Andrzej P. DOBROWOLSKI, Ewelina MAJDA[1]

Military University of Technology (1)

Application of homomorphic methods of speech signal processing in speakers recognition system

Abstract. *The paper presents the problem of automatic speaker recognition system. Automatic recognition of speaker is a process designed to determine, whether a particular statement belongs to the speaker. The speech signal is a carrier of both physiological and behavioral features. No two individuals sound identical, because their vocal tract shapes, larynx sizes, and other parts of their voice production organs are different. This paper contains a description of the speech signal analysis algorithms, designed based on normalized real cepstrum. The authors have attempted to select the optimal set of parameters describing the speaker. The study has primarily focused on assessing applicability of the cepstral analysis of speech signal. In addition results of experiments are presented using a PCA method.*

Streszczenie. *W prezentowanym referacie poruszono problematykę systemu rozpoznawania mówcy. Automatyczne rozpoznawanie mówcy jest procesem mającym na celu określenie, czy dana wypowiedź należy do określonego mówcy. Sygnał mowy jest nośnikiem zarówno cech fizjologicznych, jak i behawioralnych. Nie ma dwóch identycznie brzmiących osób, ze względu na fakt występujących różnic w budowie krtani, głośni, traktu wokalnego oraz innych organów artykulacyjnych u każdego człowieka. Praca zawiera opis algorytmów analizy sygnału mowy opracowanych w oparciu o rzeczywiste cepstrum. Dzięki tej technice multiplikatywny związek pobudzenia i traktu głosowego zastąpiony zostaje związkiem addytywnym, co znacznie upraszcza separację obu składników. Autorzy podjęli się próby wyboru optymalnego zestawu cech charakteryzujących danego mówcę. Badania koncentrowały się przede wszystkim na ocenie użyteczności analizy cepstralnej sygnału mowy. Dodatkowo uzyskane wyniki eksperymentów przedstawiono przy pomocy metody PCA.* ***(Zastosowanie homomorficznych metod przetwarzania sygnału mowy w systemach rozpoznawania mówcy)***

Keywords: speech signal, feature extraction, cepstral analysis.
Słowa kluczowe: sygnał mowy, ekstrakcja cech, analiza cepstralna.

Introduction

Automatic recognition of individuals is used in all systems that provide services or proprietary information, especially when a high degree of security of these systems is required. The growing demand for such systems not only contributes to the development of broadly defined biometric techniques, but also telecommunication and the Internet. An unquestionable advantage of the recognition systems based on individual characteristics of voice, is the fact that these attributes cannot be lost, or forgotten. In everyday contacts identification of people based on their voice characteristics is an easy task. Universality and naturalness of this phenomenon causes that in general we do not realize what qualities of speech are taken into account in this natural process and only an attempt to transfer this function to technical equipment makes us aware about the full range of intractable problems. It does not mean that sensitivity and accuracy of our senses (especially hearing) is unattainable for technical devices - on the contrary, any physical quantity characterizing the speech signal can be determined much more accurately than do our natural analyzers. However, people can make better use (at least till now) of the full information contained in the voice signal. It is so, because in humans the sense of hearing and the nervous system are highly specialized and trained in receiving and analyzing the speech signals, but, unfortunately, the processes associated therewith are not fully understood. Aside from information about the content of speech, any verbal statement carries also information related to the internal structure of its source. The speech signal is a carrier of both physiological and behavioral features, so it is one of the biometric parameters that ensure a high degree of differentiation. These inter-individual differences reflect individual characteristics of the speaker's voice. They result from differences in the construction of the organ of articulation (voice path) in different people, habits acquired on learning to speak and the degree of mastery of a given language. In practice, there are less or more links between a biometric characteristic and such features of a speaker such as his/her sex, age, health, mood, background, or the native language. For these reasons, voice analysis is the subject of studies undertaken by specialists in many fields, but in spite of decades of research, the speech signal should be considered very complex and difficult to exhaustive (i.e. analogous to the analysis performed by the sense of hearing) interpretation.

ASR Procedure

Automatic recognition of speakers, referred to as auto-sensing of voice, is a process of executing a series of decision rules on measurable features of a speech signal to determine whether a statement may be linked to a particular speaker or a set of speakers. Any such system can be grouped according to various criteria, depending on the type of decision associated with recognition, the assumptions related to the openness of a set of involved people and depending on the text of speech. There are two fundamentally different procedures: identification and verification of a speaker. Identification of a speaker is a process of decision-making, which involves confirmation of identity of the speaker, and is based only upon the characteristics of the speech (without declaring his/her identity). On the other hand, verification of a speaker is a process of decision-making, using characteristics of the speech signal to determine whether the speaker of the speech is, in fact, the person whose identity he/she declares. In general, the procedure for identification of persons can be divided into three phases (Fig. 1). The pre-processing block is responsible for receiving the signal from the microphone and its initial processing, involving also quality enhancement of the signal. The second stage involves analysis of the speech signal, in order to obtain parameters carrying information about the individual characteristics of the voice of the speaker, regardless on the speech content. The final stage of classification is based on similarity of obtained parameters of the signal sample to their previously acquired references (in the so-called teaching process) for particular persons. The outcome of the system is a binary decision either to recognize identity of the speaker, or to reject it. A simplified diagram of the speaker recognition procedure is shown in Figure 1 [1].

For any speaker recognition system, the most critical step is to arrange an adequate set of parameters which would allow carrying out the recognition procedure. The basic requirement for such a set of voices is to ensure discrimination between different individuals based on values and repeatability of the parameters for various phrases expressed by the same person. A better parameter is considered the one, the value of which is exactly reproducible (or very similar) for various expressions of the same speaker and relatively different for expressions of other speakers. In order to extract relevant parameters from a speech signal, the signal has to be parameterized, which is critical for effectiveness and reaction rate of the entire speaker recognition system. On dealing with a vast number of various parameters one should seek for some method of selecting the optimal (most discriminating) set of parameters describing the signal.

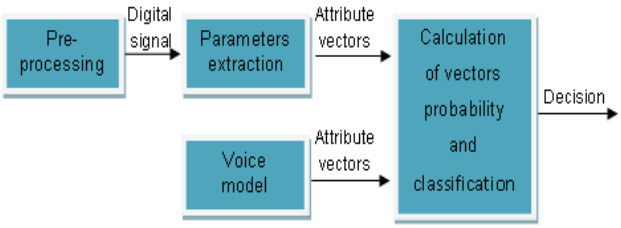


Fig. 1. Diagram of speakers identification procedure

Method for describing speech signal

The primary and basic form, in which the speech signal is present, is its temporal form. This form contains all the elements necessary for analysis and recognition, but they are rather inconvenient. This is due to a large redundancy of information contained in this form. By reference of the "human" way of speech signal analysis, a significant number of computer methods is based on spectral analysis, which is a fundamental and routinely used method to describe this kind of signals, as well as a starting point for more advanced methods for parameterization.

The glottis and the voice path, including in particular the mouth and nose as well as the tongue and lips are involved in the process of generating the speech signal. Vocal folds, commonly referred to as the vocal cords, and more specifically their edges, commonly referred to as the vocal ligaments play the essential role in the speaking (and breathing) process. The gap between these ligaments is called the vocal slit and together with the adjacent vocal folds it forms the glottis. During quiet breathing and during the articulation of voiceless speech elements the ligaments are separated and air flows freely through the vocal slit. On pronouncing voiced sounds, the ligaments, as a result of nerve impulses reaching them, are alternately closing and opening under the pressure of air. The gap between the vocal folds that may be seen with the naked eye is an optical illusion caused by inertia of human sight, which is not able to register phases of opening and closing the glottis, quickly following one another. Observation in slow-motion shows that the ligaments constrict rhythmically until the glottis is completely closed. The laryngeal sound generation process is sometimes referred to as phonation (vocalization).

The frequency of opening and closing the ligaments that determines pitch of voice depends on their length, thickness and tension (and these depend on gender and age). Pitch of voice and more specifically its fundamental frequency changes during speech due to natural intonation. In case of male voice the average frequency is 100-130 Hz, while the average frequency of female voice reaches a value of 200-260 Hz [2]. Fundamental frequency in speech varies from 60 to 200 Hz for males and from 180 to 400 Hz for females. In case of vocal the range of fundamental frequency, referred to as the scale of voice, is much broader. According to the nomenclature of the basic classification of music, there are three basic male voices: bass (73-294 Hz), baritone (98-392 Hz) and tenor (123-494 Hz) and three female voices: alt (165-652 Hz), mezzo-soprano (195-784 Hz) and soprano (247-900 Hz).

The air flow pumped by the glottis is modified during passage through the voice path, the amplitude-frequency characteristics of which is determined by several peaks called controls. Frequencies of these peaks are instantaneous resonant frequencies of the voice path resulting from the current state of the articulation process. Assuming that for quasi-stationary fragments of speech the voice path is a linear system, constant in time, the speech signal could be represented as a combination of pulse stimulation generated in the glottis and the impulse response of the voice path.

Since the Fourier transform of evenly sliding Dirac pulses

$$\mathrm{III}(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_0) \tag{1}$$

is also the sum of Dirac pulses

$$\omega_0 \cdot \sum_{m=-\infty}^{\infty} \delta(\omega - m\omega_0); \qquad \omega_0 = \frac{2\pi}{T_0} \tag{2}$$

then the spectrum of a laryngeal sound is a series of pulses, wherein the spacing in frequency is $F_0 = 1/T_0$, provided that the pulse spacing in time is T_0.

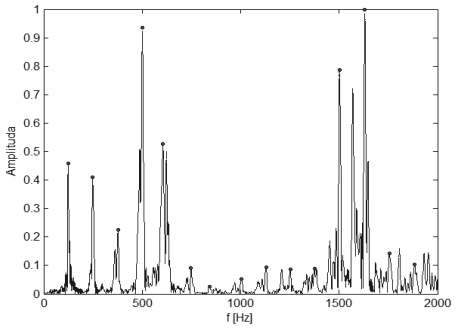
Adoption of a linear model of the voice path, in which stimulation interweaves with the impulse response of a filter in the time domain, allows – by reference to (1) and (2) - to conclude that the spectrum of the voiced fragments of speech is the product of over a distance of F_0 on the axis of the Dirac pulse (an idealized spectrum of pulses emitted by the glottis) and transmittance of voice path. In theoretical considerations, the finite when the glottis is open during phonation is taken into account in the form of additional term in the voice path transmittance. During the practical tests of the speech signal, parts of the signal are cut out using the windowing function, the spectrum of which is interweaved with the spectral Dirac's impulses and, consequently, a spectrum of the window, duplicated on each impulse, appears in the place of the expected Dirac's pulse, as illustrated in Figure 2.

Figure 3 shows the amplitude spectrum of „a" phone uttered by a man and a woman. As it could be easily noted, and as strictly confirmed by a preliminary study, by using the spectral amplitude it is easier to distinguish between spoken sounds than between the speakers. Important information that discriminate speakers is the fundamental frequency of sounds, which - as is obvious when one compares expression of a man and a woman - could possibly serve as a differentiating parameter. However, for comparison between two men, for example, this information is of little utility, especially since the fundamental frequency fluctuates according to intonation of a sentence.

Figures 2 and 3 clearly show periodicity of the spectrum resulting from laryngeal sound pulses, so one can calculate the inverse Fourier transform from the module of the spectrum and on the basis thereof determine the basic period of laryngeal stimulation. However, as the amplitude of the signal is modulated by the function of transferring the voice path, it is preferable to calculate first the logarithm of the module of the spectrum, and then subject it to a reverse

Fourier transform. This way, the multiplicative relationship between stimulation and the voice path is replaced by an additive relationship which greatly simplifies the subsequent separation of the two components. The presented reasoning leads directly to homomorphic processing methods, in particular to the concept of cepstrum.

a)



b)

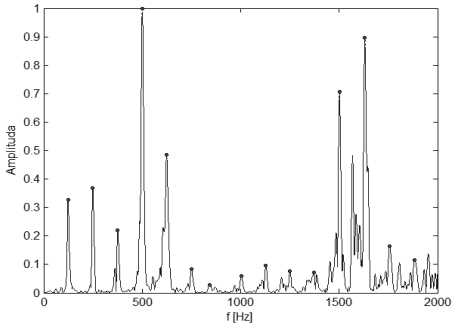
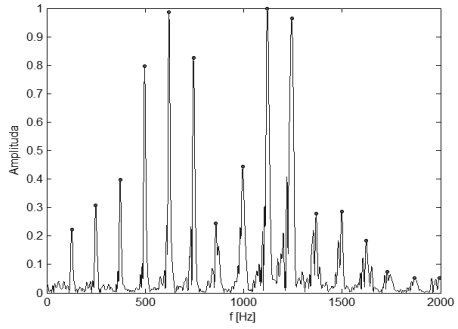


Fig. 2. The spectrum of *e* phone, male voice a) rectangular window, b) Hanning window

a)



b)

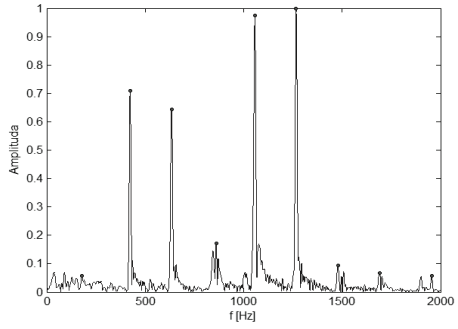


Fig. 3. The spectrum of *a* phone: a) male voice b) female voice; a Hanning window was applied

Principles of the cepstral analysis

One of the specific parameterization methods is the cepstral analysis that is based on so-called homomorphic technique. A complex cepstrum is defined as follows:

$$c_z(t) = \mathcal{F}^{-1}\left\{ \ln\left(\mathcal{F}\{x(t)\}\right) \right\} \tag{3}$$

As in the case of speech signal, the basic information is contained in the amplitude of its spectrum, and calculation of the complex logarithm is associated with complications arising from the necessity of ensuring continuity of phases, in practice one usually determines the so called real cepstrum, formally defined as follows:

$$c(t) = \mathcal{F}^{-1}\left\{ \ln\left(\left|\mathcal{F}\{x(t)\}\right|\right) \right\} \tag{4}$$

which for discrete signals, may be reduced to the following form:

$$c(n) = IDFT\left(\ln\left(\left| DFT\left(x(n)\cdot w(n)\right)\right|\right)\right) \tag{5}$$

and finally

$$\begin{aligned} c(n) &= \frac{1}{N}\sum_{m=0}^{N-1} C(m)\mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} = \\ &= \frac{1}{N}\sum_{m=0}^{N-1} \ln\left(\left|\sum_{n=0}^{N-1} x(n)\ w(n)\mathrm{e}^{-\mathrm{j}2\pi\frac{mn}{N}}\right|\right)\mathrm{e}^{\mathrm{j}2\pi\frac{mn}{N}} \end{aligned} \tag{6}$$

Due to the periodicity of the Fourier transform kernel, the logarithm of the amplitude spectrum module $C(m)$ is periodic and simultaneously it meets the equation

$$C(-m) = C(N-m) \tag{7}$$

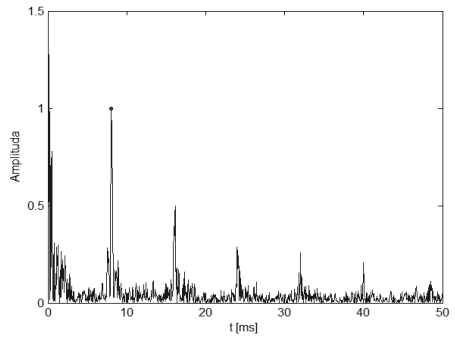
Hence, it is an even function (symmetry with respect to the axis 0y), and therefore, only cosine (even) functions appear in its expansion. As a result, it is meaningless, whether in the last step one uses a simple or inverse Fourier transformation, or simply a cosine transformation. This allows for easy interpretation of the real cepstrum as a spectral logarithm scale amplitude. On analyzing the amplitude spectrum of the speech signal one can easily seen that it is composed of a rapidly changing factor arising from the stimulation and a slowly changing one that modulates the amplitude of successive pulses resulting from the stimulation. Interpretation of the spectrum amplitude logarithm is similar, but the slowly changing component is not multiplied by the amplitudes of individual pulses from stimulation. Instead, it adds to them. Calculation of the spectrum of such signals shows that the low frequency waveforms associated with the transmittance of the voice path are close to zero on the pseudo-time axis, and pulses associated with laryngeal sound begin roughly around the laryngeal signal period and repeat periodically.

Real cepstra that correspond to spectrums on Figure 3 are shown on Figure 4. Information related to the voice path transmittance is focused around zero time, and therefore, one should look for concise information on *what is being said* in this area. On the other hand, for the time period above the laryngeal sound, information about what is being said is minimized, and the only legible information is that concerning the laryngeal sound. Because the laryngeal sound is closely connected to anatomy of the larynx and

glottis, so it is also a good carrier of individual information. In case of analysis aimed at speaker recognition, the classical method for cepstral reconvolution is to remove the undesired ingredient by resetting cepstrum samples for pseudo-time around zero.

Suitability of the real cepstrum for the purpose of speaker recognition can be easily noted by visually analyzing the waveforms shown on Figure 4 - information about pronounced phone are blurred, while differences between different speakers are clearly visible.

a)



b)

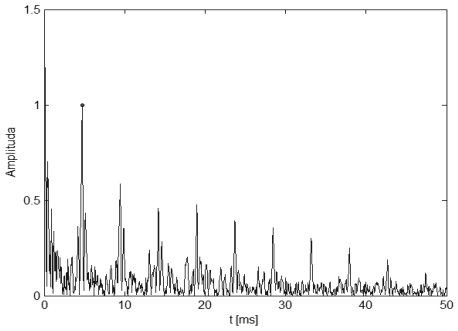


Fig. 4. Real cepstrum modules of *a* phones: a) a male voice, b) female voice, Hanning window has been applied

Result of study

Promising results of preliminary experiments using cepstral analysis of the speech signal allows extending studies of the speech signal based on a normalized real cepstrum. Therefore, it has been attempted to extract a set of 10 distinctive cepstral features from each time slice of voided speech. Due to the fact that important information related to the speaker is contained only in the so-called. voiced parts of speech, the recorded phonetic material included phones *a*, *e*, *i*, *o*, and *u*, repeated 3 times by each of the participants. The study group has consisted of 5 men and 5 women. The recordings were made under room conditions with a universal desktop MT383 microphone, a computer's sound card and Matlab software. The signals were sampled at a frequency of 22,050 Hz and amplitude resolution was 16 bits.

Amplitude spectra have been determined for the gathered input data, using *fast Fourier transformation* (FFT) and Matlab software. A base 2 FFT algorithm is a very efficient procedure for determining the discrete Fourier transform (DFT), provided that the number of the input signal is a power of 2. Therefore, in order to be able to use calculation capability of FFT, one should provide a number of input samples equal to an integral power of 2 for each of the analyzed signals. A 65536-point FFT has been adopted, assuming that such number of samples would allow for sufficiently dense graininess of the resultant spectrum. In order to minimize the leak effect, a Henning window has been applied, as defined by the formula:

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right) \tag{8}$$

Only the amplitude spectrum of a signal that carries information useful for the automatic speaker recognition systems has been used in the analysis. Furthermore, each analyzed spectrum has been normalized and its projected section has been restricted to one half of the sampling frequency due to symmetricalness of DFT. The next step involved calculation of the real cepstrum, using the formula (6).

The fundamental frequency, being an inverse of the first maximum in the cepstrum and the values on *n*-1 successive maxima have been chosen as characteristic features of the speaker's voice. Since *n* = *9*, than 8 successive maxima of the normalized real cepstrum have been covered by the analysis. The sets of cepstral features have been averaged for each speaker, based on a single recorded expression consisting of 5 vowels. Additionally the sets have been completed by the standard deviation of the basic value. The complete set of distinctive features is defined by the following functions:

$$\begin{cases} F_{av} = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} F_j, \quad F_j = \dfrac{1}{N_j T_p} \\ \sigma = \sqrt{\dfrac{\displaystyle\sum_{j=1}^{N}(F_j - F_{av})^2}{N-1}} \\ c_i = \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} c_j, \qquad i = 1,2,...,8 \end{cases} \tag{9}$$

where: F_j – fundamental frequency for each subsequent analyzed segment of speech (for the same speaker), N_j – number of the sample corresponding to the first maximum of the cepstrum for each subsequent analyzed segment of speech (for the same speaker), T_p – sampling period, N – number of analyzed segments for the same speaker, F_{av} – average value of the fundamental frequency, σ – standard deviation of the fundamental frequency, c_j – value of a feature for each subsequent analyzed segment of speech (for the same speaker), c_i – values of 8 successive features for each speaker.

A 10-dimensional attribute vector, which is VoicePrint of analyzed speaker has been obtained as a result of this analysis. According to the assumptions, 3 separate attribute vectors have been obtained for each speaker. Due to the large amount of information contained in the input data, which in our case are vectors describing the characteristics of each of the participants, it has been decided to use the PCA method in order to reduce the size. Principal Component Analysis (PCA) is a statistical method specified by a linear transformation $y=\boldsymbol{W}x$ transforming a stationary stochastic process in the form of a vector x, in such a way that the space with the reduced output retains the key information about the process. In other words, PCA transforms a large amount of information contained in the mutually correlated input data into a set of statistically independent components arranged according to their validity. PCA is an example of unsupervised learning. The

task of such a system is to describe the observed data (extracted attribute vectors) on the basis only of themselves. As a result, PCA has produced two major components, which are linear functions of the original variables. The results of PCA transformation for the 10 speakers are shown on Figure 5.

On analyzing the results one can first note that PCA transformation allows the used to distinguish the gender of a speaker. Women are noticeably grouped on the left side of the plane, while men rank on the right-hand side of it.

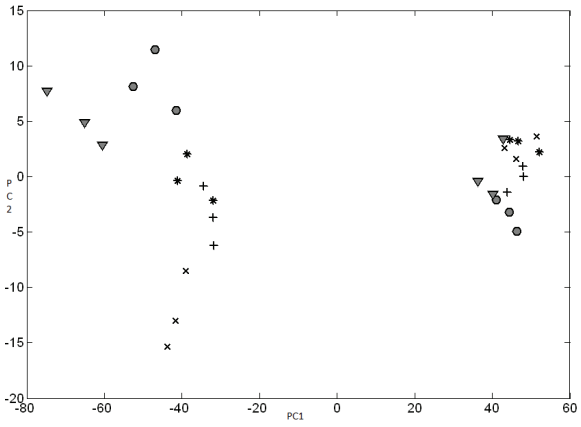


Fig. 5. The results of PCA transformation (left – female voices, right – male voices)

Figure 6 shows the enlarged parts of the surface of Figure 5 - separately for men and women (the scale of both drawings is different.)

On analyzing the results one can conclude that PCA transformation allows for separation of individual speakers from each other, which is clearly visible on the presented graphs. Furthermore, it should be noted that the average values of the cepstral parameters have been calculated for a small number of segments, which had a direct impact on the relatively large scatter of points corresponding to particular individuals.

Conclusion

The conducted experiments have allowed for a positive assessment of the usefulness of the cepstral analysis for parameterization of the speech signal. A clear separation of the individual speakers was observed after applying the PCA transformation. Practically, signals corresponding to particular persons concentrate within separate areas on the plane. However, both for men and women one can see a slight overlap of two speakers. It is caused by a small number of averages for each participant. Therefore, it could be expected that a greater number of data should even more clearly emphasize the common features and if a set of parameters within a segment may have a little correlation with the number of speakers, then after averaging the correlation should be significantly higher. At present, the authors are attempting to establish a stable criterion of segment assessment that would ensure a correct generation of features.

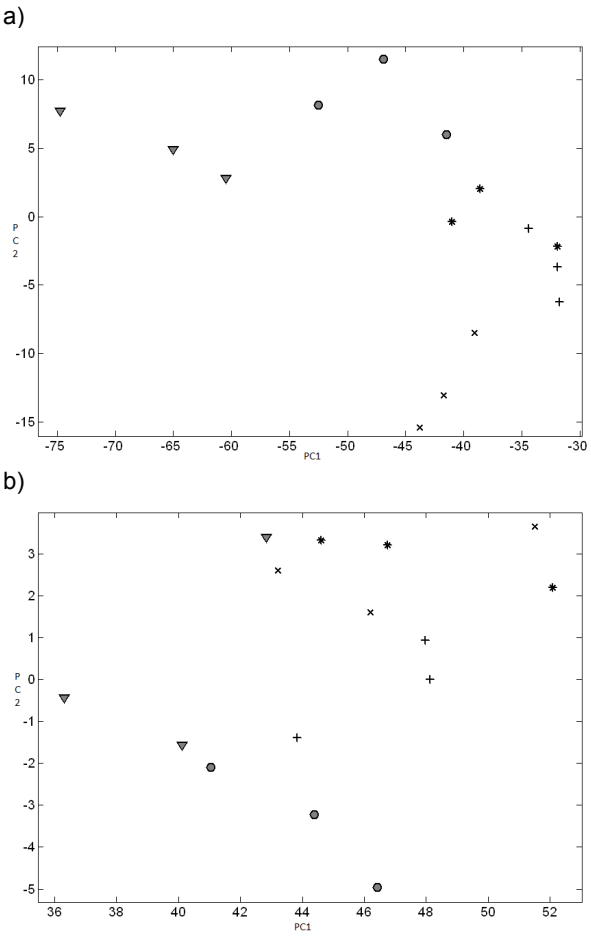


Fig. 6. Result of PCA transformation for women (a) and men (b)

"*This work is supported by the Polish Ministry of Science and Higher Education in the years 2010-2012 as a development project.*"

REFERENCES

[1] Feng L., *Speaker recognition*, Kgs, Lynby, 2004

[2] Rabiner L., Juang B. H., *Fundamentals of speech recognition*, PTR Prentice-Hall, 1993

[3] Ferras M., Leung C., Barras C., Gauvain J. L., Comparison of Speaker Adaptation Methods as Feature Extraction for SVM-Based Speaker Recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, VOL. 18, NO. 6, 2010, pp. 1366-1378

[4] Ming J., Hazen T., Glass J. R., Reynolds D. A., Robust Speaker Recognition In Noisy Conditions, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, 2007, pp. 1711-1723

[5] Zhou G., Mikhael W. B., Speaker identification based on adaptive discriminative vector quantisation, *Vision, Image and Signal Processing, IEE Proceedings*, vol. 153, no. 6, 2007, pp. 754 – 760

***Authors**: Andrzej P. Dobrowolski, Ph.D, D.Sc., Ewelina Majda, M.Sc. Military University of Technology, Faculty of Electronics, Institute of Electronic System, 2 Kaliskiego street, 00-908 Warsaw, tel. +48 22 6837534, E-mail: Andrzej.Dobrowolski@wat.edu.pl, Ewelina.Majda@wat.edu.pl.*