artificial neural networks,
speech recognition,
gender recognition

Jerzy SAS[1], Aleksander SAS[2]

GENDER RECOGNITION USING NEURAL NETWORKS AND ASR TECHNIQUES

The paper presents the simple technique of speaker gender recognition that uses MFCC features typically applied in automatic speech recognition. Artificial neural network is used as a classifier. The speech signal is first divided into 20 ms frames. For each frame, Mel-Frequency Cepstral Coefficients are extracted and the created feature vector is provided into a neural network classifier, which individually classifies each frame as male or female sample. Finally, the whole utterance is classified by selecting the class, for which the sum of corresponding neural network outputs is greater. The advantage of the method is that it can be easily combined with speech recognition, because both processes (gender recognition and speech recognition) are based on the same features. This way, no additional logic and no extra computational power is needed to extract features necessary for gender recognition. The method was experimentally evaluated using speech samples in English and in Polish. The comparison with other methods described in literature based on other feature extraction methods shows the superiority of the proposed approach, especially in cases where the recognition is carried out in noisy environment or using poor audio equipment.

1. INTRODUCTION

Automatic acquisition of attributes characterizing the person interacting with a computer makes it possible to conduct the dialog in the way that is best suited to predicted needs of the user. One of key features of (initially anonymous) interacting user is her/his gender. Automatic gender recognition can be considered as a method from the domain of biometry. In medical applications it is the source of important information of the interacting patient, especially in voice-controlled network applications, where e.g. emotional state of the patient may be important feature necessary to more precisely diagnose the patient or provide necessary aid. As pointed out in [7], detection of emotional state can be in turn aided by the knowledge about patient gender. In typical applications based on *automatic speech recognition* (ASR), speaker gender can be used to select appropriate acoustic model in order to improve ASR accuracy. In most practical applications, the gender needs to be known at the beginning of the automatically controlled dialog. Therefore, it should be possible to recognize the gender using very short sample of speech, usually a single utterance which duration is in the order of 2-4 seconds. The gender recognizer should be also insensitive to changing voice recording conditions and to background noise. In some rare cases it can be assumed that the population of speakers subject to gender recognition is known (i.e. there exist sample of speech of all speakers being recognized).

The problem of gender recognition obviously is not new and there are many attempts to automatic gender recognition described in available literature. Authors report the recognition accuracy in the range

[1]Institute of Informatics, Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland, email: jerzy.sas@pwr.wroc.pl.
[2]Institute of Computer Science, University of Wroclaw, Joliot-Curie 15, 50-383 Wroclaw, Poland (student).

80-99%. However, it should be pointed out that described tests were carried out using different data sets, so their results cannot be directly compared. Approaches to gender recognition can be divided into two groups, depending on the type of features used in the gender classifier. The first group contains methods where used features are derived from the entire energy spectrum of the acoustic signal, arranged in the way specific to ASR. In the second group we have methods based on determination of *fundamental frequency* (usually denoted by $F0$) and its formant frequencies. It was observed that due to differences in vocal tract geometry between men and women, in the population of males these frequencies are lower than in the population of women. Simple thresholding of estimated format frequency leads to relatively accurate discrimination between males and females ([8]).

The method presented in [9] identifies the fundamental frequencies and its first formant specific for vowels in training speech samples. The average fundamental and formant frequencies are calculated separately for males and females using selected speech fragments corresponding to vowels. The recognition consists in finding shorter Euclidean distance of the feature vector extracted from the utterance being recognized to vectors specific for male and female classes. Unfortunately, authors did not present the achieved accuracy of their method. Exemplary data shown in the article make it possible to estimate the accuracy at about 97%. The recognition method is simple but creation of training data is troublesome, because it requires extraction of speech samples corresponding to spoken vowels.

Similar method is described in [4]. Only fundamental frequency extracted from voiced part of speech is used in the classification process. The fundamental frequency is determined by finding the global maximum in the autocorrelation function of speech signal. The reported gender recognition accuracy is in the interval 90-95%.

Robust fundamental frequency estimation may be difficult in the case of noisy environment. In such situation, much more complex methods can be proposed to improve $F0$ estimation (as. e.g. described in [3]). The alternate way is to abandon $F0$ estimation and alternatively utilize the wider information of the speech signal spectral distribution. If the recognition algorithm is trained using samples taken in the acoustic environment similar to this one appearing during recognition, then the influence of background noise can be compensated. The article [1] presents a method of gender recognition, where Mel-Frequency Cepstral Coefficients (MFCC) are used as features. Authors consider feature vectors extracted for voiced vowels where the distinction of male/female voice is most significant. For selected vowels pronounced specifically to Hindi language, the accuracy of gender recognition reaches 100%. The authors however do not propose the method for extraction of segments corresponding to selected vowels from the continuous speech audio stream.

The combination of $F0$-based and MFCC-based classification is also possible. Authors of [2] applied a two-stage technique. At the first stage the trial is made to classify the speaker merely based on the estimated fundamental frequency. If the estimated frequency is far from the zone where male/female frequencies overlap then, it is assumed that the gender is recognized unambiguously. Otherwise the second stage is entered, where Gaussian mixture models created for male/female MFCC features distribution are used. The authors claim their method achieves 98.6% recognition accuracy.

Quite different approach is taken in [5]. The classification algorithm described there uses three features: short time energy, zero crossing rate and energy entropy. The essential classification algorithm applies the set of fuzzy rules prepared by an expert. The same features are also provided to the neural network which is used as the second component classifier. Final recognition is obtained by merging results of fuzzy logic and neural network classifiers. Obtained accuracy is at the order of 80%, so it is much lower than the accuracy obtained in other methods based of spectral features.

Although numerous approaches to gender recognition are presented in literature, relatively little attention is paid to application of *artificial neural networks* (ANN) in gender recognition. In this paper the method of gender recognition is investigated that employs artificial neural network as a classifier, where the feature vector consists of MFCC features. MFCC feature extraction technique is parametrized with a number of parameters that may affect the recognition accuracy. The aim of this work is to find optimal parameterization of MFCC feature extraction algorithm so as to maximize the recognition accuracy and to evaluate the gender recognition accuracy based on short utterances spoken in Polish.

The approach proposed and investigated in this paper makes use of *Mel-Frequency Cepstral Coefficients* (MFCC) extracted exactly in the same way as is typically being done for the sake of ASR. As a result, the same features may be used both for gender recognition and for ASR. Artificial neural network is used as a classifier. In experiments carried out, HTK package ([12]) was used for MFCC feature extraction. The advantage of the proposed method consists in making no assumptions related to the nature of voice creation process (hence there is no need to determine any parameters specific for male/female population except for MFCC features). Because the method is based on the training data acquired in specific acoustic conditions then the method can be easily adapted to particular environment of its application.

The paper is organized as follows. In Section 2, the principles of MFCC feature extraction are briefly recalled. The next section explains the complete gender recognition method based on MFCC features and ANN. The competitive method based merely on $F0$ estimation is also described there. Section 4 presents experimental results obtained by application of the presented algorithm to Polish and English speech samples. Finally, Section 5 contains conclusions and recommendations for further works.

2. MFCC FEATURES EXTRACTION

Mel Frequency Cepstral Coefficients proved to be very effective in automatic speech recognition. Our experiences in Polish speech recognition show that MFCC-based ASR accuracy is higher that in case of other competitive feature extraction techniques as *linear predictive coefficients* (LPC) or *perceptual linear prediction* (PLP) coefficients. MFCC features were successfully applied also in other speech-related problems, e.g. in automatic detection of stuttering ([11]). Therefore it can be expected that MFCC features will be also effective in relatively simpler problem of gender recognition. The features are derived from temporal spectral energy distribution of the sound signal and take into account human sound perception. MFCC features extraction begins with segmenting PCM sound stream into the sequence of equally spaced short time *frames*. The typical scheme was applied here, where frames are 20 ms long and they overlap by 10 ms. Each frame is then processed independently.

The sequence of PCM samples constituting a frame are multiplied with ordinary Hamming window. Let $(s_1, s_2, ..., s_N)$ denote the sequence of PCM samples in the frame. Corresponding Hamming-windowed samples s'_n are computed as:

$$s'_n = (0.54 - 0.46cos(\frac{2\pi(n-1)}{N-1}))s_n \tag{1}$$

In the next step, Fourier transform is applied to the sequence of samples $(s'_1, s'_2, ..., s'_N)$, resulting in estimation of energy distribution of the temporal speech signal. The frequency interval covered by the derived signal spectrum depends on PCM sampling frequency according to Nyquist criterion. To preserve conformance with used ASR system, the sampling frequency is 44.1 kHz, hence the spectrum band is limited to 22.05 kHz. Let $(a_1, a_2, ..., a_M)$ denote the magnitudes of the signal fractions corresponding to frequencies $(f_1, f_2, ..., f_M)$ resulted from Fourier analysis. The frequencies are mapped onto the Mel scale and binned according to uniformly distributed filters in Mel scale. In this way, Mel coefficients are obtained.

Mel coefficients are obtained by mapping frequency appearing the audio signal onto the Mel scale, which takes into account the specific feature of human auditory system corresponding to the ability to distinguish tones slightly differing in frequency. The distinguishing ability decreases with the tone frequency. The Mel transformation squeezes the frequency scale with increasing frequency according to the formula:

$$f_{Mel}(f) = 1127log_e(1 + f/700) \tag{2}$$

By applying Mel mapping, new energy distribution is obtained. The energy in Mel scale is then binned into the number of bins corresponding to equally spaced triangular filters in Mel scale. The acoustic

sub-band (f_{min}, f_{max}) is covered by K filters, where k-th filter is defined by the following formula:

$$T_k(f_{Mel}) = \begin{cases} 0 & \text{if } f_{Mel} < cf_k - w, \\ (f_{Mel} - cf_k + w)/w & \text{if } cf_k - w \leq f_{Mel} \leq cf_k, \\ (cf_k + w - f_{Mel})/w & \text{if } cf_k \leq f_{Mel} \leq cf_k + w, \\ 0 & \text{if } f_{Mel} > cf_k + w \end{cases}, \tag{3}$$

where $w = (f_{max} - f_{min})/K$ is the filter width, $cf_k = f_{min} + (k/K)(f_{max} - f_{min})$ is the central frequency of the k-th filter and K is the assumed number of filters in the filter bank. Mel coefficients $(m_1, m_2, ..., m_K)$ are computed as:

$$m_k = \sum_{i=1}^{M} a_i T_k(f_{Mel}(f_i)). \tag{4}$$

Finally, in order to decorrelate individual features, the log bin coefficients $log(m_k)$ obtained in the previous step are again subject to Fourier transform. In this way, cepstral coefficients $(c^{(1)}, c^{(2)}, ..., c^{(J)})$ are computed.

2.1. MFCC FEATURES ENHANCEMENTS

The process of MFCC features extraction depends on number of parameters. In particular the, following parameters affect the feature extraction process:

- the acoustic sub-band (f_{min}, f_{max}) that is covered by the filter bank,
- the number K of filters in the filter bank,
- application of spectral mean and variance normalization.

Their impact on gender recognition accuracy is analyzed in Section 4.

The normalization of cepstral features mean and variance improves ASR recognition significantly, in particular in cases of varying acoustic conditions, in which the voice is captured. Our previous experiments in ASR in Polish, using speaker-independent acoustic models showed that application of features mean and variance normalization reduces ASR word error rate relatively by approximately 25%. Similar improvement can be expected in the case of gender recognition. The normalized feature $c_t'^{(i)}$ corresponding to the original cepstral feature $c_t^{(i)}$ at the time t is calculated as:

$$c_t'^{(i)} = \frac{c_t^{(i)} - \mu_i}{\sigma_i}. \tag{5}$$

where

$$\mu_i = \frac{\sum_{t=1}^{T} c_t^{(i)}}{T}, \sigma_i = \sqrt{\frac{\sum_{t=1}^{T} (c_t^{(i)} - \mu_i)^2}{T}}. \tag{6}$$

Additionally, for the sake of ASR, the MFCC feature vector can be extended with additional components approximating first and second derivatives of original MFCC features. In the case of ASR, these additional components significantly improve speech recognition accuracy. There seems to be no rationale for inclusion of these extensions in the feature vector constructed for gender recognition. However, in the experimental evaluation, the impact of derivatives inclusion in the feature vector was also verified. The approximation of the first derivative of a feature $c_t^{(i)}$ at the t-th frame in the sequence is calculated as:

$$d_t^{(i)} = \frac{\sum_{r=1}^{R} r(c_{t+r}^{(i)} - c_{t-r}^{(i)})}{\sum_{r=1}^{R} r^2} \tag{7}$$

In the experiments described in this work, the window size R was set to 2. The approximation of the second derivative $e_t^{(i)}$ of the feature $c_t^{(i)}$ is calculated using the similar formula, which now is applied

to coefficients $d_t^{(i)}$. The complete feature vector X_t extracted from t-th frame of audio signal is finally obtained by concatenating c, d and e features:

$$X_t = (c_t^{(1)}, c_t^{(2)}, ..., c_t^{(J)}, d_t^{(1)}, d_t^{(2)}, ..., d_t^{(J)}, e_t^{(1)}, e_t^{(2)}, ..., e_t^{(J)}). \tag{8}$$

In experiments described in later sections, other methods used in ASR for feature extracted were compared with MFCC features. Two used feature types: linear predictive coefficients (LPC) and perceptual linear prediction (PLP) were compared. The methods of these features extraction is widely described in available literature, e.g. in [6]. In the experiments described here, the implementation of LPC and PLP feature extraction in HTK package ([12]) was used.

3. GENDER RECOGNITION USING MFCC AND ARTIFICIAL NEURAL NETWORKS

The gender recognition is carried out for individual utterances, typically of the duration of 3-5 seconds. Usually, the recorded utterance starts and ends with moments of silence (more precisely - ambient noise). Silence fragments obviously bring no information about the speaker gender and should be rejected. It can be observed that the amplitude of speech fragments corresponding to voiced phonemes (most informative for gender recognition) have higher amplitudes than fragments corresponding to fricatives and external noise. Therefore simple formula can be applied to reject leading/trailing silence. It can be safely assumed that signal/noise ratio in the utterances subject to recognize is at least 10dB. Hence, the initial/terminal fragment is rejected until the amplitude reaches 30% of the maximal amplitude observed in the utterance and remains on this minimal level at least for 0.05 sec. This shortest duration is selected using the observation that typical speech rate is in the interval 7-15 phonemes/sec. Let X_t denote the feature vector extracted from the t-th 20 ms frame of PCM utterance signal. In order to improve the recognition accuracy at the level of individual frames, two-sided context is considered for each frame. The context of the radius r consists of $2r$ feature vectors extracted from surrounding frames. The actual feature vector used to recognize t-th frame is obtained by concatenating $2m + 1$ successive MFCC vectors:

$$X'_t = (X_{t-r}, X_{t-r+1}, ..., X_t, ..., X_{t+r}) \tag{9}$$

Feature vectors X'_t are fed to appropriately trained ANN. The ANN used in experiments described later is a typical feedforward, fully connected neural network with a single hidden layer, trained with the back-propagation algorithm. Outputs of the ANN correspond to classes in the recognition problem (males and females). The sigmoid function was used as the neuron activation function.

ANN yields values of its two outputs $O_M^{(t)}$ and $O_F^{(t)}$ for each frame being presented to the network inputs. Average values of ANN outputs over all frames $(X_1, ..., X_T)$ are computed:

$$O_F = \frac{\sum_{t=1}^{T} O_F^{(t}}{T}, O_M = \frac{\sum_{t=1}^{T} O_M^{(t)}}{T}. \tag{10}$$

Finally this gender is recognized for which the corresponding average value is greater. The results of recognition of individual frames can be also combined using formulas other than (10), for example the products of $O_M^{(t)}$ and $O_F^{(t)}$ can be compared or the final recognition can be based just on the number of frames where $O_M^{(t)} > O_F^{(t)}$. In experiments carried out the summation formula (10) gave the best results. Details are presented in Section 4.

For the sake of comparison, the competitive gender recognition method based merely on the localization of $F0$ frequency in the audio spectrum was also implemented and tested. The method applies the Bayes classifier that utilizes the approximated probability distribution functions of $F0$ in male and female populations. The pdf's are approximated using training data. For each frame extracted from training utterances the $F0$ frequency is approximated. According to [9] it was assumed that the fundamental frequency $F0$ for both males and females comes from the interval 70-280 Hz. The $F0$ frequency is found in the following way based on Internet publication [8]:

- apply discrete cosine transform to the frame,

- find the frequency $f_x \in (70, 280)$ corresponding to the maximal amplitude in DCT results,
- test if found f_x is not a second harmonic of the true $F0$: if $f_x > 150$ Hz and the amplitude at $f_x/2$ is a local maximum then assume $F0 = f_x/2$,
- test if found f_x is not the third harmonic: if $f_x > 210$ Hz and the amplitude at $f_x/3$ is a local maximum then assume $F0 = f_x/3$,
- otherwise assume that found f_x is exactly $F0$ fundamental frequency.

Using $F0$ frequencies determined for individual frames of training utterances, histograms approximating the probability distributions $f_M(F0)$ and $f_F(F0)$ for males and females can be constructed. The experimental probability distributions obtained from the training data set described in Section 4 are shown in Fig. 1.

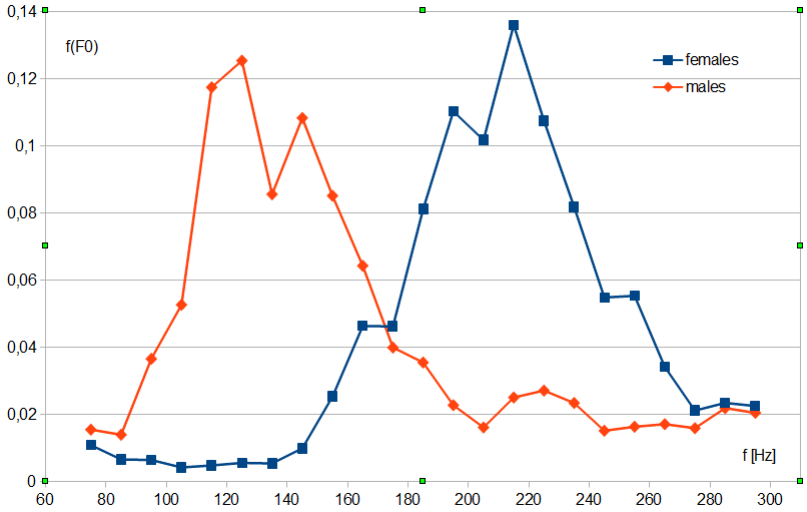

Fig. 1. F0 frequency distribution in male and female populations.

Having $F0$ pdf's estimated, the typical Bayes classifier can be applied. By assuming that prior probabilities of males and females p_M and p_F are both equal to 0.5, the decision formula at the level of an individual frame selects male if:

$$p(M|F0) = \frac{f_M(F0)p_M}{f(F0)} > p(F|F0) = \frac{f_F(F0)p_F}{f(F0)}, \tag{11}$$

where $f(F0)$ is joint probability distribution of $F0$ and can be omitted in the decision formula. Similarly as in the case of MFCC features-based method, the results of recognition at the level of individual frames need to be combined, so as to obtain the final recognition of the whole utterance. Three methods were tested, based on: a) summing obtained $p(M|F0_t)$ and $p(M|F0_t)$, b) multiplying these probabilities or c) counting winning frames for males and females. Best accuracy was obtained using the multiplicative formula. The experiments described in the next section show that despite of its simplicity, the classification based just on single feature $F0$ exhibits good gender recognition accuracy.

4. EXPERIMENTAL EVALUATION

In order to evaluate the practical usefulness of the proposed approach, the series of experiments were carried out. The first experiment was aimed at finding the optimal neural network configuration. Other feature extraction methods (LPC, PLP) were tested as well and obtained frame recognition accuracy was compared with the one obtained using MFCC features. This configuration was selected for further experiments, which resulted in the most accurate recognition at the level of individual frames. In the next experiment the final gender accuracy recognition was evaluated for three alternative methods of combining the results of individual frame recognition into the ultimate recognition based on the whole

Table 1. Frame recognition accuracy for various frame context radii and various numbers of neurons in the hidden layer - normalized MFCC features case.

Frame context radius	$h = 1.0$	$h = 2.0$	$h = 3.0$	$h = 4.0$	$h = 6.0$
0	0.63	0.67	0.70	0.72	0.73
1	0.66	0.69	0.73	0.77	0.76
2	0.71	0.71	0.73	0.75	0.75

Table 2. Frame recognition accuracy for various frame context radii and various numbers of neurons in the hidden layer - unnormalized MFCC features case.

Frame context radius	$h = 1.0$	$h = 2.0$	$h = 3.0$	$h = 4.0$	$h = 6.0$
0	0.62	0.64	0.64	0.65	0.66
1	0.62	0.64	0.68	0.71	0.71
2	0.66	0.68	0.71	0.72	0.70

utterance. Finally, the accuracy of the best variant of the method elicited in the previous experiment was compared with the popularly applied approach based solely on the $F0$ frequency.

As the experimental data, the set of speech samples in Polish was used. The training set consisted of samples uttered by 60 speakers - 30 females and 30 males. The testing set consisted of speech samples of 61 speakers, 29 of them being females. All speakers age was in the range 19 - 55 years. The utterances were recorded with 44.1 kHz sampling frequency and 16-bit resolution. The duration of utterances in training and testing sets was in the range 2 - 8 sec. The total duration of training utterances was 2721 sec. HTK package ([12]) elements (HCopy and HList programs) were used for feature extraction.

4.1. ANN OPTIMIZATION FOR INDIVIDUAL FRAMES CLASSIFICATION

This experiment was aimed at finding the neural network structure that maximizes the accuracy of recognition of individual frames. ANNs with single hidden layers were applied. Two variables were considered: a) the radius r of the frame context (as described in Section 3) and b) the number of neurons in hidden layer. The context radii $r = 0, 1, 2$ were considered. For each frame, MFCC features were extracted according to the procedure described in Section 2. For context radii 0, 1, and 2 the total number of features determining the number n_I of inputs of the neural network was:

- 39, 117 and 195 for the case, where first and second derivatives are included and
- 13, 39 and 65 otherwise.

As experiences of other ANN application indicate, the number of neurons in the hidden layer should be related to the number of neurons in input and output layer. In the application considered here, the accuracy was tested for the number of neurons in the input layer obtained by multiplying the number of ANN inputs by the multiplier $h \in \{1.0, 2.0, 3.0, 4.0, 6.0\}$ which results in the number of hidden layer neurons $1.0n_I, 2.0n_I, 3.0n_I, 4.0n_I, 6.0n_I$. The results are presented in Table 1

The data in Table 1 were obtained for MFCC features normalized, according the procedure described in Section 2-1. For comparison, in Table 2 the results for the same configurations but for MFFC features computed without normalization are presented. It can be observed that the accuracy obtained for normalized MFCC features are significantly better. This variant of feature extraction is therefore selected for further experiments.

The differences caused between various configurations defined by the context radii and varying number of neurons in the hidden layer are not significant. The best results were obtained for context radius equal to 1 and for the number of neurons in the hidden layer corresponding to the multiplier value $h = 4.0$. This configuration was used for further experiments. Quick convergence to the minimal recognition error was observed in ANN training process in the case of all configurations. The network did not exhibit the tendency to overtrain, so the training termination condition seems not to be critical. The training can be terminated after 150 training epochs. The plot in Fig. 2 presents the dependency of the mean square error of ANN outputs and the frame classification accuracy in the verification set on the number of training epochs.

Table 3. Comparison of frame recognition accuracies obtained for various feature extraction methods.

Derivatives included	LPC	PLP	MFCC
no	0.71	0.74	0.77
yes	0.70	0.75	0.77

Table 4. Final gender recognition accuracy obtained using MFCC-based and F0-based features for various frame-level results combination methods - speech samples in Polish.

Combination method	MFCC	F0
additive	0.982	0.931
multiplicative	0.971	0.939
winning class counting	0.960	0.872

Finally, using the optimal ANN configuration found in previous experiments, we evaluated the impact of inclusion of feature derivatives on the frame recognition accuracy. The frame recognition accuracy obtained using MFCC features were also compared with the accuracy obtained using LPC and PLP features. Results are presented in Table 3.

MFCC features outperform significantly LPC. They are also insignificantly better than PLP. Inclusion of feature derivatives into the feature vector did not provide any increase of accuracy in case of winning MFCC features. As the result of the experiments carried out at the individual frame recognition level, we conclude that gender recognition should be based on normalized MFCC features without their derivatives.

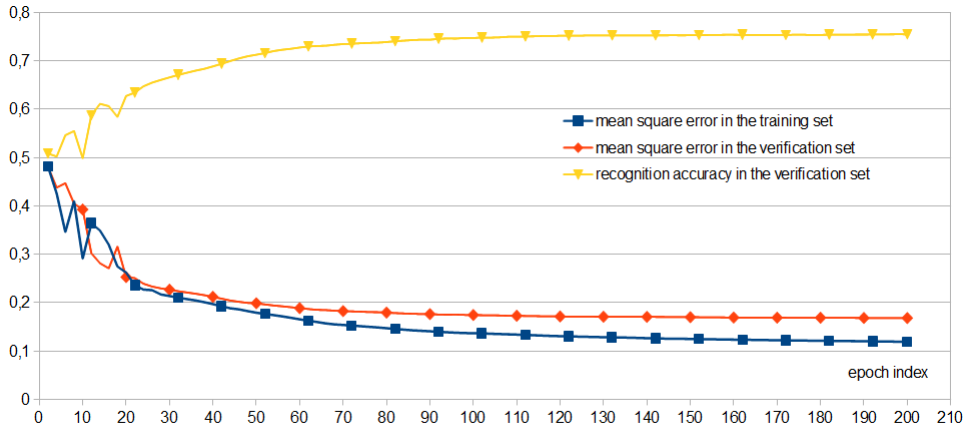

Fig. 2. Mean square error and accuracy convergence in ANN training process.

4.2. EVALUATION OF GENDER RECOGNITION METHODS FOR COMPLETE UTTERANCES

In this experiment, the accuracy of gender recognition based on the whole short utterance was evaluated. The results of individual frames recognition must be combined to provide the ultimate recognition, taking into account the whole utterance. Three combination methods were proposed in Section 3: a) summation, b) multiplication and c) winning class counting. These combination methods were applied to two competitive algorithms: MFCC-based and F0-base, described in Section 3. The obtained accuracies of gender recognition in the test utterance set are presented in Table 4.

It can be observed that application of MFCC features leads to much higher gender recognition accuracy comparing to $F0$-based method. In the case of MFCC features and ANN frame classifier, the best combination method is the additive one. In the case of Bayes classifier applied to $F0$ feature, the best results were obtained using he multiplicative combination formula.

Finally, MFCC-based and $F0$-based algorithms were applied to speech samples recorded by males and

females in English. We used the speech sample database publicly available at http://www-mmsp.ece.mcgill.ca/Documents/Data/. It contains over 1400 short utterances spoken by 24 persons, half of them being females. In contrast to our own speech samples used in test described above, the utterances in TSP database are recorded in very clear acoustic conditions without noticeable background noise and with no reverberation. The gender recognition accuracy obtained in TSP database was: 0.984 using MFCC features and 0.951 using $F0$ feature. The result is very close to the one obtained for Polish speech in the case of MFCC-based method. The performance of $F0$-based method is significantly higher in the case of clear acoustic conditions. It may indicate that $F0$-based method is more prone to acoustic quality of recorded samples than MFCC-based one.

5. CONCLUSIONS, FURTHER WORKS

The experiments described in the previous section show that MFCC features used in the artificial neural network classifier assure high accuracy of gender recognition based on short utterances. The proposed approach outperforms the competitive one, based on the analysis of the fundamental frequency $F0$, which is typically higher in females than in males. Additionally, the same MFCC features can be used in ASR. In this way, accurate and computationally inexpensive gender recognition can be applied to improve ASR accuracy be selecting appropriate acoustic model best suited for unknown speaker.

In future, the combination of both gender recognition algorithms described in Section 3 can be considered. The gender recognition algorithm provides also the confidence coefficients of the made decision. After normalizing them, so that they sum to 1.0, they can be used as the linear interpolation coefficients for acoustic model mixture in ASR, where models for males and females can be combined in proportions dependent on the unknown speaker voice similarity to typical male or female voice.

BIBLIOGRAPHY

[1] DEIV S., BHATTACHARAYA M., Automatic Gender Identification for Hindi Speech Recognition, International Journal of Computer Applications, New York, 2011, Vol. 31, No. 5, pp. 1-8.
[2] HU Y., WU D., NUCCI A., Pitch-based Gender Identification with Two-stage Classification. Security and Communication Networks, New York, 2012, Vol. 5, Issue 2, pp. 211-225.
[3] HUANG F., LEE T., Pitch Estimation in Noisy Speech Using Accumulated Peak Spectrum and Sparse Estimation Technique. IEEE Transactions on Audio, Speech and Language Processing, 2013, Vol. 21, No. 1, pp. 99-109.
[4] KARWAN J., SAEED K., A New Algorithm for Speech and Gender Recognition on the Basis of Voiced Parts of Speech. In: N. Chaki and A. Cortesi (Eds.): CISIM 2011, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 113-120.
[5] KUNJITHAPATHAM M. et al., Gender Classification in Speech Recognition using Fuzzy Logic and Neural Network. The International Arab Journal of Information Technology, September 2013, Vol. 10, No. 5, pp. 477-485.
[6] MAKHOUL J., Linear Prediction: A Tutorial Review. Proceedings of the IEEE, 1975, Vol. 63, No. 4, pp. 561-589.
[7] MOHD I., SHAHIN A., International Journal of Speech Technology, Springer, Berlin, Heidelberg, 2013, Vol. 16, pp. 133-141.
[8] PARKER R., Real-Time Kinect Player Gender Recognition using Speech Analysis, Internet publication, (http://www.radfordparker.com/papers/gender.pdf).
[9] RAKESH K., DUTTA S., SHAMA K., Gender recognition using speech processing techniques in LabView. International Journal of Advances in Engineering & Technology, 2011, Vol. 1, Issue 2, pp. 51-63.
[10] TING H., YINCHUN Y., ZHAOHUI W., Combining MFCC and Pitch to Enhance the Performance of the Gender Recognition. Proceedings of 8th Int. Conf, on Signal Processing ICSP-2006, Benjing, 2006, Vol. 1 (IEEE digital publication).
[11] WISNIEWSKI M., KUNISZYN-JOZWIAK W., Automatic detection and classification of phoneme repetitions using HTK toolkit. Journal of Medical Informatics and Technologies, 2011, Vol. 17, pp. 141-148.
[12] YOUNG S., EVERMAN G., The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2009.