Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream

Jan SILOVSKÝ [1], *Jan NOUZA* [1, 2]

[1] SpeechLab, Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czechia
[2] Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic, Prague, Czechia

jan.silovsky@tul.cz, jan.nouza@tul.cz

Abstract. *This paper presents a set of techniques for classification of audiosegments in a system for automatic transcription of broadcast programs. The task consists in deciding a) whether the segment is to be labeled as speech or a non-speech one, and in the former case, b) whether the talking person is one of the speakers in the database, and if not, c) which gender the speaker belongs to. The result of the classification is used to extend the information provided by the transcription system and also to enhance the performance of the speech recognition module. Like the most of the state-of-the-art speaker recognition systems, the proposed one is based on Gaussian Mixture Models (GMM). As the number of the database speakers can be large, we introduce a technique that speeds up the identification process in significant way. Furthermore, we compare several approaches to the estimation of GMM parameters. Finally, we present the results achieved in classification of 230 minutes of real broadcast data.*

Keywords

Speaker recognition, Gaussian mixture models, broadcast speech transcription

1. Introduction

There is a growing interest in media mining systems, namely those that can process also audio data, such as TV and radio news, political debates, talk-shows, etc. The main goal is to transcribe records of spoken data. However, the information about who is speaking is also important.

For the voice identification of talking persons several basic approaches have been developed during the last 20 years. Unfortunately, many of them cannot be used directly in automatic broadcast transcription (ABT) systems. The reason is that speaker identification in an ABT task is complicated by several factors:

- Broadcast stream contains not only speech, but also music and other sounds;
- Automatic segmentation of audiostream is not always able to detect exactly the moments where speech begins and when a new speaker starts to speak;
- Broadcast speech varies widely with microphones, transmission channels and background noise;
- The number of speakers that occur in TV and radio programs is very large (almost unlimited), the task must be solved as recognition within an open set.

There is also another aspect that must be taken into account. The goal of the speaker recognition module in an ABT system is not only to output the name of the most probable person but also to provide information that is essential for the proper function of the consequent task, which is speech recognition. As the state-of-the-art recognisers employ speaker-adapted acoustic models, a wrongly identified speaker name will cause that an incorrect model is used. In most cases it will lead to the degradation of the recogniser's performance.

2. Task and Its Analysis

The task we want to solve can be defined as follows:

Let us have a segment of acoustic data, which was cut out of a broadcast stream by a segmentation routine e.g. that described in [1]. We want to decide whether the segment contains speech and if yes who was the speaker. If the speaker cannot be identified securely we want to know at least his or her gender, because that information is essential for the speech recognition module.

If we analyze all the possible situations we get the diagrams depicted in Fig. 1 and Fig. 2, where MaleA, MaleB and FemaleA are registered speakers and MaleX is an unknown speaker. It's obvious that the overall performance is affected by several circumstances. Therefore, we distinguish and evaluate the following rates:

- Speech/non-speech error rate R_{SNSE} – corresponds to the situations when speech segments were labeled as non-speech (situation 3) or vice versa (4).

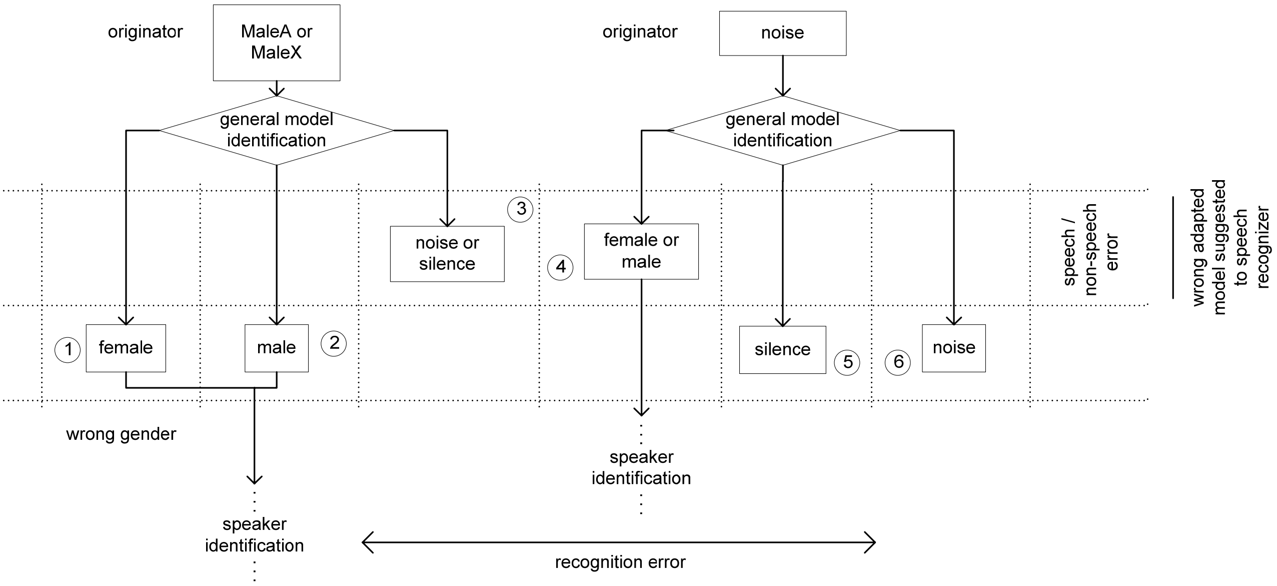

Fig. 1. Illustration of situations which can occur within general model identification.

- Gender recognition error rate R_{GDE} – reflects the situations when speech segments were correctly recognized as speech but a male speaker was recognized as female (1) or vice versa.
- Speaker identification error rate R_{SIE} – is related to the situations when a speaker from the database is wrongly identified (9 or 10).
- Equal error rate R_{EER} – it is known that the number of incorrectly rejected (8) and incorrectly accepted (9 or 11) identities varies with the verification threshold. The equal error rate corresponds to such a threshold value that makes the number of false acceptances equal to the number of false rejections.
- Global recognition error rate R_E – unlike all the previously mentioned (partial) rates, this rate reflects the overall system performance from the user's point of view. The following three results are regarded as correct: a) a non-speech segment is labeled as non-speech and is properly classified into given sub-categories (6), b) a speech segment of an unknown speaker is labeled as speech, his/her gender is correctly recognized and the verification module rejects the identity proposed by the speaker identification module (12b), c) a segment belonging to a reference speaker is labeled as speech and the speaker is correctly identified and verified (7).
- Wrong speech adaptation model rate R_{SAE} – this rate reflects the (dis)ability of speaker recognition to facilitate speech recognition by utilizing speaker adapted models. This rate also evaluates the overall performance, though in a less strict way. The following situations are considered correct: a) a non-speech segment recognized as non-speech (5 or 6), regardless of the noise sub-categories (since non-speech segments are not processed in further steps), b) for an unknown speaker, his/her gender is found correctly (10b), c) for the reference speaker, his/her identity is identified but not necessarily verified. This is still acceptable for the proper function of the model adaptation module.

Since our proposed solutions should enhance the overall performance of the complete ABT system, the most relevant rates are R_E and R_{SAE}. However, all the other rates provide information that has its practical value and therefore we evaluate them as well.

3. Speaker Recognition Using GMM

In [2] we studied and compared methods based on both VQ (Vector Quantization) and on GMM (Gaussian Mixture Model). In this paper we will report only the latter approach because it yields better results. It should be also noted that we use the same technique for identifying noise and non-speech segments as well as for recognizing individual speakers and therefore, in further text, the term 'speaker recognition' should be considered in this more general view.

For n-dimensional feature vector x, the Gaussian mixture density used for the likelihood function is defined as

$$P(x|\lambda)=\sum_{l=1}^{L} w_l P_l(x) \ . \tag{1}$$

The density is a weighted linear combination of L unimodal Gaussian densities $P_l(x)$. λ represents speaker model, parameterized by mixture weights w_l, mean vectors μ_l and covariance matrices Σ_l (in general full, but most often only

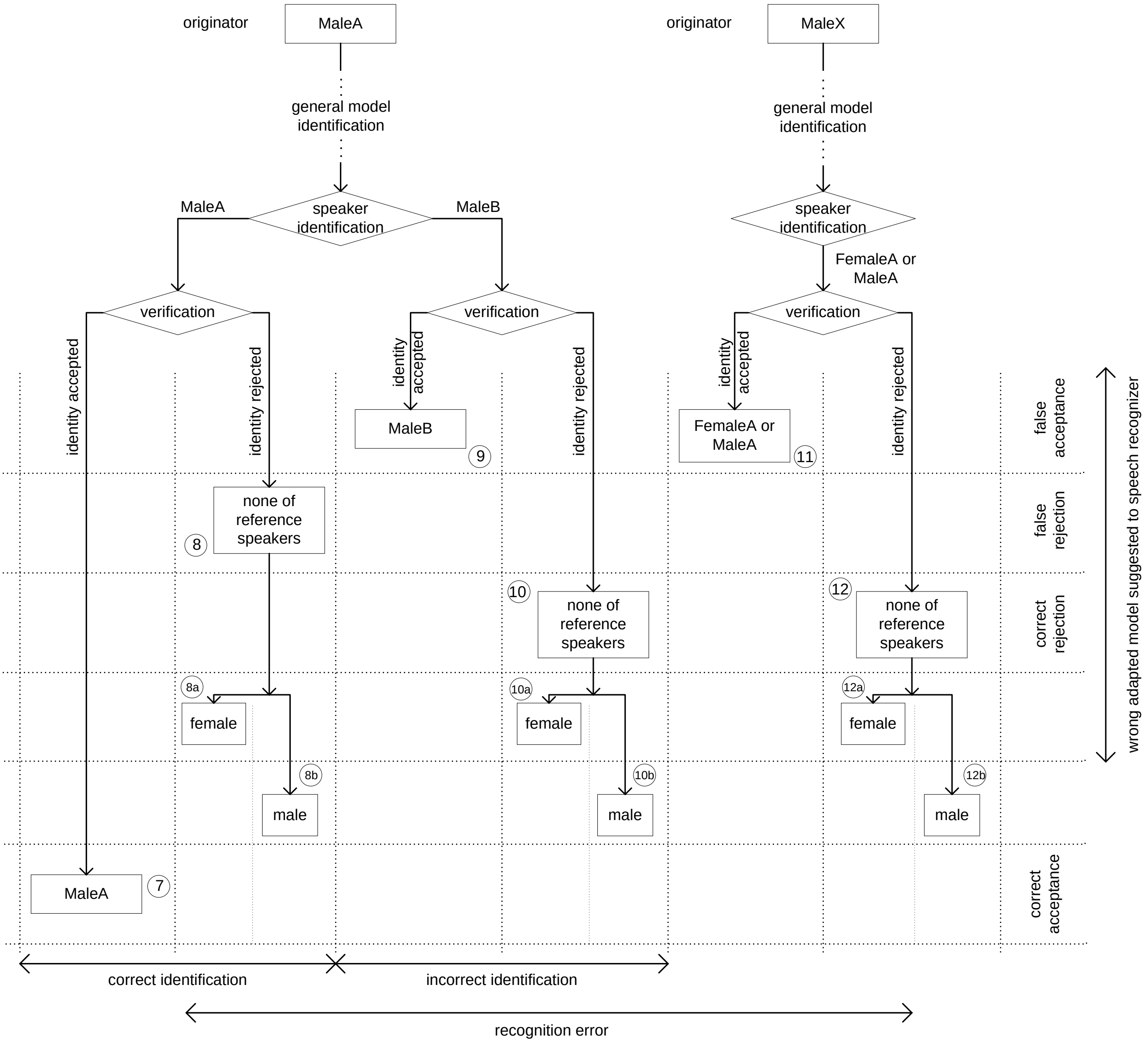

Fig. 2. Illustration of situations which can occur within speaker identification.

diagonal). Density $P_l(x)$ is defined as

$$P_l(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_l}} \exp\left(-\frac{1}{2}(x-\mu_l)' \Sigma_l^{-1} (x-\mu_l) \right) . \quad (2)$$

3.1 Speaker Identification

Let $X = \{x_1 \ldots x_T\}$ be a sequence of feature vectors representing a parameterized signal. We suppose mutual independence of feature vectors of X and then we can compute log-likelihood as

$$P(X|\lambda) = \sum_{t=1}^{T} \log P(x_t|\lambda) . \quad (3)$$

When applying the maximum log-likelihood classifier, the reference speaker s^* is proclaimed as *originator* according

$$s^* = \arg\max_s P(X|\lambda^s). \quad (4)$$

3.2 Speaker Verification

The decision whether to accept or reject the proposed identity s^* is based on the log-likelihood ratio test. A universal background model (UBM) [3] trained on data pooled from many speakers is employed to represent acoustic space of imposters. Identity s^* is accepted if

$$\frac{1}{T}\left(P\left(X|\lambda^{s^*}\right) - P\left(X|\lambda^{UBM}\right)\right) > \theta \ , \quad (5)$$

where θ is the verification threshold, otherwise is rejected. If multiple UBMs are used (e.g. those tailored to male and female voices), a suitable model employed for normalizing verification score can be chosen according to [4]

$$P\left(X\middle|\lambda^{UBM}\right)=\max\left\{P\left(X\middle|\lambda^{1}\right),P\left(X\middle|\lambda^{2}\right)\right\}. \qquad (6)$$

4. Proposed System and Its Further Improvement

The scheme of the proposed system is depicted in Fig.3. First, identification of a general model is performed. Four general models were trained: noise, silence, male and female voices. Such a set of models allows for performing speech/non-speech decision as well as gender identification at one step. Next, if the segment is recognized as speech (i.e. male or female voice), speaker identification starts. The task of open-set identification is split into successive tasks of close-set identification and verification of the proposed identity. If the verification step rejects the proposed speaker, the gender identified in the first step is declared as the result of recognition.

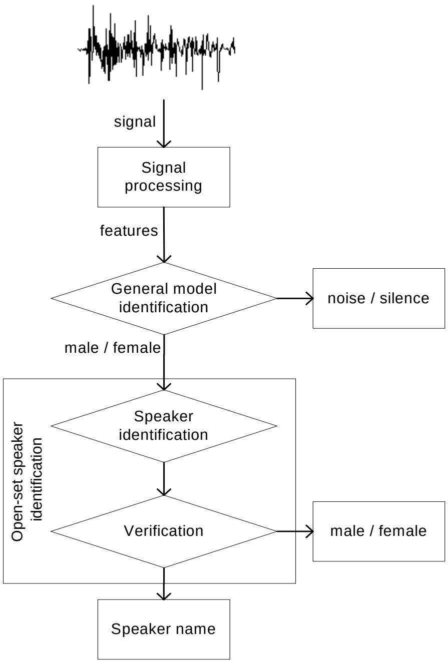

Fig. 3. Scheme of the proposed speaker recognition module.

An important question is: how to set the verification threshold. Speech recognition yields best results when utilizing proper speaker adapted (SA) models. For an unknown speaker, a proper gender dependent (GD) model is still significantly better than the speaker independent (SI) one. On the other side, speech recognition usually degrades if it uses a model adapted for an improper speaker. It was already mentioned that errors caused by false acceptance of a wrongly identified speaker and false rejection of the correctly identified speaker are equal from the user's point of view. But from the speech recognition point of view, false rejection will not harm recognition, irrespective whether the speaker was identified correctly or not, because the GD model is employed for speech recognition. That is why it should be better to set a rather higher verification threshold. Yet, in the evaluation experiments described further, we set it in compliance with the equal error rate (EER), in order to allow some comparisons.

4.1 Baseline System

In the baseline scenario, both speaker and general GMMs were trained using the Maximum Likelihood (ML) method and the EM (Expectation-Maximization) algorithm. All had 256 components. Feature vectors were formed from 12 static MFCCs (zeroth cepstral coefficient c_0 was excluded). These 12 features were a subset of the 39 MFCCs used for speech recognition. No limit was applied to the maximum amount of training data for speaker GMMs. For training the male and female general models, the limit was set to maximum of 200 s per speaker, in order to avoid model biasing. Model likelihoods were computed over all frames of each signal segment according to eq. (3).

4.2 Silence Frame Removal

For some speakers, silence can make significantly long parts of their utterances but it does not contain any speaker characteristic information. Thus, many state-of-the-art systems aim at removing silence frames. We adopted the approach based on unsupervised learning of a bi-Gaussian model [5]. Due to distinct nature of speech and silence frames, one Gaussian should represent speech frames, the other silence ones. For this classification task, complete feature vectors (39 MFCCs) were used. It should be noted that the Bi-Gaussian model has to be employed during both training and recognition and new speaker GMMs and UBMs have to be retrained with only speech frames.

4.3 Reduction of Frame Flow

Observing the fact, that segment lengths differ in wide range (from 1 s up to 100 s), we studied the possibility to reduce the computational cost by evaluating the likelihood over a limited number of frames of the segment, obviously without any notable degradation of recognition rate. The most promising approach consists in computing the likelihood over F frames equally distributed across the entire segment length T, according to

$$P\left(X\middle|\lambda\right)=\sum_{f=1}^{F}\log P\left(x_{f\frac{T}{F}}\middle|\lambda\right). \qquad (7)$$

If the number of frames in segment is lower than F, likelihood $P(X|\lambda)$ is computed in standard way (eq. (3)).

4.4 Verification with MAP-Made SA Models

In several recently published papers, slight improvements in speaker verification were reported when models had been adapted from UBM by maximum a posteriori (MAP) method. Unlike the maximum likelihood method, the MAP supposes, that model parameters are variables with a priori known distribution. In our case, this a priori information is represented by a well trained UBM model. A speaker model adapted by the MAP method is derived from the gender-specific UBM [3].

4.5 Two-Level Classification

Combination of classical GMM likelihood evaluation and majority voting rule for single frames [6] could be utilized for speaker identification. Each frame can be considered to be a subject of an independent classifier. In the standard GMM likelihood computation, probabilities from these classifiers are combined by multiplication or summation in log domain according to eq. (3). However, it was observed [6] that if a sequence of feature vectors contains frames lying on the tail of distribution defined by the GMM, then even a few of such vectors can dominate and have negative impact on the final likelihood. Frame voting is used to assure a more equal contribution of frame classifiers to the final result. According to the maximum probability approach, each frame casts a vote to a particular speaker according to

$$s_t^* = \arg\max_s P\left(x_t \middle| \lambda^s\right) . \quad (8)$$

The votes are collated and the speaker with the highest number of votes is the winner. As the number of available votes is equal to the number of frames, it is convenient to perform the frame voting scheme only for a limited number of speakers in order to avoid spreading of the votes among many speakers. Thus, frame voting is performed as the second pass of speaker identification only for *N* best speakers selected by classical GMM likelihood evaluation.

5. Experimental Evaluation

5.1 Database Used for Training and Testing

The system was evaluated on our own large database of Czech broadcast programs (news, debates, talk-show etc.). It contains more than 31 hours of speech data from more than 800 speakers and it has been collected during the last 5 years. The amount of data from individual speakers differs in large scale (from several seconds to more than 1000 s), which is typical for broadcast programs. Records were sampled at 16 kHz, 16 bits parameterized into 39 classic MFCC features.

Minimal amount of 75 s training data was set as criterion for including a person into the speaker database. This threshold was passed by 306 subjects, mainly news presenters, radio and TV reporters and major politicians.

The test part contained 230 minutes of broadcast news records from 255 speakers. Most of them were listed in the database, but not all. The data was further split into a development set (75 minutes), which was used to estimate free parameters (such as the verification threshold), and evaluation set (155 minutes).

5.2 Performance of Baseline System

Table 1 shows all evaluated rates that were achieved for the baseline system. The results proved that the proposed general models were able to provide reliable identification of speech and non-speech segments with error rate lower than 1 %. Also gender identification yielded similarly good results.

speech / non-speech error rate R_{SNSE}	**0.78%**
gender recognition error rate R_{GDE}	**0.99%**
speaker identification error rate R_{SIE}	**10.03%**
equal error rate R_{EER}	**12.50%**
recognition error rate R_E	**19.18%**
wrong speech adapted model suggestion rate R_{SAE}	**7.34%**

Tab. 1. Baseline system results.

5.3 Effect of Silence Frame Removal

To our surprise, the most important rate R_E increased after we applied the silence frame removal technique, as shown in Table 2. It happened even though the bi-Gaussian model was able to mark the silent frames rather reliably. Most probably, the number of speech frames that were also removed, was larger than we expected and this caused the small degradation. It shows us that we have to focus on better application of the technique.

Silence frames	R_{SNSE} [%]	R_{GDE} [%]	R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
kept	0.78	1.00	10.04	12.50	19.18	7.34
removed	0.49	1.09	11.01	12.89	20.94	7.34

Tab. 2. Influence of silent frames removal.

5.4 Effect of Frame Flow Reduction

The test records varied much in their length as it is usual in broadcast speech. Minimum was 48 frames, maximum about 5500 frames, with average length of 900 frames. Experiments were performed for the *F* value in eq. (7) ranging from 5 to 600.

Table 3 and Figure 4 summarize the achieved results. They show that the R_E rate increases rapidly only for $F < 100$. Larger number of frames has almost no impact on its value, they only bring additional computation costs. As conclusion from this experiment, we decided to use $F=150$ as the fixed value for the further experiments. This assured the R_E rate of 19.08 % and made the recognition process 6 times faster.

5.5 Verification with MAP Speaker Models

The fast scoring verification technique described in [3] was applied in this experiment. Only GMM means were adapted for new models, weights and covariance matrices were taken from the UBM. The impact of the adaptation relevance factor was analyzed and is shown in Table 4. It is evident that the evaluated rates are almost independent on the adaptation relevance factor.

frame count F	R_{SNSE} [%]	R_{GDE} [%]	R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
5	1.86	5.23	47.96	27.12	49.32	22.02
10	1.17	2.50	26.98	16.99	33.66	11.45
20	0.78	1.99	19.05	13.24	24.07	9.30
30	1.08	1.90	15.02	14.45	24.46	9.10
40	0.68	1.19	13.93	12.79	21.62	7.63
50	0.59	1.49	11.62	12.65	21.04	7.24
75	0.78	1.00	12.32	13.89	20.35	7.63
100	0.59	1.09	10.74	12.40	19.37	7.24
125	0.68	1.19	10.56	12.90	19.96	7.63
150	0.78	1.19	10.21	12.50	19.08	7.05
200	0.68	1.09	10.39	12.85	19.47	6.95
250	0.68	1.09	9.86	12.40	19.86	7.83
300	0.78	0.90	9.68	13.77	18.88	7.14
400	0.78	1.00	9.51	13.77	19.57	8.02
600	0.78	1.00	10.21	12.50	19.08	7.53
unlimited	0.78	1.00	10.04	12.50	19.18	7.34

Tab. 3. Dependence of evaluated rates on frame flow.

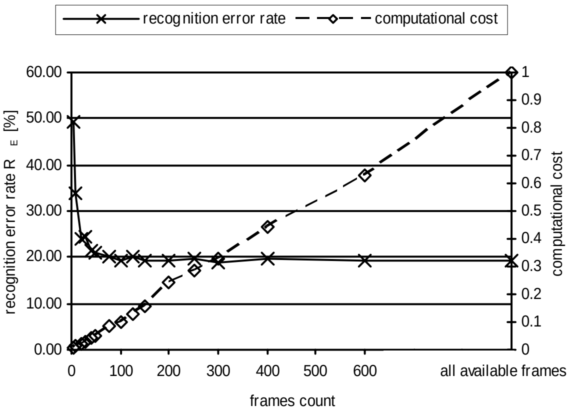

Fig. 4. Recognition rate and normalized computational cost as a function of frame flow.

adaptation relevance factor	R_{EER} [%]	R_E [%]	R_{SAE} [%]
2	18.59	24.85	10.08
6	18.59	25.15	10.27
10	18.59	25.34	10.57
16	18.79	25.15	10.47

Tab. 4. Dependence of rates affected by MAP utilization on adaptation relevance factor.

However, more important is our observation that verification employing standard ML trained models outperforms verification using the MAP adapted models, which is does not comply with literature and our expectation. Figure 5 depicts corresponding DET curves for both techniques. One possible reason is the widely varying amount of data available for each speaker and used for UBM adaptation. In the broadcast corpus the amount of data available for some speakers exceeds more than 10 times the amount available for the others. In this way, our target application differs from those referred in literature.

Models adapted using less data remain more similar to UBMs and thus the normalized verification score obtained for the correct speaker is lower than for the speaker with model adapted with large amount of data. Basically, the normal distribution of verification score matching both correct and incorrect speakers becomes more spread and more overlapping. This complicates estimation of the verification threshold and causes worse results. The R_E rate reached within this experiment 24.85 %.

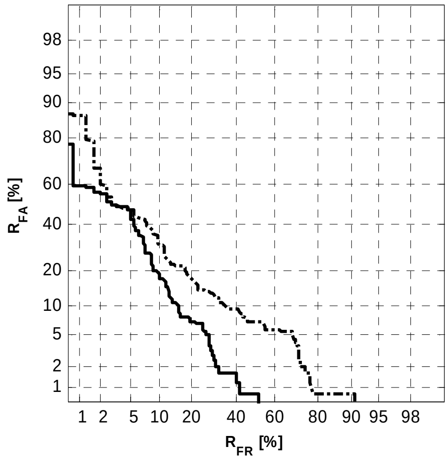

Fig. 5. DET curve for verification using models trained by ML method – solid line and MAP method – dash dot line.

5.6 Two-Level Classification

In the first pass, likelihoods were evaluated in standard way (with *F=150*). The majority frame voting rule was employed in the second pass for the *N* best speakers. Preliminary tests showed that *N=5* was the optimal value both from the performance and costs points of view. All speech frames were utilized in the second pass. Table 5 shows only the rates affected by this modification, the other rates remained unchanged. We can see that slight improvement of the R_E rate was achieved (18.69 %).

R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
11.44	12.05	18.69	6.56

Tab. 5. Evaluated rates for two-level classification.

5.7 Effect for Speech Recognition

So far we focused mainly on the improvement of the speaker recognition rate as it would be evaluated from user's point view. However, the described scheme contributes to significant improvement of speech recognition, too. If a speaker is correctly recognized and verified, his/her SA model can be used. But even though he or she is rejected (either correctly or wrongly), the ordered list of the closest speakers is very helpful for the advanced SA methods described in [7]. Here, we can benefit also from the improved R_{SAE} rate, whose best value achieved 6.56 %. The significance of applying the most proper acoustic models is demonstrated in Table 6, where we can compare WER values for speaker-independent and speaker-adapted (by method [7]) models for different broadcast programs.

program	SI models	SA models	WER relative reduction [%]
radio news	19.45	15.03	22.7
TV news	22.96	19.04	17.0
parliament debates	26.80	20.74	22.6

Tab. 6. The WER [%] for different tasks after application of speaker models adaptation.

6. Conclusions

In this paper we propose a set of techniques that deals with the problem of acoustic segment classification followed by speaker recognition and verification. We describe a modular scheme that has been successfully implemented and recently has become an essential part of a broadcast transcription platform, the first one developed for Czech.

We have proposed several modifications of existing strategies. Some led to significant improvements of speaker recognition, namely the frame flow reduction technique and the two-pass approach that combines the classic ML classifier with frame voting. On the other side, the proposed silence frame removal technique has not brought expected effects nor the MAP training procedure, at least so far. In the paper, we try to find explanation why these recently popular techniques failed in our case.

Using quite large population of 306 reference speakers and in quite complex conditions of broadcast records, we were able to achieve global recognition rate at 81 % level. Moreover, in more than 93 % cases the speaker recognition module proposed correct strategy for utilizing the proper speaker adaptation scheme and thus contributed to more than 20 % WER relative reduction in the speech recognition module.

Acknowledgements

The research was supported by the Grant Agency of the Czech Academy of Sciences (grant 1QS108040569).

References

[1] ŽĎÁNSKÝ, J., NOUZA, J. Detection of acoustic change-points in audio records via global BIC maximization and dynamic programming. In *Interspeech 2005*. Lisboa (Portugal), 2005 p. 669–672. ISSN 1018-4074.

[2] SILOVSKÝ, J. *Speaker Recognition in Broadcast Streams*. Diploma thesis (in Czech). Liberec: TU of Liberec, 2006.

[3] REYNOLDS, D.A., QUATIERI, T.F., DUNN, R.B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing 10*, 19-41, 2000.

[4] PADRTA, A., RADOVÁ, V. On the background model construction for speaker verification using GMM. *Text, Speech and Dialogue TSD 2004. Lecture Notes in Artificial Intelligence 3206.* Springer-Verlag: Berlin, Heidelberg, 2004, p. 425–432. ISBN 3-540-23049-1, ISSN 0302-9743.

[5] MARIÉTHOZ, J., BENGIO, S. *An Alternative to Silence Removal for Text-Independent Speaker Verification*. IDIAP-RR 03-51, 2003.

[6] NARAYANASWAMY, B. *Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities*. A Report in Candidacy for the Degree of Master of Science. Carnegie Mellon University, May 2005.

[7] ČERVA, P., NOUZA, J., SILOVSKÝ, J. Two-step unsupervised speaker adaptation based on speaker and gender recognition and HMM combination. In *Interspeech 2006*. Pittsburg (USA): 2006.

About Authors...

Jan SILOVSKÝ was born in Liberec in 1982. He joined the SpeechLab team at the Technical University of Liberec in 2005 as a M.E. student. His research work is focused on speaker recognition.

Jan NOUZA was born in 1957. He got his Master and PhD degrees at the Faculty of electrical engineering at the CTU in Prague. Since 1987 he has been with the Technical University in Liberec, since 1999 in the position of professor at the Department of Electronics and Signal Processeing. In 1993 he founded the SpeechLab, a research team that is involved in speech recognition, speaker recognition and other voice technology oriented research. He partly works also for the IREE ASCR in Prague and in 2006 he has been the guest professor at ETH in Zurich.