Interspeech 2018
2-6 September 2018, Hyderabad, India

Acoustic analysis of whispery voice disguise in Mandarin Chinese

Cuiling Zhang[1,2], *Bin Li*[3], *Si Chen*[4], *Yike Yang*[4]

[1]Southwest University of Political Science & Law, Chongqing, China
[2]Chongqing Institutes of Higher Education Key Forensic Science Laboratory, Chongqing, China
[3]City University of Hong Kong, Hong Kong S.A.R., China
[4]The Hong Kong Polytechnic University, Hong Kong S.A.R., China

cuiling-zhang@forensic-voice-comparison.net, binli2@cityu.edu.hk, sarach.chen@polyu.edu.hk, yi-ke.yang@connect.polyu.hk

Abstract

This paper investigates the auditory and acoustical characteristics of whispery disguised voice, and compares the patterns with those of normal (non-disguised) voices. It also evaluates effects of whispery disguise on forensic voice comparison. Recordings of eleven male college students' normal voices and whispery disguised voices were collected. All their normal and whisper speech was acoustically analyzed and compared. The parameters including average syllable duration, intensity, vowel formant frequencies, and long term average spectrum (LTAS) were measured and statistically analyzed. The effect of whispery voice disguise on speaker recognition by auditory perception and an automatic system were evaluated. Correlation and regression analyses were made on the parameters of whispery voice and normal voice. These simple regression models can be used for parameter compensation in forensic casework.

Index Terms: whisper, voice disguise, forensic speaker recognition, Chinese

1. Introduction

Voice disguise is a deliberate action by a speaker who wants to change his or her voice for the purpose of falsifying and concealing true identity [1][2]. There are electronic and non-electronic disguise types. The former relies on electronic devices to modify voice features and the latter is a deliberate deviation from habitual articulation, for example, whisper, falsetto, feigned foreign accent, change of speaking rate, or pinched nostrils. These disguises are easily manipulated and can create great obstacles for forensic speaker recognition, misguiding human listeners and degrading performance of automatic speaker recognition [2].

Of all voice disguise types mentioned above, whisper is a relatively common type used in kidnappings, threatening and harassing telephone calls. Though lack of vocalization, whispering conveys linguistic information pertaining "secret" or "threatening" clearly. Such efficiency and the difficulty in perceptual identification make whisper one of the most favorite voice disguise type found in criminal cases.

However, relatively few studies have investigated whispery voice disguise. [3][4] investigated the effects of whispers on criminal speaker identification by listening and reported whispering of suspect's voice significantly influenced identification performance on both suspect-present and suspect-absent lineups. [5][6] assessed effectiveness of forensic automatic speaker recognition systems and reported failure or significant issues during system evaluation using whispery voices. There are also abundant non-forensic studies that compared the difference between whispery and modally phonated voices, reporting higher vowel formant frequencies in whisper than in phonated equivalents in male and females [7-10].

This paper investigates the auditory and acoustic features of whispery disguised voice in Mandarin Chinese, focusing on acoustic differences between whispered-disguise and normal voice, and the influence of whispery disguise on forensic speaker recognition. It also examines effects of different strategies adopted by speakers to achieve disguise effect using whispery voices.

2. Speech materials and data analysis

The recordings for this study were selected from a database of disguised voices collected in 2004 which included speech recordings from 11 male college students using normal voices and nine different kinds of disguised voices including whisper [11]. As there were more male suspects and criminals than females ones in forensic cases recorded in [11], our study focused on and tested only speech samples produced by adult male speakers.

The speakers were students from China Criminal Police University, ranging from 21 to 24 years old. They spoke Standard Chinese as their first language. All speakers were asked to read ten sentences which are typical in kidnapping cases using their normal voices and subsequently using whisper to disguise their voices.

Recordings were made in a quiet room using a portable digital recorder (Sony ICD-P520). The recordings were made at sampling rate 16 kHz with 16 bit quantization.

All speakers' normal voice and whispery voice were analyzed by aural perception and acoustic comparison using Praat software [12]. Acoustic parameters including syllable duration, intensity, central frequencies of the first four formants of five representative monophthongs (/i/, /a/, /u/, /ɣ/, /y/), and long term average spectrum (LTAS) were measured and statistically compared.

10.21437/Interspeech.2018-2598

3. Results and discussion

The acoustical differences between whispery and normal voices are distinct: no periodic pulse and fundamental frequency in whispery voice, which exhibits noise chaos similar to those of fricatives but with clear formant structures. The formants of whispering are weaker, obscurer, but distinctively higher than those of normal voices, which can be due to the narrowing of the vocal tract in the false vocal fold regions and weak acoustic coupling with the subglottal system [25]. Fig.1 and Fig.2 present waveforms and spectrograms of speech samples in whispery voice and normal voice by a same speaker.

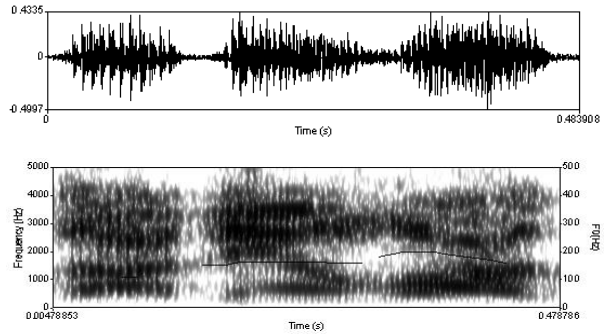

Fig.1. Waveform (top) and spectrogram (bottom) of speech by speaker ZW in normal voice.

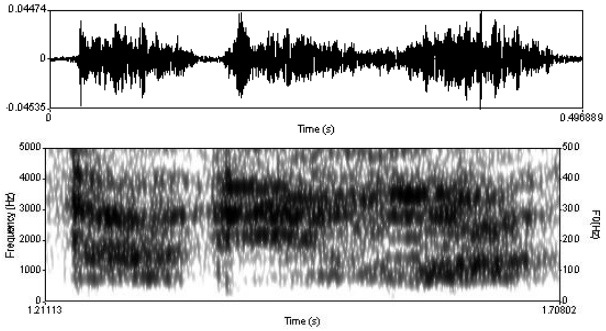

Fig.2. Waveform (top) and spectrogram (bottom) of speech by speaker ZW in whispery voice.

3.1. Syllable duration

The syllable duration is an alternative representation of articulation rates. As shown in Table 1, the total mean of average syllable duration in whisper across all speakers is 226ms, meaning 4.42 syllables per second, whereas in normal voices, the rate is 5.05 syllables per second. This indicates that the articulation in whispering tends to be slower than in normal phonation.

Paired T-tests of the average duration reveal that there is significant difference ($p<0.001$) between these two phonation types. Articulation in whisper is 14% slower than that in normal voice, but the pattern is not consistent across all speakers (two sped up slightly). There are also variances in intra-speaker variation in whisper and normal voice. Generally, the intra-speaker variation of whispery voice (total sd.=28) is smaller than that of normal voice (total sd=29), but is statistically negligible.

Table 1: *Average syllable durations of whispery and normal voices (ms) (N: Normal, W: Whisper).*

Speaker	Mean		St. Dev.		ΔMean	
	W	N	W	N		
HHK	218	186	63	26	32	17%
JF	327	267	33	42	60	22%
GL	214	209	25	25	5	2%
SZQ	179	180	24	30	-1	-1%
WLT	207	185	21	15	22	12%
WYS	201	212	34	32	-11	-5%
XB	222	191	30	22	31	16%
XH	226	203	21	26	23	11%
YT	252	183	24	52	69	38%
ZSP	246	206	18	31	40	19%
ZW	195	155	18	14	40	26%
TOTAL	226	198	28	29	28	14%

3.2. Intensity

Average intensity was calculated by sentences in two phonation conditions for all speakers.

Table 2: *Average intensity of whispery and normal voices (dB) (N: Normal, W: Whisper).*

Speaker	Mean		St. Dev.		ΔMean	
	W	N	W	N		
HHK	47	56	2	3	-9	-16%
JF	45	57	3	3	-12	-21%
GL	48	55	4	4	-7	-13%
SZQ	44	57	3	4	-13	-23%
WLT	42	60	3	3	-18	-30%
WYS	48	61	4	5	-13	-21%
XB	44	65	3	3	-21	-32%
XH	46	57	3	4	-11	-19%
YT	52	60	3	3	-8	-13%
ZSP	43	53	2	3	-10	-19%
ZW	47	66	3	3	-19	-29%
Total	46	59	3	3	-13	-22%

As shown in Table 2, there is a consistent tendency that all speakers' intensity in whisper decreased compared to that in their normal voices. Paired T-tests across sentences and speakers reveal significant difference ($p<0.001$) between whisper and normal voice. The total intensity level of whisper is 13 dB (22%) lower than normal voice. Besides, our speakers show similar extent of intra-speaker variation of intensity in both phonation conditions (total s.d.=3dB).

3.3. Formant frequency

Formant patterns, formant trajectories, and transitions within syllables or between syllables are all similar in normal phonation and in whisper. But formants in whisper seem higher than those in normal voice. Mean frequencies of the first four formants of five monothongs (/i/, /a/, /u/, /ɣ/, /y/) in both whispery voice and normal voice were measured and compared.

We measured all tokens of 5 monosyllabic words (/ni214/, /na35/, /şu35/, /tşɣ55/, /єy214/) in all sentences, results of which are listed in Table 3. Paired T-tests of the average formant frequencies across vowels and speakers show significant differences between phonation conditions for all four formants ($p=0.001$ for F4, and $p<0.001$ for F1, F2, and

F3). The formant frequencies of whispered vowels are higher than those vowels in normal voice, especially for F1 and F2 (75% increase for F1, 30% for F2, 5% for F3, and 4% for F4). But the patterns of between-whisper-normal-voice difference vary across vowel types: /u/, /ɣ/, and /i/ show greater extents of variations than /a/ and /y/ with /a/ showing the smallest variation especially on F3 and F4. The increments from normal to whispery voices are: 63% for /u/, 35% for /ɣ/ 31% for /i/, 13% for /y/, and 6% for /a/ respectively. Then, among the four formants, F1 shows the greatest extent in increment in general, as illustrated in Fig.5.

Table 3: *Mean formant frequencies of 5 vowels of whispery voice and normal voice (Hz).*

	F1		F2		F3		F4	
	W	N	W	N	W	N	W	N
i	951	440	2230	2186	2860	2831	3602	3487
a	1058	901	1603	1495	2594	2588	3582	3579
u	968	394	1573	798	2633	2518	3420	3292
ɣ	853	546	2096	1192	2655	2512	3410	3396
y	880	419	1985	1872	2543	2374	3368	3211

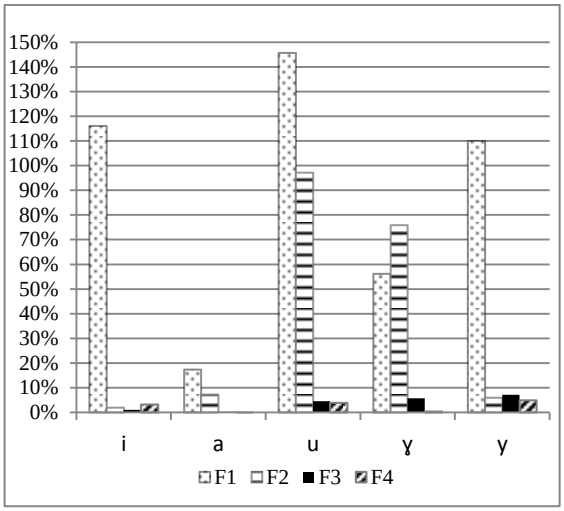

Fig.3. *Increments in formant frequency of 5 vowels in whisper and normal voice across all speakers.*

3.4. Long Term Average Spectrum (LTAS)

LTAS is a parameter describing the resonance characteristics of a speaker. We analyzed LTAS using CSL3700 to calculate its mean, root mean square (RMS), min and max for all speakers. Table 4 lists LTAS of whispery voices and other statistics including differences in mean (ΔMean) and root mean square (ΔRMS) between whisper and normal voices.

The data show consistent patterns for all speakers that there is significant energy drop of mean and RMS, by 109% and 44% respectively, from normal phonation to whisper. Paired T-tests across all speakers reveal significant differences ($p<0.001$) between whispery and normal voices. Though the patterns show variations across speakers, the shape of LTAS changes in whisper and normal voices of the same speaker are basically similar. Fig.4 and Fig.5 illustrate the LTAS difference between speech samples in whispery voice and normal voice by two speakers. So, it is proposed that the contours of LTAS curves should be more informative and reliable for speaker recognition than actual energy values in whisper examination.

Table 4: *LTAS of whispered sentences (dB).*

Speaker	Mean	RMS	ΔMean		ΔRMS	
GL	2.94	6.17	-4.88	-62%	-5.83	-49%
HHK	0.41	8.68	-4.65	-92%	-3.11	-26%
JF	-4.01	8.19	-13.88	-141%	-5.42	-40%
SZQ	-3.44	6.79	-14.04	-132%	-7.21	-52%
WLT	-4.04	8.15	-10.65	-161%	-3.7	-31%
WYS	2.41	6.33	-12.47	-84%	-9.88	-61%
XB	-2.57	6.45	-13.37	-124%	-8.3	-56%
XH	-0.7	8.08	-7.23	-111%	-4.15	-34%
YT	3.43	9.73	-6.28	-65%	-3.04	-24%
ZSP	-3.55	6.71	-4.9	-363%	-2.42	-27%
ZW	0.11	6.91	-13.6	-99%	-11.11	-62%
Total	-0.82	7.47	-9.63	-109%	-5.83	-44%

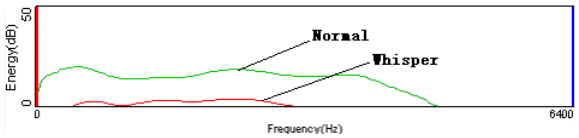

Fig.4. *LTAS contours of whisper and normal voices from speaker JF.*

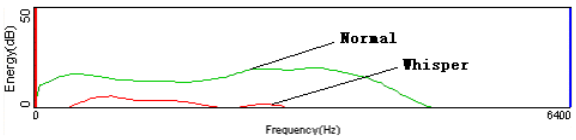

Fig.5. *LTAS contours of whisper and normal voices from speaker SZQ.*

4. Influence on forensic speaker recognition

Auditory speaker recognition test was made on both whispery voice and normal voice. We chose 3 sentences for each speaker in normal phonation and whisper, to form 10 normal-normal voice pairs and normal-disguised voice pairs. For each type of the pairs, five were same-speaker pairs and the others different-speaker pairs. Ten male students other than the speakers were selected as listeners, in which 5 listeners were familiar with the speakers and the others were not. They listened and compared each pair, decided on whether the voice pair were from the same speaker (Yes, No, No conclusion).

The results show that for the listener group who was familiar with the speakers the correct speaker identification rate of normal-normal voices and normal-whispery voices are 100% and 56% respectively. While for the listener group not familiar with the speakers, the accuracy rates for normal-normal voices and normal-whispery voices were 96% and 78% respectively. We then tested on the performance of a forensic automatic speaker recognition system [11]. For normal voices the correct recognition rate is 100%. Their similarity rates for all 20 speakers range from 0.68 to 1. However, whisper cases all failed in the test, with the similarity rates ranging from 0.25 to 0.79 (the threshold is 0.80).

In order to explore the parametric correlation between whispery and normal voices, we ran the correlation and liner regression analysis. Statistics show that there is no significant correlation on LTAS and the fourth formant between whispery and normal voice. For other parameters the correlation and regression is significant. This indicates certain parametric compensation method should be promising if the offender adopted whispery disguise and the suspect spoke in normal voice. Table 5 shows the correlation coefficients and corresponding regression equations on several parameters between normal and whispery voice (Y: normal voice, X: whispery voice).

Table 5: *Results of correlation and regression analysis between normal and whispery voices.*

Parameter	R	*p*	Regression equation
Syllable duration	0.612	<0.001	Y＝83.047＋0.507X
Intensity	0.408	<0.001	Y＝34.611＋0.527X
F1	0.290	0.032	Y＝379＋0.171X
F2	0.639	<0.001	Y＝160＋0.753X
F3	0.376	0.005	Y＝1666＋0.336X

5. Conclusions

Whispery voice is often used in criminal cases because it involves less effort in execution yet is effective in conveying secretiveness or threatens without revealing identities. Whispering results in significant auditory and acoustic variations, bringing difficulties and challenges to forensic speaker identification.

Our study shows that there are significant differences between speakers' normal voice and whispery voice for all phonetic parameters including syllable duration (articulation rate), intensity, formant frequencies of vowels, and LTAS. Most speakers in our experiments slowed down in speaking when they whispered. Their intensity levels were lowered, but formants of vowels especially F1 and F2 raised significantly. LTAS figures also confirm that whispery voice contains significant energy drop of mean and RMS, but that the contour of LTAS curves should be more useful for speaker recognition.

Speakers vary in the extent of between-whisper-normal-voice differences. This seems to suggest that they may have adopted individualized strategies when using whisper as a disguise of voices to conceal their identities. So, caution has to be taken in forensic casework that speakers' variations are possible in addition to general tendencies.

Auditory speaker identification tests show that for the normal-normal voice pairs, listeners who are familiar with the speakers show higher accuracy rates than those who are not. However, when whisper is involved in voice pairing, familiarity with speakers' voices does not seem to facilitate matching any more. The automatic speaker recognition test confirms further the extreme difficulty that whisper poses to our system.

Correlation and regression analysis shows relatively strong parametric correlation between whispery and normal voice. The simple linear regression is significant for syllable duration, intensity, and the first three formants of vowels. In forensic casework these simple models or more sophisticated statistical models should be useful for parametric compensation between whispery voice (from offender) and normal voice (from suspect), as well as for other voice disguise types.

6. Acknowledgements

This research was supported by the National Social Science Foundation of China Key Program (Grant No. 16AYY015), Southwest University of Political Science and Law research funding (2015-XZRCXM003), and Chongqing Social Enterprise and People's Livelihood Guarantee Scientific and Technological Innovation Special Research and Development Key Project (cstc2017shms-zdyfX0060), and partially by City University of Hong Kong Grants 7004740 and 6000653.

7. References

[1] R. D. Rodman, Computer recognition of speakers who disguise their voice, the International Conference on Signal Processing Applications and Technology ICSPAT2000.

[2] H. Hollien, *Forensic Voice Identification*, Academic Press, New York, 2002.

[3] A. Reich, K. Moll, J. Curtis, Effects of selected vocal disguises upon spectrographic speaker identification, *J. Acoust. Soc. Am.* 60 (1976) 919–925.

[4] T. L. Orchard, A. D. Yarmey, The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification, *Appl. Cognitive Psych.* 9 (1995) 249–260.

[5] H. J. Künzel, J. Gonzalez-Rodriguez, J. Ortega-García, Effect of voice disguise on the performance of a forensic automatic speaker recognition system, *ODYSSEY04 - The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp 153–156.

[6] C. Zhang, T. Tan, Voice disguise and automatic speaker recognition, *Forensic Sci. Int.* 175(2007)118–122.

[7] M. Matsuda, K. Kasuya, Acoustic nature of the whisper, in: *Proceedings of the Eurospeech '99*, 1999, pp. 133–136.

[8] I. Eklund, H. Traunmüller, Comparative study of male and female whispered and phonated versions of the long vowels of Swedish, *Phonetica.* 54 (1997) 1–21.

[9] Li, B. and Guo, Y., Mandarin tone contrast in whisper, *Proceedings of the Tonal Aspects of Language*, *ISCA*, Nanjing, 2012, pp66-69.

[10] Li, B. and Zhang, C. Production and perception of Mandarin tones with whisper. *Chinese Journal of Rehabilitation*. 31(6), 2016: 408-411.

[11] C. Zhang, Acoustical Study on Disguised Voices, PhD dissertation, Nankai University, Tianjin, China, 2005.

[12] P. Boersma, D. Weenink, Praat: doing phonetics by computer (Version 4.2), http://praat.org/, 2004.