Piotr STARONIEWICZ[1]

17. EFFECT OF DELIBERATE AND NON-DELIBERATE NATURAL VOICE DISGUISE ON SPEAKER RECOGNITION PERFORMANCE

17.1. Introduction

The recognition of the speaker from his or her speech is widely used in modern biometric systems, such as secure access control, transaction authentication and forensics. Regardless of whether a voice is recognised by a human being or, as it is becoming increasingly common, by an automatic system, its measurable parameters are subject to change, which can lead to an incorrect recognition. These changes in voice parameters are due to the fact that, unlike biometric identifiers such as DNA, a fingerprint or an iris, which are highly permanent over time, the human voice undergoes significant changes due to ageing, emotions and many other deliberate and unintended factors. These factors, regardless of their origin, are a deviation, transformation or distortion of a 'normal voice' and are therefore treated as voice disguises [1, 4, 5, 6, 8, 13].

It is possible to classify voice disguise according to two independent dimensions: deliberate – non-deliberate and technical – natural (sometimes also called electronic – non-electronic) [3, 10, 12]. The types of disguises according to the above classification are presented in Table 1.

[1] Department of Acoustics, Multimedia and Signal Processing, Wroclaw University of Science and Technology, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland, piotr.staroniewicz@pwr.edu.pl

Table 1

Types of Voice Disguises

Type of Disguise	Technical	Natural
Deliberate	Usage of a device or computer software which digitally processes the speech signal to modify its parameters (e.g. changing fundamental frequency of the speaker's voice etc.).	Deliberate speaker's manipulation on speech production organs leading to a significant change of pronunciation naturalness.
Non-deliberate	Distortions dependent on the properties of the telecommunication channel (i.e. the frequency band in telephony, applied speech encoding technique, handset variations etc.).	Changes in voice parameters caused by a temporary influence of illness, drugs, alcohol, the speaker's physical condition, emotional state or even aging.

As the deliberate one, we take a voice disguise that depends on the will of the speaker. In order to hide his or her identity or to impersonate another person [7], the speaker deliberately changes the parameters of his voice. To achieve this, the speaker can use technical tools (e.g. electronic devices or software applications) to convert his or her voice. It is therefore a process of physical transformation of the characteristics of the voice, leaving, of course, the semantic information of the speech preserved. Voice conversion technology is currently used in many applications such as speech synthesis, language learning and entertainment, and can also be used to try and deceive the speaker recognition system, e.g. by pitch or vocal tract modification or scaling [6]. As far as non-technical methods (we call them 'natural') are concerned, we should mention here a wide range of methods which allow the speaker, without additional tools or devices, to modify the prosodic, phonetic and phonation features of speech or to deform parts of the speech organ (e.g. by clenching teeth or plugging the nose) [10, 12].

The recognised parameters of the speaker's voice are also subject to changes that are not controlled by the speaker, and we are then dealing with non-deliberate voice disguises. Some non-deliberate disguises have technical (or electronic) reasons. These are mainly distortions and a degradation of speech resulting from the telecommunication channel, i.e. telephone transmission, microphones used, etc. [11] Non-deliberate natural disguises are mostly caused by changes that affect the normal functioning of our body, such as: aging, illness (e.g. hoarseness, breathing problems), emotional state, fatigue and drowsiness or finally intoxication (e.g. alcohol, drugs) [3].

In the presented work, the influence of natural, i.e. not resulting from transmission conditions or the use of additional electronic devices or methods, voice disguise on the effectiveness of voice recognition will be discussed.

17.2. Natural voice disguises

Table 2

Different Types of Deliberate and Non-deliberate Natural Voice Disguises

Natural voice disguise	Types	Examples
Non-deliberate	Ageing	Anatomical and physiological changes throughout life.
	Intoxication	Speech under the influence of alcohol or drugs.
	Emotional state	Speech under emotional arousal, e.g. one of the so-called Big Six: anger, sadness, happiness, fear, disgust, surprise.
	Illness	Hoarseness, laryngitis, vocal cord nodules, etc.
	Change in condition	Sleepiness, fatigue, impact of external conditions (i.e. temperature, vibration, acceleration, loud or annoying noise etc.).
Deliberate	Phonation	Pitch changes (raised or lowered), 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.
	Deformation	Objects in or over the mouth, pinched nostrils, lips protrusion, holding of the tongue.

Selected types of natural voice disguises that affect the parameters of the produced speech and cause difficulties in recognizing the speaker correctly are listed in Table 2.

One of the natural voice disguises that we are always dealing with, for example, in the case of biometric systems, is the ageing effect. Continuous anatomical and physiological changes due to ageing cause in the produced speech: voice tremor, slower articulation, laryngeal tension, air loss or imprecise consonants [3]. Research into the impact of the effects of ageing is difficult and sometimes debatable, as it requires repeated recording of the voices of a group of the same speakers over a long period of time.

The effect of intoxication (i.e. alcohol or drugs) is also confirmed by comparative studies of sober and intoxicated people. It is difficult to speak of universal relationships, because the influence of intoxication depends largely on personal characteristics or the amount of alcohol in the blood, but it is most often manifested in the speech produced through some of its parameters, such as the slowing of speaking rate and changes of the fundamental frequency distribution.

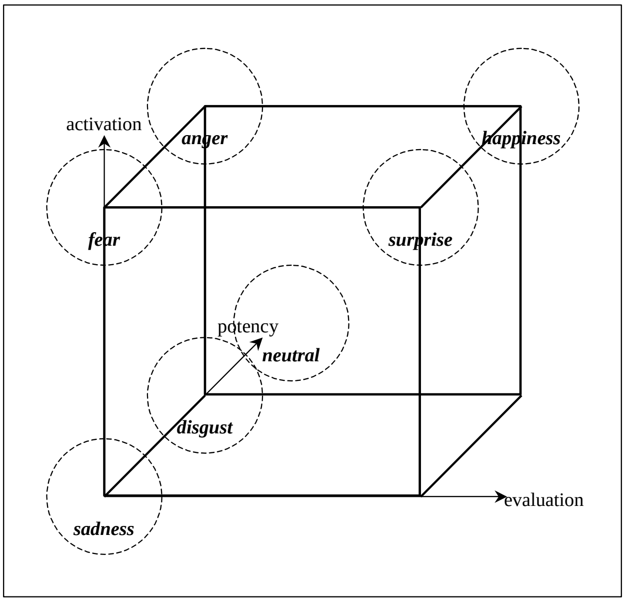

Fig. 1. Six basic emotional states and neutral state on three dimensions of emotional space
Rys. 1. Sześć podstawowych stanów emocjonalnych oraz stan neutralny w trójwymiarowej przestrzeni emocjonalnej

Another important unintended natural factor that can significantly affect the effectiveness of a speaker's recognition is the emotional variability of the speaker's voice. The main source of problems and complications during work on emotional speech is the lack of precise definition of emotions and rules for their classification. The literature describes them as emotional dimensions such as potentiality, activation and evaluation or as discrete concepts such as anger, fear, joy. Discreet concepts that are easy to understand by speakers are usually chosen when recording simulated emotions. Despite the lack of restrictions on the number of emotions, there is a "general agreement" on the so-called "big six" - six basic emotional states that are understandable in all cultures: anger, sadness, happiness, fear, disgust, surprise and neutral state [9].

The above figure (Fig. 1) shows in a simplified way the location of six basic emotional states in the space of three dimensions: potentiality, activation and evaluation.

Diseases affecting the speaker's voice may have a significant impact on the effectiveness of the speaker's recognition. Hoarseness, laryngitis or vocal cord nodules may significantly affect the spectral characteristics of the speech produced. This phenomenon is also used in acoustic diagnostics to detect certain health conditions.

The last non-deliberate natural voice disguise category proposed in the taxonomy presented in Table 2 is ‘change in the speaker’s condition’. This category refers to factors not previously mentioned which may also affect the voice. It includes such extreme physical and perception effects as loud noise (Lombard effect), pain, high or low temperature, vibration, acceleration, physical exhaustion, etc. Tiredness and sleepiness, which also affect the speaker's voice, are also included in this category. Automatic voice-based sleepiness detection is made for people such as traffic dispatchers, pilots or drivers, and is assessed, for example, according to the Stanford sleepiness scale (SSS).

There is a wide range of non-electronic intentional (deliberate) voice disguise methods. Their purpose may be to imitate the voice of another person for stage entertaining purposes. In this case the impersonator often tries to imitate the body language and non-verbal expressions of the copied person. In terms of speech parameters, the impersonator attempts to exaggerate the most prominent features. Another objective that is much more common (e.g. with criminals) is to try to mask or distort the parameters of the speech signal in order to prevent the correct verification of the speaker's identity. It is difficult to systematise deliberate natural disguise methods. The literature proposes a division into four main categories: phonation, phonemic, prosodic and deformation techniques.

The phonation techniques include all the methods which involve abnormal glottal activity such as: raised pitch (falsetto), lowered pitch, creaky voice (glottal fry), whisper etc.

The phonemic techniques refer to abnormal allophone use, which appears when the speaker uses a dialect or a foreign accent, feigns a speech defect or mimics someone.

The prosodic techniques concern the intonation issues. Stress placement, intonation changes, speech segments lengthening or shortening or speech tempo are the exemplary natural prosodic disguise techniques.

When the forced physical changes in the vocal tract take place, we can encounter the deformation techniques. The typical deformation techniques are: various objects put by the speaker in or over the mouth, tongue holding, pulled cheeks, pinched nostrils or clenched jaws.

It is very difficult to carry out meaningful tests of the effectiveness of automatic speaker recognition for all types of non-deliberate natural voice disguises. The recording of speakers necessary for testing certain types, such as intoxication, is difficult to carry out under controlled conditions and sometimes unethical. Similarly, it can be difficult to record types of masking such as illness or ageing. They require the same group of speakers to be interviewed over a long period of time. The research presented below focuses on selected types of natural voice masking: emotional states (for six basic emotional states: anger, sadness, happiness, fear, disgust and surprise) as representatives of the natural, non-deliberate voice disguises. As the natural, deliberate disguises two techniques were selected for each of the four types of masking (e.g. phonation, phonemic, prosodic and deformation).

17.3. Speaker recognition scores for chosen natural voice disguise techniques

17.3.1. Speaker recognition for non-deliberate voice disguise with speaker's emotional state

There are three types of sound bases designed to recognise emotions: natural, forced (provoked) and simulated [9].

Natural databases are created by recording natural emotions in response to authentic situations. For example, these are statements made by pilots during the flight, journalists from accident scenes, participants in TV shows. There are not many natural databases because it is very difficult and time-consuming to obtain them. In addition, the low number of speakers, the changing acoustic conditions of the recording and the lack of control over the content of the statements mean that these bases are often not used in automatic emotion recognition systems.

The recording of a forced emotion is made by subjecting a group of people to specific stimuli that provoke a particular emotional state. These stimuli include multimedia presentations, computer games, films, stressful situations and psychotropic measures. A big disadvantage of these bases, in terms of usefulness in systems of automatic recognition of emotions, is the fact that not all recorded people have to react in the same way, with the same emotional state, to a given stimulus. In addition, the degree and strength of an emotional expression can vary greatly from speaker to speaker.

The great advantage of the third type of bases is that they are much easier to obtain. Actors, professional speakers or amateurs take part in the games. These people are asked to read a certain text with clearly defined categories of emotion. The material may consist of single words, sentences as well as longer fragments of the text. The content of a speech is usually emotionally neutral, so that speakers can use it to present several different emotional states. Unfortunately, the differences between the emotions played out and the real ones are not known. There is a risk that speakers may over-emphasise or under-emphasise the characteristics of a particular emotional state, or simulate an emotional utterance in a way that is not true and is not reflected in real emotions. This, in turn, may result in automatic emotion recognition systems that are designed on such a basis and are not highly effective for natural speech.

Table 3

Sentences with Emotionally Neutral Content Used in Database

No	Sentence in Polish	SAMPA transcription	English translation
1	Jutro pójdziemy do kina.	jutro pujdz’emy do kina	Tomorrow we’ll go to the cinema.
2	Musimy się spotkać.	mus’imI s’e~ spotkats’	We have to meet.
3	Najlepsze miejsca są już zajęte.	najlepSe miejsca s~ juZ zaje~te	The best seats are already taken.
4	Powinnaś zadzwonić wieczorem.	povinnas’ zadzvonits’ vietSorem	You should call in the evening.
5	To na pewno się uda.	to na pevno s’e~ uda	It must work out.
6	Ona koniecznie chce wygrać.	ona kon’etSn’e xtse vIgrats’	She simply must win.
7	Nie pij tyle kawy.	n’e pij tIle kavI	Don’t drink so much coffee.
8	Zasuń za sobą krzesło.	zasun’ za soba~ kSeswo	Put the chair back.
9	Dlaczego on nie wrócił.	dlatSego on n’e vruts’iw	Why hasn’t he come back.
10	Niech się pan zastanowi.	n’ex s’e~ pan zastanovi	Think about it.

Ten phonetically balanced, emotionally neutral colloquial speech sentences were used (based, among others, on the basis of attendance dictionaries of the Polish language) (Table 3). The group of speakers consisted of 13 people, 6 women and 7 men, each of whom recorded a few repetitions of 10 sentences in 7 states (6 emotions and the neutral state). In total, 2351 statements, 1168 for male and 1183 for female voices were recorded. The average duration of a single speech was about 1 second. After a preliminary assessment, the recordings of questionable or poor quality were rejected. In the end, 2118 statements were divided into a set of teaching and testing sessions to allow for later testing of automatic emotional recognition systems.

The methods of automatic recognition of emotional states use parameters related to the pitch (F0), energy, spectral waveforms and time. The examples of pitch functions for female voice are presented in Fig. 2. As far as anger is concerned, regardless of the research material (both spontaneous and acting speech), high F0, wide range of voice pitch, high energy and fast tempo are characteristic for this emotion. Happiness increases the pitch and range of changes F0. Increased speech speed and intensity are also noted. Studies describing the feeling of sadness report a lowered or normal basic frequency and its narrow range of changes. A slower rate of speech is also reported. For fear, an increased average frequency F0 and an increased range of changes are recorded. For disgust, the previous studies indicate a smaller range of F0 changes.

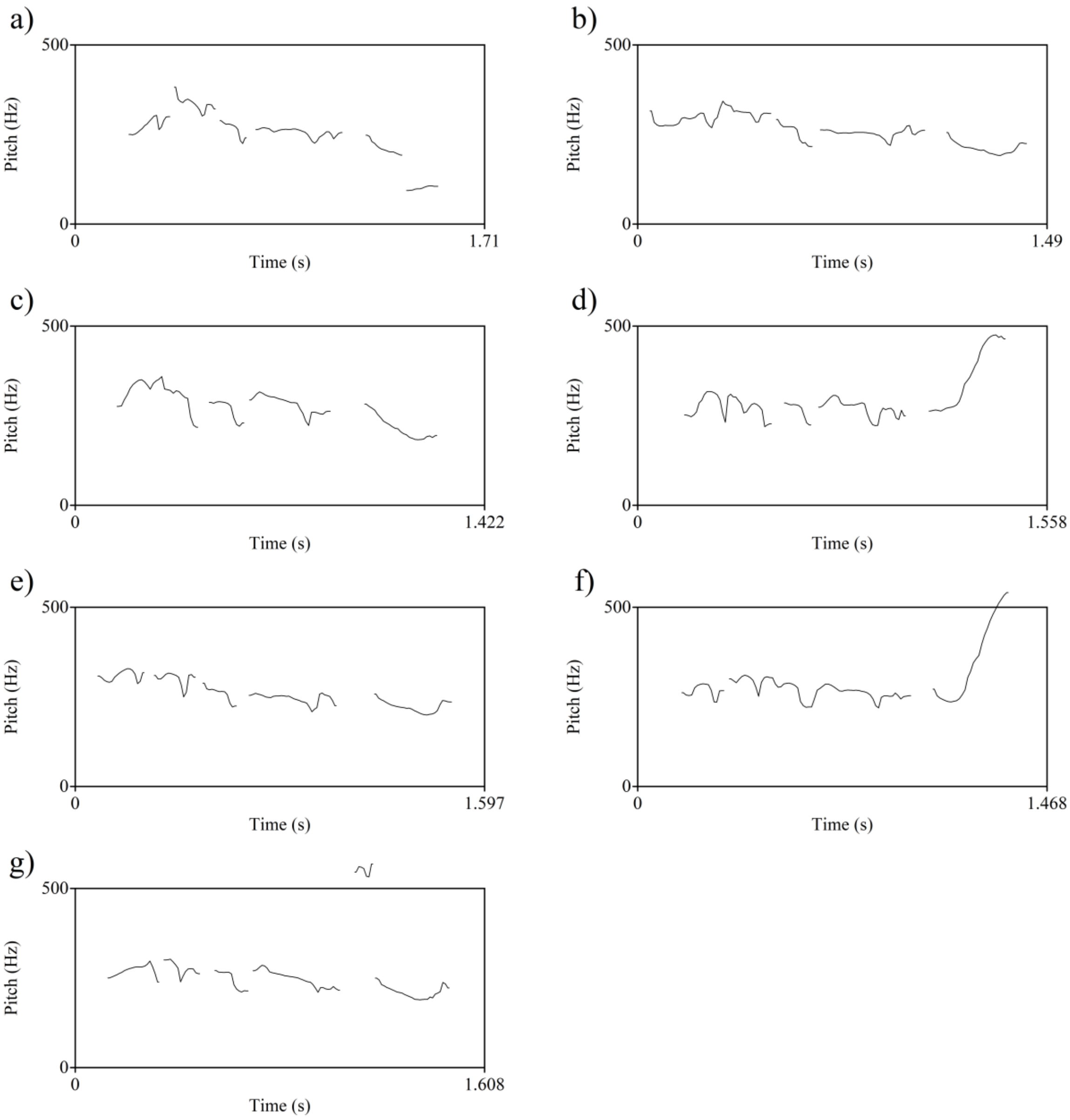

Fig. 2. Pitch functions of the same utterance (in Polish: 'Jutro pójdziemy do kina', SAMPA transcription: /jutro pujdz'emy do kina/) for the same female speaker pronounced: a) naturally (emotionally neutral state), b) happiness, c) anger, d) fear, e) sadness, f) surprise, g) disgust

Rys. 2. Przebiegi częstotliwości podstawowej tonu krtaniowego dla wypowiedzi głosu żeńskiego o tej samej treści („Jutro pójdziemy do kina", transkrypcja SAMPA: /jutro pujdz'emy do kina/): a) stan neutralny, b) radość, c) gniew, d) strach, e) smutek, f) zdziwienie, g) zdegustowanie

17.3.2. Speaker recognition for deliberate voice disguise

The database created to test the impact of deliberate voice disguise consisted of sixteen speakers - eight women and eight men. A set of 12 sentences in Polish was arranged for each person. Some of them were used in the training set and some in the testing set. Five-six sentences from a given set were used for the training set, which was one sound file and six sentences as the samples for the testing set (each sentence as a separate file). The training sets lasted about 20-30 seconds. The sentences constituting the base were selected Polish proverbs or other phonetically rich sentences selected from the Corpus base of the Polish language. Each sentence was read in a normal voice (without any modifications) and in sequence by means of specific voice modifications. Two masking (disguise) techniques were selected for each of the four main types (8 altogether):

- Phonation techniques – Lowered pitch and raised pitch.
- Prosodic techniques – Lowered pronunciation tempo and raised pronunciation tempo.
- Phonemic techniques – American accent and whisper.
- Deformation techniques – Pinched nostrils and clenched jaws.

Differences between spectral images for different deliberate disguise methods are presented in Fig. 3 and Fig. 4. Both presented spectrograms (Fig. 3) and pitch functions (Fig.4) were made for one speaker (female voice) and the utterances of the same content. Significant differences between the received spectral images of the natural speech for most of the disguise methods can be observed. This is particularly visible for techniques such as whisper (no automatic pitch detection) and raised pitch (significant increase in the mean value of F0).

17.3.3. Automatic speaker recognition system used for voice disguise tests

The tests of automatic speaker recognition were carried out in a voice recognition system using MFCC (Mel Frequency Cepstral Coefficients) parameterisation and GMM (Gaussian Mixture Models) classification, commonly regarded as one of the most effective solutions at present [2].

In order to perform parameterisation, the speech signal was first preemphasised. The purpose of the preemphasis is to highlight the higher frequencies of the speech signal spectrum that are attenuated during the articulation process. After the Hamming window was used, the Fast Fourier Transform (FFT) was determined. The FFT module was multiplied by a bank of Mel filters to smooth out and obtain an envelope of the spectrum on the auditorium scale. After the transition to dB, a discrete cosine transform was used as the final step of the parameterization procedure, giving cepstral coefficients. The cepstral coefficients were then centred, which was achieved by subtraction of the average cepstral vector and that reduced the possible influence of a slow-changing noise convoluted into a signal. Finally, the polynomial approximation of the first and second order derivatives was added to the vectors of the parameters in order to better reflect the dynamic information in the signal. Thus, the obtained vectors of the parameters were given to the classification procedure.

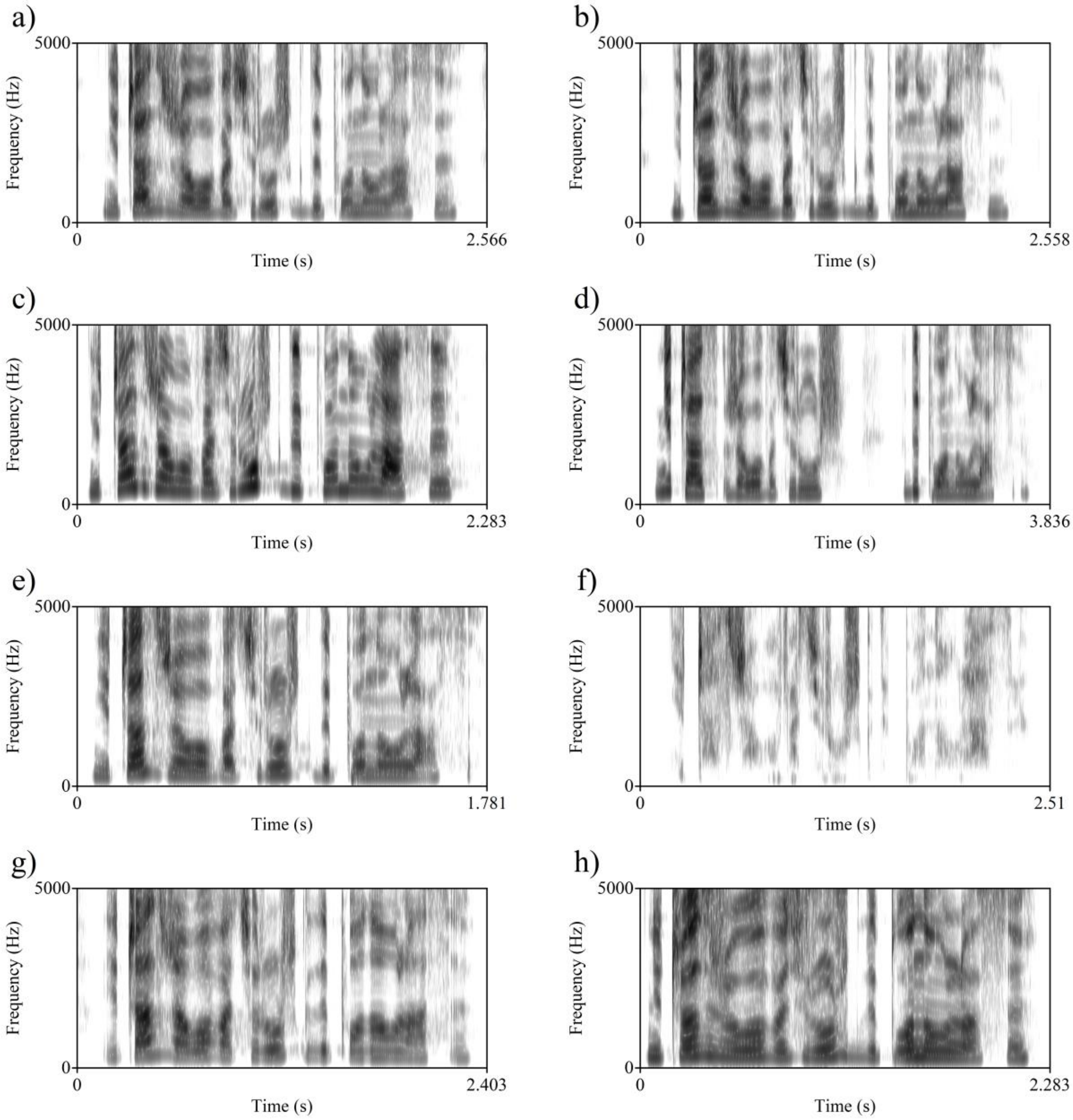

Fig. 3. Spectrograms of the same utterance (in Polish: ‘Nie czas żałować róż gdy płoną lasy’, SAMPA transcription: /n’e tSas Zawovats’ ruZ gdI pwono~ lasI/) for the same female speaker pronounced: a) naturally, b) lowered pitch, c) raised pitch, d) lowered pronunciation tempo, e) raised pronunciation tempo, f) whisper, g) pinched nostrils, h) clenched jaws

Rys. 3. Spektrogramy dla wypowiedzi głosu żeńskiego o tej samej treści („Nie czas żałować róż gdy płoną lasy”, transkrypcja SAMPA: /n’e tSas Zawovats’ ruZ gdI pwono~ lasI/): a) wypowiedź naturalna, b) obniżony ton głosu, c) podniesiony ton głosu, d) obniżone tempo wypowiedzi, e) podwyższone tempo wypowiedzi, f) szept, g) zatkany nos, h) zaciśnięte zęby

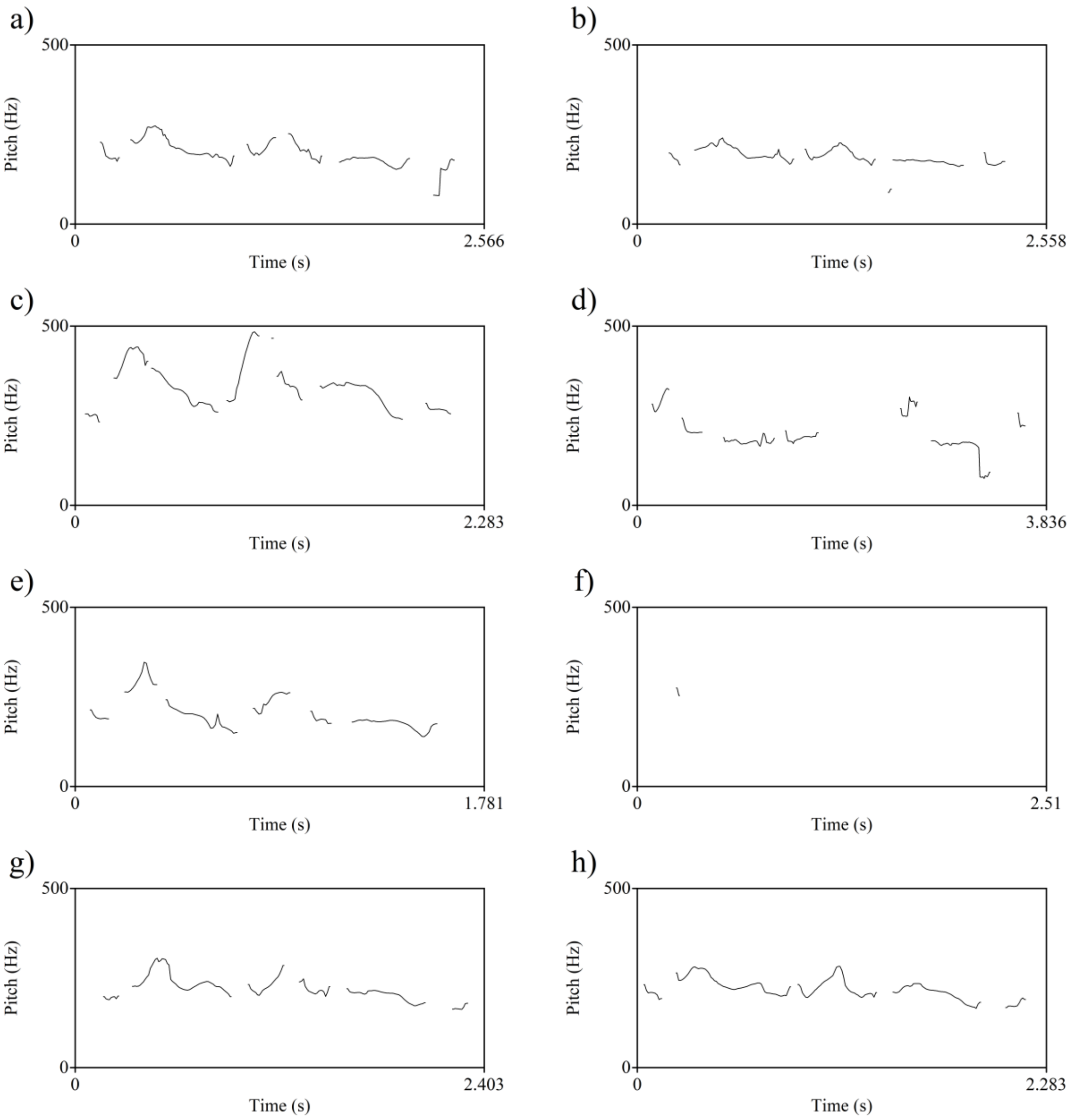

Fig. 4. Pitch functions of the same utterance (in Polish: ‘Nie czas żałować róż gdy płoną lasy’, SAMPA transcription: /n’e tSas Zawovats’ ruZ gdI pwono~ lasI/) for the same female speaker pronounced: a) naturally, b) lowered pitch, c) raised pitch, d) lowered pronunciation tempo, e) raised pronunciation tempo, f) whisper, g) pinched nostrils, h) clenched jaws

Rys. 4. Przebiegi częstotliwości podstawowej tonu krtaniowego dla wypowiedzi głosu żeńskiego o tej samej treści („Nie czas żałować róż gdy płoną lasy”, transkrypcja SAMPA: /n’e tSas Zawovats’ ruZ gdI pwono~ lasI/): a) wypowiedź naturalna, b) obniżony ton głosu, c) podniesiony ton głosu, d) obniżone tempo wypowiedzi, e) podwyższone tempo wypowiedzi, f) szept, g) zatkany nos, h) zaciśnięte zęby

As mentioned above, in the presented research, the statistical modelling was carried out using Gaussian Mixture Models. The final step of the verification process is a decision consisting of comparing the probabilities obtained from a comparison between the requested speaker models and the input speech signal with the decision threshold. The desired speaker is accepted if the probability is higher than the threshold level, otherwise it is rejected. The selection of the decision threshold is a complex problem in the process of verifying the speaker. Standardisation techniques are introduced for this purpose. In the process of verifying speakers, two types of misrecognition may occur: false acceptance (i.e. acceptance of the voice of a fraudster as the voice of one of the clients) and false rejection (rejection of the voice of a client as not belonging to him or her). Both the number of false acceptance errors and false rejection errors depend monotonically on the value of the decision threshold. A pair of these errors mark the point at which the system works. Determining the optimal working point (or equivalent decision threshold values) is a compromise between the two figures. The Equal Error Rate (EER) represents a point of work where the false acceptance and the false rejection are equal. The EER rarely corresponds to the actual point of work, but it is nevertheless a popular measure of the system's ability to separate customers and fraudsters from each other. The obtained EER values are presented in Table 4.

17.4. Results and discussion

A summary of the obtained results is presented in Table 4. The reference [9] uses a description of emotional states in a multidimensional space where it is possible to place them on the basis of dimensions such as: evaluation, activation or potentiality (Fig. 1). Activation, in particular, turns out to be a very important parameter significantly influencing the results of voice recognition. This fact is reflected in the obtained values of the EER (Equal Error Rate) (Table 4) where, apart from a good result for the neutral state (average value of about 0.7%), quite good results were obtained for sadness and disgust (about 1%). Significantly worse results were obtained for emotional states with strong activation, i.e. surprise, happiness, etc., often with EER values as high as over 2%.

As predicted, much worse results in recognising the speakers were achieved for the intended voice disguise techniques. The highest error values were obtained for the cases when the voice verification system was tested with samples of speakers with a raised pitch (EER value as high as 58.33%). This ‘falsetto disguise’ allowed for a much greater masking effect on the speaker's identity than the other tested techniques, even compared to whisper, where a very high result was also obtained (EER of 38.26%). The results of the EER for all the other deliberate voice disguise techniques did not exceed 20%. The examined prosodic techniques (i.e. lowered pronunciation tempo and raised pronunciation tempo) proved to be the least misleading for the tested voice recognition system (EER values of 6.32% and 10.49% respectively). The lowest score of EER 7.29% was obtained for lowered pitch. Such a result may be surprising, considering that the highest score was obtained for a raised pitch. Although in theory the two techniques are similar, it is much easier for speakers to raise the laryngeal tone significantly during the articulation process than to lower it.

Table 4

Equal Error Rate (EER) Results for Chosen Natural Voice Disguises

Natural voice disguise	Types	Techniques	EER results
Non-deliberate	None (emotionally neutral state)		0.70%
		Happiness	2.05%
		Anger	1.75%
		Fear	1.85%
		Sadness	0.95%
		Surprise	2.25%
		Disgust	1.25%
Deliberate	None (natural speech without any deliberate disguise)		0.00%
	Phonation	Lowered pitch	7.29%
		Raised pitch	58.33%
	Prosodic	Lowered pronunciation tempo	6.32%
		Raised pronunciation tempo	10.49%
	Phonemic	American accent	18.75%
		Whisper	38.26%
	Deformation	Pinched nostrils	16.88%
		Clenched jaws	14.68%

The results presented above refer to classical automatic voice verification systems using the parameterisation of MFCC and GMM classification. Automatic speaker recognition systems have recently been used more and more willingly in forensic applications, often in combination with classical phonemic-acoustic methods or even on their own. However, before they completely replace the work of forensic experts, their limitations and weaknesses should be recognised, especially with regard to voice disguise methods. Automatic methods still do not take into account the whole range of information that forensic experts analyse, such as the type and range of vocabulary used by the speaker, the structure of the speech he or she is building, the way he or she breathes, and finally the condition of the speaker.

Building a reliable model of a speaker requires the collection of speech samples of each person for all the emotional states studied with sufficient duration. Such an assumption has limited the possible tests to those based on simulated emotions. Despite the fact that the presented tests were not conducted on natural emotions, the obtained results confirmed the fact that the emotional characterisation of a speech may have a significant impact on the effectiveness of voice recognition.

The tests carried out have shown a significant impact of natural voice masking methods on the results of automatic voice recognition, especially with regard to deliberate methods. When considering the deliberate methods, it is also important to bear in mind the crucial, even tenfold in values of EER, differences between the tested techniques.

Bibliography

1. Alegre F., Soldi G., Evans N., Fauve B., Liu J.: Evasion and Obfuscation in Speaker Recognition Surveillance and Forensics, [in:] Proc. International Conference on Biometrics and Forensics (IWBF) (ed.): IEEE, 2014.
2. Bimbot F., Bonastre J.F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz D., Reynolds D.A.: A tutorial on text-independent speaker verification, EURASIP J. Appl. Signal Process., vol. 2004, 430-451.
3. Farrus M.: Voice Disguise in Automatic Speaker Recognition, ACM Computing Surveys, Vol. 51, No. 4, Article 68, July 2018.
4. Kajarekar S.S., Bratt H., Shriberg E., de Leon R.: A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition, [in:] Proc. Speaker and Language Recognition Workshop, 2006 (ed.): IEEE Odyssey 2006.
5. Kunzel H.J., Gonzales-Rodriguez J., Ortega-Garcia J.: Effect of voice disguise on the performance of a forensic automatic speaker recognition system, [in:] Proc. IEEE Odyssey – The Speaker and Language Recognition Workshop, 2004.
6. Krzosek-Piwowarczyk I., Komosa O., Maciejko W.: Kryminalistyczna identyfikacja mówcy maskującego głos, Problemy Kryminalistyki 280 (2) 2013 39-52.
7. Majewski W., Staroniewicz P.: Imitation of Target Speakers by Different Types of Impersonators, [in:] Analysis of Verbal and Nonverbal Communication and Enactment, (ed.): Springer LNCS vol. 6800, 104-112, 2011.
8. Perrot P., Aversano G., Chollet G.: Voice disguise and automatic detection: review and perspectives, [in:] Progress in nonlinear speech processing, 101-117, (ed.): Springer 2007.
9. Staroniewicz P.: Considering basic emotional state information in speaker verification, [in:] Proc. 4th International Conference on Biometrics and Forensics (IWBF) IEEE 2016.
10. Staroniewicz P.: Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition, [in:] Proc. of Joint Conf. - Acoustics, Ustka 2018, 1-4 (ed.): IEEE 2018.
11. Staroniewicz P.: Test of robustness of GMM speaker verification in VoIP telephony, Archives of Acoustics 2007, vol. 32, nr 4, suppl. 187-192.
12. Rodman R.D., Powell M.S.: Computer Recognition of Speakers Who Disguise Their Voice, [in:] Proc. of the International Conference on Signal Processing Applications and Technology 2000 (ed.): (ICSPAT 2000) Dallas, TX, October 2000.
13. Zhang C., Tan T.: Voice disguise and automatic speaker recognition, Forensic Science International 175 (2008) 118-122.