ARCHIVES OF ACOUSTICS
32, 4 (Supplement), 173–178 (2007)

COMPARISON OF SUBJECTIVE AND OBJECTIVE SPEAKER RECOGNITION UNDER VOICE DISGUISE CONDITIONS

Wojciech MAJEWSKI

Wrocław University of Technology
Wybrzeże Wyspiańskiiego 27, 50-370 Wrocław, Poland
e-mail: wojciech.majewski@pwr.wroc.pl

(*received July 15, 2007; accepted October 16, 2007*)

An experiment was performed in order to compare the results of subjective and objective speaker recognition under voice disguise conditions. The experimental material consisted of the key sentence "To jest akustyka" (Eng. "This is acoustics") spoken several times by Polish male speakers in a natural mode and under voice disguise conditions. In the subjective method the utterances were grouped in pairs and presented to the listeners whose task was to make a decision whether a given pair of speech samples was produced by one speaker or by two different speakers. In the objective method two parametrical representations of speech (LPC coefficients and Km reflection coefficients) were utilized and a computer program for automatic speaker verification was applied. For normal speech both methods exhibited a very good effectiveness of speaker recognition and the results for the subjective method were a bit lower (98.9% in comparison to 99.4% for the objective method). Under voice disguise conditions, however, the mean results for the subjective method (92.3%) were substantially better than for the objective one (77.2%).

Keywords: speaker recognition, voice disguise.

1. Introduction

Recognition of speakers based on their utterances has many different applications. Among the most important are forensic applications. In legal proceedings and inquiries there is often a need to find out if a recorded speech sample of an unknown criminal was produced by any of the suspects. Criminals frequently want to cover their identity and try to disguise their voices or mimic a voice of some other speaker [1]. Only few reports on speaker recognition under voice disguise conditions are available [2–5]. Disguised speech may be typically found in situations when a blackmailer or kidnapper makes a call to his/her victim and expects his voice is being recorded. The overall occurrence of voice disguise is in such cases over 50% and it increases to almost 70% in the case of blackmailing [5]. The problem is very serious, since voice disguise has a deteriorating

effect on speaker recognition scores [1], regardless of the method utilized for speaker recognition [6].

To cope with this problem, in the case of automatic speaker recognition, an experiment on automatic speaker identification under voice disguise conditions was performed [7], utilizing different methods of speech signal parametrization and different recognition algorithms. Unfortunately, this experiment did not solve the problem because under the adopted measuring conditions (steady segments of vowels extracted from the key word "logarytm") the recognition scores were not encouraging. The conclusion from that study was that the steady segments of vowels of proven usefulness for automatic speaker recognition under a normal mode of speech production [6] are not sufficient for automatic speaker recognition under voice disguise conditions because in such a case interspeaker variations may be smaller than the intraspeaker ones. Thus, a subsequent experiment [8] was performed utilizing the same speech material, but instead of automatic speaker identification an aural-perceptual speaker verification was carried out. The idea to compare the performance of a machine with the performance of a human being under the same measuring conditions was not realized because the trials to make subjective judgments of a speaker identity on the basis of steady segments of vowels failed. Thus, in contrast to the previous study [7] in the experiment the subjective judgments of the speaker identity were based on the whole word "logarytm". The subjective speaker verification scores (70–90%) were much higher than 40–50% scores obtained in the automatic speaker identification experiment. A conclusion resulted from the comparison of these two studies that under voice disguise conditions the perceptual method provides a better result than the automatic one, needs to be checked under exactly the same measuring conditions and this was the main purpose of the present study.

2. Experimental procedure

2.1. Phonetic material

As a phonetic material a key sentence "To jest akustyka" (eng. "This is acoustics") was selected since it contains five of six Polish vowels considered as good carriers of individual voice features. This sentence was spoken 10 times by each of 20 male Polish speakers under four different conditions, i.e. in a natural mode of speech production and under the following voice disguise conditions:

- low pitch,
- whisper,
- pencil between the front teeth.

These three conditions have been selected because they may be easily applied and they are often used by the criminals to cover they identity.

The recordings were made in an ordinary room by means of a standard computer microphone ("Logitech") and Sound Blaster 128PCL card. The signal from the microphone was sampled at a rate of 11.025 kHz and digitised with a 16-bit resolution. All speech samples were normalized to one level and recorded in *wav format*.

2.2. Subjective tests

Experiments on aural-perceptual speaker verification were performed by a group of six listeners of normal hearing, who were comparing subsequent pairs of key utterances. The utterances were reproduced under average acoustic conditions (ordinary room) by means of a loudspeaker system consisting of an amplifier (Unitra Fonica PN9013) and a loudspeaker column (Tonsil, 60 W). The listeners were placed in 2–3 m distance from the loudspeaker. The task of the listeners, who did not have any previous experience in subjective listening tests, was to make a decision whether a given pair of the stimuli (one from the reference set and one from the test set) was produced by one speaker or two different speakers and to write down their decisions on answer sheets. Next the answer sheets were checked with regard to the numbers of false rejection error α and false acceptance error β.

For a given speaker the reference set consisted of 10 repetitions of natural speech. The test set consisted of the same 10 repetitions of natural speech and 10 speech samples randomly chosen from the speech samples produced by the remaining speakers. Since each sample from the reference set was compared with a speech sample from the test set, 200 pairs of speech samples were obtained. In these 200 pairs 100 pairs contained the speech samples produced by the same speaker and 100 pairs contained the speech samples produced by different speakers. An advantage of such an arrangement was that a granularity of α and β error was the same and equal to 1%. Since the voices of 20 speakers were examined, the total number of speech pairs combinations for given speaking conditions was 4000 $(200 \times 20 = 4000)$. Since four different speaking conditions were examined, the total number of examined speech pairs was 16 000.

Each of the subjects took part in eight listening sessions of roughly three hours in duration each. After a short training, each subject made during each session 2000 judgments with short breaks for a rest, usually after listening to 400 pairs of stimuli. Each session concerned the comparisons of two sets of stimuli coming from a definite manner of speech production. Within these 2000 judgments, 1000 concerned the pairs of speech samples produced by the same speakers and 1000 – the pairs produced by different speakers.

The sets of stimuli in the listening sessions were arranged in the following order:

Session 1 and 2 – normal speech vs. normal speech.

Session 3 and 4 – normal speech vs. “pencil between the front teeth” disguise.

Session 5 and 6 – normal speech vs. “low pitch” disguise.

Session 7 and 8 – normal speech vs. ”whisper” disguise.

Thus, the comparisons of speech samples concerning the easiest case (normal speech vs. normal speech) were carried out at the very beginning and the most difficult case (normal speech vs. whisper) at the very end.

The experimental procedure for the subjective speaker recognition was executed within the diploma work [9] supervised by the author of the present paper.

2.3. Objective tests

The experiments on objective speaker verification were carried out by means of a Voice Print computer program [10] that has been worked out at Wrocław University of Technology. This program utilizes orthogonal prediction coefficients of a speech signal. On the basis of the learning sequence consisting of 10 repetitions of normal speech samples for each speaker, the recognition system created 20 reference classes of voices from the mean vectors of individual voice features. The testing sequence consisting of 10 normal speech samples for each of 20 speakers was utilized to select the optimal parameters of the recognition system and to set up a verification threshold to obtain equal error rate (i.e. $\alpha = \beta$), which, however, was not always feasible because the granularity of false rejection error α was equal to 10% (for a given speaker only 10 speech samples were available), while the granularity of false acceptance error β was equal to 0.53% (since in this case 190 speech samples of the remaining 19 voices were available).

Under the voice disguise conditions, the testing sequence consisted of 200 disguised speech samples (20 speakers times 10 repetitions for each speaker) for each of the three methods of voice disguise. Each vector from the testing sequence was compared with the reference vector of a given speaker. If the distance between the two vectors under comparison was smaller than the verification threshold, the system accepted a given speech sample as belonging to a given speaker. In the opposite case a given speech sample was rejected. On the basis of α and β errors a verification effectiveness VE was calculated from the following formula, applied also in the evaluation of subjective tests:

$$\mathrm{VE} = 100 - (\alpha + \beta)/2 \quad [\%]. \tag{1}$$

The calculations of speaker verification effectiveness were carried out for two methods of speech signal parametrization: LPC coefficients and Km reflection coefficients.

The experimental procedure for the objective speaker recognition was executed within the diploma work [11] supervised by the author of the present paper.

3. Results

The results of the speaker verification effectiveness, averaged over the speakers and for the subjective method also over the listeners, are presented in Table 1. Similarly, the averaged verification errors are presented in Table 2.

From the data presented in Table 1 it may be seen that for normal speech samples (no disguise) both methods exhibited a very high effectiveness of the speaker verification and the results for the subjective method were a bit lower (98.9% in comparison to 99.4% for the objective method). Under the voice disguise conditions the mean results for the subjective method were still high (92.3%), while for the objective method the overall mean dropped to 77.2%.

Similar trend may be observed in the verification errors presented in Table 2. Under the normal conditions (no disguise) the errors for both methods are very small, from 0.0

to 1.4 %. Under the voice disguise conditions the errors are much higher, ranging from 2.9% to 17.6% for the subjective method and from 13.7% to 31.5% for the objective one.

Table 1. Speaker verification effectiveness (in %).

Disguise method	Verification effectiveness in percent			
	Subjective tests	Objective tests		
		LPC	Km	Mean
No disguise	98.9	99.5	99.3	99.4
Pencil	96.0	74.0	81.0	77.5
Low voice	95.6	78.2	80.7	79.5
Whisper	85.3	76.6	72.8	74.7
All disguises	92.3	76.3	78.2	77.2

Table 2. False rejection error α and false acceptance error β (in %).

Disguise method	Verification errors in percent					
	Subjective tests		Objective tests			
			LPC		Km	
	α	β	α	β	α	β
No disguise	1.0	1.1	0.0	1.0	0.0	1.4
Pencil	2.9	5.1	31.5	21.4	19.0	19.1
Low voice	4.2	4.6	30.0	13.7	22.0	16.6
Whisper	11.9	17.6	26.0	20.9	28.5	25.9
All disguises	6.3	9.1	29.2	18.7	23.2	20.5

4. Conclusions

A comparison of the results obtained in the adopted measuring conditions by the two methods under investigation indicates that under the voice disguise conditions the subjective method provides better results of speaker verification than the objective one, in spite of the fact that for normal speech both methods were equally good.

Since the objective method may be easy and fast applied, while the subjective method is very tedious and time consuming, it seems to be advisable in forensic applications to start the verification procedure with the objective method and carry out the subjective tests only when the automatic speaker verification system provides large false rejection errors α and false acceptance error β. A special attention should be given to false acceptance error β since in forensic applications this error may lead to a condemnation of an innocent person.

References

[1] HOLLIEN H., *The Acoustics of Crime*, Plenum Press, New York 1990.

[2] HOLLIEN H., MAJEWSKI W., DOHERTY E. T., *Perceptual identification of voices under normal and disguise speaking conditions*, Journal of Phonetics, **10**, 139–148 (1982).

[3] REICH A. R., DUKE J. E., *Effects of selected voice disguises upon speaker identification by listening*, J. Acoust. Soc. Amer., **66**, 1023–1028 (1979).

[4] KÜNZEL H. J., *Effects of voice disguise on speaking fundamental frequency*, Forensic Linguistics – Int. J. Speech, Language and the Law, **7**, 2, 149–179 (2000).

[5] MASTHOFF H., *A report on a voice disguise experiment*, Forensic Linguistics – Int. J. Speech, Language and the Law, **3**, 1, 50–64 (1996).

[6] MAJEWSKI W., BASZTURA C., *Integrated approach to speaker recognition in forensic applications*, Forensic Linguistics – Int. J. Speech, Language and the Law, **3**, 1, 50–64 (1996).

[7] MAJEWSKI W., MAZUR–MAJEWSKA G., *Automatic speaker recognition under voice disguise conditions*, [in:] Proc. 17 ICA, vol. IV, 62–63, Rome 2001.

[8] MAJEWSKI W., *Aural-perceptual speaker verification under voice disguise conditions*, [in:] Proc. Subjective and Objective Assessment of Sound, CD Rom, Poznań 2004.

[9] WYSTUP Ł., *Subjective and objective speaker recognition under voice disguise conditions. Part 1. Subjective experiments* [in Polish], Master's Thesis, Wrocław University of Technology, 2006.

[10] NICIARZ S., *AVR system based on LPC parameters* [in Polish], Masters' Thesis, Wrocław University of Technology, 1999.

[11] MARSZAŁEK D., *Subjective and objective speaker recognition under voice disguise conditions. Part 2. Objective experiments* [in Polish], Master's Thesis, Wrocław University of Technology, 2006.