INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 3, PP. 615-620
Manuscript received June 18, 2024; revised July, 2024. doi: 10.24425/ijet.2024.149587

Subjective tests of speaker recognition for selected voice disguise techniques

Piotr Staroniewicz

***Abstract*—Research work on the effectiveness of voice disguise techniques is important for the development of biometric systems (surveillance) as well as phonoscopic research (forensics). A speaker recognition system or a listener can be deliberately or non-deliberately misled by technical or natural methods. It is important to determine the impact of these techniques on both automatic systems and live listeners. This paper presents the results of listening tests conducted on a group of 40 people. The effectiveness of speaker recognition was investigated using selected natural (chosen from four groups of deliberate natural techniques: phonation, phonemic, prosodic and deformation) and technical (pitch shifting, GSM coding) voice disguise techniques. The results were related to the previously obtained outcomes for the automatic method of verification carried out using a classical speaker recognition system based on MFCC (Mel Frequency Cepstral Coefficients) parameterisation and GMM (Gaussian Mixture Models) classification.**

***Keywords*—speaker recognition; forensics; biometrics; voice desguise**

I. Introduction

VARIOUS types of voice disguise are encountered in situations where it is difficult to recognise a person from their voice. In most cases, this involves remote communication, e.g. a telephone call. Telephone crimes are often committed, mostly against elderly people, where the caller claims to be someone close to them (e.g. a family member). Voice disguise is also used for another type of crime such as criminal threats and ransom demands. In these cases, criminals often presents their demands using a voice recording. However, they must protect themselves against a possibility of their voice being recognised by the police, for example. Here again, a much favoured solution is to use one of many voice masking techniques. Criminals use various types of voice modulators and are thus able to retain their anonymity and make the work of the services much more difficult. The aim of this thesis is to investigate the impact of voice masking techniques using subjective tests. The tests were conducted on a group of 40 people.

II. Objectives

In modern biometric systems speaker recognition is now widely used. These include applications, such as secure access control, transaction authentication and forensics. Both in applications where the voice is recognised by a human and in increasingly common automated systems, the measurable parameters of the speech signal are subject to changes, which can lead to incorrect recognition of the speaker [1]. Other biometric identifiers, such as DNA, fingerprint or iris, are highly persistent over time, while the human voice undergoes significant changes due to ageing, emotion and many other intentional and unintentional factors, resulting from at least the encoding or transmission of the speech signal [2]. Regardless of their origin, these factors represent a distortion of the voice considered 'normal' for an individual and are therefore known as voice masking or alternatively, as voice disguise techniques [3-8].

TABLE I
SYSTEMATISATION OF VOICE DISGUISE TECHNIQUES

Type of voice disguise	Technical	Natural
Deliberate	The deliberate use of a device or computer software that digitally processes the speech signal in order to modify its parameters (e.g. pitch-shifter software that changes the fundamental frequency of the speaker's voice, etc.).	The intentional manipulation of the speaker's speech production organs that results in a significant change in the naturalness of pronunciation.
Non-deliberate	Unintentionally introduced speech distortion dependent on the coding method used or the characteristics of the telecommunications channel (i.e. frequency band in telephony, speech coding technique used, earphone varieties, etc.).	Not intentionally introduced changes in voice parameters as a result of the temporary effects of illness, drugs, alcohol, the speaker's physical state, emotional state or even ageing.

Voice disguising techniques can be classified according to two independent categories [9]. The first subdivision is that of intentional and non-intentional techniques, whereas the second division is between technical and natural techniques (sometimes also referred to as the division between electronic and non-electronic methods). The kinds of disguises according to the above classification are shown in Table I. Deliberate voice disguising techniques are those where, at the intent of the speaker trying to hide his or her identity or to impersonate another person [10], the speaker deliberately changes the parameters of their voice. To accomplish this, the speaker may use technical tools to convert his or her voice, such as electronic devices or computer applications. While keeping the semantic

Author is with Wroclaw University of Science and Technology, Poland (e-mail: piotr.staroniewicz@pwr.edu.pl).

information of the speech signal, a physical transformation of the voice characteristics is performed.

Non-deliberate voice disguising occurs when the speaker's voice undergoes changes that are not controlled by the speaker. Unintentional technical disguises are mainly the distortion and degradation of speech resulting from the telecommunications channel [11]. In contrast, unintentional natural distortion is most often caused by changes that affect the normal functioning of the speaker's body. Factors such as ageing, diseases affecting speech organs, emotional state, fatigue and drowsiness or the influence of intoxicants should be mentioned here [12].

The deliberate natural disguising of the speaker's voice may have two possible purposes: one may be to mask the identity of the speaker and the other to imitate another speaker [10]. Among the natural deliberate voice disguising techniques, we encounter numerous examples and this variety of techniques can be grouped into four main types [9]: phonation (raised or lowered pitch, whisper, inspiratory speech, screeching), phonemic (foreign accent, dialect, feigning speech defect, imitating), prosodic (intonation changes, stress placement, pronunciation tempo, changes in the length of speech segments) and deformational techniques (objects in or over the mouth, pinched nostrils, lips protrusion, holding of the tongue).

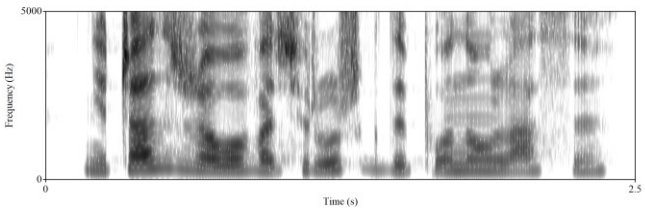
The previous studies [13] were performed for an automatic voice verification system and selected natural disguise techniques: lowered pitch (phonation), raised pitch (phonation), lowered pronunciation tempo (prosodic), raised pronunciation tempo (prosodic), American accent (phonemic), whisper (phonation), pinched nostrils (deformation), clenched jaws (deformation). The speaker's voice was masked best for the phonation techniques, where the highest EER (Equal Error Rate) values were obtained. For the raised pitch technique, this was 58.33% and for whisper, 38.26%. The error values for the other masking techniques were significantly lower (all below twenty per cent). The automatic system was most difficult to deceive with prosodic techniques. For the lowered pronunciation rate technique, it was 7.29%, while for the raised pronunciation rate it was 10.49%.

It was decided to test how previously studied masking techniques would be able to mislead listeners. In the subjective study the number of natural masking techniques was reduced, abandoning all the prosodic techniques that were of least importance in the earlier automatic tests and in the preliminary tests with groups of listeners showed little significant difference to the unmasked voice. Instead, the selected technical voice masking methods were added to the listening tests for comparative purposes.

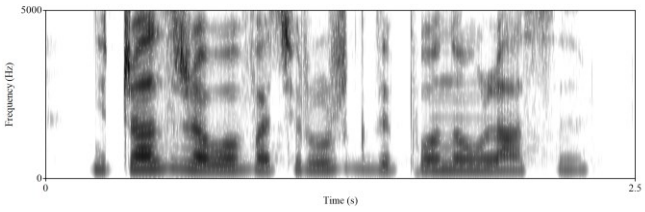
III. METHODS

The appropriate preparation of an acoustic database containing relevant recordings of speech samples is usually a key stage in voice recognition research [14,15]. An acoustic base containing natural voice masking techniques was used for the study. It consisted of 16 speakers - eight women and eight men. The database contained recordings of masking techniques such as: lowering and raising the tone, slower and faster speech, American accent, whispering, nasal speech, speech through clenched teeth. Six utterances with different content, containing each of the above-mentioned masking techniques were recorded for each speaker. Six different utterances of the normal voice were also created. For the subjective listening tests, recordings containing normal voice, raised and lowered tone, whisper, nasal speech and speech through clenched teeth were selected from this database. In the remaining masking techniques, the effect of voice masking was so small that it was decided to omit these examples.

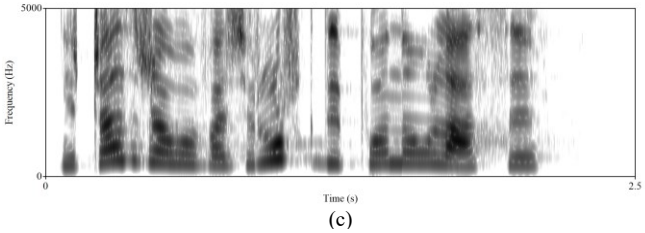
Fig.1 shows the example spectrograms for a selected speaker (female voice) and the utterances of the same content spoken with different natural voice masking techniques that were used in the acoustic signal database being tested. Fig.2 summarises the automatically detected fundamental frequency waveforms for the same signals. Noteworthy are the large differences apparent in the spectral characteristics relative to the natural speech, especially for the two phonation techniques of voice masking, which are whisper and raised pitch (Fig.1 c, Fig.1g, Fig.2 c, Fig.2 g). With the whispered speech, understandably, the presence of laryngeal tone frequency was not detected.



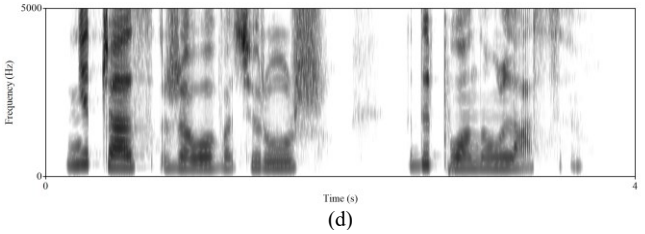
(a)



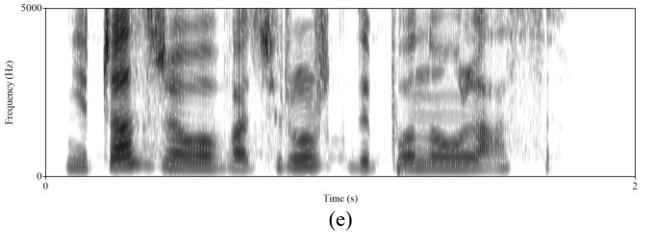
(b)



(c)



(d)



(e)

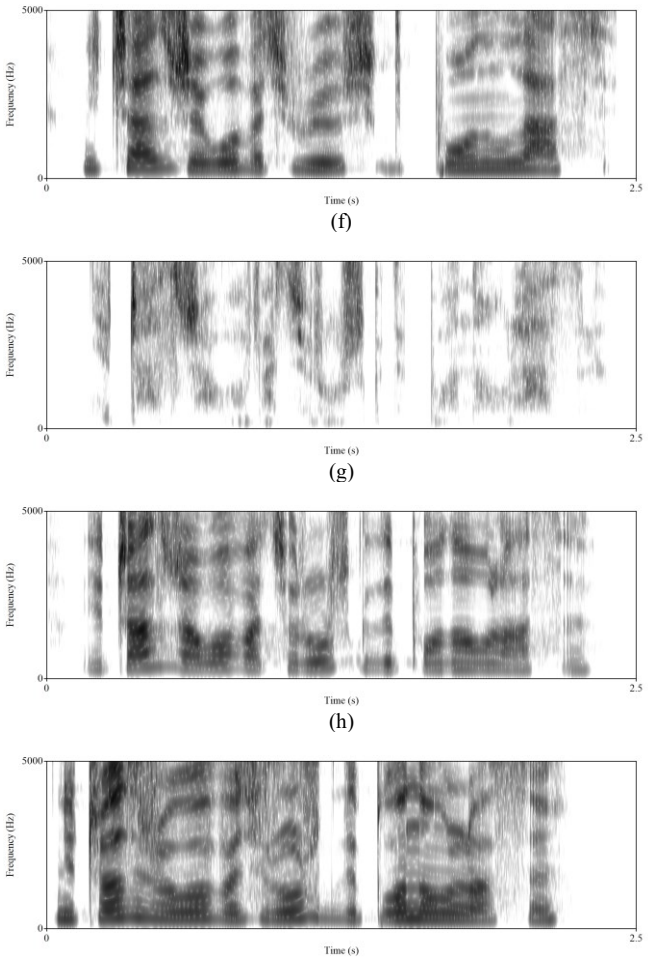
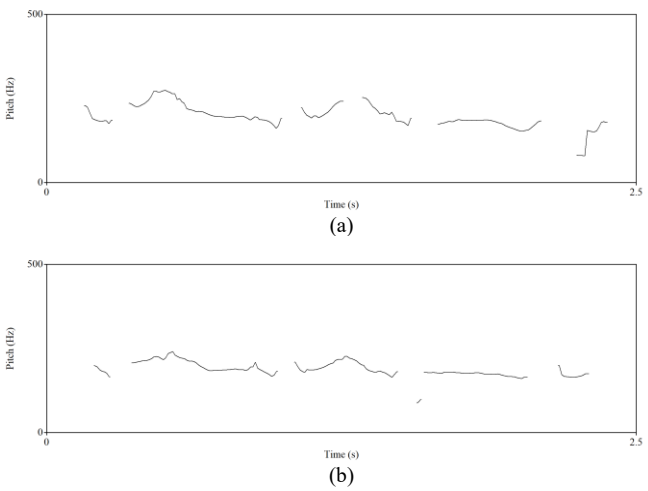


Fig. 1. Spectrograms of utterance ‘Nie czas żałować róż gdy płoną lasy’ (in Sampa notation: ‘n’e tSas Zawovats’ ruS gdy pwonow lasy’) of one female speaker: (a) no disguise, (b) lowered pitch, (c) raised pitch, (d) lowered pronunciation tempo, (e) raised pronunciation tempo, (f) American accent, (g) whisper, (h) pinched nostrils, (i) clenched jaws.



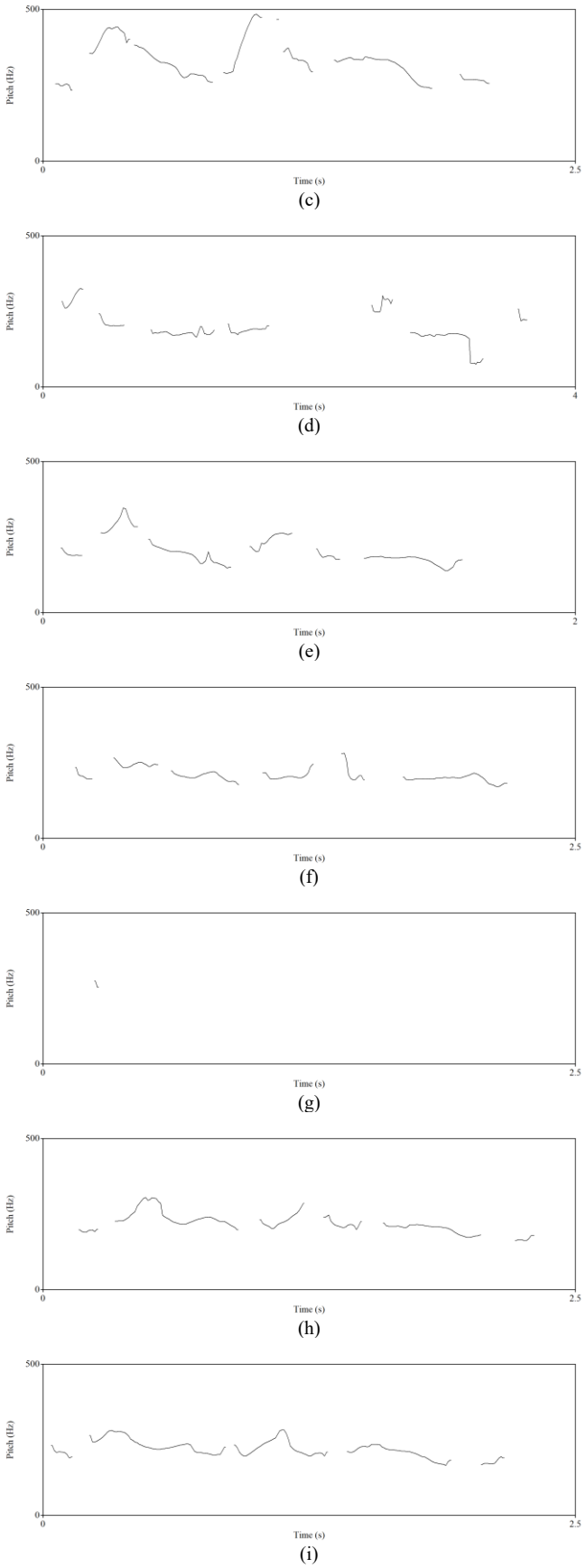


Fig. 2. Automatic pitch waveform detection for utterance ‘Nie czas żałować róż gdy płoną lasy’ (in Sampa notation: ‘n’e tSas Zawovats’ ruS gdy pwonow lasy’) of one female speaker: (a) no disguise, (b) lowered pitch, (c) raised pitch, (d) lowered pronunciation tempo, (e) raised pronunciation tempo, (f) American accent, (g) whisper, (h) pinched nostrils, (i) clenched jaws.

The database was expanded to include the technical voice masking methods. The recordings modified with the pitch-shifter programme were used for it. The in-house software implementing the PSOLA (Pitch Synchronous Overlap and Add) algorithm was applied. For lowering the tone, the fundamental tone frequency was lowered by 40% for women and 30% for men. Raising the fundamental tone frequency was done by 50% for women and 60% for men. This resulted in a heavy distortion of the voice, but still did not lead to a degradation of speech intelligibility (checked with listening tests). The next voice masking techniques chosen were GSM 6.10 coding and AMR coding (bit rate 12.20 kbit/s). Both techniques are examples of unintentional masking and are based on the Algebraic- Code-Excited Linear Predictive method. Once the specific recordings had been selected, an audio file was assembled, which was used as a listening test at a later stage.

The test consisted of 30 examples. Each example contained 2 recordings. In the first recording, the listeners heard a natural voice, followed by a second recording about 3 seconds apart, containing one of the selected voice masking methods. It was up to the listeners to determine whether the exact same speaker was present in both recordings or whether the speaker was different. To facilitate the responses, an answer sheet was developed which required a 'yes' or 'no' mark next to each example, as well as stating the age, gender and whether the listener had a musical training. There was a cue between examples to move on to the next example and a pause of 3-4 seconds. When editing the audio material, the sound levels were aligned, so that they were equal. The next step was to prepare a presentation. An already finished sound file was added in the background and the presentation was made in such a way that, as the example changed, its number was shown on the computer screen. Everything was done to prevent the listener from stopping the test or returning to previous examples.

The results of the subjective tests presented in the paper were compared with the objective tests previously carried out using GMM (Gaussian Mixture Models). The GMM system [17,18], which is currently one of the most common and effective likelihood functions, was used for speaker verification. The GMM classifier achieves good performance in limited situations.

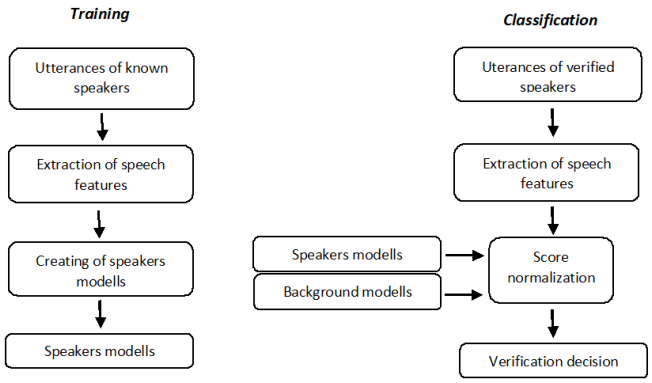


Fig. 3. Scheme of the speaker verification system.

Figure 3 shows a schematic of the training and classification phase. The training started with speech feature extraction procedures. The most standard speech parameterisation methods were used: pre-emphasis, windowing with a Hamming function (with a length of 23 ms) and extraction of Mel Frequency Cepstral Coefficients (MFCC) vectors [19]. The MFCC coefficients were then centred by subtracting the cepstral mean vector (CMS) and reducing the proportion of slowly varying splicing noise. 15 filters were use in mel scale transitions. The dynamic information was incorporated into the feature vectors, using the first and second derivatives (their polynomial approximations). Finally, the speaker verification step was performed using GMM. Due to the limited size of the test database, the maximum likelihood method with 24 unimodal Gaussian distribution was applied.

IV. Results and Discussion

Table II presents the masking methods tested with their labels used further on in the paper, as well as the information on the method of masking. Mainly the natural, intentional phonation and deformation masking methods were tested. Phonetic and prosodic methods were found to be less important in the subjective tests - it is difficult to deceive the listener with them. In addition to natural methods, the technical methods of voice masking were tested for comparison. These included two, mostly non-intentional masking techniques (GSM and AMR coding) and two intentional techniques related to shifting the fundamental frequency of the laryngeal tone (using pitch-shifter software).

Figure 4 shows the percent success rate of the correct speaker verification by the listeners. It should be noted that 'whispering' is the most difficult technique among the natural voice masking techniques from the point of view of automatic methods, as indicated by the previous studies [2,13]. The results shown in Figure 1 indicate that this was no longer such a big problem for the listeners and they did very well in recognising speakers masking their voice in this way (correct recognition at 85%). Listeners did relatively well with most of the natural techniques tested (obtaining the lowest score of 58.3% for the "clenched teeth" technique). The best masking effect was obtained for technical methods using pitch-shifter software. For these

Table II
Applied Natural and Technical Voice Disguise Techniques

Disguise technique	Marking	Type of technique
Whisper	WH	Natural, phonation
Raised pitch	RP	Natural, phonation
Lowered pitch	LP	Natural, phonation
Pinched nostrils	PN	Natural, deformation
Clenched jaws	CJ	Natural, deformation
AMR	AMR	Technical, non-deliberate
GSM	GSM	Technical, non-deliberate
Pitch-shifter lowered pitch	PSL	Technical, deliberate
Pitch-shifter raised pitch	PSR	Technical, deliberate

techniques, only 45% correct responses were obtained when the speaker's tone was lowered and even 34% when the tone was raised.

Table IV presents the detailed results of the speaker's voice recognition performance for each group of the listeners. The total number of the listeners was 40, of whom 23 were male and 17 female. Six men and five women had a musical background. The previous research conducted on various aspects of voice recognition (e.g. emotion recognition in the voice [16]) has shown that both the gender of the listeners and their musical education can be important factors influencing the outcome of listening tests. Musical education was understood to be the completion of at least a first-level music school. Of course, there are people with high musical abilities without musical education, but it is difficult to find another objective criterion to indicate this.

The results collected in Table IV clearly showed the great impact of having a musical education on the listeners. Both women and men with such education were much more successful in identifying speakers. Those without such knowledge only in isolated cases scored as well as the listeners with such knowledge.

Listeners were also asked about their age. The listening group consisted of people with normal hearing, aged between 20 and 50 years. In the study group, there was no effect of age on the speaker's voice recognition performance.

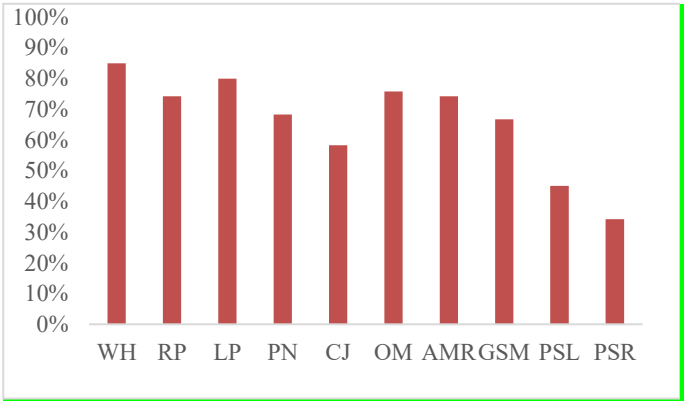


Fig. 4. Effectiveness of subjective correct voice verification for selected voice disguise techniques.

The results obtained for the listening subjective tests of the speaker voice verification were compared to the previously performed objective tests with the automatic voice verification using MFCC parameterisation and GMM classification. The comparison refers only to the natural voice masking techniques. The tests were performed on the same acoustic database containing the natural voice masking techniques as described previously. The results, including the EER (Equal Error Rate) for the automatic method and the total error (total false acceptance and false rejection) for the subjective listening method are summarised in Table III. Note that the automatic system was most easily deceived by phonation methods (whisper or raised pitch). These masking methods did not result in such significant errors in the listeners' recognition of the speaker's voice. The listeners found the deformation techniques (clenched jaws, pinched nostrils) much more problematic than the phonation techniques.

TABLE III
COMPARISON OF SPEAKER VERIFICATION FOR SELECTED NATURAL VOICE DISGUISE TECHNIQUES FOR AUTOMATIC AND SUBJECTIVE TESTS

Disguise technique	Equal Error Rate for automatic speaker verification	Error for subjective speaker verification
Whisper	38.3 %	15.0 %
Raised pitch	58.3 %	25.8 %
Lowered pitch	7.3 %	20.0 %
Pinched nostrils	16.9 %	31.7 %
Clenched jaws	14.7 %	41.7 %

V. CONCLUSIONS

In the paper, the results of the subjective tests of the speaker voice recognition performed for the selected natural (whisper, raised pitch, lowered pitch, pinched nostrils and clenched jaws) and technical (AMR, GSM, Pitch-shifer raising and lowering the fundamental frequency) voice masking techniques were presented.

The previous research indicates that, of the natural voice masking techniques, the phonation methods such as whisper and raised pitch are the most misleading to the automatic speaker recognition systems. In the subjective listening tests, of the natural techniques, it was not the phonation techniques, but the deformation techniques, such as pinched nostrils and clenched jaws that showed the most significant error in the speaker recognition. Also, the differences in the recognition performance between the natural methods themselves are not as significant as in the case of the automatic voice verification. As it was to be expected, the use of technical methods allows the personal characteristics of the speaker's voice to be masked much better than the natural methods allow.

It was found that the people with musical training were significantly better at detecting attempts to mask a speaker's voice than the people without such training. This indicates the relevance of using trained subjects, after the appropriate training, demonstrating the right aptitude for phonoscopic testing and expertise used in forensic science.

Due to the relatively small listening group, the results of the research presented in this paper are still of a preliminary character and are planned to be extended. The study examined the natural intentional methods of voice masking, and it is planned to extend the tests to the non-intentional methods related to changes in the speaker's state, such as emotional state, fatigue, illness or external factors. In addition, it is planned to extend previously performed objective studies using GMM classification [13] to newer methods such as deep neural networks (DNNs).

TABLE IV
DETAILED RESULTS OF THE SPEAKER'S VOICE RECOGNITION PERFORMANCE FOR EACH GROUP OF THE LISTENERS

	Voice disguise techniques	Women without musical education	Men without musical education	Women with musical education	Men with musical education	Together
Natural voice disguise	Whisper	80.6 %	78.4 %	93.3 %	100 %	85.0 %
	Raised pitch	58.3 %	76.5 %	93.3 %	83.3 %	74.2 %
	Lowered pitch	61.1 %	84.3 %	93.3 %	94.4 %	80.0 %
	Pinched nostrils	63.9 %	58.8 %	93.3 %	83.3 %	68.3 %
	Clenched jaws	72.2 %	47.1 %	80.0 %	77.8 %	58.3 %
	Objects in mouth	36.1%	72.6 %	80.0%	88.9 %	75.8 %
Technical voice disguise	AMR	55.6 %	62.8 %	93.3 %	88.9 %	74.2 %
	GSM	44.4 %	70.6 %	66.7 %	66.7 %	66.7 %
	Pitch-shifter – lowering	75.0 %	43.1 %	46.7 %	50.0 %	45.0 %
	Pitch-shifter - rising	61.1%	43.1 %	13.3 %	22.2 %	34.2 %

REFERENCES

[1] F. Alegre, G. Soldi, N. Evans, B. Fauve, J. Liu, "Evasion and Obfuscation in Speaker Recognition Surveillance and Forensics", Proc. International Conference on Biometrics and Forensics (IWBF), IEEE, 2014. http://dx.doi.org/10.1109/IWBF.2014.6914244

[2] P. Staroniewicz, "Effect of the deliberate and non-deliberate natural voice disguise on speaker recognition performance", Acoustics, acoustoelectronics and electrical engineering/ed. Franciszek Witos, Gliwice, Wydawnictwo Politechniki Śląskiej, 2021. pp. 312-325. (Monografia - Politechnika Śląska; nr 888), 2021. https://dx.doi.org/10.34918/80139

[3] M. Farrus, "Voice Disguise in Automatic Speaker Recognition", ACM Computing Surveys, Vol. 51, No. 4, Article 68, July 2018. https://doi.org/10.1016/j.forsciint.2007.05.019

[4] I. Krzosek-Piwowarczyk, O. Komosa, W. Maciejko, „Kryminalistyczna identyfikacja mówcy maskującego głos", Problemy Kryminalistyki 280 (2) 2013 39-52.

[5] S. S. Kajarekar, H. Bratt, E. Shriberg, R . de Leon, "A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition", Proc. Speaker and Language Recognition Workshop, 2006, IEEE Odyssey 2006. https://doi.org/10.1109/ODYSSEY.2006.248123

[6] H. J. Kunzel, J. Gonzales-Rodriguez, J. Ortega-Garcia, "Effect of voice disguise on the performance of a forensic automatic speaker recognition system", Proc. IEEE Odyssey – The Speaker and Language Recognition Workshop, 2004.

[7] P. Perrot, G. Aversano, G. Chollet, "Voice disguise and automatic detection: review and perspectives", Progress in nonlinear speech processing, pp. 101-117, (ed.): Springer 2007. https://doi.org/10.1007/978-3-540-71505-4_7

[8] C. Zhang, T. Tan, "Voice disguise and automatic speaker recognition", Forensic Science International 175 (2008) 118-122. https://doi.org/10.1016/j.forsciint.2007.05.019

[9] R. D. Rodman, M. S.Powell, "Computer Recognition of Speakers Who Disguise Their Voice", Proc. of the International Conference on Signal Processing Applications and Technology 2000 (ICSPAT 2000) Dallas, TX, October 2000. https://api.semanticscholar.org/CorpusID:16980245

[10] W. Majewski, P. Staroniewicz, "Imitation of Target Speakers by Different Types of Impersonators", Analysis of Verbal and Nonverbal Communication and Enactment, Springer LNCS vol. 6800, 104-112, 2011. https://doi.org/10.1007/978-3-642-25775-9_10

[11] P. Staroniewicz. "Test of robustness of GMM speaker verification in VoIP telephony", Archives of Acoustics 2007, vol.32, nr 4, suppl. pp.187-192. https://acoustics.ippt.pan.pl/index.php/aa/article/download/1408/1225

[12] J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller, "Applying Multiple Classifiers and Non-Linear Dynamic Features for Detecting Sleepiness from Speech", Neurocomputing 84, pp. 65-75, 2012. https://doi.org/10.1016/j.neucom.2011.12.021

[13] P. Staroniewicz, "Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition", Proc. of Joint Conference - Acoustics, Ustka 2018, pp.1-4 (ed.): IEEE 2018. https://doi.org/10.1109/ACOUSTICS.2018.8502372

[14] S. Brachmański "Speech signal noise reduction in forensic audio analysis", Proc. 56 OSA 15-18.09.2009, pp.135-140, 2009.

[15] A. B. Dobrucki, S. Brachmański "Test signals used in electroacoustics and speech technology", Proc. Signal processing, algorithms, architectures, arrangements and applications, SPA 2017, 20-22.09.2017 IEEE 2017. https://doi.org/10.23919/SPA.2017.8166828

[16] P. Staroniewicz, "Considering basic emotional state information in speaker verification", Proc. 4th International Conference on Biometrics and Forensics (IWBF) IEEE 2016. https://doi.org/10.1109/IWBF.2016.7449689

[17] D. A. Reynolds, R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. Speech and Audio Proc. Vol.3(1), pp. 72-83. 1995. https://doi.org/10.1109/89.365379

[18] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker Verification Using Adapted Gausian Mixture Models", Digital Signal Processing 10, pp. 19-41, 2000. https://doi.org/10.1006/dspr.1999.0361

[19] F. Bimbot, J. F. Bonastre, C. Fredouille, G.Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, D. A. Reynolds, "A tutorial on text-independent speaker verification", EURASIP J. Appl. Signal Process., vol.2004, pp.430-451, 2004. https://doi.org/10.1155/S1110865704310024