DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL 68 (2017), NO1, 3–12

GMM-based speaker age and gender classification in Czech and Slovak

Jiří Přibil, [*] Anna Přibilová, [**] Jindřich Matoušek [***]

The paper describes an experiment with using the Gaussian mixture models (GMM) for automatic classification of the speaker age and gender. It analyses and compares the influence of different number of mixtures and different types of speech features used for GMM gender/age classification. Dependence of the computational complexity on the number of used mixtures is also analysed. Finally, the GMM classification accuracy is compared with the output of the conventional listening tests. The results of these objective and subjective evaluations are in correspondence.

Keywords: GMM classifier, spectral and prosodic features of speech, speaker gender and age classification

1 Introduction

Automatic identification of age and gender of a person from his/her speech has gained increasing importance in recent years because of its necessity in various commercial, medical, and forensic applications. Generally, human-computer interaction systems should adapt to different needs of the users [1]. In the same way as we adapt our speaking style to the person we are talking to [2], the automatic dialogue system might change speed of its speech synthesis according to the identified user's age [3]. The interactive voice response system can also select an appropriate background music played while waiting for an operator [4]. Assisted living and smart home systems can also be improved by automatic adaptation to different user needs. Forensic applications dealing with such cases as kidnapping or threatening calls can profit from speaker age estimation in identifying suspected criminals [5]. No less important speaker age estimation is in voice biometry [6].

A pioneering work in automatic speaker age estimation [7] used only two adult groups of 43 elderly and 43 non-elderly speakers with training data of 34 speakers and testing data of 9 speakers, Gaussian mixture models (GMMs) of 32 mixtures with diagonal covariance matrices, features of mel frequency cepstral coefficients (MFCCs), delta MFCCs, and delta power giving about 90 % identification. A much more elaborate research [2] divided speakers into 7 classes (children up to 13 years, young males/females up to 19 years, adult males/females up to 64 years, and senior males/females) with 80 speakers in each class and training data of 44 utterances per speaker. Using MFCC features and support vector machines (SVMs) with a GMM supervector the overall precision of about 75 % was achieved. The same speech database structure with the age boundaries 15, 24, and 54, the prosodic features, the formant features (mean, standard deviation), and its first derivative together with 128-mixture GMM of 12 MFCCs yielded over 60 % classification accuracy [4]. Almost the same boundaries with the exclusion of a child's voice were used in the support vector machine (SVM) classifier [3] giving up to 5 % better classification results than the baseline system. Fusion of several GMM and SVM subsystems using acoustic and prosodic features improved age classification by 4 % and gender classification by 6 % [8]. A similar approach including also voice quality features was presented in [1] for the age boundaries of 13, 19, and 54 giving again 7 classes with similar improvement. The relative improvement of mean absolute error of around 5 % was obtained by using the method of i-vectors applied on telephone conversations [5]. A pitch-range based feature set [9] achieved overall accuracy of more than 60 % in age and gender classification using the SVM classifier. The artificial neural net in combination with i-vectors [10] reached the relative improvement of 4.5 % in comparison with the support-vector regression baseline. In spite of different age boundaries between the classes, the majority of researchers use 7 classes consisting of 3 classes for each gender and a separate child's voice class as the gender is not discriminable for the speech of children before puberty.

This paper describes our experiment with using the GMM approach for automatic classification of the speaker age category. The proposed two-level GMM classifier can

[*] Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia; New Technologies for the Information Society (NTIS), Faculty of Applied Sciences, University of West Bohemia, Plze, Czech Republic, Jiri.Pribil@savba.sk

[**] Slovak University of Technology in Bratislava, Faculty of Electrical Engineering and Information Technology, Ilkovičova 3, 812 19 Bratislava, Slovak Republic, Anna.Pribilova@stuba.sk

[***] Department of Cybernetics, New Technologies for the Information Society (NTIS), Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic, jmatouse@kky.zcu.cz

DOI: 10.1515/jee-2017-0001, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X © 2017 FEI STU

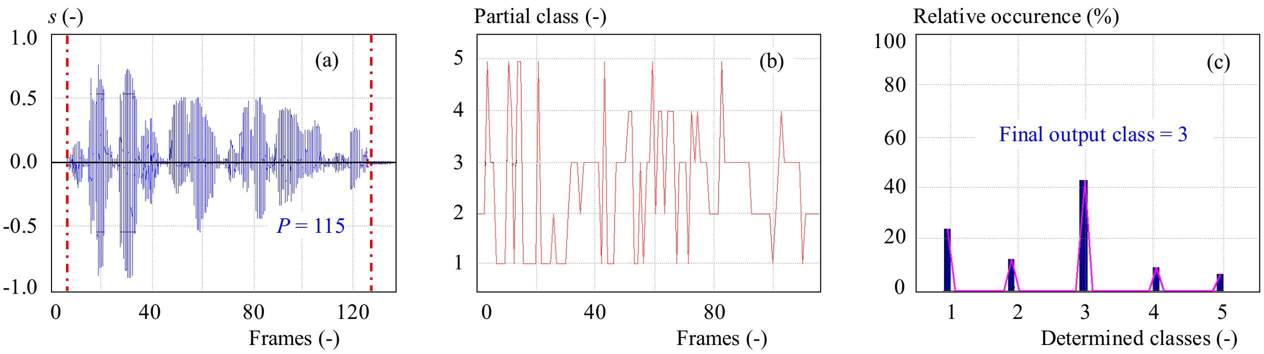

Fig. 1. Example of accumulated score determination: input sentence (male adult speaker) (a) — only P frames with the significant energy are processed, (b) — currently achieved partial classes within the working length of P frames, (c) — histogram of score values for $N = 5$ output classes with finally determined $m_{ACC} = 3$ for the whole sentence

detect four age categories (child, young, adult, senior) as well as discriminate gender for all but children's voices. For the GMM evaluation, the basic and supplementary spectral properties including the supra-segmental parameters were determined from the processed sentences and used in the input feature vectors. The paper next analyses and compares influence of the number of used GMM mixture components and different types of spectral features and supra-segmental parameters on the GMM gender/age classification accuracy, as well as the stability of the resulting gender/age speaker classification process depending on the time duration of the currently classified sentence. In addition, the comparison of the computational complexity (CPU times for the phases of the GMM creation, training, and classification) is performed. Finally, the obtained gender/age classification accuracy values are compared with the results achieved by the conventional listening test which is a standard method for evaluation of naturalness and quality of the synthetic speech [11,12].

2 Subject and method

2.1 GMM-based age classification method

The Gaussian mixture models can be defined as a linear combination of multiple Gaussian probability distribution functions of the input data vector. For GMM creation, it is necessary to determine the covariance matrix, the vector of mean values, and the weighting parameters from the input training data. The maximum likelihood function of GMM is found using the expectation-maximization iteration algorithm [13] controlled by the number of mixtures NGMIX and the number of iteration steps. Most GMM classifiers use the final score of the model given by the maximum overall probability for the corresponding class

$$m^* = \arg\max_{1 \le n \le N} score(T, n)\,, \tag{1}$$

where the $score(T, n)$ represents the probability value of the GMM classifier for the models trained for the current n-th class in the evaluation process, N is the number of output classes, and T is the input vector of the features obtained from the tested sentence. This relatively simple and robust approach cannot achieve the best recognition accuracy in all cases. Therefore, for final decision about the classified speaker age in our experiment the method was extended by the calculation of the accumulated score expressed as

$$m_{ACC} = \arg\max_{1 \le i \le M} \bigcup_{p=1}^{P} \left(m^*(i,p) \equiv i\right) \tag{2}$$

where $m^*(i,p)$ represents the value calculated by (1) for the current p-th frame, P is the number of the processed frames in the sentence, and the union operator represents the occurrence rate of the i-th class – see an example in Fig. 1.

Practical realization is carried out in the experimental two-level architecture of the GMM age classifier as shown in Fig. 2. First, the feature vectors are extracted from the input tested sentences. The speaker gender identification block uses the GMM models that were created and trained on the data of the feature vectors obtained from the sentences of the used speech corpora structured in dependence on the speaker gender (child/female/male). In the classification phase, the input feature vectors from the tested sentences are used to obtain the scores that are then sorted by the absolute size and quantized to N levels corresponding to N output classes. The second level of developed GMM classifier consists of two classification blocks with the GMM models created and trained on the data of each age class (young/adult/senior) separately for the previously determined speaker gender. The obtained individual values of $score(T, m)$ are further used for calculation of the accumulated score m_{ACC} that is next subject to M-level discrimination giving the resulting age class.

2.2 Determination of speech features

In the area of GMM-based speaker [14] and emotional voice classification [15] the most commonly used spectral features are mel frequency cepstral coefficients together with energy and prosodic parameters. The speech feature

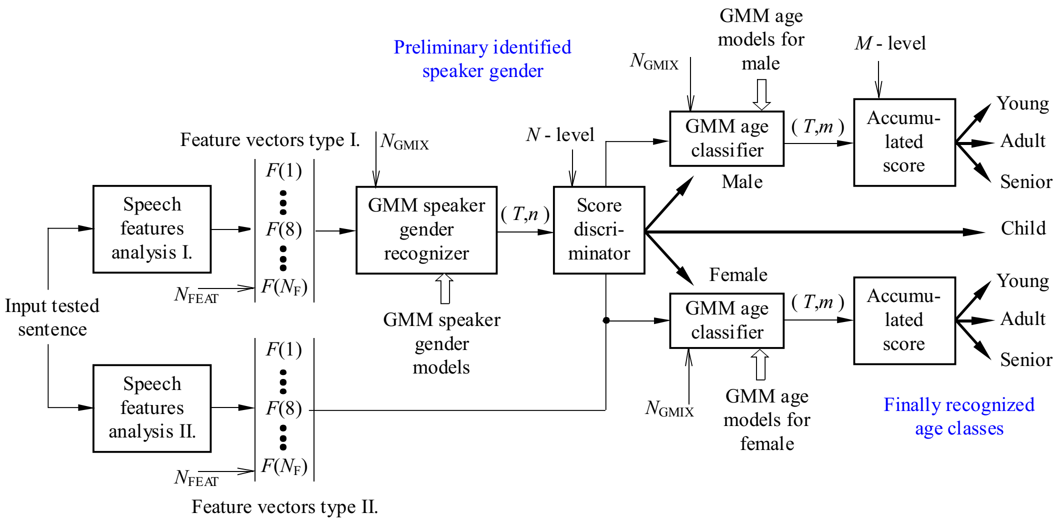

Fig. 2. Block diagram of the two-level architecture of the developed GMM age classifier

analysis in our experiment arises from the previous research [16, 17]. First, the weighted frames of the input sentence are used to calculate the energy contour from the first cepstral coefficient. The beginning and the ending frames of the sentences with the energy lower than the threshold are removed to eliminate very low amplitudes at the onset and the decay of the analysed speech signal. After this limitation, the remaining P frames are used for determination of the fundamental (pitch) frequency F0, and other speech spectral and prosodic properties. However, this approach does not give satisfactory results in itself [17]. Better final classification accuracy can be achieved for $P-2K$ representative statistical values computed from $W_{\rm AVER} = 2K+1$ values around the i-th value, where

$$K+1 \le i \le P-K\,, \tag{3}$$

evaluated for $1 \le K \le \frac{P}{2}-1$, see the detailed block diagram in Fig. 3.

The feature vectors of the length $N_{\rm FEAT}$ obtained in this way are stored in the feature databases that are subsequently used for creation and training of the GMM models as documents the summary block diagram in Fig. 4.

The basic as well as the supplementary spectral features are determined from the smoothed magnitude or power spectrum envelope obtained from the analysed speech frame. The basic spectral properties are expressed by the additional statistical parameters:

- the spectral centroid ($S_{\rm centr}$) being a centre of gravity of the power spectrum and representing an average frequency weighted by the values of the normalized energy of each frequency component in the spectrum;
- the spectral spread ($S_{\rm spread}$) parameter representing the dispersion of the power spectrum around its mean value,
- the spectral skewness ($S_{\rm skew}$) being a measure of the asymmetry of the spectrum around its mean value (for negative skewness the data are spread out more to the left; for positive skewness the data are spread out more to the right),
- the spectral kurtosis ($S_{\rm kurt}$) being a measure of peakedness or flatness of the spectral shape relative to the normal distribution for which it is 3 (or 0 after subtraction of 3),

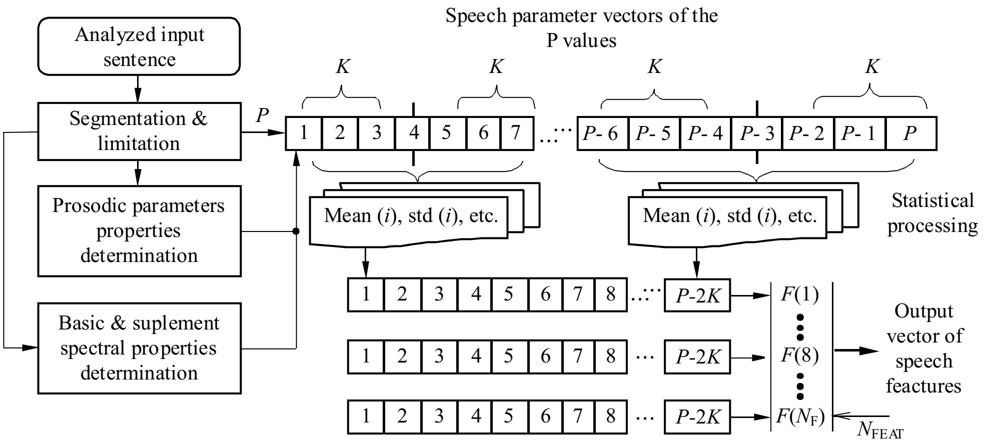
Analyzed input sentence
Speech parameter vectors of the P values
K
K
K
K
Segmentation & limitation
P
1 2 3 4 5 6 7 ... P- 6 P- 5 P- 4 P- 3 P- 2 P- 1 P
Prosodic parameters properties determination
Mean (i), std (i), etc.
Mean (i), std (i), etc.
Statistical processing
Basic & suplement spectral properties determination
1 2 3 4 5 6 7 8 ... P-2K
F(1)
F(8)
F(N_F)
Output vector of speech feactures
N_FEAT

Fig. 3. Detailed block diagram of determination of the feature vectors from the speech signal

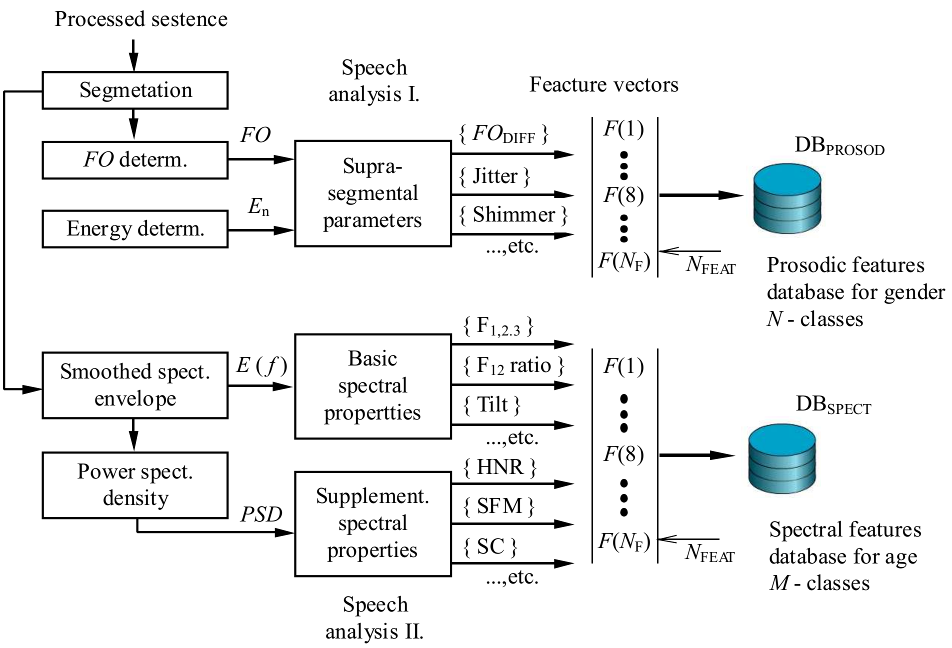

Fig. 4. Summary block diagram of the feature database creation from the speech spectral properties and supra-segmental parameters

Table 1. Structure of the Czech and Slovak speech corpus for the GMM age classification

Category	No of speakers	No of sentences	DUR_{mean} (s)	$F0_{mean}$ (Hz)	Age_{mean} (years)
Child	7	177	3.2	315	10
$Young_{(Male)}$	7	81	3.1	165	23
$Young_{(Female)}$	7	81	3.0	215	22
$Adult_{(Male)}$	6	118	3.5	125	43
$Adult_{(Female)}$	6	98	3.4	165	40
$Senior_{(Male)}$	6	109	3.4	80/150*)	60
$Senior_{(Female)}$	6	97	3.6	240/150*)	62

*) The senior speech F0 distribution has two local maxima.

- the spectral flatness (S_{flat}) providing an indication of the degree of the signal periodicity determined as a ratio of the geometric and the arithmetic mean values of the power spectrum.

Cepstral coefficients $\{c_n\}$ obtained during the process of cepstral analysis give information about spectral properties of the human vocal tract [19], so they can also be used in the feature vector for GMM classification. As the supplementary spectral features the following parameters are determined:

- the first two formant frequencies (F_1, F_2) and their ratio (F_1/F_2),
- the spectral decrease (tilt – S_{tilt}) representing the degree of fall of the magnitude or power spectrum. calculated by the least-mean-squares linear regression method,
- the harmonics-to-noise ratio (HNR) as another indicator of the overall periodicity of the speech signal by the ratio between its periodic and aperiodic components,
- the spectral entropy (SE) quantifying a degree of randomness [18] of spectral probability density represented by the normalized frequency components of the spectrum (for structured speech it is lower; for non-structured speech it is higher). The supra-segmental features are determined from the pitch frequency (F0) contour in the following steps:
- calculation of the virtual F0 (VF0) contour obtained by cubic interpolation in unvoiced parts of the speech signal,
- determination of F0_{mean} values from the VF0 contour and calculation of the linear trend (LT) by the mean square method,
- calculation of the differential microintonation signal F0_{DIFF} by subtraction of F0_{mean} and LT values from the corresponding VF0 contour and detection of zero crossings to determine F0_{ZCR},
- calculation of the absolute jitter (J_{abs}) values as the average absolute difference between consecutive pitch periods L measured in samples,
- calculation of the shimmer as the relative amplitude perturbation (AP_{rel}) from peak amplitudes detected within the n-th voiced speech frame.

3 Material, experiments and results

The main speech database consists of short declarative sentences in Czech and Slovak originating from stories and audio books uttered by professional speakers or actors in a neutral state (excluding the child category). Original MP3 stereo audio signals sampled at 44.1 kHz were converted to 16-bit PCM mono audio signals resampled to 16 kHz. The main speech corpus is divided into three basic speaker gender categories: male, female, and child. Adult voices are subdivided into another three age categories so the resulting seven age classes are stored:

a) Child – up to 12 years,

Table 2. Structure of the used feature sets for the GMM age classification

Set	Feature type	Statistical value
P0	{HNR, spectral decrease, centroid, SFM, SE, $\mathrm{F0}_{\mathrm{DIFF}}$, jitter, and shimmer}	{min, rel. max, min, mean, std, median}
P1	{spectral spread, decrease, centroid, SFM, HNR, $\mathrm{F0}_{\mathrm{DIFF}}$, $\mathrm{F0}_{\mathrm{ZCR}}$, jitter, and shimmer}	{mean, median, std, rel. max, min, max}
P2	{F_1/F_2, spectral decrease, centroid, HNR, SFM, SE, $\mathrm{F0}_{\mathrm{DIFF}}$, jitter, and shimmer}	{mean, std, median}
P3	{$F_1, F_2, F_1/F_2$, spectral decrease, HNR, SFM, SE, $\mathrm{F0}_{\mathrm{DIFF}}$, jitter, and shimmer}	{skewness, kurtosis, std, mean, median, rel. max, max}
P4	{$c_1 - c_4$, spectral decrease, centroid, SFM, SE, $\mathrm{F0}_{\mathrm{DIFF}}$, jitter, and shimmer}	{skewness, kurtosis, mean, std, median}

Table 3. Influence of the number of applied Gaussian mixtures on the speaker gender classification accuracy in (%) for the 1st level of the developed GMM classifier

Gender/N_{GMIX}	8	16	32	48	64	128
Child	89.8	96.1	94.9	96.0	94.9	94.4
Male (Y+A+S)	95.3	97.7	99.3	98.6	99.2	96.9
Female (Y+A+S)	64.0	84.4	90.4	91.7	98.1	94.8
Mean (std)	83.1 (16.7)	92.7 (7.3)	94.8 (4.5)	95.4 (3.5)	97.4 (2.2)	95.4 (1.4)

Table 4. Influence of the used window length W_{AVER} for the speech feature determination on the GMM gender classification accuracy in (%) for the 1st level of the GMM classifier

Gender/W_{AVER}	11	21	31	41
Child	90.40	94.92	94.90	93.81
Male (Y+A+S)	96.86	99.18	99.22	97.53
Female (Y+A+S)	94.29	98.15	98.07	96.39
Mean (std)	94.8 (4.5)	95.4 (3.5)	97.4 (2.2)	95.4 (1.4)

b) young male/female – up to 25 years,

c) adult male/female – up to 55 years,

d) senior male/female – over 55 years.

The frame length for speech signal analysis was set according to the speakers' mean F0 values as 24/20/12 ms for the male/female/child voices. The pitch frequency contour was created with the help of the PRAAT program [19] by the autocorrelation method with experimentally chosen pitch frequency ranges as follows: $250 \div 400$ Hz for children's voices, $55 \div 200$ Hz for male voices, and $105 \div 300$ Hz for female ones. The basic information and determined parameters of the used speech material are given in Table 1.

In accordance with the previous research results [12,16, 17] the length of the input feature vector was set to $N_{\mathrm{FEAT}} = 16$. The feature sets for GMM creation, training, and classification contain the speech features determined from the spectral envelopes (centroid, flatness, skewness, kurtosis, spread, and tilt), the supplementary spectral parameters (HNR, F_1/F_2, SE), and the suprasegmental parameters (F0, jitter, and shimmer). The basic statistical parameters — mean values and standard deviations (std) — were used as the representative values in the feature vectors for GMM evaluation of the spectral features. In the case of cepstral coefficients $c_1 \sim c_4$ the histograms of distribution were also calculated and the extended statistical parameters (skewness and kurtosis) were subsequently determined from these histograms. On the other hand, median values, ranges of values, std, and/or relative maximum and minimum were used for implementation of the supra-segmental speech parameters in the feature vectors as shown by their detailed structure in Table 2.

Two basic comparison experiments were performed within the research described in this paper. The first one was the objective automatic classification of the speaker age category using the GMM-based method together with the detailed comparison of the obtained results using the gender classifier (first level) called "1Lg", the age classifier (second level) called "2Lmf", and the whole two-level GMM architecture denoted as "12Lall". The second comparison experiment comprised the subjective speaker age classification using the conventional listening test carried out manually. For optimization of GMM classification the analysis and comparison was aimed at investigation of

- Influence of the number of applied mixtures of the Gaussian probability density functions on the resulting classification accuracy analysed for $N_{\mathrm{GMIX}} = \{8, 16, 32, 48, 64, \text{ and } 128\}$ – see the summarized values for the 1st classification level in Table 3, and graphical representation in Fig. 5 for the 2nd level (joined values for male and female).
- Influence of the chosen window length W_{AVER} for computing the representative statistical values used in the input feature vectors on the resulting classification accuracy analysed for $K = \{5, 10, 15, 20\}$ – see the numerical results for both classification levels in Tables 4 and 5.
- Influence of different types of speech features (basic and supplementary spectral, and prosodic) in the input vector on the GMM classification accuracy evaluated for the sets P0$\sim$P4 – see summary mean values in

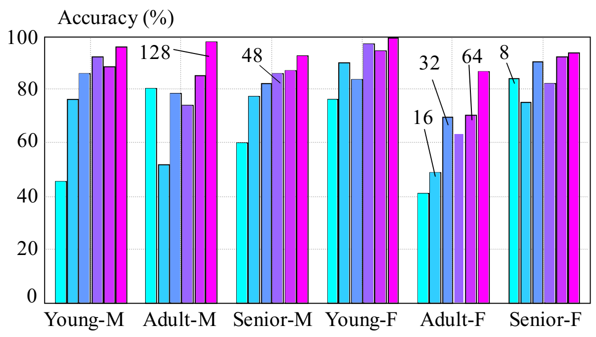

Fig. 5. Influence of the number of applied Gaussian mixtures on the speaker age classification accuracy for the 2nd level (joined values for males and females)

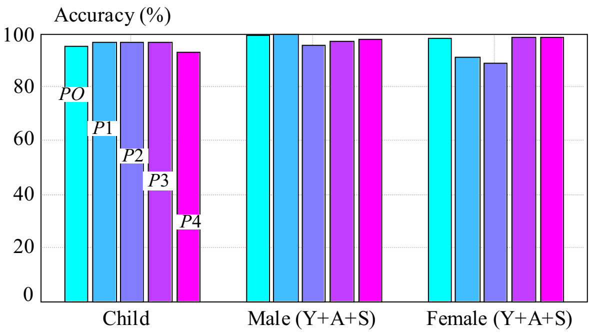

Fig. 6. Influence of different types of the input feature sets P0 ∼ P4 on the speaker gender classification accuracy for the 1^{st} level

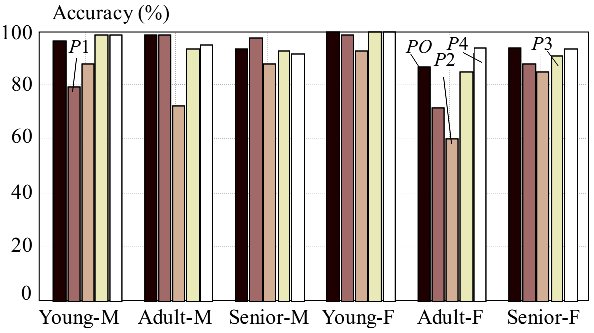

Fig. 7. Influence of the number of applied Gaussian mixtures on the speaker age classification accuracy for the 2nd level (joined values for males and females)

Table 6 and detailed bar-graphs in Figures 6 and 7 for the 1^{st} and the 2^{nd} classification level.

- Stability of the resulting gender/age speaker classification process depending on the time duration of the currently classified sentence – see Figures 8 and 9.
- Comparison of computational complexity: CPU times for the phases of the GMM creation, training, and classification in dependence on the used number of mixtures; presented in detail by Table 7.

The comparison was performed per structure parts of the developed GMM classifier: for the speaker gender – 1^{st} level, and for the speaker age (2^{nd} level – integrated for male and female voices) as documented by the detailed results in the form of the 2D confusion matrices in Fig. 10. Finally, the obtained gender/age classification accuracy values of the whole two-level architecture were compared with the results achieved by the listening test method – see the summary results in Table 8 for the "Full corpus" (using all the sentences from the testing GMM corpus) and the "Limited corpus" (using only the sentences having been evaluated in the listening test). For all the presented comparison results, the used parameter setting in tables or figure captions is: $K = 10$ for the W_{AVER} computation, $N_{\text{GMIX}} = 128$, the feature set P0, if not defined otherwise. The resulting gender/age classification accuracy was calculated from the number X_a of sentences with correctly determined gender/age class and the total number N_u of tested sentences as $(X_a/N_u) * 100(\%)$. Evaluation of the values for the confusion matrix was carried out in a similar way. To obtain the speaker independence, the data k-fold cross-validation method [13] was applied during the training and the testing process. The groups were divided in such a way that always 5 original speakers were used for training. For the speech corpus structure given by Table 1, it means that the groups were divided by the ratio of $5 : 2$ (five for training, two for testing/classification) for child and young classes (both genders together), or by the ratio of $5 : 1$ for adult and senior classes (male and female separately). A simple diagonal covariance matrix of the GMM was applied in this classification experiment due to its lower computational complexity. The basic functions from the Ian T. Nabney "Netlab" pattern analysis toolbox [20] were used for creation of the GMM models, data training, and classification. The computational complexity was tested on the UltraBook with the following configuration: processor Intel(R) Intel i5-4200U at 2.30 GHz, 8 GB RAM, and Windows 8.1.

Subjective evaluation was realized by the conventional listening test called "*Determination of the speaker age category*" located on the web page
http://www/lef.um.savba.sk/scripts/itstposl2.dll.
This listening test program was accessible for listeners always in the specific time interval (in this case it was in the period from September 25 to December 20, 2015) so that afterwards the obtained results were evaluated and processed. Twenty two listeners (6 women and 16 men) took part in our listening test experiment composed of 10 evaluation sets, each using 7 sentences selected randomly from the speech corpus, so 70 sentences were evaluated in total. For each sentence there was a choice from seven possibilities: "*Child*", "*Young*" Male/Female, "*Adult*" Male/Female, and "*Senior*" Male/Female. These choices comprise four age categories (child, young, adult, senior) as well as discrimination of gender for all but children's voices.

4 Discussion of obtained results

From the basic description of the used speech material introduced in Table 1 follows that the senior speech F0

Table 5. Influence of the used window length ($2K+1$) for the speech feature determination on the GMM gender classification accuracy in (%) for the 2nd level of the GMM classifier

W_{AVER} length/	Male speakers			Female speakers		
age category	Young	Adult	Senior	Young	Adult	Senior
$K = 5(W_{\text{AVER}} = 11)$	91.63	89.52	90.03	93.06	86.90	87.72
$K = 10(W_{\text{AVER}} = 21)$	96.30	98.31	96.12	100	87.75	93.81
$K = 15(W_{\text{AVER}} = 31)$	95.90	99.01	95.01	99.82	88.16	90.45
$K = 20(W_{\text{AVER}} = 41)$	87.86	91.13	91.43	93.46	81.69	84.07

Table 6. Comparison of the mean speaker age and gender classification accuracy in (%) including the standard deviation (in parentheses) for different types of feature sets P0∼P4

Classif. level/Feature set	P0	P1	P2	P3	P4
1Lg	97.5 (0.98)	95.9 (4.51)	93.7 (4.23)	96.2 (2.22)	96.7 (3.04)
2Lmf	94.7 (4.69)	88.7 (11.5)	80.7 (12.2)	93.3 (5.59)	95.3 (3.43)

Table 7. Comparison of the computational complexity (CPU time in (s)) for different number of used mixtures

Phase&level/N_{GMIX}	8	16	32	48	64	128
GMM creation: 1Lg	47.0	74.3	133.9	188.2	249.5	530.3
2Lmf	30.8	55.9	104.2	152.6	191.6	379.6
12Lall	77.8	130.2	238.1	240.8	341.1	900.9
Classification[A]: 1Lg	0.57 (0.19)	0.61 (0.25)	0.64 (0.26)	0.69 (0.27)	0.75 (0.36)	0.90 (0.45)
2Lmf	0.72 (0.27)	0.73 (0.27)	0.75 (0.28)	0.77 (0.29)	0.78 (0.30)	0.89 (0.33)
12Lall	1.3	1.3	1.4	1.4	1.6	1.8

[A] Mean values per sentence including the standard deviation values in (s) (in parentheses)

Table 8. Final comparison of the speaker's age classification accuracy in (%) based on the two-level GMM approach (full corpus: all sentences, limit. corpus: only sentences of the listening test) and the standard listening test

Classif. method/	Child	Male speakers			Female speakers		
age category		Young	Adult	Senior	Young	Adult	Senior
GMM-full corpus	90.40	95.06	92.37	93.20	100	86.73	88.66
GMM-limit. corpus	100	86.67	80.00	93.33	93.33	73.33	93.33
Listening test	100	66.67	80.65	68.75	63.93	66.13	76.19

distribution has two local maxima for males as well as for females. This interesting result of our analysis might be a reason for rather contradicting phonetic research of the aged voice. Generally, typical of "old" voices by the listeners' opinion is lower pitch [21]. On the other hand, relatively high pitch of the geriatric voice was reported [22]. Another studies differentiating between male and female voices state that the mean fundamental frequency increases in older males while it remains constant or decreases in older females [23]. It means that the mean F0 will not be very useful for distinguishing of senior voices. The F0 range included in the speech feature vector helps more as the vocal ageing is accompanied with the reduced pitch range [24].

The obtained results of the first auxiliary task concerning the influence of the used number of Gaussian mixtures on the classification accuracy shows, that the maximum of 128 mixtures does not bring significant increase of the classification accuracy. It is most visible in the case of the 1st level of the tested GMM classifier (compare the mean values in Table 3). Greater influence of the number of used mixtures is observed in the 2nd classification level — compare bar-graph values in Fig. 5. In both cases, the minimum accuracy can be found when the small number of mixtures (of 8) was used. The subsequent evaluation of the influence of different window lengths W_{AVER} for the speech feature determination has shown a local maximum of the achieved gender/age classification accuracy for K in the interval from 10 to 15 with slight differences for the male and the female voices (see the values in Tables 4 and 5 separately for both classification levels). It can be explained by the fact that for $K > 15$ the representative values in the feature vectors are too averaged — so they cannot reflect local changes in the speech signal. Thus, the value $K = 10$ ($W_{\text{AVER}} = 21$) was chosen for next experiments.

The next auxiliary experiment is aimed at finding the suboptimal structure of the input feature set for the whole

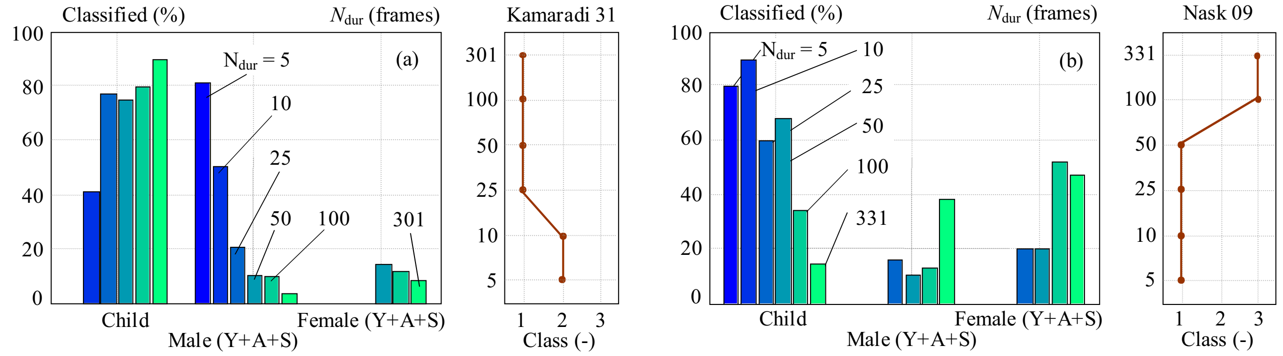

Fig. 8. Detailed comparison of the gender classification for different number of frames N_{DUR} (the bar-graph of the classification accuracy and the output class trajectory) of the analysed sentence for (a) — children and (b) — female voices; output classes: 1=Child, 2=Male, 3=Female, 1st classification level.

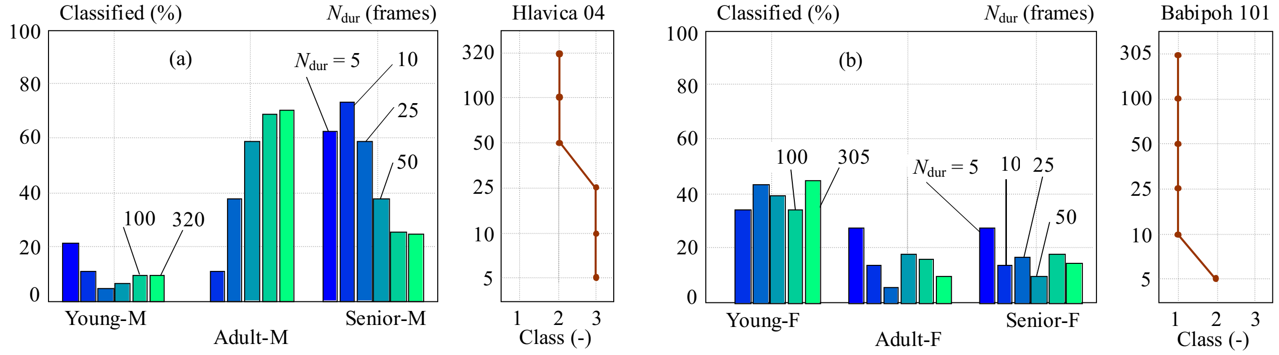

Fig. 9. The bar-graph of the classification accuracy together with the output class trajectory for different number of frames N_{DUR} of the analysed sentence for (a) — adult male and (b) — young female voices; output classes: 1=Young, 2=Adult, 3=Senior, 2nd classification level

two-level GMM gender/age classifier. The gender classification accuracy values are relatively balanced as documented by Table 6. Some types of speech features are not sufficient for this identification task, especially those based on formant frequencies; see the worst results for the set P2 (94 % for the gender classification accuracy and 81 % for the age classification accuracy in the 2nd level). On the other hand, using the first four cepstral coefficients (their skewness and kurtosis values — see the structure of the feature set P4 in Table 2) helps to achieve the best speaker age accuracy of 95 %. The applied number of mixtures has a great influence on the computational complexity (the measured CPU time) for creation and training of the GMM models but its impact on duration of the identification phase is negligible. In the case of GMM creation the use of the maximum number of 128 mixtures increases the CPU time more than 11 times when compared with 8 mixtures — see values in Table 7. According to these results, the settings of initial parameters for the whole GMM classifier were chosen as follows: $N_{\mathrm{GMIX}} = 64$ with using the feature set P0 for the 1st level and $N_{\mathrm{GMIX}} = 128$ together with the feature set P4 for age classification at the 2nd level. From our stability experiment follows that the length limitation of the processed speech signal does not play an essential role when the minimum number of about 50 speech frames is processed (as documented by the graphs of the class stability determination in Figures 8 and 9). It holds for both levels of the proposed GMM speaker gender/age classifier currently developed for testing of a continuous speech (*ie* sentences – not isolated words).

The realized comparison of the speaker's age classification accuracy based on GMM approach and standard listening test shows that the children voice was detected with approx. 100 % in both comparison experiments — see the bar-graph in Fig. 11a. In contrast to previous results, the adult voice (male as well as female) was identified worse using the GMM classifier. The listening test results are consistent with the finally achieved low values of the mean recognition accuracy — 63 % for the young-female and 67.8 % for the young-male categories in contrast to 100 % accuracy for the child category. From the listeners' feedback information follows that it was very easy to detect the children's voices, on the other hand, the young female voices as well as the young male voices were often confused with the adult ones as documented by the confusion matrix in Fig. 11b.

5 Conclusion

The performed experiments have successfully confirmed that the proposed two-level architecture of the

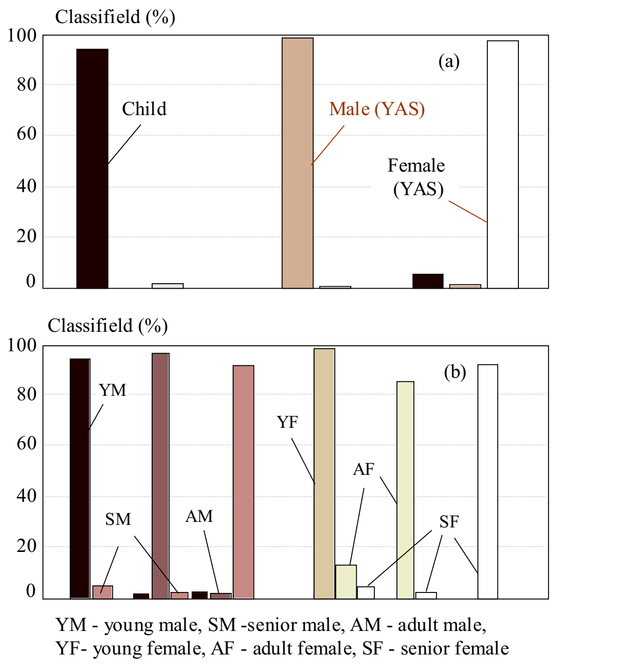

Fig. 10. 2D confusion matrices of the GMM classification per function levels – the best obtained results: (a) — of the speaker gender – 1st level, (b) — of the speaker age (2nd level – integrated for male and female voices)

speaker gender/age GMM classifier is correct and the system is practically functional. A critical issue is a correct function of the first block (the gender type recognizer) as the age class determination block operates with different models trained for the male and the female voices. In the case of wrong determination (gender class exchange) it occurs that the resulting age classification accuracy is decreased.

The second aim of this work — to find an alternative approach to the standard listening tests for evaluation of speech quality — was successfully attained. It is important especially in the cases when the audible differences of the evaluated sentences were too small or hardly recognizable by listeners, or there existed a problem with their collective realization, *etc.* In addition, the distributed listening test execution brings one principal disadvantage of having no feed-back information about the evaluators' hearing conditions, due to their individual connection to our internet server. They may use earphones, headphones, or even external loudspeakers, although we suppose that majority uses headphones. The main advantage of the developed GMM-based classification system is that it works automatically without human interaction and the obtained results can be numerically matched as the objective comparison criterion.

From the detailed analysis of the computational complexity (CPU time) follows that in the current realization of the GMM age classifier (in the Matlab environment) the identification phase of the tested sentence runs more than 1 second, so this processing must be run off-line. It is assumed that after the program optimization and implementation in a higher programming language such as C++, C #, or Java, the real-time processing will be available. In near future, we will try to test the GMM evaluation using larger speech databases as well as the ones spoken in other languages (English, German), and finally we plan to test the performance of the proposed GMM-based gender and age classification method using a speech corpus consisting of signals under low SNR conditions (*eg* the well-known NTIMIT database) or to train the model on the noise-free database and adapt it to the noise conditions in a similar way as for speech recognition [25].

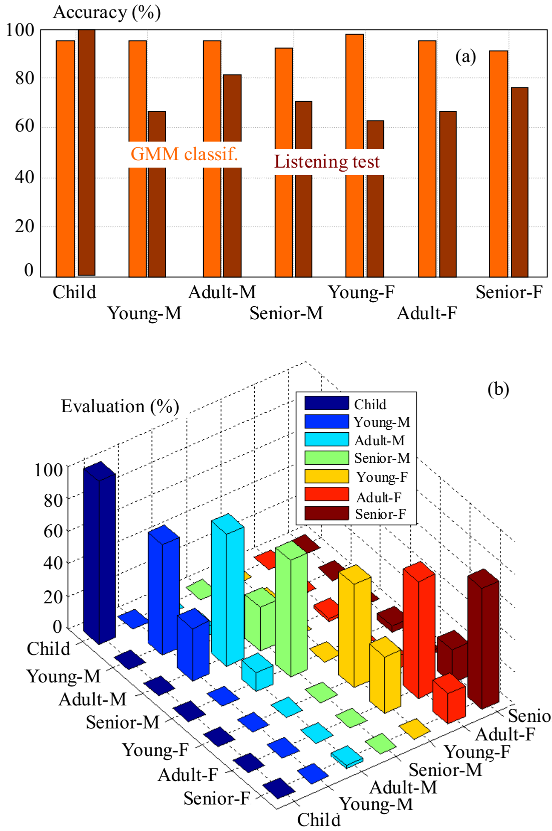

Fig. 11. The bar-graph comparison of the mean GMM gender/age recognition accuracy together with results of (a) — listening tests, (b) — 3D confusion matrix of the speaker age category evaluation by the listening test

Acknowledgment

The work has been supported by the Scientific Grant Agency of the Slovak Academy of Sciences and the Ministry of Education, Science, Research, and Sports of the Slovak Republic (VEGA 2/0001/17 and VEGA 1/0905/17) and and the Czech Science Foundation (GA CR), project No. GA16-04420S.

The authors would also like to express thanks to all the people who participated in the listening test.

References

[1] M. Li, K. J. Han and S. Narayanan, "Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level In-

formation Fusion”, *Computer Speech and Language*, vol. 27, 2013, 151–167.

[2] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt and E. Nöth, ”Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 31 March – 4 April 2008, 1605–1608, Las Vegas, NV: IEEE.

[3] G. Dobry, R. M. Hecht, M. Avigal and Y. Zigel, ”Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on Acoustic Speech Signal”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, 2011, 1975–1985.

[4] C. Van heerden, E. Barnard, M. Davel, C. Van der Walt, E. Van dyk, M. Feld and C. Müller, ”Combining Regression and Classification Methods for Improving Automatic Speaker Age Recognition”, *IEEE International Conference on Acoustics, Speech, and Signal Processing,* 14-19 March 2010, 5174–5177, Dallas, TX: IEEE.

[5] M. H. Bahari, M. Mclaren, H. Van Hamme and D. A. Van-Leeuwen, ”Speaker Age Estimation Using i-Vectors”, *Engineering Applications of Artificial Intelligence*, vol. 34, 2014, 99–108.

[6] M. Fairhurst, M. Erbilek and M. Da Costa-Abreu, ”Selective Review and Analysis of Aging Effects in Biometric System Implementation”, *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 3, 2015, 294–303.

[7] N. Minematsu, M. Sekiguchi and K. Hirose, ”Automatic Estimation of One’s Age with His/her Speech Based upon Acoustic Modeling Techniques of Speakers”, *IEEE International Conference on Acoustics, Speech, and Signal Processin*, 13-17 May 2002, I-137–I-140, Orlando, FL, USA: IEEE.

[8] H. Meinedo and I. Trancoso, ”Age and Gender Classification using Fusion of Acoustic and Prosodic Features”, *Interspeech 2010*, 26-30 September 2010, Makuhari, Japan, 2822–2825.

[9] B. D. Barkana and J. Zhou, ”A new Pitch-Range Based Feature Set for a Speaker’s Age and Gender Classification”, *Applied Acoustics*, vol. 98, 2015, 52–61.

[10] A. Fedorova, O. Glembek, T. Kinnunen and P. Matějka, ”Exploring ANN Back-Ends for i-Vector Based Speaker Age Estimation”, *Interspeech 2015*, 6-10 September 2015, Dresden, Germany, 3036–3040.

[11] D. Tihelka, M. Grůber and Z. Hanzlíček, ”Robust Methodology for TTS Enhancement Evaluation”, *Text, Speech and Dialogue*, I. Habernal, V. Matoušek, 2013, LNAI 8082, Berlin Heidelberg, Springer, 442–449.

[12] J. Přibil, A. Přibilová and J. Matoušek, ”Experiment with GMM Based Artefact Localization in Czech Synthetic Speech”, *Text, Speech, and Dialogue* (TSD), P. Král, V. Matoušek, LNAI 9302, Springer, 2015, 23–31.

[13] D. A. Reynolds, R. C. Rose, ”Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models”, *IEEE Transactions on Speech and Audio Processing*, vol. 3, 1995, 72–83.

[14] A. Venturini, L. Zao and X. Coelho, ”On Speech Features Fusion, α-Integration Gaussian Modeling and Multi-Style Training for Noise Robust Speaker Classification”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, 2014, 1951–1964.

[15] M. Shah, C. Chakrabarti and A. Spanias, ”Within and Cross-Corpus Speech Emotion Recognition Using Latent Topic Model-Based Features”, *EURASIP Journal on Audio, Speech, and Music Processing*, 2015, vol. 4, 2015, 1–17.

[16] J. Přibil, A. Přibilová and D. Ďuračková, ”Storytelling Voice Conversion: Evaluation experiment using Gaussian mixture models”, *Journal of Electrical Engineering*, vol. 66, 2015, 194–202, DOI: 10.1515/jee-2015-0032/.

[17] J. Přibil and A. Přibilová, ”GMM-Based Evaluation of Emotional Style Transformation in Czech and Slovak”, *Cognitive Computation*, 2014, DOI: 10.1007/s12559-014-9283-y.

[18] B. Božilović, B. Todorović and B. M. Obradović, ”Text-Independent Speaker Recognition Using Two-Dimensional Information Entropy”, *Journal of Electrical Engineering*, vol. 66, no. 3, 2015, 167–173.

[19] P. Boersma and D. Weenink, ”Praat: Doing Phonetics by Computer” (Version 5.4.22) [Computer Program], Retrieved 8 October 2015, from http://www.fon.hum.uva.nl/Praat.

[20] I. T. Nabney, ”Netlab Pattern Analysis Toolbox”, Copyright (1996-2001), Retrieved February 16, 2012, from http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab.

[21] S. E. Linville, ”Source Characteristics of Aged Voice Assessed from Long-Term Average Spectra”, *Journal of Voice*, vol. 16, no. 4, 2002, 472–479.

[22] R. J. Baken, ”The Aged Voice: A New Hypothesis”, *Journal of Voice*, vol. 19, no. 3, 2005, 317–325.

[23] J. D. Harnsberger, R. Shrivastav, W. S. Brown, H. Rothman and H. Hollien, ”Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age”, *Journal of Voice*, vol. 22, no. 1, 2008, 58–69.

[24] J. D. Harnsberger, W. S. Brown, R. Shrivastav and H. Rothman, ”Noise and Tremor in the Perception of Vocal Aging in Males”, *Journal of Voice*, vol. 24, no. 5, 2010, 523–530.

[25] G. Gosztolya and T. Grósz, ”Domain Adaptation of Deep Neural Networks for Automatic Speech Recognition via Wireless Sensors”, *Journal of Electrical Engineering*, vol. 67, no. 2, 2016, 124–130.

Received 23 June 2016

Jiří Přibil (Ing, PhD), born in 1962 in Prague, Czechoslovakia. He received his MSc degree in computer engineering in 1991 and his PhD degree in applied electronics in 1998 from the Czech Technical University in Prague. At present, he is a senior scientist at the Department of Imaging Methods Institute of Measurement Science, Slovak Academy of Sciences in Bratislava. His research interests are signal and image processing, speech analysis and synthesis, and text-to-speech systems.

Anna Přibilová (Assoc Prof, Ing, PhD) received her MSc and PhD degrees from the Faculty of Electrical Engineering and Information Technology, Slovak University of Technology (FEEIT SUT) in 1985 and 2002, respectively. Since 1992 she has been working as a university teacher at the Radioelectronics Department, and in 2014 she has become an associate professor at the Institute of Electronics and Photonics of the FEEIT SUT in Bratislava. The main field of her research and teaching activities is audio and speech signal processing.

Jindřich Matoušek (Assoc Prof, Ing, PhD) received his MSc and PhD degrees from the Faculty of Applied Sciences (FAS), University of West Bohemia (UWB), Pilsen, Czech Republic in 1997 and 2001, respectively. Since 1999 he has been working as a researcher at the Department of Cybernetics FAS UWB, and since 2012 he also has been working as member of a research team of the New Technology for Information Society (NTIS) centre at UWB. In 2009 he became an associate professor at FAS UWB. The main field of his research and teaching activities is computer speech processing, especially speech synthesis.