Tone features in whispered Chinese *

LI Xueli[1, 2**] and XU Boling[1]

(1. State Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China; 2. School of Information Science and Engineering, Shandong University, Jinan 250100, China)

Received June 16, 2004; revised September 6, 2004

Abstract Study on the whispered tone is important to speech recognition and conversion in whispered Chinese. In this paper, the characteristics of whispered speech are introduced and the tone features in whispered Chinese are discussed. There is no fundamental frequency in the whispered speech, so other features, such as the amplitude envelope, duration, glottal area, lip area, formant, and vocal tract length, are extracted and their contributions to the automatic tone recognition are compared. From the experiments with six simple Chinese whispered vowels in four tones, it is proved that loudness-weighted 32 Mel-frequency bands log-amplitude envelopes and duration can be used as the main tone features in the whispered Chinese tone recognition. The average tone recognition rate approaches that of the human perception level.

Keywords: whispered speech, tone feature, loudness-weighted, Mel-frequency band, amplitude envelope, sound duration.

Whispering is a natural but unusual mode of speech. There is no vocal cord vibration in the whispered speech and it has low sound power. People can communicate with whispered speech without any difficulty. Because the whispered speech provides knowledge about the coding of prosodic information when fundamental frequency (F_0) is not available, the whispered tone is of interest to phoneticians all over the world. In addition, with the wide use of mobile phones, the need for private communication in public is increasing. For privacy, people whisper to the phone. Sometimes, it is desired that the whispered speech can be converted into normal speech. Reconstruction of normal speech could also be used by the aphonic individuals as a voice prosthesis[1]. Reconstruction of normal speech from whispered speech needs the whispered tone information in the tonal language. However, some problems about whispered tone have not been solved up to now. For example, which acoustic cues contribute to the whispered tone recognition? How much contribution can every tone feature offer? These problems should be solved firstly before the reconstruction of normal speech from whispered speech. Therefore, the study on the whispered tone is of great importance.

Chinese is a tonal language. Every syllable has the initial, final and tone. There are four lexical tones which are commonly labeled in sequence from Tone 1 to Tone 4. The tonality of normal speech is characterized mainly by the contours of its F_0. In whispered Chinese, there is no F_0, but it can be uttered with four tones by speakers and its tone can be perceived by receivers. Even in the isolated whispered Chinese, the tone can be identified by auditors. The average tone perception score of 62.1% can be achieved by the perception tests[2, 3]. The tone perception score of Tone 3 is the highest, and the score of Tone 2 is the lowest among the four tones. At present there are no reports about research on the whispered tone recognition by the computer and on the contribution of every whispered tone feature in the automatic tone recognition. The tone recognition is important to both whispered speech recognition and reconstruction of normal speech from whispered speech in Chinese. So the tone features in the isolated whispered Chinese are studied by using an automatic tone recognition system.

In this paper, the whispered tone features are discussed and a new feature for the automatic whispered tone recognition is proposed.

1 Characteristics of whispered speech

When people talk in the whispery mode, the glottis is half opened and the turbulent flow created by exhaled air passing through this glottal constriction provides a noise source of sound. There is no obvious difference between the normal and whispered consonant when comparing their spectrograms, but there

* Supported by the National Natural Science Foundation of China (Grant Nos. 60272037 and 60340420325)

** To whom correspondence should be addressed. E-mail: lixueli@hotmail.com

are great differences between the normal and whispered vowel. There is no vocal cord vibration, so the whispered vowel is not quasi-periodic and has no F_0, but the formants still exist. It is the narrowing tract in the false vocal fold regions and the weak acoustic coupling with the subglottal system that change the vocal tract transfer function, so the first two formant frequencies (F_1 and F_2) shift to higher frequency and the formant bandwidths expand[4]. On the other hand, due to the noise source, the whispered speech has lower sound power and longer duration than the normal speech. Figure 1 gives the waveforms and spectrograms of the normal and whispered vowel /a/.

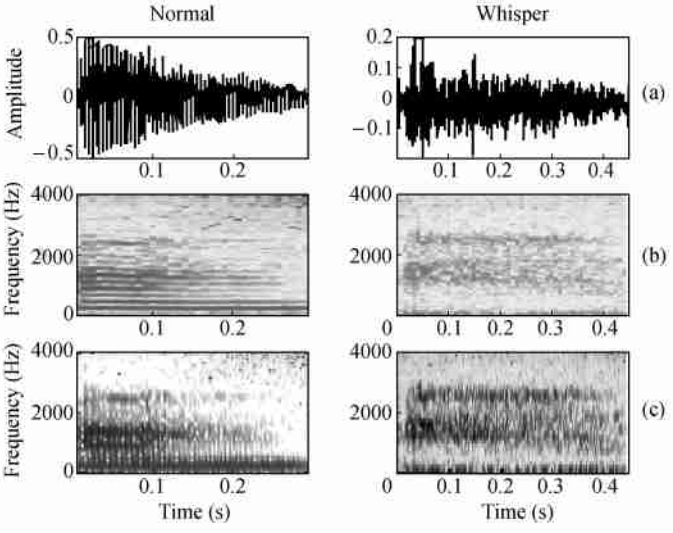

Fig. 1. Comparisons between the normal and whispered vowel /a/. (a) Waveform; (b) narrow-band spectrogram; (c) wide-band spectrogram.

2 Tone features extraction and comparison

Since F_0 is used to differentiate the meanings of various lexical items with identical consonants and vowels in the tone language, and F_0 is absent in whisper, it may be difficult to explain how native speakers of tone languages get the full information on tones from whispered speech. There are a number of possible explanations for this phenomenon: special maneuvers of the whisper, substitute of temporal envelope cues or stress features, and the aid of semantic context[5]. By observing the spectrograms / laryngoscopic videos and doing the perception tests, some researchers thought that the whispered tone might be characterized by the amplitude envelope, duration, glottal area, vocal tract length, and formant[3-5]. However, no report on the contributions of these features has been published. In this paper, these features are extracted and used in the whispered tone recognition. Their performances are compared, so the best feature for whispered tone recognition is obtained.

2.1 Tone feature extraction

The extraction methods of the amplitude envelope, duration, glottal area, formant, and vocal tract length are as follows:

The amplitude envelope (AE) is extracted by half-wave rectification and low-pass filtering with the cutoff frequency of 500 Hz.

The whispered speech is segmented into frames of 22.5 ms with the Hamming window. The duration is the total frame number of the speech. For the isolated whispered Chinese, the duration of Tone 3 is the longest and the duration of Tone 4 is the shortest among the four tones.

The vocal tract area (A_m) is obtained from the 10th-order linear predictive coding (LPC) model[6]

$$A_m = \frac{1-K_m}{1+K_m}A_{m-1}, \qquad (1)$$

where K_m is the reflection coefficient, A_1 is the area near the glottis and it is used as the glottal area, and A_{10} is the area near the lip and it is defined as the lip area.

The formant tracks are obtained by the improved dynamic programming formant tracking algorithm[4]. The second formant frequency track (F_2) and the

difference between the first two formant frequencies (DF_{21}) are used as the tone features.

The vocal tract length (VTL) is estimated with the vocal tract area and formant value by the Wakita algorithm[7]. The computation is complicated.

The tone pattern has nothing to do with the absolute values of these features. It is only related to their contour shapes except the duration. So the discrete orthogonal Legendre coefficients a_j ($j = 2, 3, 4$)[8] of these features are chosen as the tone features:

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f\left(\frac{i}{N}\right)\phi_j\left(\frac{i}{N}\right), \quad j = 2, \cdots, 4, \tag{2}$$

where N is the duration, and $\phi(\,)$ is the Legendre polynomial.

Because the tone information is decided mainly by the final unit, only features in the final unit are considered. The final unit can be obtained by the initial/final segmentation method[9]. In order to avoid the effect of the initial and the trailing part of the final, the features in the 0.2—0.8 part of the final are extracted.

These features are input into the vector quantization recognizer to train and recognize, respectively.

2.2 Experiment

Six simple Chinese whispered vowels /i/, /ü/, /a/, /o/, /e/ and /u/ with four tones were uttered 34 times by a woman in the lab with some machine noises. The four speech data of every vowel with four tones were used as the training set, and the other data were used as the recognizing set. These data are used in all the experiments. The features mentioned above were extracted, trained and recognized, respectively.

The tone recognition rates of all the features are listed in Table 1 and Table 2. It can be concluded that the contributions to the whispered tone recognition from the largest to the smallest are amplitude envelope, vocal tract length, vocal tract area, and formant change; and the duration is helpful to the tone recognition, especially for Tone 3 and Tone 4. The recognition rate of the amplitude envelope combined with duration is highest, which is 49.4%, but it is lower than the human perception rate, 62.1%. So the tone feature should be improved further.

Table 1. The tone recognition rates of all the features without duration (%)

Amplitude envelope (AE)	Glottal area (A_1)	Lip area (A_{10})	2nd formant (F_2)	Formant difference (DF_{21})	Vocal tract length (VTL)
34.2	27.1	25.8	23.6	23.8	31.4

Table 2. The tone recognition rates of all the features with duration (%)

Amplitude envelope (AE)	Glottal area (A_1)	Lip area (A_{10})	2nd formant (F_2)	Formant difference (DF_{21})	Vocal tract length (VTL)
49.4	46.1	43.8	38.1	38.9	47.0

3 Improved tone feature

In the study of chimaeric sounds[10], it is found that the perceptual importance of the envelope increases with the number of frequency bands, while that of the fine structure diminishes. So the multi frequency bands envelope feature might be helpful to the whispered tone recognition.

In addition, the tone is suprasegmental phoneme which is correlated with the human mentality and hearing. So the psychological and auditory feature should be taken into consideration. The multi frequency bands envelope feature can be obtained by the Mel-scale band-pass filters, which divide the frequency bands according to the human psychological sense. From the equal loudness contours, it can be seen that the frequency response of the human auditory system is emphasized in the frequency range of 3 kHz—5 kHz. So the speech spectrum should be weighted by an IIR loudness-weight filter to emphasize this sensitive frequency band[11]. The function of IIR loudness-weight filter is given as

$$H(z) = 1.3562\,\frac{1-0.731z^{-1}-0.3362z^{-2}+0.6404z^{-3}-0.2448z^{-4}-0.113z^{-5}+0.0546z^{-6}}{1-0.5716z^{-1}-0.1191z^{-2}+0.1491z^{-3}-0.0549z^{-4}-0.0085z^{-5}-0.0009z^{-6}}. \tag{3}$$

So the loudness-weighted multi Mel-frequency bands log-amplitude envelopes are used as the tone features in the whispered Chinese tone recognition. The band number is decided by the experiment. The recognition results of different band numbers are shown in Fig. 2. In consideration of the recognition rate and computational complexity, the loudness-weighted 32 Mel-frequency bands log-amplitude envelopes are chosen as the whispered tone features. In addition, the duration is added to improve the recognition rate. The vector quantization is used as the recognizer.

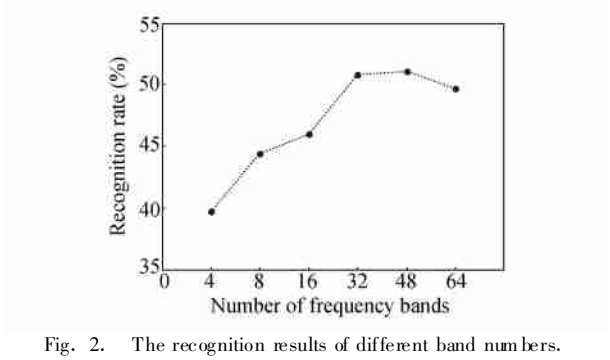

Fig. 2. The recognition results of different band numbers.

The tone recognition results of six vowels with the loudness-weighted 32 Mel-frequency bands log-amplitude envelopes and duration are given in Fig. 3. The average tone recognition rates are 48.9%, 56.1%, 83.9%, and 70.6% for Tone 1, Tone 2, Tone 3, and Tone 4, respectively. The average tone recognition rate for all the tones is 64.9%, and it has approached that of the human perception level. It is proved that loudness-weighted 32 Mel-frequency bands log-amplitude envelopes and duration are the main features in the automatic whispered tone recognition.

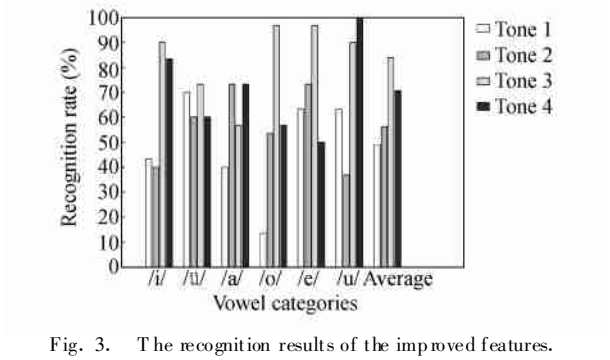

Fig. 3. The recognition results of the improved features.

4 Conclusion

In this paper, the characteristics of whispered speech are introduced and the whispered tone features are discussed. Because there is no F_0 in the whispered speech, other features, such as the amplitude envelope, duration, glottal area, lip area, formant, and vocal tract length, are extracted and their contributions to the tone recognition are compared. By the isolated Chinese tone experiments for speaker dependent, it is proved that the contributions to the whispered tone recognition from the largest to the smallest are amplitude envelope, vocal tract length, vocal tract area, and formant change; and the loudness-weighted 32 Mel-frequency bands log-amplitude envelopes and duration, which simulate the psychological and auditory characteristics of human, can be used as the main tone features in the whispered tone recognition. The average tone recognition rates in the isolated whispered Chinese approach that of the human perception level. The preliminary results may be helpful to the studies on Chinese whispered speech recognition and conversion.

Without F_0 in whispered speech, people tend to exaggerate the movements of the speech organs to make the whispered speech more intelligible, so the function of the amplitude envelope for the tone recognition is strengthened, and the amplitude envelope becomes one of the main tone features. The recognition rate is lower than that of normal speech, which is caused by the characteristics of the isolated whispered speech. The study of the continuous whispered tone recognition might improve the recognition rate. This will be done in the future work.

References

1 Morris W. R. and Clements A. M. Reconstruction of speech from whispers. Medical Engineering & Physics, 2002, 24(8): 515—520.

2 Liang Z. A. The auditory perception of Mandarin tones. Acta Physiologica Sinica (in Chinese), 1963, 26(2): 85—91.

3 Sha D. Q., Li X. L. and Xu B. L. Study on characteristics of the tones in whispered Chinese. Audio Engineering (in Chinese), 2003, 11: 4—7.

4 Li X. L. and Xu B. L. The formant comparison between Chinese whispered and voiced vowels. In: Proceedings of the 18th International Congress on Acoustics, Kyoto, Japan, 2004, 3291—3294.

5 Gao M. and Esling J. H. Articulatory features of tones in whispered Chinese. In: Proceedings of 15th International Congress of Phonetic Sciences, Barcelona, Spain, 2002, 2629—2632.

6 Yang X. J., Chi H. S. and Tang K. Digital Processing of Speech Signals (in Chinese), 1st ed. Beijing: Publishing House of Electronics Industry, 1995, 86—87.

7 Wakita H. Normalization of vowels by vocal-tract length and its application to vowel identification. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1977, 25(2): 183—192.

8 Chen S. H. and Wang Y. R. Vector quantization of pitch information in Mandarin speech. IEEE Transaction on Communications. 1990, 38(9): 1317—1320.

9 Li X. L., Ding H. and Xu B. L. Entropy-based initial / final segmentation for Chinese whispered speech. ACTA Acustica (in Chinese), 2005, 30(1): 69—75.

10 Smith M. Z., Delgutte B. and Oxenham J. A. Chimaeric sound reveal dichotomies in auditory perception. Nature, 2002, 416: 87—90.

11 Jiang W. J., Lin Y. R. and Wei G. A weighted method for noisy speech recognition based on loudness property. Pattern Recognition and Artificial Intelligence (in Chinese), 2001, 14(2): 166—170.