DE GRUYTER OPEN

Journal of ELECTRICAL ENGINEERING, VOL 69 (2018), NO2, 138–147

Evaluation of speaker de-identification based on voice gender and age conversion

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

Two basic tasks are covered in this paper. The first one consists in the design and practical testing of a new method for voice de-identification that changes the apparent age and/or gender of a speaker by multi-segmental frequency scale transformation combined with prosody modification. The second task is aimed at verification of applicability of a classifier based on Gaussian mixture models (GMM) to detect the original Czech and Slovak speakers after applied voice de-identification. The performed experiments confirm functionality of the developed gender and age conversion for all selected types of de-identification which can be objectively evaluated by the GMM-based open-set classifier. The original speaker detection accuracy was compared also for sentences uttered by German and English speakers showing language independence of the proposed method.

K e y w o r d s: GMM classifier, spectral and prosodic features of speech, speaker gender and age classification

1 Introduction

Data de-identification is a process of data transformation that ensures their anonymity and thus enables personal privacy protection. Methods of automatic de-identification of personal data have been mostly elaborated in the area of medicine but globalization of the world and the information technologies broadened the need for de-identification of general multimedia data in the form of the text, image, video, or speech. From these modalities the speech is the least investigated in the context of de-identification [1] though the methods of conversion between the voice of the source and the target talker have been used for more than three decades [2].

Probably the first attempt to de-identify the voice was combination of spectral transformation and prosodic modification by the pitch-synchronous overlap and add (PSOLA) algorithm [3]. A more elaborate approach transforms the source speech of different persons in such a way that it sounds as if uttered by the same target speaker. The output diphone-based synthetic speech with the consistent duration characteristics regardless of the input speaker and further extrapolation of the features derived from the mel-frequency cepstral coefficients (MFCC) gave 100% de-identification rate by the voice identification system based on Gaussian mixture models (GMM) and 87.5% for the phonetic speaker identification system [4]. The target speech synthesis was also used by [5] where the intelligibility of the speech de-identified by two methods was compared resulting in 33% average word error rate for the speech synthesis system based on hidden Markov models (HMM) and 21% for the diphone time-domain PSOLA synthesis system. The phase vocoder and the standard vocal tract length normalization were used to conceal the gender of the speaker [6] showing that for such a de-identification system the preceding gender recognition is necessary. Speaker de-identification gender conversion was investigated also in [7] where the spectral amplitude scaling was combined with the piecewise linear transformation and the linear modification of the fundamental frequency (F_0) giving 96.9% de-identification accuracy by the speaker identification in the i-vector space. The pre-calculated voice transformations based on GMM mapping and harmonic plus stochastic models with the target synthetic HMM-based voice were used for successful de-identification in the open set comparable to the closed-set de-identification 87.4% open-set *vs* 91% closed-set de-identification rate. The spectral envelope conversion based on the time-varying frequency warping combined with GMMs was used in [8] where the most appropriate target speaker is selected by adding the module of GMM-based or i-vector speaker identification. Comparison of de-identification for the unauthorized subject and re-identification for the authorized one gives higher de-identification accuracy (90%) than the identification accuracy of the speech re-transformed back to the original speaker (87%).

In voice de-identification it is not advisable to transform a person's voice to another person's voice so that this target person would not be falsely identified. The transformed voice should not resemble any known voice, so we decided not to apply training on the target voice. In our experiment, we use simpler solution based on a spectral transformation by frequency scale warping accompanied with a modification of prosody parameters. Parameters of spectral and prosodic transformations are

* Institute of Measurement Science, Slovak Academy of Sciences, Bratislava, Slovakia, Jiri.Pribil@savba.sk, ** Slovak University of Technology in Bratislava, Faculty of Electrical Engineering and Information Technology, Bratislava, Slovakia Anna.Pribilova@stuba.sk, *** Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Plzeň, Czech Republic, jmatouse@kky.zcu.cz

DOI: 10.2478/jee-2018–0017, Print (till 2015) ISSN 1335-3632, On-line ISSN 1339-309X © 2018FEI STU

estimated according to the results of phonetic research about age-related and gender-related changes in speaker characteristics. Cepstral speech analysis is used to obtain the original speakers' spectral properties which are modified to get a de-identified speech. It is resynthesized by the source-filter speech synthesizer with cepstral parameterization of the impulse response of the vocal tract model [9]. Comparison of the obtained results with those of the other authors using different methods for speech parameterization, transformation, and resynthesis, is difficult to be done. Successfulness of different speaker and voice transformation methods can be evaluated by objective or subjective criteria. The subjective ones consist of the conventional listening tests. The objective approaches comprise evaluation by numerical comparison of the determined differences between the spectral envelopes, the spectral distances, etc. Such evaluation approaches are often applied in the automatic speaker recognition yielding the final evaluation in the form of a recognition score with predominant use of GMMs [10].

This work was motivated by two aims: to evaluate the de-identification potential of the introduced transformation method and to find the alternative to the standard listening tests. It is important in the cases when the test is relatively difficult (only small audible differences could be heard) or when there is a problem with its practical collective realization, etc. The main advantage of the proposed GMM-based system is its automatic functioning without human interaction and the possibility of direct numerical matching of the obtained results using the objective comparison criterion. In practice, for a task of this type the used speaker verification systems usually reject speakers previously unknown to the system. In accordance with general knowledge, also our GMM-based classifier for re-identification of a de-identified speaker was designed as an open-set system.

The paper first describes the de-identification method which changes the fundamental frequency and the frequencies in the first four formant areas for conversion of the voice gender (male-to-female, female-to-male) and the voice age (younger, older). This step is followed by the functionality verification of the developed GMM-based original speaker recognizer. The main recognition experiment consists of three basic tasks: evaluation of the influence of the cepstral parameterization and the speech signal reconstruction without modification on the original speaker detection accuracy, detailed analysis for each of the four de-identified voice types, and analysis of reversibility of the used de-identification method by the inverse transformation after the first conversion to four gender/age types. These experiments were first done on the sentences from the database of Czech and Slovak stories and audio books (CZ&SK-DB) performed by professional actors. The obtained results were finally compared with those achieved employing the sentences in a neutral speaking style from the Berlin Database of Emotional Speech (Emo-DB) in German language [11] and the Texas Instruments and Massachusetts Institute of Technology (TIMIT) database in English language [12].

2 Subject and method

2.1 Applied method of voice de-identification

In this paper we propose and apply four combinations of spectral and prosodic modifications to de-identify the speaker's voice. The prosodic transformation comprises modification of the fundamental frequency and the energy of the speech signal. The spectral transformation is realized by expansion or compression of the speech spectral envelope depending on the wanted increase or decrease of formant frequencies. Apart from the formant frequencies also their bandwidths are changed in the same direction, which is consistent with physical properties of the human vocal tract [13]. Short-time signal processing in equidistant time frames is performed first by the fast Fourier transform (FFT) applied to each frame normalized by the Hamming window. Then application of the inverse fast Fourier transform on the log spectrum gives the real cepstrum. Its limitation in the quefrency domain is transformed by inverse FFT back to the frequency domain to get the log spectral envelope. The frequency scale warping is created in such a manner that the new spectral value at the frequency f_{in} is equal to the original value of the unwarped spectrum at the corresponding frequency f_{out}. For $f_{\text{out}} < f_{\text{in}}$ the spectral value at f_{in} corresponds to the spectral value at lower frequency f_{out}. Consequently, the spectral envelope is expanded and the formant frequencies are raised. For $f_{\text{in}} < f_{\text{out}}$ the opposite situation arises, the spectral envelope is compressed and the formant frequencies are lowered. For this reason, the shift of the nth formant between the de-identified and the original voice ΔF_n % at the frequency f_{in} is evaluated as

$$\Delta F_n(f_{\text{in}}) = \frac{f_{\text{in}} - f_{\text{out}}}{f_{\text{in}}} 100. \tag{1}$$

The greatest difference in the identity of the speaker is perhaps the difference in the gender. Although there are gender-related differences in the fundamental frequency of the vocal fold vibration, for the perception of the gender more important are the resonances of the vocal tract because female speakers are perceived as female even with the speaking fundamental frequency in the typical male voice range [14]. A recent cross-cultural study comprising 700 men and women [15] has presented relationships between the human body morphology and the voice parameters, among them also the fundamental frequency F_0 and the first four formants (F_1 to F_4). The ratio of the mean values of these frequencies between females and males yields 1.84 for F_0, 1.11 for F_1, 1.22 for F_2, and 1.17 for both F_3 and F_4. From the published data we can also derive mean border frequencies of the formant regions: the highest F_1 for women ($F_{1\max} = 595\text{Hz}$), the lowest F_2 for men ($F_{2\min} = 1380\,\text{Hz}$), the highest F_2 for women ($F_{2\max} = 2048\,\text{Hz}$), the lowest F_3 for

Table 1. Parameters for voice de-identification; m_{F0} and m_{En} are output-to-input F_0 and energy ratios, f_s =16 kHz

Transformation type/parameter	m_{F0} (-)	ΔF_1 (%)	ΔF_2 (%)	ΔF_3 (%)	f_c (Hz)	m_{En} (-)
$M \rightarrow F$	1.84	+11	+22	+17	8000 (f_s)	1.2
$F \rightarrow M$	0.54	–10	–18	–15	8000 (f_s)	0.56
Older	0.8	–20	none	none	1380 (F_{2min})	1.1
Younger	1.25	+25	none	none	1380 (F_{2min})	0.86

men ($F_{3\mathrm{min}} = 2460\,\mathrm{Hz}$), and the highest F_4 for women ($F_{4\mathrm{max}} = 4317\,\mathrm{Hz}$).

Identity of the speaker might be concealed also by changing the apparent age of the speaker. The effect of ageing on the speech of the same male and female persons was explored by [16]. This phonetic research showed that with increasing age there is a decrease in both F_0 and F_1 for a female speaker and a decrease followed by an increase for a male speaker at the age of about 85, while there were no age-dependent differences for F_2 and a weak, but not significant increase of F_3 in the majority of the examined speakers. From the presented results follows that F_0 as well as F_1 fall by about 10 % in the age increase by 20 years and by about 20 to 30 % in the span of 40 years.

All these data related to the pitch frequency and the formants will be useful for design of the multi-segmental function for the transformed frequency scale. The segments of this function are of two types: the first ones correspond to the frequency regions of the formants F_n with the constant formant shift ΔF_n defined by (1) related to a ratio of mean values of the formant frequencies between the target and the original voice; the segments of the second type represent connection between the adjacent formant regions by interpolation to prevent abrupt changes in formant shifts throughout the whole frequency range.

In the first segment the input frequencies are shifted by ΔF_1

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = \left(-\frac{\Delta F_1}{100} + 1\right) f_{\mathrm{in}} \quad \text{for} \quad 0 \le f_{\mathrm{in}} \le F_{1\mathrm{max}}. \tag{2}$$

The second segment connects the boundaries of the F_1 and F_2 areas

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = f_{\mathrm{out}}(F_{1\mathrm{max}}) + \frac{\left(-\frac{\Delta F_2}{100} + 1\right) F_{2\mathrm{min}} - f_{\mathrm{out}}(F_{1\mathrm{max}})}{F_{2\mathrm{min}} - F_{1\mathrm{max}}} f_{\mathrm{in}} \quad \text{for} \quad F_{1\mathrm{max}} < f_{\mathrm{in}} \le F_{2\mathrm{min}}. \tag{3}$$

The third segment covers the F_2 area with the shift F_2

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = \left(-\frac{\Delta F_2}{100} + 1\right) f_{\mathrm{in}} \quad \text{for} \quad F_{2\mathrm{min}} < f_{\mathrm{in}} \le F_{2\mathrm{max}}. \tag{4}$$

The fourth segment represents the connection between the F_2 and F_3 areas

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = f_{\mathrm{out}}(F_{2\mathrm{max}}) + \frac{\left(-\frac{\Delta F_3}{100} + 1\right) F_{3\mathrm{min}} - f_{\mathrm{out}}(F_{2\mathrm{max}})}{F_{3\mathrm{min}} - F_{2\mathrm{max}}} f_{\mathrm{in}} \quad \text{for} \quad F_{2\mathrm{max}} < f_{\mathrm{in}} \le F_{3\mathrm{min}}. \tag{5}$$

The F_3 and F_4 areas with almost the same shift ΔF_3 are concatenated

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = \left(-\frac{\Delta F_3}{100} + 1\right) f_{\mathrm{in}} \quad \text{for} \quad F_{3\mathrm{min}} < f_{\mathrm{in}} \le F_{4\mathrm{max}}. \tag{6}$$

In the last but one segment the formant shift gradually decreases to zero at its end

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = f_{\mathrm{out}}\,(f_{\mathrm{br}}) + \frac{f_c - f_{\mathrm{out}}(f_{\mathrm{br}})}{f_c - f_{\mathrm{br}}} f_{\mathrm{in}} \quad \text{for} \quad f_{\mathrm{br}} < f_{\mathrm{in}} \le f_c, \tag{7}$$

where f_{br} is the breakpoint frequency (F_{4max} for gender conversion or F_{1max} for age conversion), and f_c is the cut-off frequency with the zero formant shift. In the whole last segment the formant shift is zero

$$f_{\mathrm{out}}(f_{\mathrm{in}}) = f_{\mathrm{in}} \quad \text{for} \quad f_c < f_{\mathrm{in}} \le \frac{f_s}{2}, \tag{8}$$

where f_s is the sampling frequency.

The input parameters for these equations are resumed in Table 1. The proposed de-identification system is gender-independent so modification designated $M \rightarrow F$ transforms the voice of a male to a female, and the voice of a female to a high-pitched female or a child. Similarly, $F \rightarrow M$ modification changes the voice of a female to a male, and the voice of a male to a low-pitched male. The system is also age-independent so *Younger* conversion may be applied to a young voice and *Older* conversion to an old one. The transformation functions of the frequency scale together with the frequency-dependent formant shifts are shown in Fig. 1. For speech resynthesis

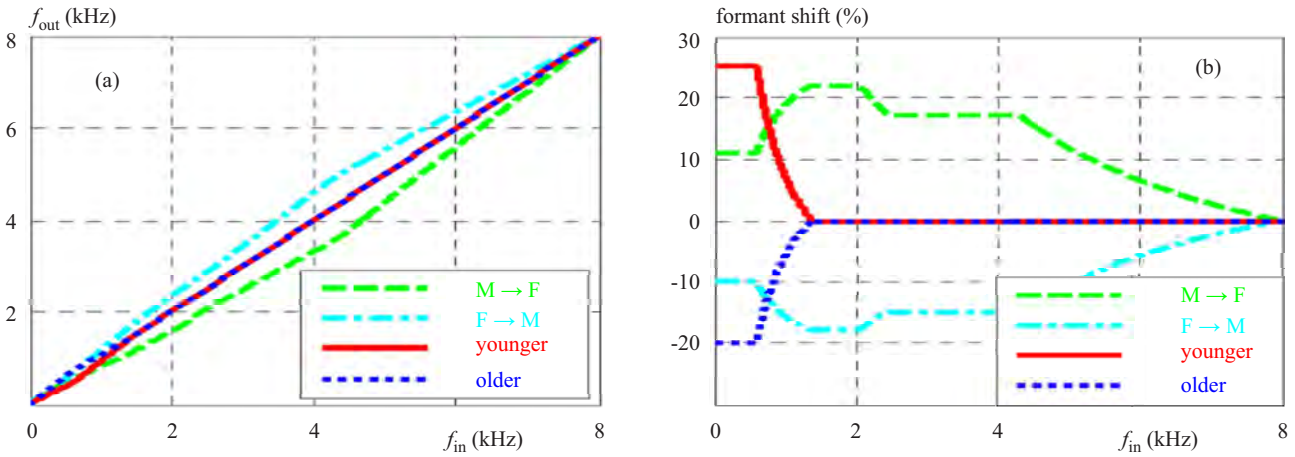

Fig. 1. (a) – frequency scale transformation function, and (b) – the corresponding formant shift using the data from Table 1

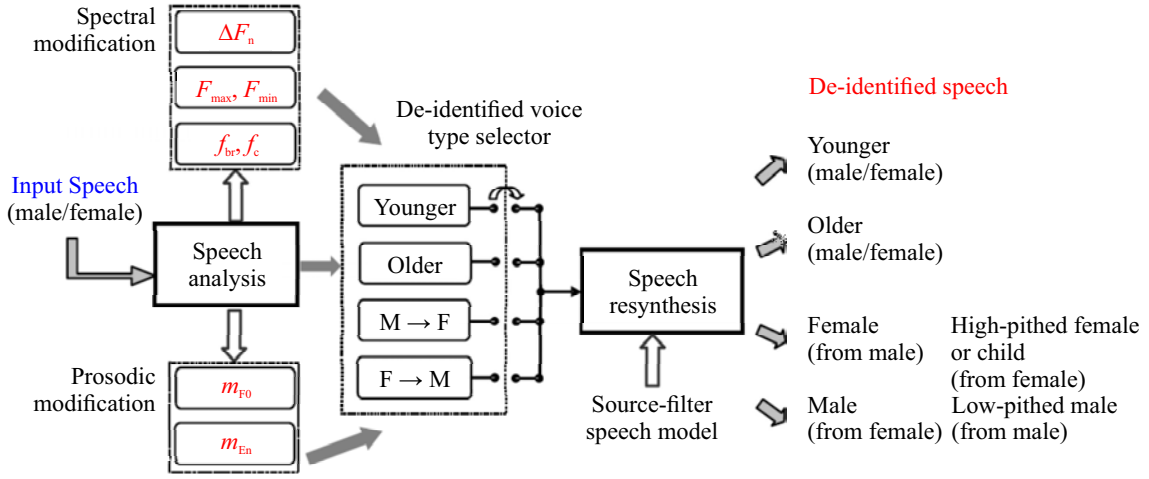

Fig. 2. Block diagram of the applied method of voice de-identification

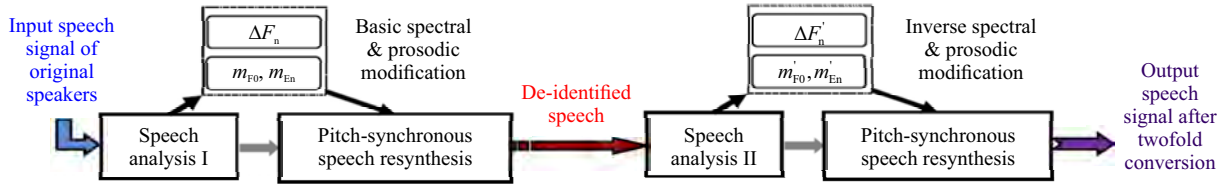

Fig. 3. Block diagram of the two-step voice transformation process used for testing of de-identification reversibility

of the converted voices the source-filter model with cepstral parameterization of the vocal tract transfer function was used in all cases [9]. The structure and the function of the proposed de-identification system can be seen in Fig. 2. Reversibility of the used de-identification method can be tested by processing of the de-identified speech by the transformation inverse to the first one, *ie* instead of the ratios m_{F0}, m_{En}, and the formant shifts ΔF_n shown in Tab. 1 the following input parameters are applied

$$m'_{F0} = \frac{1}{m_{F0}}, \quad m'_{En} = \frac{1}{m_{En}}$$
$$\Delta F'_n = \left(\frac{1}{1 + \frac{\Delta F_n}{100}} - 1 \right) 100 \tag{9}$$

This two-step voice transformation process for testing of de-identification reversibility is illustrated in Fig. 3.

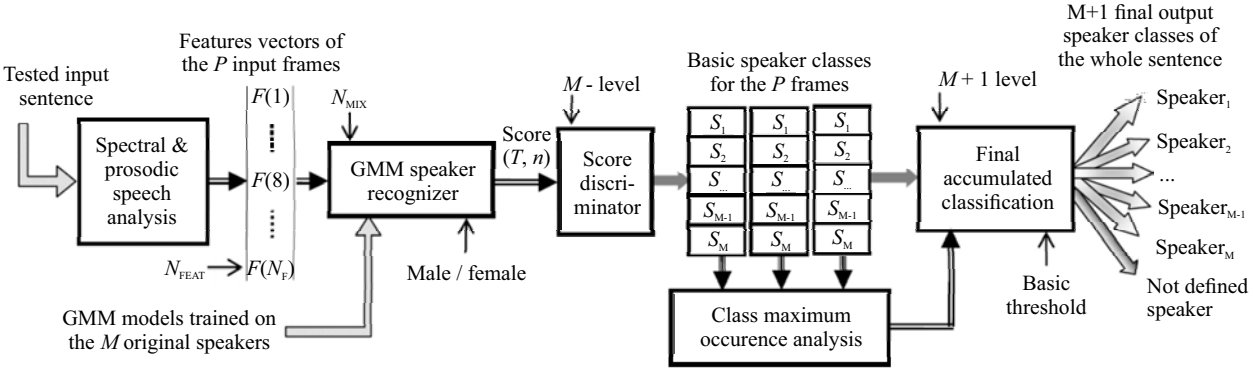

Fig. 4. Block diagram of the GMM-based classifier for identification of the original speaker from de-identified speech

2.2 Description of GMM - based speaker re-identification method

Principally, GMMs represent a linear combination of multiple Gaussian probability distribution functions of the input data vector [10]. The covariance matrix and the vector of means together with the weighting parameters must be determined from the input training data. For the mixture of Gaussians the use of maximum likelihood gives no closed-form analytical solution which would be an ideal case.

Therefore, the expectation-maximization (EM) iteration algorithm is used for maximizing the likelihood functions [17]. Initial parameters for the EM algorithm are the number of mixtures and the number of iterations. In general, elements of feature vectors could be correlated so rather a high number of mixture components and a full covariance matrix would be necessary for sufficient approximation. On the other hand, a GMM with a diagonal covariance matrix is usually used in speaker identification [10] due to its lower computational complexity. The GMM classifier returns the probability score that the tested utterance belongs to the GMM model. In the standard realization of the GMM classifier, the resulting class is given by the maximum overall probability of all obtained scores corresponding to M output classes. This relatively simple and robust approach is also used in our proposed original speaker identifier – see the principal block diagram in Fig. 4. In the first step, the speech spectral and prosodic features are determined from the tested utterance of the original or the de-identified speaker. The analysed input sentence is processed in frames so finally we obtain P feature vectors corresponding to the number of frames in the input sentence. The feature vectors obtained in this way are further processed in the speaker recognizer block using the GMM models that were trained on the data of the feature vectors from the original speakers' sentences. The output scores are next processed to determine the maximum overall probability in the discriminator block where one of M output classes is assigned to each of P vectors. Then, the class distribution based on histograms is constructed and the maximum occurrence is determined. The final classification block works with $M + 1$ output classes – a virtual class called *Not defined* is added to the basic closed set of M speakers to create an open-set speaker identifier. The classification strategy is based on a consideration that when the class distribution is relatively flat (without a dominant class), the whole tested sentence is classified as *Not defined*.

MFCC coefficients together with prosodic parameters are often used in speaker recognition systems [18] which can also be used to check the de-identification performance. However, because of our previous good experience, different types of speech features comprising basic and supplementary spectral properties complemented with supra-segmental parameters were used in this experiment for GMM creation, training, and classification. These speech features have been successfully used in the GMM-based evaluation of emotional style transformation [19], the text-independent original speaker recognition from the emotionally converted speech [20], and the speaker gender and age classification after voice conversion [21].

The speech signal analysis is performed as follows: After segmentation of the input sentence the fundamental frequency is determined. From the windowed speech frames the smoothed spectral envelope and the power spectral density are computed for further determination of spectral and prosodic features. The basic spectral features comprising the first two formant positions with their ratios and the first four cepstral coefficients are used here together with the statistical features determined from the spectral envelope: tilt, spread, and centroid. The used supplementary spectral features are harmonic-to-noise ratio, spectral flatness, and spectral entropy [22]. Amongst the prosodic parameters, the differential F_0 contour, zero-crossing rate, jitter, shimmer, etc., are determined. Every vector of P speech features is subsequently processed to obtain N_{FEAT} representative statistical values used in the GMM classification process.

3 Material, experiments and results

3.1 Used speech material

The main speech corpus CZ&SK-DB used for GMM training and testing consists of declarative sentences originating from stories and audgio books uttered without emotional arousal by 10 male and 10 female professional actors in Czech and Slovak languages with the duration from 0.5 s to 8.5 s. It is divided into two parts: the first, consisting of 89/79 sentences from M1-5/F1-5 speakers, includes the basic speaker classes in all comparison experiments. This basic part of the speech database was originally collected for the experiment with GMM-based speaker gender and age classification after voice conversion, presented in [21]. The extendged second part of the whole CZ&SK-DB database consists of another 85 sentences from 5 male and 5 female speakers in the classes M6-10/F6-10 which were used for testing of the *Not defined* speaker group class. Apart from categorization by the gender, the speakers whose speech is stored in this database can also be categorized by the age into the young/adult/senior voice classes. The basic speakers' parameters the gender/age type, the mean F_0 value, and the total duration of all included sentences are presented in Table 2. The sentences of the basic M1-5/F1-5 classes were processed using four types of voice de-identification methods: $M \rightarrow F$, $F \rightarrow M$, *Younger*, and *Older*, further called $Vtran1-4$. Subsequently the transformation using the inverse parameters (9) with the same algorithm as the forward transformation was applied (called $Ivtran1-4$). Finally the original speech signal without modification was re-synthesized for comparison (called $Resyn0$). The structure of the speech corpus, the composition of the speakers, and the recording time durations in the CZ&SK-DB are similar to those in the German Emo-DB (100 sentences with total duration of about 400/350 s for male/female voices) being a free public database from which only the sentences in a neutral speaking style uttered by 5 male and 5 female speakers were extracted. The TIMIT speech database in English is often used for recognition/identification comparisons. It consists of sentences in a neutral style only, so it is taken for the second comparison with the CZ&SK-DB. As in the case of the Emo-DB, we chose 100 sentences uttered by 5 males and 100 sentences by 5 female speakers with the total duration of 500/400 seconds for male/female voices. The speech material was processed per frames with the duration selected in correlation with the speaker's mean F_0.

3.2 Description of performed experiments and obtained results

The functionality of the proposed age and gender de-identification is verified by the developed GMM-based original speaker recognizer. Then, the main comparison is performed consisting of the GMM re-identification of the original speaker from the sentences with the applied voice de-identification. For the best performance of the GMM-based original speaker recognizer, two classification tasks were performed:

(1) Ability of the tested GMM classifier to detect the speakers who had been included in the basic groups of the originals (sentences $Orig1-5$ from the speakers $M1-5/F1-5$),

(2) Analysis of the sentences $Norig1-5$ from the speakers who had not been included in any of the basic groups of the originals ($M6-10/F6-10$) and should be classified into one resulting output class called *Not defined*.

The main original speaker re-identification experiment was based on analysis of sentences with gender and age transformation applied on the original male and female speakers. The reversibility of the de-identification method for the authorized user knowing the transformation parameters was tested by the inverse transformation following the forward voice transformation and further classification of the resulting speech signal. For comparison, each of the four types of de-identified signals was inverse-transformed using all four types of voice de-identification methods. First, only the sentences resynthesized without modification were tested and classified to know the influence of the speech reconstruction method and its cepstral parameters on the original speaker detection accuracy. Five types of analysis and comparison were carried out:

- detailed analysis of the GMM original speaker detection accuracy from the sentences with applied sole resynthesis ($Resyn1-5$ for male/female voices)-see a summary bar-graph for both genders in Fig. 5(a),
- analysis of the original speaker detection accuracy for each of the four transformed voice types (merged with

Table 2. Male and female speakers' age types, mean F_0, number of sentences, and their mean durations (CZ&SK-DB speech corpus)

Voice/Speaker	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Age type-M	young	adult	senior	senior	adult	adult	adult	adult	senior	young
$F_{0\,\mathrm{mean}}(Hz)$-M	133	117	132	98	103	88	119	112	97	111
Number/Dur(s)	30/90	20/80	15/60	9/55	15/100	20/80	15/85	21/120	12/70	17/90
Age type-F	young	adult	adult	adult	adult	senior	senior	adult	senior	young
$F_{0\,\mathrm{mean}}$ (Hz)-F	228	215	177	198	207	170	165	195	150	217
Number/Dur(s)	18/60	30/70	8/55	8/60	15/60	15/60	18/80	17/75	15/65	20/95

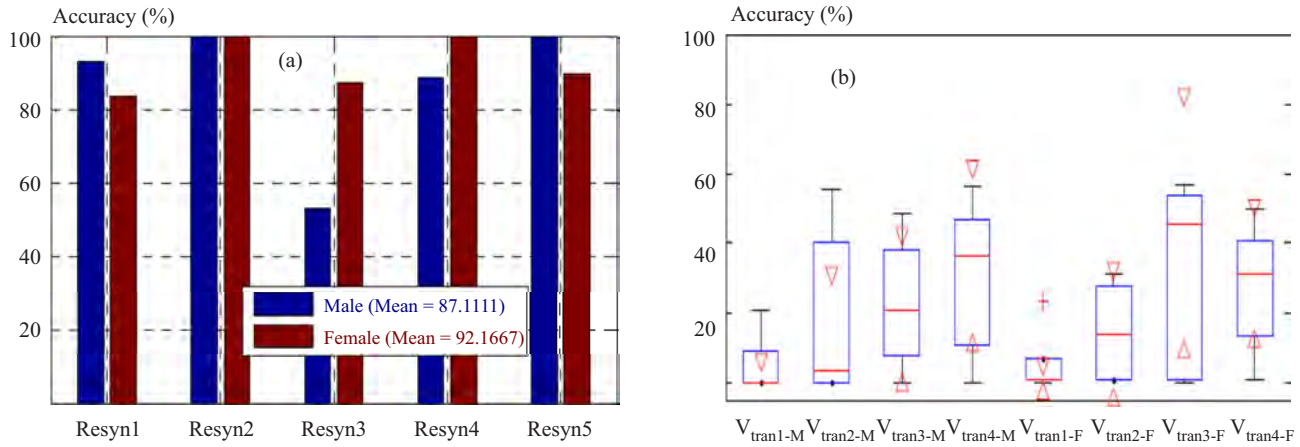

Fig. 5. Graphical results of the main original speaker re-identification experiment: (a) – summary bar-graph for male/female voices after applied sole resynthesis, (b) – summary box-plot of basic statistical parameters of the original speaker detection accuracy for male/female voices

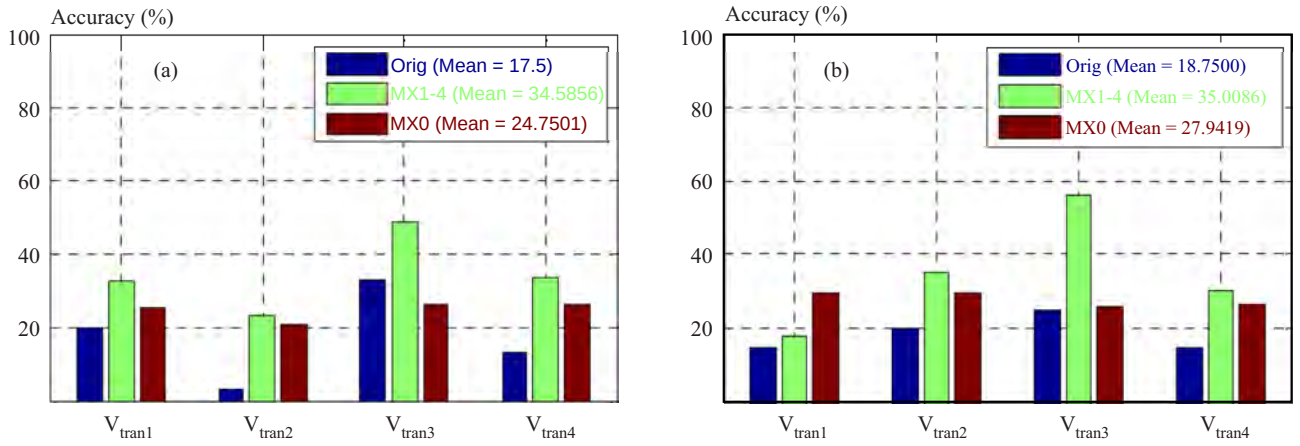

Fig. 6. Influence of parameters of transformations $Vtran1-4$ on the resulting original speaker detection accuracy: bar-graph for the speaker (a) – M1 and (b) – F5: used speaker models: original ($Orig$), modified by individual transformations ($MX1-4$) and by all transformations together ($MX0$)

Table 3. Partial results of the main original speaker re-identification experiment for male speaker M2 using the transformations $Vtran1-4$ and the sole resynthesis $Resyn0$

Transformation	Type/Evaluated in %		
	Class2	Class4	not defined
$Vtran1$	90	10	0
$Vtran2$	100	0	0
$Vtran3$	50	45	5
$Vtran4$	100	0	0
$Resyn0$	100	0	0

Class1,Class3, Class5 – all zeros

Table 4. Partial results of the main original speaker re-identification experiment for male speaker M2 using the transformations $Vtran1-4$ and the sole resynthesis $Resyn0$

Transformation	Type/Evaluated in %		
	Class2	Class5	not defined
$Vtran1$	30	30	40
$Vtran2$	17	80	3
$Vtran3$	47	47	6
$Vtran4$	90	10	0
$Resyn0$	100	0	0

Class1,Class3, Class4 – all zeros

the sole resynthesis for comparison)-see partial results for the speakers M2/F2 in Tables 3 and 4, and basic statistical parameters of the results for both genders presented in the form of a box-plot graph in Fig. 5(b),

- analysis of the resulting original speaker detection accuracy using GMMs re-trained with additional data of speech parameters from the corresponding transformation ($MX1-4$) and from all transformations together ($MX0$)-see bar-graph comparison of partial results for speakers M1/F5 in Fig. 6 and summarised mean values for all speakers in Table 5,
- analysis of reversibility of the used de-identification method by inverse transformation after conversion to four gender/age types-see partial results of the original speaker identification accuracy for male and female speakers in Tables 6 and 7, and summary results for

Table 5. Mean original speaker detection accuracy in % with standard deviation for different models; summarized for $Vtran1-4$

GMM model type/ Speaker group	Male M1-5	Female F1-5	Total
Original	18.4 (12.6)	18.8 (16.7)	18.6
Modified by			
MX1-4	34.6 (5.3)	40.9 (9.7)	37.8
MX0	22.7 (1.7)	25.1 (1.8)	23.9

four gender/age transformations together with the inverse ones and the resynthesis only (for comparison) for all five male/female speakers in Table 8,

- final comparison of the original speaker detection accuracy for the utterances from CZ&SK-DB with the results obtained for the utterances from Emo-DB and TIMIT-see mean values of the achieved accuracy summarized in Table 9.

The speaker detection accuracy from among of the five basic classes of the original speakers $Orig1-5$ and the sixth class *Not defined* for the $Norig$ speaker was calculated in a simple way for each of the output classes from X_A sentences with the correctly identified original speaker and the total number N_U of the tested sentences as $(X_A/N_U)\times 100$ (%). The basic functions from the Ian T. Nabney $Netlab$ pattern analysis toolbox version 3.3 [23] were used for implementation of the GMM functions. The current realization of GMM original speaker detection was implemented in the Matlab environment, so the re-identification phase could be real-time processed after optimization and final use of the Matlab compiler.

4 Discussion of obtained results

The performed auxiliary analysis shows the importance of proper selection of input features for the GMM-based original speaker detection as well as correct setting of a basic threshold for the open-set classification. The best performance of the proposed GMM classifier was achieved with the feature vectors formed by 16 speech features containing the first four cepstral coefficients together with the spectral and basic prosodic speech features. The optimum basic threshold for the open-set classification was higher by 20% comparing to the basic level given calculated as P/M. The best classification accuracy was achieved with 128 mixtures in combination with the full covariance matrix in spite of higher computation complexity than for lower number of Gaussian mixtures and the standard simple diagonal covariance matrix.

The results obtained by testing the sentences of the sole resynthesis showed that determination of F_0 was critical. For the precise and correct speech parameterization the resynthesized signal is very similar to the original speech and the accuracy of the original speaker detection approaches 100%-see Fig. 5(a). Otherwise, the original speaker detection may decrease to the value lower than 50% due to the process of speech signal parameterization

Table 6. The best results of the original speaker detection accuracy obtained from the analysis of reversibility of the used de-identification method $Ivtran1-4$ together with the sole resynthesis $Resyn0$ for the male speaker M5 (Class5)

Transformation	Type/Evaluated in %	
	Class5	not defined
$Ivtran1$	87	13
$Ivtran2$	93	7
$Ivtran3$	100	0
$Ivtran4$	93	7
$Resyn0$	100	0

Class1 to Class4 – all zeros

Table 7. The best results of the original speaker detection accuracy obtained from the analysis of reversibility of the used de-identification method $Ivtran1-4$ together with the sole resynthesis $Resyn0$ for the male speaker F4 (Class5)

Transformation	Type/Evaluated in %		
	Class1	Class4	not defined
$Ivtran1$	15	63	22
$Ivtran2$	0	100	0
$Ivtran3$	0	100	0
$Ivtran4$	0	75	25
$Resyn0$	0	100	0

Class2, Class3, Class5 – all zeros

Table 8. Mean original speaker detection accuracy in % with standard deviation for transformations, their inversions, and resynthesis

Speaker/ transformation type	$M\to F$ & $F\to M$	$F\to M$ & $M\to F$	Younger & Older	Older & Younger	Resynthesis only
Male Orig1-5	67.2 (11.8)	73.1 (15.9)	88.9 (12.4)	62.2 (19.9)	87.1 (12.5)
Female Orig1-5	73.6 (22.2)	75.1 (22.3)	78.0 (20.5)	77.1 (18.4)	92.2 (13.1)
Total	70.4	74.1	83.45	69.65	89.65

Table 9. Mean original speaker detection accuracy in % with standard deviation for de-identification types and speech databases

Database/ transformation type	$M \rightarrow F$	$F \rightarrow M$	Younger	Older	Total
CZ&SK male	5.1 (8.6)	17.9 (13.8)	21.9 (18.2)	28.7 (22.7)	18.4
CZ&SK female	5.2 (9.5)	14.1 (13.3)	29.9 (26.8)	26.1 (18.0)	18.8
Emo-DB male	6.2 (4.8)	6.6 (7.6)	20.1 (18.3)	3.0 (1.3)	8.9
Emo-DB female	11.4 (18.9)	12.7 (16.4)	17.8 (18.5)	15.1 (19.8)	14.3
TIMIT male	2.6 (4.2)	12.6 (16.2)	18.6 (17.2)	12.5 (17.5)	11.6
TIMIT female	13.9 (15.7)	10.2 (16.7)	20.9 (24.6)	10.8 (18.2)	13.9
Total	7.40	12.35	21.53	16.03	

itself. It can heavily distort main results of the original speaker detection from the speech with the applied gender/age conversion. In correspondence with our assumption, modification of GMMs by additional data from the transformed speech caused increase of the resulting original speaker detection accuracy and equalizing of differences between the results for different transformations as documented in Fig. 6. This effect is greater using additional data from all four voice transformations. If only the data from the current transformation are used, the differences are partly preserved; however, the accuracy is twofold. Both these effects are negative for our purposes-the original speaker should not be detected from the well de-identified speech. Next, as we expected, the two-fold cepstral analysis and parameterization for the forward and inverse transformation combination degrades the resulting speech synthesis to some extent. The worst results of the original speaker detection accuracy were obtained for the gender transformation $Ivtran1$ (67/74% for male/females speakers) with the greatest spectral changes. The best results (very close to the pure resynthesis) corresponded to the age transformation $Ivtran$. Furthermore, the original speaker identification accuracy was decreased not due to exchange between speaker classes but due to wrong detection as the virtual class *Not defined* – compare numerical values in Table 6 and 7. Finally, the summary mean original speaker detection accuracy is still comparable with the one obtained from the applied pure resynthesis-see Tab. 8.

The precise setting of the speech parameters is essential for the satisfactory original speaker detection accuracy from the sentences with applied four speaker de-identification types. The lower the achieved original speaker detection accuracy, the higher the successfulness of the performed speaker de-identification. Table 9 shows its values lower than 6 % for the $M \rightarrow F$ transformation using the CZ&SK-DB (both voices), the TIMIT (male voice), and for the *Older* transformation using the Emo-DB (male voice). On the other hand, the effectiveness of the applied age/gender transformation is low for relatively high original speaker detection accuracy (over 50%) as shown in Fig. 5(b) in the case of the $Younger$ transformation using the female voice F2 from CZ&SK-DB. From these findings we can conclude that the applied voice changes must respect the basic vocal tract parameters and the speech characteristics of the original speaker used in the de-identification process.

5 Conclusion

From the main point of view, the basic motivation task of finding the alternative to the standard listening tests was fulfilled. The proposed and tested evaluation method based on GMM approach is functional and produces usable results. The main disadvantage of the human evaluation lies in its subjectivity, lack of reproducibility (different obtained results for repeated tests), and dependence on environment conditions. The human subjective evaluation is more complex, *eg* there is also great influence of the current emotional state, the simultaneously perceived visual stimuli, etc. It is always different from automatic approaches based on statistical processing of speech features (spectral, prosodic, *etc*). On the other hand, the stable automatic classification process (with well-trained GMM models, correctly selected and used speech features, sufficient volume of the processed speech data, and so on) gives the same results in the repeated experiments independently of environment conditions.

The obtained results of the basic original speaker detection experiment were in correspondence with the proposed working hypothesis about lowering of the achieved GMM recognition score for greater degree of de-identification by spectral and prosodic speech modifications. However, the specificity of this research and comparison lies in the fact that we primarily use the speech material composed of the sentences uttered by the original Czech and Slovak speakers. The parameters used for the gender and age conversion were proposed upon the phonetic research [13-15]. The basic comparison of the original speaker detection using the German and English speech showed irrelevance of the language for determination of the re-identification accuracy.

In future we plan to test successfulness of our de-identification method with the help of automatic speaker verification/identification systems based on statistical approach based on universal background model (UBM)

models or other favoured methods such as support vector machines, HMM, *etc.* Although the GMM-UBM has often been used as a baseline to be compared with, in our research only a very small number of speakers can be used, so such a comparison cannot be done. Therefore we must first collect a larger speech databases in Czech and Slovak as well as in other languages.

Acknowledgements

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, VEGA 1/0905/17) and the Czech Science Foundation (GA16-04420S).

References

[1] S. Ribaric, A. Ariyaeeinia and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey", *Signal Processing: Image Communication*, 2016, 47, 131–151.

[2] A. Sayadian and F. Mozaffari, "A novel method for voice conversion based on non-parallel corpus", *International Journal of Speech Technology*, 2017, 20, (3), 587–592.

[3] H. Valbret, E. Moulines and J. P. Tubach, "Voice transformation using PSOLA technique", *Speech Communication*, 1992, 11, (2-3), 175–187.

[4] Q. Jin, A. R. Toth, T. Schultz *et al*, "Voice convergin: Speaker de-identification by voice transformation", *Proc. 2009 IEEE Int. Conf. Acoustics*, Speech, and Signal Processing (ICASSP 2009), Taipei, Taiwan, April 2009, pp. 3909–3912.

[5] T. Justin, V. Štruc, S. Dobrišek *et al*, "Speaker de-identification using diphone recognition and speech synthesis", *Proc. 11th IEEE Int. Conf. and Workshops Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–7.

[6] M. Faundez-Zanuy, E. Sesa-Nogueras and S. Marinozzi, "Speaker identification experiments under gender de-identification", *xperiments under gender de-identification. Proc. 49th Annual IEEE Int. Carnahan Conf. Security Technology* ICCST 2015, Taipei, Taiwan, September 2015, pp. 309–314.

[7] C. Magarinos, P. Lopez-Otero, L. Docio-Fernandez *et al*, "Reversible speaker de-identification using pre-trained transformation functions", *Computer Speech and Language*, 2017, 46, pp. 36–52.

[8] M. Abou-Zleikha, Z. -H. Tan, M. G. Christensen *et al*, "A discriminative approach for speaker selection in speaker de-identification systems", *Proc. 23rd European Signal Processing Conf. (EUSIPCO 2015)*, Nice, France, August 2015, pp. 2102-2106.

[9] R. Vích, J. Přibil and Z. Smékal, "New cepstral zero-pole vocal tract models for TTS synthesis", *Proc. IEEE Region 8 EUROCON'2001; vol. 2*, Section S22-Speech Compression and DSP, Bratislava, Slovakia, July 2001, pp. 458–62.

[10] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Transactions on Speech and Audio Processing*, 1995, 3, 72–83.

[11] F. Burkhardt, A. Paeschke, M. Rolfes *et al*, "A database of German emotional speech", *Proc. 9th European Conf. Speech Communication and Technology (INTERSPEECH 2005)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[12] P. Klosowski, A. Dustor and J. Izydorczyk, "Speaker verification performance evaluation based on open source speech processing software and TIMIT speech corpus", *P. Gaj et al* , Communications in Computer and Information Science 522 (Springer International Publishing Switzerland, 2015), pp. 400–409.

[13] M. Fleischer, S. Pinkert, W. Mattheus *et al*, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall", *Biomechanics and Modeling in Mechanobiology*, 2015, 14, (4), 719–733.

[14] M. P. Gelfer and Q. E. Bennett, "Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender", *Journal of Voice*, 2013, 27, (5), 556–566.

[15] K. Pisanski, B. C. Jones, B. Fink *et al*. "Voice parameters predict sex-specific body morphology in men and women", *Animal Behaviour*, 2016, 112, 13–32.

[16] U. Reubold, J. Harrington and F. Kleber, "Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers", *Speech Communication*, 2010, 52, (7-8), 638–651.

[17] C. M. Bishop, "Pattern Recognition and Machine Learning", *Springer,*.

[18] G. Muhammad and K. Alghathbar, "Environment recognition for digital audio forensics using MPEG-7 and mel cepstral features", *Journal of Electrical Engineering*, 2011, 62, (4), 199–205.

[19] J. Přibil and A. Přibilová, "GMM-based evaluation of emotional style transformation in Czech and Slovak", *Cognitive Computation*, 2014, 6, (4), 928–939.

[20] J. Přibil and A. Přibilová, "Comparison of text-independent original speaker recognition from emotionally converted speech", A. Esposito *et al*, Smart Innovation, Systems and Technologies 2016, 48, pp. 137–149.

[21] J. Přibil an d A. Přibilová, J. Matoušek, "GMM-based speaker age and gender classification in Czech and Slovak", *Journal of Electrical Engineering*, 2017, 68, (1), 3–12.

[22] B. Božilovic, B. M. Todorovic and M. Obradovic, "Text independent speaker recognition using two-dimensional information entropy", *Journal of Electrical Engineering*, 2015, 66, (3), 169–173.

[23] I. T. Nabney, "Netlab Pattern Analysis Toolbox, Release 3", http://www.aston.ac.uk/eas/research/groups/ncrg/resources /netlab/downloads, accessed 2 October 2015.

Received 14 November 2017

Jiří Přibil (Ing, PhD), born in 1962 in Prague, Czechoslovakia, received MSc degree in computer engineering in 1991 and 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. Research interests: signal and image processing, speech analysis and synthesis, and text-to-speech systems.

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

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