Iterative Unsupervised GMM Training for Speaker Indexing

Martin PARALIČ, Roman JARINA

Department of Telecommunications, University of Žilina, Univerzitna 1, 010 26 Žilina, Slovakia

paralic@fel.uniza.sk, jarina@fel.uniza.sk

Abstract. *The paper addresses a novel algorithm for speaker searching and indexation based on unsupervised GMM training. The proposed method doesn't require a predefined set of generic background models, and the GMM speaker models are trained only from test samples. The constrain of the method is that the number of the speakers has to be known in advance. The results of initial experiments show that the proposed training method enables to create precise GMM speaker models from only a small amount of training data.*

Keywords

Speaker indexing, training, iteration, log. likelihood, Gaussian Mixture Model (GMM).

1. Introduction

With the increasing availability of archived audio and video material an increasing need for efficient and effective means of searching and indexing through this digital content comes. One of the demanding issues is indexing speakers for faster and more convenient information retrieval or browsing through multimedia archives (e. g. TV news archives, discussions, voice mails, audio/video conversations, etc.) [6], [8], [11], [12]. This paper addresses the speaker indexing based on acoustic information analysis. Obviously, audio content based analysis is required for dealing with both audio–only archives as well as video archives [11], [12].

An output of the speaker segmentation and indexation process is metadata with information about segment boundaries and relevant labels such as speaker's index or name. Thus the speaker based searching in a multimedia document is turned to the task of searching the speaker's label in the metadata.

Such speaker segmentation and indexing is a difficult task in absence of a priori information about speakers and the speaker modeling has to be done on the fly. It becomes even more challenging when the number of speakers in a conversation is not known. Thus conventional techniques based on GMM, well known from the area of automatic speaker identification or verification, which require a great amount of training data with a priori knowledge about speakers, cannot be directly applied to speaker segmentation and indexing.

Previous research works on speaker indexing such as [2], [4], [5], [9] suggest projecting each utterance into a speaker space defined by anchor (or universal background) models which are a set of predetermined reference speaker–independent models. Each utterance is then represented by a vector of distances between the utterance and each anchor model. The distance of a speaker segment is calculated using the anchor model set, and a model with minimum distance is selected. Then the selected model is adapted via Bayesian adaptation scheme to create a particular speaker's model.

Several speaker indexation methods based on Bayesian Information Criterion (BIC) were also proposed [1], [3], [8]. But the BIC approach works well only if the speech segments are sufficiently long (i. e. speakers turns are not very frequent).

In our paper, we propose an alternative method of training GMM speaker's model without any generic background or anchor models. In the first phase, we use a very low number of training observations to gain a course initial model (iGMM) followed by preliminary indexation of the data. In the second phase, we use selected parts from these preliminary indexed segments as training data for a more complex model. An aim of the selection process is to exclude incorrectly indexed segments from the training process. The method is inspired by the work of Adami et al. [1], but in contrast to our approach, the authors in [1] always used all segments for training the BIC model and evaluated their method only on 2 speakers task.

2. Gaussian Mixture Model

We assume the audio waveform is transformed into a parametric form. The common parameterization method is the transformation to Mel Frequency Cepstral Coefficients (MFCC). Then each MFCC vector represents one observation of the signal. Here each speaker, which is defined by one complete Gaussian mixture, represents one class of data. For illustration, a feature vector distribution of one

speaker is shown in Fig. 1 (only the first two dimensions are depicted). These vectors are modeled by 3 components Gaussian Mixture Model (GMM).

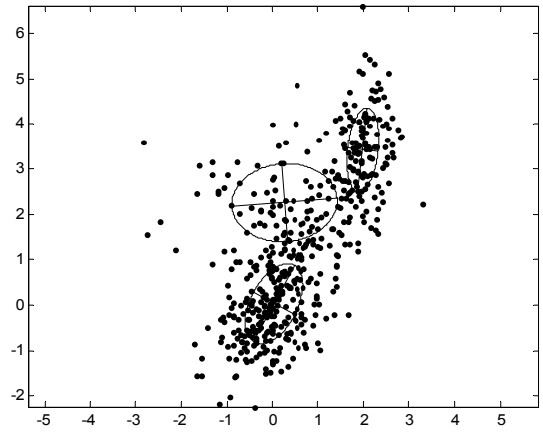

Fig. 1. Sample picture of 2 dimensional features for 3 components Gaussian mixture.

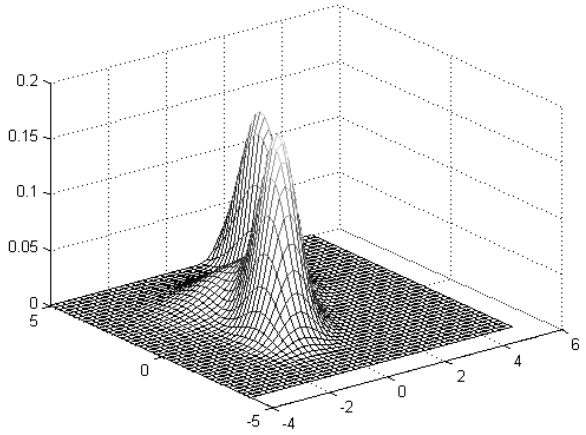

Fig. 2. Surface plot of PDF function for 2 – D features using 3 components Gaussian mixture.

In the GMM, each Gaussian component is represented by the mean, the variance and the weight of each component. Dimensions of the mean vectors and variances are the same as dimensions of the features (e. g. MFCC). An example of the 3–component GMM distribution is shown in Fig. 2. The probability density function (PDF) for d–dimensional observations is defined as follows

$$P(\boldsymbol{x} \mid \omega_k) = \frac{1}{\sqrt{2\pi}^{d} \left|\boldsymbol{\Sigma}_k\right|^{1/2}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \quad (1)$$

where $\boldsymbol{x}$ is the observation vector, ω_k is the k–th component, d is the dimension, $|\boldsymbol{\Sigma}_k|$ is the determinant of covariance matrix, $\boldsymbol{\mu}_k$ is the mean vector.

The probability of observation $\boldsymbol{x}$ in dependence on the mixture model M is computed as follows

$$P(\boldsymbol{x} \mid \mathrm{M}) = \sum_{k=1}^{K} w_k P(\boldsymbol{x} \mid \omega_k) \quad (2)$$

where K is the number of components, w_k is the component weight.

In the case of one–component mixture, the mean vector is computed as the mean of the feature vectors. The variance vector is formed from the main diagonal of the covariance matrix that is computed from the observations (i. e. feature vectors). For the k–component mixture $k > 1$ it is necessary to divide the observations into k–groups, where each is described by the mean and variance vector. Usually Expectation Maximization (EM) and Maximization Likelihood Estimation (MLE) algorithms are used for such procedure.

Given a parameterized family D_Θ of PDF's associated with a known PDF, denoted as f_Θ, we may draw a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of n values from this distribution and then using f_Θ we may compute the probability density associated with the observed data [2]

$$f_\Theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \Theta). \quad (3)$$

As a function of Θ with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ fixed, the likelihood function is

$$L(\Theta) = f_\Theta(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \mid \Theta). \quad (4)$$

Since Θ is not observable, the method of maximum likelihood uses the values of $\hat{\Theta}$ that maximizes $L(\hat{\Theta})$ as an estimate of Θ.

3. Proposed Method – Selective Iterative Training Algorithm

As explained above, each temporal segment is assumed to content only one voice of only one speaker. GMM creation commonly requires a sufficient amount of training data. A small number of speech observations yields very inaccurate GMM modeling.

Here we describe the proposed method, called Selective Iterative Training Algorithm (SITA), which enables to overcome the problem with an insufficient amount of training data for GMM creation.

The common GMM with a small number of components (1–2), which is trained from only a small number of observations (less than 3 seconds of speech training data) will be referred as the course initial Gaussian Mixture Model (iGMM) in the text. The proposed algorithm uses iGMM as initialization for further iterative training. The whole procedure is depicted in Fig. 3.

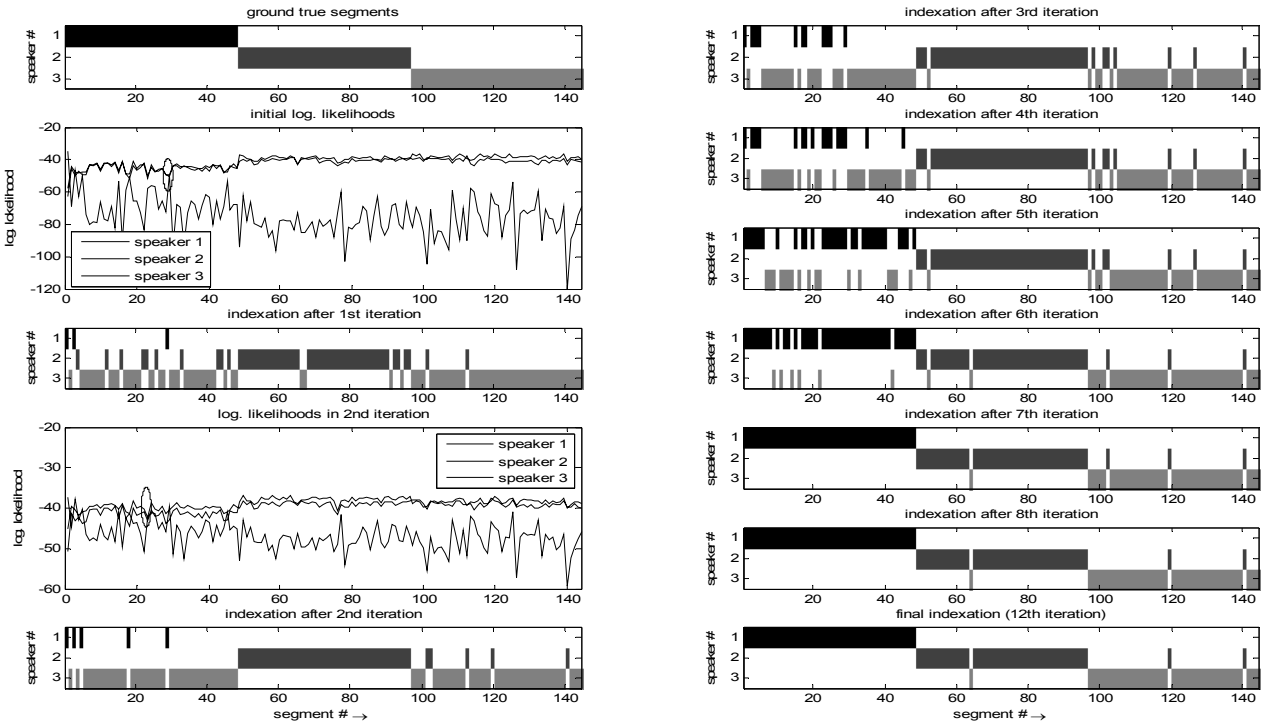

Fig. 3. An example of succesfull indexing of the stream by the proposed selection iterative algorithm.

The proposed algorithm follows a semi–automatic approach. In the first stage, iGMM's are used for initial indexation of the audio stream as follows:

- For each speaker, any segment uttered by the given speaker is selected. Thus it is assumed at least a short segment is a priori assigned to each of the speakers.
- These representative segments are taken for training iGMMs by conventional EM method.
- Logarithmical likelihoods of the iGMMs are computed for each segment. Each segment is assigned by the index of the model according to maximal log.–likelihood. Thus, each index represents one particular speaker.

Logarithmical likelihood of the *s*-th speaker is computed as follows

$$L_s(\mathrm{SE}_i \mid \mathrm{M}_s) = \frac{1}{N_i}\sum_{j=1}^{N_i} \log P(\boldsymbol{x}_{ij} \mid \mathrm{M}_s) \tag{5}$$

where SE_i is *i*-th segment, N_i is the number of observations in SE_i, $\boldsymbol{x}_{ij}$ is *j*-th observation vector in the *i*-th segment, M_s is the component mixture model of the *s*-th speaker.

We assume that the number of correctly indexed segments is equal or greater than the number of the segments allocated incorrectly. This assumption is based on preliminary experimental results. Some of them are shown in Fig. 4, and explained in chapter 5.

In the second stage, an iterative GMM re–training is performed as seen in Fig. 3. All segments (utterances) allocated to the same speaker during the initialization step are used to improve training of the new speaker model. In the first iteration, all relevant utterances are used for training. Usually some of the segments are incorrectly indexed, and such segments may decrease precision of the new GMM. Hence, the negative effect of such incorrectly indexed segments has to be suppressed during further iterative GMM training.

We assume that log.–likelihood of an incorrectly indexed utterance is lesser than log.–likelihood of the segments correctly assigned. The decision if a segment is assigned correctly is based on the following assumptions in case of two speakers:

- If the segment was correctly assigned in the training process, likelihood computed using Equation (5) of a correct speaker is increased and likelihood of an incorrect speaker is decreased. So distance between both likelihoods grows.
- On the other hand if that segment was incorrectly assigned in the training process, likelihood of the correct speaker is decreased as likelihood of the incorrect speaker is increased. So distance between both likelihoods decreases.

The aim is to find such kind of segment following the minimal distance between likelihoods.

$$\arg\min_i \left| L_1(\mathrm{SE}_i \mid \mathrm{M}_1) - L_2(\mathrm{SE}_i \mid \mathrm{M}_2) \right| \tag{6}$$

where *i* is the index of the wrong segment assumed; L_1, L_2 are log. likelihoods of the segment for each speaker.

In case of the multiple speakers (more than 2 speakers), we can replace distance measure with computation of the variance of all log.–likelihoods for a single segment

$$\arg\min_i \left[\operatorname{var} L(\mathrm{SE_i} \mid \mathrm{M_s})\right], \mathrm{SE_i} \not\subset \mathrm{VSE} \tag{7}$$

where VSE is the set of segments excluded from the training process in the previous iterations (in the first iteration all segments are used). So we assume that the segments with very–low variance are probably incorrectly assigned.

Hence in the following iterations, the segment with the minimal variance of log.–likelihoods is excluded. The excluded segment is no longer used for the training. Obviously the segment cannot be excluded if the only one available training segment is left. Exclusion is only in the stage of training process $\mathrm{SE}_i \rightarrow$VSE, where the excluded segment becomes a member of the set of excluded segments VSE.

The following conditional rules are applied to end the GMM training process:

- The maximum number of the iteration exceeds.
- No relevant changes happen after a few iterations.

Some advantages of the proposed selective iterative training algorithm, as are demonstrated by the experiment, are as follows:

- Improving the course GMM.
- Indexation of multiple speakers is also enabled.
- A small number of training observations is required in the initial step.

This method has also some drawbacks as follows:

- The number of speakers must be a priori known.
- A short segment of each speaker has to be manually labeled in advance.

Fig. 3 shows an example of iteration results.

4. Evaluation

To reliably evaluate a precision of the speaker indexation, several problems arise. One of the problems is the determination of the correct boundaries of a speech segment. An error–rate of automatic segmentation methods affects also the following indexing. Segmentation errors can affect GMM convergence during the proposed selective iterative training. Manual segmentation and annotation is also influenced by a subjective decision of human being. To overcome the segmentation problems, the recordings containing only one speaker's voice are manually merged as it is described in the experiment.

We evaluate the performance of the indexation method by the standard measures: precision and recall. The results of automatic indexation are compared against manual annotated ground true data.

The precision is the ratio of the number of relevant records retrieved to the total number of retrieved records. It is expressed as follows

$$P = \frac{A}{A + C} \tag{8}$$

where A is the number of relevant records retrieved, C is the number of irrelevant records retrieved. The recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the dataset

$$R = \frac{A}{A + B} \tag{9}$$

where A is the number of relevant records retrieved, B is the number of relevant records not retrieved. This measure can evaluate searching performance for one speaker.

To evaluate the indexation performance over all speakers in the audiostream (the audiostream is defined in chapter 5), we average the precision and the recall as follows

$$\overline{F_S} = \frac{1}{N_{SP}} \sum_{i=1}^{N_{SP}} F_i \tag{10}$$

where N_{SP} is the number of speakers in the stream, F_i is the measure F of the i–th speaker in the stream. The measure F is defined as

$$F_i = \frac{2 R_i P_i}{R_i + P_i}. \tag{11}$$

Then $\overline{F_S}$ measures were averaged over all experimental test database as follows

$$\overline{F_O} = \frac{1}{N_S} \sum_{s=1}^{N_S} \overline{F_s} \tag{12}$$

where N_S is the number of audiostreams, $\overline{F_s}$ is the measure F for the s–th audiostream.

5. Experiment

In the experiment, the following audio data sources were utilized – two large annotated speaker databases – TIMIT with 630 speakers (16 bit, mono, 16 kHz) and SpeechDat_E–SK with 1000 speakers (16 bit, mono, 8 kHz) [7]. Together approximately 30 hours of audio were used (TIMIT – 4 hours, SpeechDat_E–SK – 26 hours). Then the selected recordings of 2, 3 or 6 speakers were quasi-randomly merged into streams. Thus the audio streams compiled by such way contained the voices of 2, 3 or 6 speakers. Within one stream, only speakers with the same gender are merged (i.e. female/male discrimination,

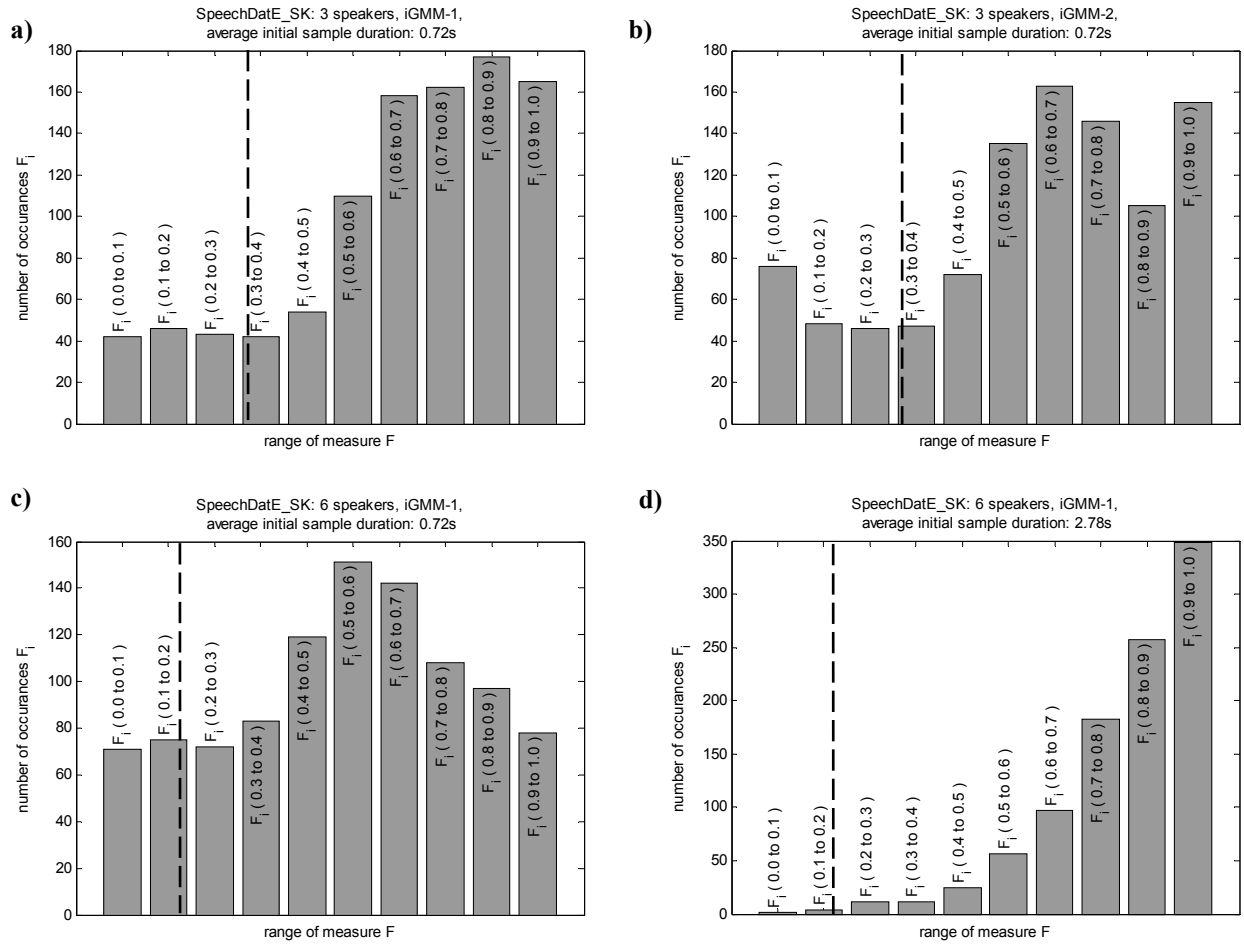

Fig. 4. Distributions of F values after initial classification (dashed line show the random classification threshold).

which obviously could simplify the task, is not considered). The silence parts were dropped out from the audio segments by thresholding short–time energy. Each stream has contained a group of a unique speaker. For example, from the database of 1000 speakers, there were created 500, 333, or 166 streams with 2, 3, or 6 speakers, respectively, in each stream. Details about the stream creation are listed in Tab. 1.

	No of streams created / average stream duration [s]		
Resource	**2-speakers experiment**	**3-speakers experiment**	**6-speakers experiment**
Timit	**315 / 61.52**	**210 / 92.28**	**105 / 184.57**
Speechadat	**500 / 479.76**	**333 / 719.71**	**166 / 1439.77**

Tab. 1. Details about the stream creation.

The segment boundaries correspond to the boundaries of the source audio files. Segment labels (for ground true) are automatically created from information about speakers in the source databases TIMIT and SpeechDat_E–SK .

As the front–end of the speaker indexation system, we used a standard parameterization method using 12 Mel Frequency Cepstral Coefficient's (MFCC). 30 ms Hamming window with 20 ms shifting was applied for analysis. Each speaker was modeled by one GMM. We tested the proposed algorithm on one– and two–component GMM (referred as GMM–1, and GMM–2, respectively). We examined a performance of the system, for various durations of the initial samples (i. e. the samples that were used to create the initial course model).

For the experiments, randomly selected audio segments where taken as the initial training samples. The samples were selected either from the single word segments (Tab. 2, the average initial sample duration is 0.72 s), or whole sentence segments. (Tab. 2, the average initial sample is 2.78 s for SpeechDat, 2.56 s for TIMIT). Remark, the audio segments correspond with original files as specified in SpeechDat and TIMIT.

For example, Fig. 4 shows the distribution of F measure, as defined by (11), after the initialization for 500 experiments to support the theoretical assumptions about initial indexation. In each experiment, the measure F is computed for each of the 3 (Fig. 4a,b) or 6 speakers (Fig. 4c,d) in the stream. Remark, F greater than 1/2, 1/3, or 1/6 in case of 2, 3 or 6 speakers streams, respectively, means

the classification better than random. Thus for instance, in case of 3 speakers discrimination by using 1 component GMM, 860 of 1000 results fulfill the given assumption (bins 4-10 in Fig. 4a). These preliminary experiments have shown the assumption that the number of correctly indexed segments is greater than the number of the segments allocated incorrectly, is fulfilled in the most cases.

Due to a short duration of the initial training samples (less than 3 seconds), only simple Gaussian model with a small number of components (1-2) as iGMM was used. Parameters of the Gaussians were adapted during the selective iterative training. In these experiments we don't upgrade the number of Gaussians in the GMM.

The indexation results on audio streams, compiled from TIMIT and SpeechDat source files respectively, are shown in Tab. 2–4. Tab. 2 shows performance of the system if 2 speakers were indexing. Tab. 3 and 4 show performance of the indexing for 3 and 6 speakers respectively. Three examples of the iterative training process are shown in Fig. 5 where precision and recall were computed after each iteration. In a very few cases the iteration process didn't converge to the optimal result (dash–dotted curve in Fig. 5). While the training samples with duration of approximately 2.56 seconds were used, only small improvement against iGMM is seen, but much higher improvement is obtained if the samples shorter than one second were processed. The proposed method also successfully adapted GMM in the multi–speaker records.

Overal measure F for 2 speakers	average duration of initial sample [s]	1 component GMM's	
		SITA $\overline{F_O}$	iGMM $\overline{F_O}$
SpeechDat _E–SK	0.72	0.85	*0.75*
SpeechDat _E–SK	2.78	0.93	*0.83*
TIMIT	2.56	0.93	*0.93*
–	–	2 component GMM's	
SpeechDat _E–SK	0.72	0.88	*0.71*
SpeechDat _E–SK	2.78	0.97	*0.93*
TIMIT	*2.56*	*0.94*	*0.92*

Tab. 2. Results for 2 speakers indexation.

Overal measure F for 3 speakers	average duration of initial sample [s]	1 component GMM's	
		SITA $\overline{F_O}$	iGMM $\overline{F_O}$
SpeechDat _E–SK	0.72	0.85	0.65
SpeechDat _E–SK	2.78	0.93	0.88
TIMIT	2.56	0.93	0.86
–	–	2 component GMM's	
SpeechDat_E–SK	0.72	0.82	0.6
SpeechDat_E–SK	2.78	0.95	0.86
TIMIT	2.56	0.92	0.88

Tab. 3. Results for 3 speakers indexation.

Overal measure F for 6 speakers	average duration of initial sample [s]	1 component GMM's	
		SITA $\overline{F_O}$	iGMM $\overline{F_O}$
SpeechDat_E–SK	0.72	0.74	*0.53*
SpeechDat_E–SK	2.78	0.89	*0.81*
TIMIT	2.56	0.87	*0.78*
–	–	2 component GMM's	
SpeechDat_E–SK	0.72	0.73	*0.47*
SpeechDat_E–SK	2.78	0.92	*0.82*
TIMIT	*2.56*	*0.86*	*0.78*

Tab. 4. Results for 6 speakers indexation.

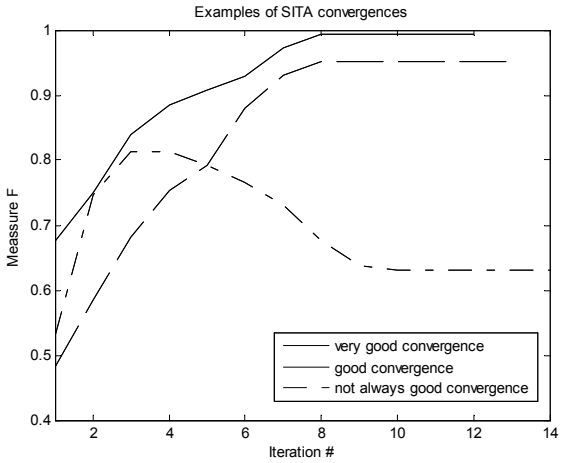

Fig. 5. Three examples of measure F convergation after a few iterations.

6. Conclusion

We proposed an alternative approach of speaker indexing to the recent approaches that requires a predefined anchor model set trained on non–target speakers. The models are created on the fly only from the test samples by an iterative way. The process of training is further sped–up by the proposed selection algorithm.

The method is based on unsupervised adaptation of the initial GMM speaker model. The iterative algorithm automatically searches appropriate segments to improve the initial course GMM that was created only from very short speech samples. The results of initial experiments show that the proposed training method enables to create precise GMM speakers models from only small amount of training data. The constrain of the method is that the number of the speakers has to be known in advance.

Our experiments were performed only on artificial audiostreams. We plan to perform next experiments on real multimedial database, which is created from broadcast streams, after the finalization of annotation. Next experiments will be focused also on adaptation and optimization of number of components in GMM towards higher order GMM.

Acknowledgments

The experiments on SpeechDat-E_SK database were carried out in the Department of Speech Analysis and Synthesis in the Institute of Informatics of Slovak Academy of Science. Technical support of this Department is gratefully acknowledged.

This work was partially supported by the Scientific Grant Agency of the Ministry of Education of the Slovak Republic under the contract No. 1/4066/07.

References

[1] ADAMI, A. G., KAJERAK, S. S., HERMANSKY, H. A new speaker change detection method for two–speaker segmentation. In *Proc. ICASSP-2002*. Orlando (Florida, USA), May 2002.

[2] AKITA, Y., KAWAHARA, T. Unsupervised speaker indexing using anchor models and automatic transcription of discussions. In *Proc. of Eurospeech 2003*. Geneva (Switzerland), 2003.

[3] DELACOURT, P., WELLEKENS, CH. J. DISTBIC: A speaker–based segmentation for audio data indexing. *Speech Communication*, 2000, no. 32 , p. 111 – 126.

[4] DUNN, R. B., REYNOLDS, D. A., QUATIERI, T. F. Approaches to speaker detection and tracking in conversational speech. *Digital Signal Processing*, 2000, no. 10, pp. 93–112.

[5] KWON, S., NARAYANAN, S. A method for on-line speaker indexing using generic reference models. In *Proc. of Eurospeech 2003*, Geneva (Switzerland), 2003.

[6] ROSENBERG, A., GORIN, A., PARTHASARATHY, S. Unsupervised speaker segmentation of telephone conversations. In *Int. Conf. on Spoken Language Processing*, vol. 1, 2002.

[7] RUSKO, M., DARŽÁGÍN, S., TRNKA, M. SpeechDat - E“ telephone speech database as an important source for basic acoustic - phonetic research in Slovak. In *ICA 2004*: part I. [electronic source], 2004, pp. 1676-1682.

[8] SIEGLER, M., JAIN, U., RAJ, B., STERN, R. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. Darpa Speech recognition Workshop*. Chantily, VA, February 1997.

[9] STURIM, D. E., REYNOLDS, D. A., SINGER, E., CAMPBELL, J. P. Speaker indexing in large audio databases using anchor models. In *Proc. ICASSP,* p. 429-432, 2001.

[10] TRITSCHLER, A., GOPINATH, R. Improved speaker segmentation and segment clustering using the Bayesian Information Criterion. In *Proc. EUROSPEECH'99*. Budapest (Hungary), September 1999.

[11] WANG, Y., LIU, Z., HUANG, J.-CH. Multimedia content analysis using both audio and visual clues. *IEEE Signal Processing Magazine*, November 2000.

[12] ZHANG, T., KUO, C. C. J. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing,* 2001, vol. 9, no. 4, pp. 441 – 457.

About Autors …

Martin PARALIČ graduated from the University of Žilina with the Ing. degree (Masters degree) in 2003. He is currently an assistant lecturer at the Department of Telecommunications of the University of Žilina, Slovakia where he is also pursuing the Ph.D. degree. His research interests are in the area of speech analysis, speaker recognition, multimedia computing and computer science.

Roman JARINA is head of the Digital Signal Processing section at the Department of Telecommunications of the University of Žilina, Slovakia. He received the Ing. degree in electronics from the University of Transport and Communications, Žilina, in 1990, and Ph.D. degree in telecommunications from the University of Žilina, in 2000.

He was a research postgraduate at the Engineering College of Copenhagen, Denmark in 1994-1995 and a research fellow at the Dublin City University, Ireland in 2000-2002. His research interests and expertise are in the areas of digital audio and speech analysis, processing and recognition, and multimedia information retrieval. He is a member of the Czech and Slovak Radioengineering Society, IET, IEEE and AES. He is also Slovak representative of the EU action COST292.