Elektrotehniški vestnik 75(4): 205–210, 2008
Electrotechnical Review, Ljubljana, Slovenija

The Impact of Audio Segmentation to Speaker Tracking in Broadcast News Data

Janez Žibert

Univerza v Ljubljani, Fakulteta za elektrotehniko, Tržaška 25, 1000 Ljubljana, Slovenija
E-pošta: janez.zibert@fe.uni-lj.si

Abstract. A system for speaker tracking in broadcast-news audio data is presented. The process of speaker tracking in continuous audio streams involves several processing tasks and is therefore treated as a multistage process. The main building blocks of such system include the components for audio segmentation, speech detection, speaker clustering and speaker identification. Our system was developed by implementing the most recent published methods in each component of the system, whereas we focused mainly on the component for audio segmentation for being considered as one of the most critical components of such systems. Two alternative approaches of speaker change detection and speaker clustering are explored and their impacts to the overall speaker-tracking performance are evaluated. The evaluation experiments were performed on broadcast-news audio data with a speaker-tracking system capable of detecting 41 target speakers. The comparison of the evaluation results of different versions of the speaker-tracking system indicates the importance of the tasks in the audio-segmentation module and provides valuable insights into how the system works.

Key words: audio segmentation, speaker clustering, audio indexing, speaker tracking,

Vpliv segmentacije na sledenje govorcev v zvočnih posnetkih informativnih oddaj

Povzetek. Predstavljen je razvoj sistema za iskanje govorcev v zvočnih posnetkih informativnih oddaj. Sistemi za obdelavo zvočnih posnetkov za pridobivanje informacije o prisotnosti in identiteti govorcev ponavadi združujejo več nalog. Osnovne naloge, ki jih rešujemo v takšnih sistemih, vključujejo segmentacijo zvočnih posnetkov, detekcijo govornih delov, združevanje segmentov po govorcih in identifikacijo govorcev na podlagi podatkov v združenih segmentih. V sistemu, ki ga opisujemo, smo združili obstoječe znanje s teh področij v enovit sistem za detekcijo in sledenje govorcev v zvočnih posnetkih informativnih oddaj. Sistem je bil zasnovan modularno, kar nam je omogočilo preučevanje delovanja posameznih modulov in njihovega sovplivanja na končno delovanje sistema. Vpliv posameznih komponent sistema na končno delovanje smo izmerili z vrsto poskusov v sistemu za sledenje 41 govorcev v zvočnih posnetkih informativnih oddaj. V članku smo se posvetili predvsem preučevanju vpliva segmentacije in razvrščanja segmentov zvočnih posnetkov. Postopek segmentacije se v takšnih sistemih običajno izvaja v začetnih fazah obdelave zvočnih posnetkov in zato pogojuje uspešnost delovanja vseh drugih komponent sistema. Izvedli smo dve metodi segmentacije, ki temeljita na Bayesovem informacijskem kriteriju, in primerjali uspešnost teh metod na končno delovanje sistema za sledenje govorcev. Pri tem smo potrdili pomembnost segmentacije v takšnih sistemih in pokazali, da učinkovita detekcija mej med akustično različnimi segmenti posredno in neposredno vpliva na celotno delovanje sistema. V primeru, kjer se segmentacija izvaja pred detekcijo govornih in negovornih delov, je vpliv segmentacije posreden, saj se pri napakah poslabša detekcija govornih delov v zvočnih posnetkih, kar vpliva na končne rezultate detekcije govorcev, pa tudi čas obdelave zvočnih posnetkov se poveča. Po drugi strani pa ima slabša detekcija mej med govorci v govornih delih zvočnih posnetkov tudi neposreden vpliv na detekcijo govorcev. Izkazalo se je namreč, da napake pri določanju mej v tem primeru povzročijo napake pri procesih združevanja segmentov po govorcih in identifikacije govorcev, ki se izvajata v zadnji fazi tvorjenja indeksa za sledenje govorcev v zvočnih posnetkih.

Ključne besede: segmentacija zvočnih posnetkov, razvrščanje segmentov po govorcih, indeksacija zvočnih posnetkov, sledenje govorcev v zvočnih posnetkih

Received 6 March 2008
Accepted 19 June 2008

1 Introduction

With the increasing availability of audio data derived from various multimedia sources comes an increasing need for efficient and effective means for searching through and indexing this type of information. Searching or tagging speech based on who is speaking is one of the basic components required for dealing with spoken documents collected in large audio-data archives, such as recordings of broadcast news or recorded meetings. In this paper we focus on development of an application for indexing and searching of speakers in audio broadcast news (BN).

The audio data of BN shows present a typical multi-speaker environment. When searching target speakers in

such an environment, the goal is to find and identify regions in the audio streams that belong to the target speakers. The task of finding such speaker-defined regions is known as a speaker diarization and was first introduced in the *Rich Transcription* project in *'Who spoke when'* evaluations, [1]. Such information is of interest for several speech- and audio-processing applications. For example, in automatic speech-recognition systems the information can be used for unsupervised speaker adaptation [2], which can significantly improve the performance of speech recognition in large-vocabulary continuous-speech-recognition systems [3]. This information can also be used for the indexing of multimedia documents, where homogeneous speaker or acoustic segments usually represent the basic units for indexing and searching in large archives of spoken audio documents, [4]. The outputs of a speaker-diarization system could also be used in speaker-identification and speaker-tracking systems, [5], which was also the case in our presented application.

This paper aims to present development of a system for audio indexing of BN news according to target speakers. The system was developed by implementing the most recent published methods in each component of the system. The main work was done to optimize the procedures in an audio segmentation module, which is considered as one of the key component of such systems. A short overview of the audio-indexing system is given in Section 2. In Section 3 we focus on audio-segmentation tasks. We provide a condensed description of the audio segmentation procedures implemented in our system and evaluated in experiments on the Slovenian audio BN database. An experimental setup and a discussion of the results are given in the last sections.

2 Speaker-based Audio Indexing

Most audio-indexing systems for the detection of speakers in continuous audio streams have a similar general architecture, [6, 7]. The baseline speaker-indexing system architecture, that was implemented in our system, is shown in Figure 1.

The main building blocks of such system include the components for speech detection, audio segmentation, speaker clustering and speaker identification. A typical combination of these components is shown in Figure 1 although it is possible to perform some of the stages jointly or in different ordering. In our case a typical ordering was followed. First, audio data are divided in a speech-detection module into speech and non-speech data. Speech regions are further processed in an audio segmentation module, while non-speech data are discarded. The audio segmentation module aims to provide segments of audio data, that contain speech from just one speaker, and associate together segments from the same speaker. The first task is performed in a speaker-change-detection component, while the second stage is done during speaker clustering. The resulting data are homogeneous segments, that are clustered together according to speakers, i.e., segments with relative labels are produced. During the final stage, each cluster of segments is labeled with a corresponding speaker-identification name, or it is left unlabeled if the speech data in the cluster do not correspond to any of the previously enrolled target speakers.

Our speaker-based indexing system [8] was built by implementing the most recent approaches in each processing task. Development of the system was carried out by examining the alternative approaches in each component of the system and measuring their impacts to the overall speaker-tracking results. In this paper we focus mostly on exploring the impacts of the audio-segmentation tasks. Thus, we tested different versions of speaker change detection and speaker clustering procedures in the audio segmentation module, while the modules for speech detection and speaker identification remained the same in all our evaluation experiments. We give a short overview of these two components in the remaining of this section, while the audio segmentation procedures are presented in more details in the next section.

Our speech detection was based on a maximum-likelihood classification with Gaussian Mixture Models (GMMs), which were trained on manually labeled training data. When audio signals are modeled with standard acoustic representations, we used a general approach for speech detection in continuous audio streams [7]. However, we implemented an alternative approach, based on phoneme-recognition features [9], which proved to be a better choice for such kind of applications [10]. The speech detection was in this case performed by a Viterbi decoding in a classification network composed of speech and non-speech GMMs.

The speaker-identification component was adopted from a speaker-verification system that was originally designed for the detection of speakers in conversational telephone speech, [11]. The speaker-verification system was based on the standard Gaussian Mixture Model – Universal Background model (GMM-UBM) approach, [12]. In addition to this, we computed a new set of features based on mel-frequency cepstral coefficients (MFCCs), which were subjected to feature warping [13] to compensate for different channel effects, and the log-likelihood scores normalization was performed at the end by applying the T-normalization technique [14].

3 Audio Segmentation

3.1 Speaker-Change Detection

We implemented two different acoustic change detection procedures, which both aim to find time-stamps in audio streams at changes between different speakers or acoustic

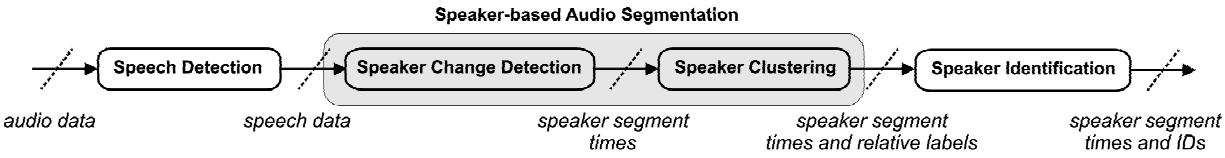

Figure 1. Main building blocks of a typical speaker-based audio indexing system.
Slika 1. Osnovne komponente sistema za samodejno indeksacijo zvočnih posnetkov po govorcih.

environments and are both based on the Bayesian Information Criterion (BIC), [15].

When the BIC is used as a model selection criterion for the audio segmentation, a problem of change detection is reformulated as a model selection task between two competing models. In this case, a part of an audio stream X, where a change point needs to be found, is presented by a sequence of frame-based acoustic feature vectors, i.e., $X = x_1, x_2, \ldots, x_N$. Additionally, it is assumed that each acoustic homogeneous segment of audio data can be modeled by a single Gaussian distribution. Then, for each point $1 < b < N$ within the data set X the BIC difference between two models can be produced:

$$\begin{aligned} \Delta_{BIC}(b;X) &= L(X;M_2) - L(X;M_1) \\ &- \lambda(C(M_2) - C(M_1)), \end{aligned} \quad (1)$$

where M_1 represents a model, when the data $x_1, \ldots, x_N$ are modeled by a single Gaussian distribution, and M_2 stands for a model, which is composed by two Gaussian distributions: one is estimated from data $x_1, \ldots, x_b$ and the other from data $x_{b+1}, \ldots, x_N$. In the case, when $\Delta_{BIC}(b;X) < 0$, the data is better modeled by a single Gaussian distribution, while $\Delta_{BIC}(b;X) > 0$ favors a representing the data by two distributions, which supports a hypothesis, that change in acoustics occurs at point b. While the first term in (1) corresponds to a difference in log-likelihoods of models M_2 and M_1, the second term presents a difference in the number of parameters of both models. The first term accounts for the quality of the match between the model and the data, while the second one is a penalty for the model complexity with λ allowing the tuning of the balance between the two terms. Based on this measure, two competing segmentation procedures were proposed and implemented in our system.

In the first segmentation procedure, a *standard approach* of finding acoustic-change detection points was followed, which was first proposed in [15] and improved in [16]. This procedure processed the audio data in a single pass while searching for change points within a window using the Δ_{BIC} measure, defined in (1). Candidates for segment boundaries were points b in an initial window X, where $\Delta_{BIC}(b;X) > 0$, and among them a point with the highest Δ_{BIC} score was selected as a change point. In this case, the window was moved to that position, while the length of the window was set to the initial size, and a computation of the Δ_{BIC} continued within the new window. If there were no change points in the initial window X (i.e. $\Delta_{BIC}(b;X) < 0$ for all points b within the window), a window was increased by additional length, and a computation of the Δ_{BIC} was redone on the extended window. These steps were repeated until there were no more data for processing. The threshold, which was implicitly included in the penalty term λ of the Δ_{BIC} score, had to be given in advance and was in our case estimated from the training data. This procedure is widely used in most of the current audio-segmentation systems [7, 1, 17].

The alternative procedure, which was also tested in our system, was based on the DISTBIC approach, [18]. A segmentation with this approach was done in two passes. In the first pass the candidate points for change detections were proposed, while in the second pass validation of these candidates was performed. For the derivation of speaker (or acoustic) turn detection points one of the distance measures is usually applied in the first pass. In our case a symmetric Kullback-Leibler (KL2) distance [19] was chosen. The candidates are computed by finding the local maximum points in the plot produced by a frame-by-frame calculation of a distance measure on two adjacent fixed-length windows of the processed audio signal, [18]. In the second pass, these points are then validated or discarded using the Δ_{BIC} measure, defined in (1). This technique tends to be less independent of the average segment size and can greatly reduce computational time of the segmentation process due to the less frequent usage of the computationally expensive BIC measure.

The outputs of this module in both cases were acoustic-change detection points, which defined basic speaker-based audio segments for further processing.

3.2 Speaker Clustering

The purpose of this stage is to associate or cluster together segments of the same speaker. In an ideal case, such clustering should be produced, where all segments of each speaker are grouped in a single cluster.

The general method, that was also implemented in our system, is to perform agglomerative clustering using a bottom-up approach with the BIC measure as a merging criterion [7, 19]. Such clustering can be described in three main steps:

1. *initialization:* $t = 1$
   each segment C_i present one cluster;

intial clustering is $\mathcal{C}_0 = \{C_i | i = 1, ..., N\}$

2. *merging procedure:*
   Repeat: $t = t + 1$
   - Among all possible pairs of clusters (C_r, C_s) in $\mathcal{C}_{t-1}$ find the one, say (C_i, C_j), such that
   $$\Delta_{BIC}(C_i, C_j) = \max \Delta_{BIC}(C_r, C_s) \quad (2)$$
   - Define $C_q = C_i \cup C_j$ and produce new clustering $\mathcal{C}_t = (\mathcal{C}_{t-1} - \{C_i, C_j\}) \cup \{C_q\}$
3. *stopping criterion:*
   The merging procedure is repeated until in $\mathcal{C}_t$ exists such pairs (C_r, C_s), for which
   $$\Delta_{BIC}(C_r, C_s) > 0.0. \quad (3)$$

In the *merging procedure* the joining of clusters was performed by searching for the maximum BIC score among all the possible pair-wise combinations of clusters. The BIC measure was the same as the one used for the speaker change detection, that is defined in (1), but it needed to be reformulated in the following way: the Δ_{BIC} score was in the previous case computed from the Gaussians, estimated from the segment of data X, and the Δ_{BIC} was computed on all time-points b, which divided the X to two parts. In the clustering case the data were constructed from the data of the current processing clusters C_r and C_s, i.e., $X = C_r \cup C_s$ and a dividing point at time b is in this case obsolete. Therefore $\Delta_{BIC}(C_r, C_s)$ was defined as $\Delta_{BIC}(C_r, C_s) := \Delta_{BIC}(.; C_r \cup C_s)$ without the time b.

The merging process was *stopped* when the highest BIC score was lower than a specified threshold, which was in our case set to 0.0.

The output of the speaker-clustering module produced relative segment labels (for example 'spk1', 'spk2', etc.), which corresponded to speaker clusters. Note that this is also a final output of the speaker-based audio segmentation module.

4 Evaluation Experiments

The presented audio-indexing system was evaluated in the task of speaker tracking on the SiBN database [20], which consisted of 32 hours of BN shows in Slovene. 20 hours were used for an estimation of all the open parameters in all the components of our indexing system, and the remaining 12 hours served for the assessment of the system performance.

Tuning of the open parameters in all the system modules corresponded to optimizing the overall speaker-tracking performance on the training data. The evaluated speaker-tracking system was capable of detecting 41 target speakers from the audio data, which included 551 different speakers. The performance of the evaluated system was assessed by including all target speakers with the addition of non-speech segments.

Two groups of experiments were conducted. In each group the impact of one component of the system was explored by comparing the final speaker-tracking performance. The overall speaker-tracking results were produced in terms of the false-acceptance (FA) and false-rejection (FR) rates computed at different operating points, and presented in the form of Detection Error Trade-off (DET) curves, [11]. This evaluation measure is commonly used for the assessment of speaker-recognition systems, where FA and FR rates are computed on the basis of speaker's utterances. In the case of speaker tracking in continuous audio streams, the FA and FR rates have to be derived on the audio segments. Thus, a generalization of the DET measure has to be applied to compute the FA and FR rates at the frame level.

4.1 Evaluation Results

Figure 2 presents speaker-tracking results from the evaluated system, where different versions of the system components were combined. The speech-detection and speaker-identification modules were the same in all the evaluations, while in the components for audio-segmentation and speaker-clustering different approaches were applied.

In Figure 2, speaker-change detection procedures are marked as *S* and speaker clustering procedures as *C*. In addition to that, the FA and FR rates in all the figures correspond to false alarm probabilities and miss probabilities, respectively.

In the first evaluation experiment, shown in Figure 2 (a), the impact of the speaker change detection procedures to the overall speaker-tracking performance was assessed. These segmentation approaches were additionally compared with a manual segmentation. The experiments were conducted in a way to measure just the impact of the audio segmentation. This was achieved by applying the same procedures in all other system components, while no speaker-clustering was performed (marked as *C:w/o* in Figure 2 (a)). Audio segmentation procedures, that were compared, were: manual segmentation (legend name *S:manual* in Figure 2 (a)), and a standard BIC segmentation (legend name *S:standBIC*) and a DISTBIC approach (legend name *S:DISTBIC*), which were both presented in Section 3.1.

As can be seen from the results, the manual segmentation outperforms the automatic versions by more than 3% across the whole range of operating points. The same phenomenon can be observed by inspecting both automatic versions of segmentation procedures. During the evaluation phases we also tested their performances in a segmentation task alone. The segmentation results, that were obtained by using the *F-measure* [21], spoke in fa-

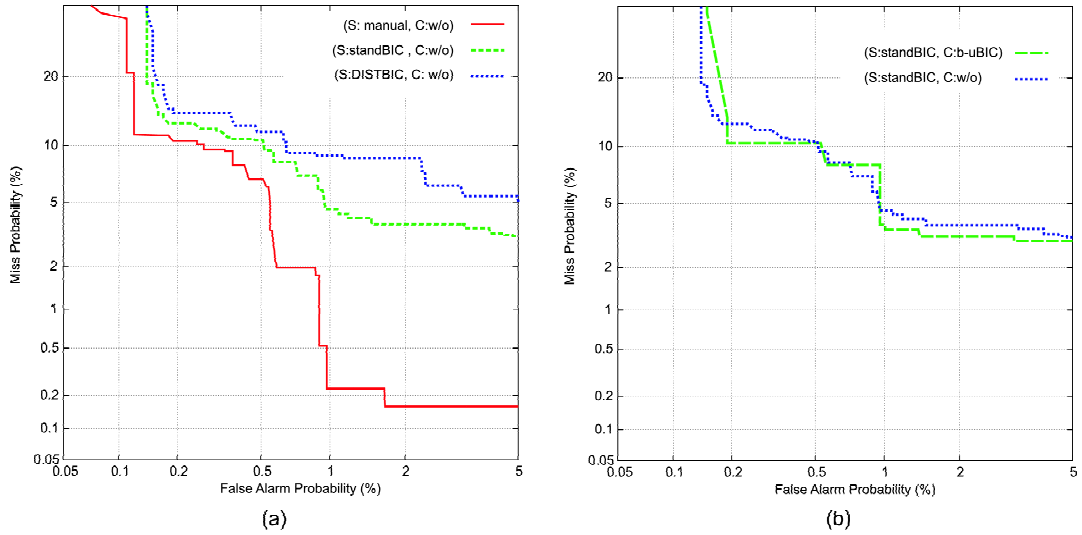

Figure 2. Overall speaker-tracking results of the evaluated system, where different audio-segmentation (a) and speaker-clustering (b) procedures were applied. Lower DET values correspond to a better performance.
Slika 2. Skupni rezultati sledenja govorcev ob izvedbi različnih verzij segmentacije (a) in razvrščanja segmentov po govorcih (b). Nižje vrednosti DET pomenijo boljše razpoznavanje govorcev.

vor of the *standBIC* approach, where the segmentation accuracy of 74% was achieved, in comparison to *DISTBIC* approach, where the segmentation accuracy was 70%. Nearly the same difference by both methods can be observed in the overall speaker-tracking results in Figure 2 (a). These segmentation results indicate the importance of the speaker-change detection task. Since this task is usually applied in the first steps of speaker-tracking systems, the errors from the segmentation have an impact on all subsequent procedures. In our case, the errors in detecting change points in continuous audio streams produced non-homogeneous segments, which caused the unreliable detection of speech/non-speech regions and the unreliable detection of target speakers as well. Accordingly, both types of errors were, therefore, additionally integrated into the overall results of the evaluated systems.

In another experiment we tried to measure the impact of speaker clustering. The results are shown in Figure 2 (b). While we used the same procedure for speaker-change detection (*S:standBIC*), we tested two systems with and without speaker clustering. In the first case, a standard bottom-up speaker-clustering approach with the BIC measure, described in Section 3.2, was implemented (referred as *C:b-uBIC* in Figure 2 (b)). The system, where no clustering was implemented, has a legend name *C:w/o*.

This evaluation aims to examine whether or not it is better to use a speaker clustering procedure in audio-indexing systems. As can be seen from the results in Figure 2 (b), there is not so much difference in the performances of the systems where clustering was applied compared to those without clustering. Our tracking results with automatic clustering show that just a marginal gain can be obtained. This indicates that in our case the speaker-tracking system can not benefit from speaker clustering.

5 Conclusion

The comparison of the evaluation results of different versions of the tasks in the audio segmentation module provides a valuable insight into how the audio indexing works and which components of the system have a greater impact on the overall system performance. The evaluation was performed on a system for speaker-based audio indexing of BN shows. We implemented two alternative approaches of the speaker-change detection procedure and the standard speaker-clustering procedure, which were all based on the BIC measure. It was found that one of the key components in such systems is the speaker-change detection procedure. If the segmentation procedure produces too many non-homogeneous segments, due to improperly detected change points in an audio stream, this causes unreliable performance of the speech-detection and the speaker-identification modules, and thus degrades the overall performance of the system. As far as the speaker clustering is concerned, we showed that it has no impact on improvement in the overall performance of the system.

6 References

[1] J. Fiscus, J. S. Garofolo, A. Le, et al., Results of the Fall 2004 STT and MDE Evaluation, *Proc. of the Fall 2004 Rich Transcription Workshop*, Palisades, NY, USA, 2004.

[2] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul, A Compact Model for Speaker-Adaptive Training. *Proc.*

of International Conference on Spoken Language Processing (ICSLP1996), Philadelphia, PA, USA, 1996, pp. 1137-1140.

[3] P. C. Woodland, The development of the HTK Broadcast News transcription system: An overview, *Speech Communications*, 37, 2002, pp. 47-67.

[4] J. Makhoul, F. Kubala, T. Leek, et al., Speech and language technologies for audio indexing and retrieval, *Proceedings of the IEEE*, 88, 2000, pp. 1338-1353.

[5] P. Delacourt, J. Bonastre, C. Fredouille, et al., A Speaker Tracking System Based on Speaker Turn Detection for NIST Evaluation, *Proc. of ICASSP 2000*, Istanbul, Turkey, June, 2000.

[6] C. Barras, X. Zhu, S. Meignier, J.-L. Gauvain, Multistage Speaker Diarization of Broadcast News, *IEEE Transactions on Speech, Audio and Language Processing, Special Issue on Rich Transcription*, 14, 2006, pp. 1505-1512.

[7] S. Tranter, D. Reynolds, An Overview of Automatic Speaker Diarisation Systems, *IEEE Transactions on Speech, Audio and Language Processing, Special Issue on Rich Transcription*, 14, 2006, pp. 1557-1565.

[8] J. Žibert, *Obdelava in analiza zvočnih posnetkov informativnih oddaj z uporabo govornih tehologij*, PhD. Thesis, University of Ljubljana, Ljubljana, Slovenia, 2006.

[9] J. Žibert, N. Pavešić, F. Mihelič, Speech/Non-Speech Segmentation Based on Phoneme Recognition Features, *EURASIP Journal on Applied Signal Processing*, 6, 2006, pp. 1-13.

[10] J. Žibert, B. Vesnicer, F. Mihelič, Novel Approaches to Speech Detection in the Processing of Continuous Audio Streams, *Robust Speech Recognition and Understanding*, M. Grimm, K. Kroschel (Eds.), I-Tech Education and Publishing, 2007, pp. 23-48.

[11] A. Martin, M. Przybocki, G. Doddington, D. Reynolds, The NIST speaker recognition evaluation - overview, methodology, systems, results, perspectives, *Speech Communications*,31, 2000, pp. 225-254.

[12] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, *Digital Signal Processing*, 10, 2000, pp. 19-41.

[13] J. Pelecanos, S. Sridharan, Feature warping for robust speaker verification, *Proc. of the Speaker Odyssey Workshop*, Crete, Greece, 2000.

[14] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, Score normalization for text-independent speaker verification system, *Digital Signal Processing*, 10, 2000, pp. 42-54.

[15] S. S. Chen, P. S. Gopalakrishnan, Speaker, environment and channel change detection and clustering via the Bayesian information criterion, *Proc. of the DARPA Speech Recognition Workshop*, Lansdowne, Virginia, USA, February 1998, pp. 127-132.

[16] A. Tritschler, R. Gopinath, Improved speaker segmentation and segments clustering using the Bayesian information criterion, *Proc. of Eurospeech 99*, Budapest, Hungary, 1999.

[17] J. Žibert, F. Mihelič, J.-P. Martens, et al., The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results, *Proc. of Interspeech 2005 - Eurospeech*, Lisbon, Portugal, 2005.

[18] P. Delacourt, C. J. Wellekens, DISTBIC: A speaker-based segmentation for audio data indexing, *Speech Communication*, 32, 2000, pp. 111-126.

[19] M. Siegler, U. Jain, B. Raj, R. Stern, Segmentation, Classification and Clustering of Broadcast News Data, *Proc. of the DARPA Speech Recognition Workshop*, Chantilly, VA, USA, 1997.

[20] J. Žibert, F. Mihelič, Development of Slovenian Broadcast News Speech Database, *Proc. of the LREC 2004*, Lisbon, Portugal, 2004.

[21] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, Strategies for Automatic Segmentation of Audio Data, *Proceedings of the ICASSP 2000*, Istanbul, Turkey, 2000.

Janez Žibert
Janez Žibert was born in 1974. He received his B.Sc. degree in mathematics from the Faculty of Mathematics and Physics in 1998 and the M.Sc. and Ph.D. degrees in electrical engineering from the Faculty of Electrical Engineering of the University of Ljubljana in 2001 and 2006, respectively. He is currently working as a research assistant at the Laboratory of Artificial Perception, Systems and Cybernetics at the University of Ljubljana. His research interests include audio-signal processing, automatic speech and speaker recognition and audio information retrieval.