed, if the two parts of the syllable is processed separately, the mean pitch used may be different which would result in a discontinuity in pitch. So, we break the input speech apart at low energy frames near the buffer boundary. According to Chinese phonology, frames with very low energy never appear inside a syllable, so this would assure one syllable be processed as a whole. This would require the circular linked buffer list to have at least three buffers, as when buffer 2 is full, there's still something left in buffer 1 while buffer 3 is accepting new input.

Breaking up the input speech with low energy segments is different from syllable segmentation as we don't need to find all syllable boundaries. The segments to be processed by the tone enhancement system can be one syllable or several consecutive syllables. The mean pitch in Eq (1) is the mean pitch of the segment to be processed, which may be syllable mean pitch or word-level mean pitch. Instead of the complex algorithm used in syllable segmentation, a single threshold on the energy curve is enough to meet the needs of this system.

In practice, speech detection is also incorporated. If the cached input is speech, the first frame whose energy is below a certain threshold before the buffer boundary is selected as a separation point. Frames after this separation point would be processed with the next buffer.

3.2. Pitch Extraction

In this paper, an improved fundamental frequency extraction method [7] was employed. Compared to the traditional autocorrelation method, this method can produce fundamental frequency curves with less error, and with fewer mistakes that a voiced segment is treated as unvoiced.

3.3. Speech Reconstruction

TD-PSOLA algorithm [8] is widely used in speech modification and synthesis. It is a time-domain method of low computation complexity which maintains high acoustic fidelity for a small duration or pitch modification. With the unified K/M pair of K=1.3 and M=1.13, this method well meets the requirements of the tone enhancement system.

3.4. Smoothing

Another problem is the discontinuity that occurs between the adjacent processed segments. An example is shown in Figure 3(a). The discontinuity will result in noise in the output speech, degrading the sound quality. Smoothing is employed to solve this problem.

The algorithm for smoothing is as follows:

1) Let the output speech signal be $s(t)$, and the discontinuity at $s(t_0)$ and $s(t_0+1)$. N extreme points are found around at each side of t_0. Let the greater absolute value of $s(t_0)$ and $s(t_0+1)$ be $|A_0|$. Here we assume $A_0 = s(t_0+1)$. (The procedure would be similar when $A_0 = s(t_0)$)

2) Let the first extreme point before t_0 with the same sign as $s(t_0+1)$ be $s(t_1)$;

3) Among the $2N$ extreme points found in step 1, let the one with the greatest absolute value and different sign with A_0 be $s(t_2)$. Let $A_m=s(t_2)$;

4) Set the value of the point at $(t_1+t_0)/2$ to $-A_m$;

5) Use half period of a sine wave (from one extreme to another extreme) to link $s(t_1)$ and $s((t_1+t_0)/2)$;

6) Use another half period of a sine wave to link $s((t_1+t_0)/2)$ and $s(t_0)$.

An example is shown in Figure 3.

3.5. Summaries

The work flow of the enhancement system is as follows:

1) Input speech is cached in the buffer list;

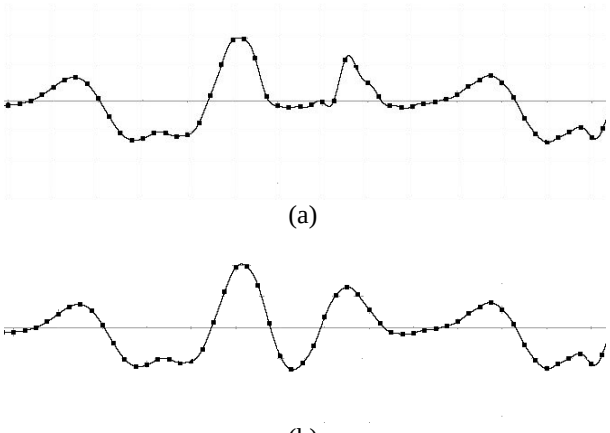
2) When a buffer is full, a separation point is determined in the buffer, and the signals not yet processed before the separation point would be processed;

3) The pitch contour of the speech segment to be processed is estimated with an automatic pitch extraction algorithm;

4) The enhanced pitch contour is determined with the unified *K/M* parameters;

5) TD-PSOLA algorithm is utilized to modify the speech using the enhanced pitch contour.

At last, the system output the enhanced speech. The signals process would be dropped from the buffer. As the system can only process full syllables or syllable sequences, the delay of the system is about the duration of 2 syllables in speech with normal speaking rate, which is about 800ms. This is also set as the length of one buffer circular linked buffer list.



(a)

(b)

Figure 3: waveform before (a) and after (b) smoothing

4. EXPERIMENT

Two experiments are conducted to verify the effectiveness of the method proposed in this paper. The first experiment is an objective test, which aims at verifying that in the result of the proposed tone enhancement model the tonal pattern is easier to be discriminated than before enhancement. The second experiment is a subjective test to verify that the model could work well for continuous speeches.

4.1. Exp1: Clustering Analysis on Tone Perception

We use a clustering method to verify the pitch curve of each tone is more separated from others after enhancement.

Agglomerative hierarchical clustering [5] is carried out separately on pitch contours of each syllable in original and enhanced

corpus using the proposed model. Pitch contours are resampled to 20 points. The clustering initiates each sample as a cluster, then recursively combines clusters which are the closest to each others, until a certain number of clusters are got. The Ward's distance is used to as the distance measure between clusters, as shown in Eq (2).

$$D_w\left(C_i, C_j\right) = \sqrt{\frac{2n_i n_j}{\left(n_i + n_j\right)}} \left\| c_i - c_j \right\|_2 \qquad (2)$$

where $D_w(C_i,C_j)$ is the Ward's distance between clusters C_i and C_j, n_i and n_j are the number of samples in C_i and C_j, $\|c_i-c_j\|_2$ is the Euclidean distance between the centroids of clusters C_i and C_j.

200 syllables (50 per tone) were used in the experiment. They are clustered into 4 clusters separately before and enhancement. The average inter-cluster distance is calculated according to Eq (3):

$$D_{avg} = \frac{2}{N(N-1)} \sum_{i \neq j} D_w\left(C_i, C_j\right) \qquad (3)$$

where D_{avg} is the average inter-cluster distance, N is the number of clusters (4 in this experiment).

The average inter-cluster distance D_{avg} increases from 37.5 to 42.3 after enhancement. The increase of cluster distance may also show the increase of perceptual distances between clusters, which further means different tones will be distinguished more easily by a listener. These results indicate the proposed tone enhancement model is feasible.

4.2. Exp2: Tone Identification Test for Continuous Speeches

Sixteen subjects were invited to participate in this test. They are hearing-impaired subjects with pure tone averages between 40 and 75 dB HL.

Four kinds of stimuli are used in this experiment, which are recordings before and after enhancement under quiet and noisy environments. Enhancement of speech is carried out with our proposed real-time enhancement model. The noise is babble noise at 0dB. The stimuli are presented at each participant's most comfortable level.

The test result is shown in Table 2.

Table 2. Tone identification correct rate of speeches that modified with unified *K/M* value under quiet and noisy backgrounds

quiet		noisy	
original	enhanced	original	enhanced
83.4	89.2	73.3	78.1

As seen in Table 2, conclusions can be made as follows:

- The unified *K/M* tone enhancing model produces an increase rate of 5.8% in quiet environment;
- The unified *K/M* tone enhancing model produces an increase rate of 4.8% in noisy environment;
- The unified *K/M* model works well on utterances in both quiet and noisy environments, so it is feasible and effective.

5. DISCUSSION AND CONCLUSIONS

In this paper, based on the experimental investigation on the acoustic features most related to tone perception, we propose a practicable tone enhancing model which employs the unified parameters independent of tonal patterns. Using this model, we further implement a real-time tone enhancement method which can avoid syllable segmentation and tonal pattern recognition. By the tone identification test for the normal and SNHL people under both quiet and noisy backgrounds, it is found that the enhanced speech gains an average 5% higher correct rate compared to original speeches. On basis of the model and experiments, we build a real-time tone enhancement system for continuous speech. The time delay of the real-time system can be controlled within 800ms, which can be further used in hearing aids to benefit the SNHL people in their daily life. The system is in trial, and it is worth believing that the system would be applied to hearing aids in the near future.

6. ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program (973 Program) of China (2013CB329304), the research funds from the National Natural Science Foundation of China (61003094, 90920302).

7. REFERENCES

[1] K.W. Grant, "Encoding voice pitch for profoundly hearing-impaired listeners", J.Acoust. Soc. Am., vol.82, pp.423–432, 1987.

[2] K.W. Grant, "Identification of intonation contours by normally hearing and profoundly hearing-impaired listeners", J.Acoust. Soc. Am., vol.82, pp.1172–1178, 1987.

[3] J.Lu, N. Uemi, and T. Ifukube, "Proposal of a digital hearing aid with emphasis function of tones in Chinese spoken language", Trans. Tech. Comm. Physiol. Acoust, H-99-44, The Acoustical Society of Japan, June 11, 1999.

[4] J.Lu, N.Uemi, G.Li. and T.Ifukube, "Tone Enhancement in Mandarin Speech For Listeners with Hearing Impairment", IEICE TRANS. INF. & SYST., VOL.E84–D, NO.5 MAY 2001.

[5] J.Jiang, J.Jia, Y.Tian, Y.Wang and L.Cai. "Tone Enhancing Model for Disyllable Words in Chinese Mandarin Speech", Applied Mathematics & Information Sciences, 6(1), pp.1-7, 2012.

[6] Y.Tian, J.Jia, J.Jiang, and L.Cai. "Tone enhancement for Chinese mandarin speech", NCMMSC 2011

[7] Y.So, J.Jia, L.Cai. "Analysis and Improvement of Autocorrelation Pitch Extraction Algorithm based on Candidate Set", Recent Advances in Computer Science and Information Engineering, 2011, Volume 5, Series: Lecture Notes in Electrical Engineering, Vol. 128, pp3693-3698, ISBN 978-3-643-25791-9.

[8] J.Jia, J.Xu, Y.Xu, L.H.Cai. "A Speech Modification based Singing Voice Synthesis System", NCMMSC 2008