We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

19

Novel Approaches to Speaker Clustering for Speaker Diarization in Audio Broadcast News Data

Janez Žibert[1] and France Mihelič[2]
[1]*Primorska Institute for Natural Science and Technology, University of Primorska*
[2]*Faculty of Electrical Engineering, University of Ljubljana*
Slovenia

1. Introduction

The growing demand to shift content-based information retrieval from text to various multimedia sources means there is an increasing need to deal with large amounts of multimedia information. The data provided from television and radio broadcast news (BN) programs are just one example of such a source of information. In our research we focus on the processing and analysis of audio BN data, where the main information source is represented by speech data. The main issues in our work relate to the preparation and organization of BN audio data for further processing in information audio-retrieval systems based on speech technologies.

This chapter addresses the problem of structuring the audio data in terms of speakers, i.e., finding the regions in the audio streams that belong to a single speaker and then joining each region of the same speaker together. The task of organizing the audio data in this way is known as speaker diarization and was first introduced in the NIST project of *Rich Transcription* in the "Who spoke when" evaluations (Fiscus et al., 2004; Tranter & Reynolds, 2006). The speaker-diarization problem is composed of several stages, in which the three main tasks are performed: speech detection, speaker- and background-change detection, and speaker clustering. While the aim of the speech detection and the speaker- and acoustic-segmentation procedures is to provide the proper segmentation of the audio data streams, the purpose of the speaker clustering is to join or connect together segments that belong to the same speakers, and this is usually applied in the last stage of the speaker-diarization process. In this chapter we focus on speaker-clustering methods, concentrating on developing proper representations of the speaker segments for clustering, and research different similarity measures for joining the speaker segments and explore different stopping criteria for the clustering that result in a minimization of the overall diarization error of such systems.

The chapter is organized as follows: In Section 2, two baseline speaker-clustering approaches are presented. The first is a standard approach using a bottom-up agglomerative clustering principle with the Bayesian information criterion as the merging criterion. In the second system an alternative approach is applied, also using bottom-up clustering, but the representations of the speaker segments are modeled by Gaussian mixture models, and for

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

the merging criteria a cross log-likelihood ratio is used. Section 3 is devoted to the development of a novel fusion-based speaker-clustering system, where the speaker segments are modeled by acoustic and prosody representations. By adding prosodic information to the basic acoustic features we have extended the standard clustering procedure in such a way that it will work with a combination of both representations. All the presented clustering procedures were assessed on two different BN audio databases and the evaluation results are presented in Section 4. Finally, a discussion of the results and the conclusions are given in Sections 5 and 6.

2. Speaker clustering in speaker-diarization systems

Speaker diarization is the process of partitioning the input audio data into homogeneous segments according to the speaker's identity. The aim of speaker diarization is to improve the readability of an automatic transcription by structuring the audio stream into speaker turns, and in cases when used together with speaker-identification systems, by providing the speaker's true identity. Such information is of interest to several speech- and audio-processing applications. For example, in automatic speech-recognition systems the information can be used for unsupervised speaker adaptation (Anastasakos et al., 1996, Matsoukas et al., 1997), which can significantly improve the performance of speech recognition in large vocabulary, continuous speech-recognition systems (Gauvain et al., 2002; Woodland, 2002; Beyerlein et al., 2002). This information can also be applied 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, (Makhoul et al., 2000). The outputs of a speaker-diarization system can also be used in speaker-identification or speaker-tracking systems, (Delacourt et al., 2000; Nedic et al., 1999).

Most speaker-diarization systems have a similar general architecture to that shown in Figure 1. First, the audio data, which are usually derived from continuous audio streams, are segmented into speech and non-speech data. The non-speech segments are discarded and not used in subsequent processing, which is done in a speech-detection module. The speech data are then chopped into homogeneous segments in an audio-segmentation module (marked as acoustic change detection in Figure 1). The segment boundaries are located by finding the acoustic changes in the signal, and each segment is, as a result, expected to contain speech from only a single speaker. The resulting segments are then clustered so that each cluster corresponds to just a single speaker. This is done in a speaker-clustering module and usually represents the final stage in speaker-diarization systems. At this stage, each cluster is labeled with relative speaker-identification names. Additionally, speaker identification or gender detection can be performed. In the first case, each of the speaker clusters can be given a true speaker name, or it is left unlabelled if the speech data in the cluster do not correspond to any of the target speakers. In the case of gender detection, each cluster gets an additional label to indicate to which gender it belongs. As such the speaker diarization of continuous audio streams is a multistage process made up of four main components: speech detection, speaker audio segmentation, speaker clustering, and speaker identification. The latest overview of the approaches used in speaker-diarization tasks can be found in (Tranter & Reynolds, 2006).

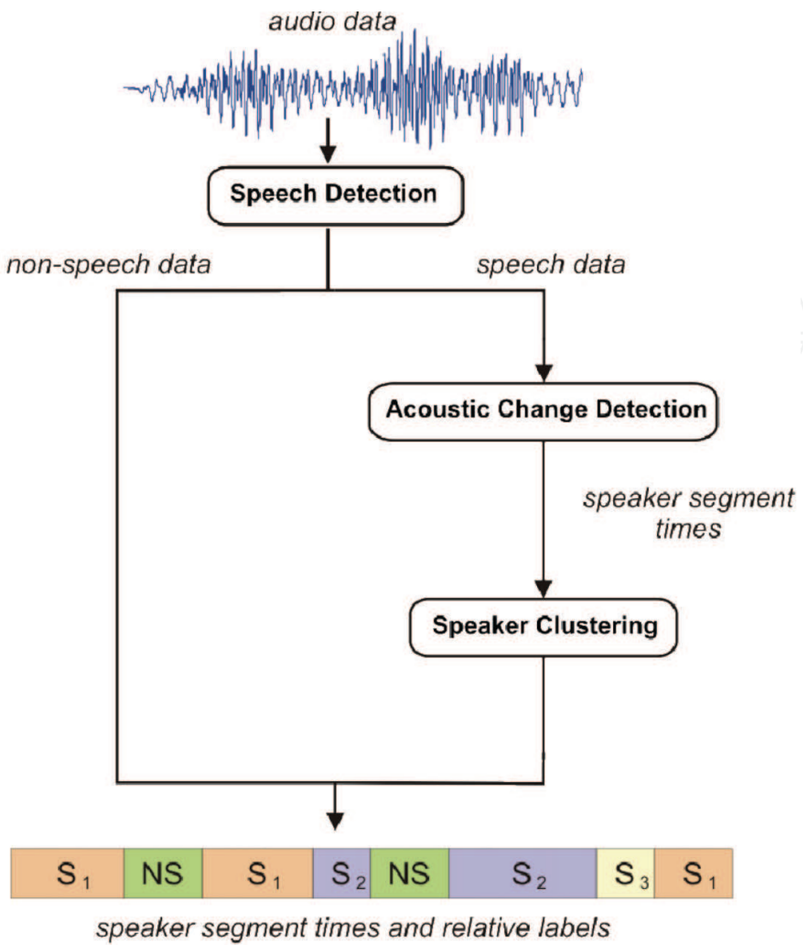

Fig. 1. The main building blocks of a typical speaker-diarization system. Most systems have components to perform speech detection, speaker- or acoustic-based segmentation and speaker clustering, which may include components for gender detection and speaker identification.

We built a speaker-diarization system that is used for speaker tracking in BN shows (Žibert, 2006b; Žibert et al., 2007). The system was designed in the standard way by including components for speech detection, audio segmentation and speaker clustering. Since we wanted to evaluate and measure the impact of speaker clustering on the overall speaker-diarization performance, we built a system where the components for speech detection and audio segmentation remained fixed during the evaluation process, while different procedures were implemented and tested in the speaker-clustering task.

The component for speech detection was derived from the speech/non-speech segmentation procedure, which was already presented in (Žibert et al., 2007). The procedure aimed to find the speech and non-speech regions in continuous audio streams represented by phoneme-recognition features (Žibert et al., 2006a). The features were derived directly from phoneme transcripts that were produced by a generic phone-recognition system. A speech-detection procedure based on these features was then implemented by performing a Viterbi decoding in the classification network of the hidden Markov models, which were previously trained on speech and non-speech data. This rather alternative approach to deriving speech-

detection features proved to be more robust and accurate for detecting speech segments (Žibert et al., 2006a; Žibert et al., 2007).

Further segmentation of the speech data was made by using the acoustic-change detection procedure based on the Bayesian information criterion (BIC), which was proposed in (Chen & Gopalakrishnan, 1999) and improved by (Tritchler & Gopinath, 1999). The applied procedure processed the audio data in a single pass, with the change-detection points found by comparing the probability models estimated from two neighboring segments with the BIC. If the estimated BIC score was under the given threshold, a change point was detected. The threshold, which was implicitly included in the penalty term of the BIC, has to be given in advance and was in our case estimated from the training data. This procedure is widely used in most current audio-segmentation systems (Tranter & Reynolds, 2006; Fiscus et al., 2004; Reynolds & Torres-Carrasquillo, 2004; Zhou & Hansen, 2000; Istrate et al., 2005; Žibert et al., 2005).

While the aim of an acoustic-change detection procedure is to provide the proper segmentation of the audio-data streams, the purpose of speaker clustering is to join together the segments that belong to the same speakers. In our system we realized three different speaker-clustering procedures, which are described in detail in the following sections.

The result of such a speaker-diarization system is segmented audio data, annotated according to the relative speaker labels (such as 'spk1', 'spk2', etc.). Each such speaker cluster can be additionally processed through the speaker-identification module to find the true identities of the speakers who are likely to be in the processing audio data (such as prominent politicians or the main news anchors and reporters in the BN data). This can be achieved by a variety of methods that can be performed during the speaker-clustering stage. However, finding the true identities of the speakers was not within the scope of this research.

2.1 Speaker clustering

The aim of speaker clustering in speaker-diarization systems is to associate or cluster together the segments from the same speaker. Ideally, this clustering produces one cluster for each speaker, with all the segments from a given speaker in a single cluster. The dominant approach used in diarization systems is called hierarchical agglomerative clustering (Theodoridis & Koutroumbas, 2003); it consists of the following steps (Tranter & Reynolds, 2006):

1. *Initialization:* each segment represents a single cluster;
2. *Similarity measure:* compute the pair-wise distances between each cluster;
3. *Merging step:*
   a. merge the closest clusters together;
   b. update the distances of the remaining clusters to the new cluster;
4. *Stopping criterion:* iterate step 3 until some stopping criterion is met.

The main issues concerning the above speaker-clustering approach include the choice of a proper similarity measure, the proper representations of the cluster data and finding a suitable stopping criterion. The audio data used for the speaker clustering is in general represented by acoustic features consisting of either mel-frequency cepstral coefficients (MFCCs) or perceptual linear-prediction coefficients (PLPCs), (Picone, 1993). The cluster data represented by these features are then usually modeled by Gaussian distributions, and the resulting similarity measures are computed as the likelihood functions from these

models (Moh et al., 2000; Reynolds & Torres-Carrasquillo, 2004; Sinha et al., 2005). The most common approach is to represent the clusters by single full-covariance Gaussian distributions, whereas for the similarity measure a Bayesian information criterion is used (Chen & Gopalakrishnan, 1999). For good performance of the clustering, the stopping criterion also needs to be properly chosen. A suitable stopping criterion should end the merging process at the point where the audio data from each speaker is concentrated mainly in one cluster, and in general it is set according to a similarity measure and cluster models that are used in the merging process of the speaker clustering.

In our research we implemented the same clustering approach, but we experimented with different similarity measures, different representations of the audio data and different cluster models. Three approaches were investigated. In the first one we followed the standard procedure of speaker clustering, based on the Bayesian information criterion. The alternative approach, which is also widely used in speaker-diarization systems and was also implemented in our study, aims to incorporate Gaussian mixture models into the speaker-clustering process. The audio data in both approaches are usually represented by a single stream of acoustic features (MFCCs, PLPs), which result in an acceptable performance of the speaker clustering in cases when the acoustic conditions do not change. But this is not the situation when dealing with BN data, since BN news is composed of audio data from various acoustic environments (different types of acoustic sources, different channel conditions, background noises, etc.). To improve speaker clustering in such cases we proposed an alternative representation of speech signals, where the acoustic parameterizations of the clusters were extended by prosodic measurements.

When speaker clustering is used as one stage in a speaker-diarization system, several improvements can be made to increase the performance of the speaker diarization, like joint segmentation and clustering (Meignier et al., 2000) and/or cluster re-combination (Zhu et al., 2005). Both methods aim to improve the base speaker-clustering results by integrating several speaker-diarization tasks together or re-running the clustering on under-clustered fragments of audio data. In our research we focused mainly on an evaluation and a development of the base speaker-clustering approaches, and did not implement any of these methods, even though they could be easily applied in the same manner as they are applied in other systems.

Also note that the presented agglomerative clustering approach is not the only possible solution for speaker clustering. This kind of approach is suitable in cases when all the audio data are available in advance. When the data need to be processed simultaneously, e.g., in the online processing of BN shows, other approaches need to be applied. The most common approach in this case is a sequential clustering, which needs to resolve the same operating issues as are present in an agglomerative clustering: what kind of data representation should be applied, how should the clusters be modeled, and what similarity measure should be used? Therefore, we decided to focus our research only on those components that are essential for the good performance of the speaker clustering, regardless of the approach that is being used.

2.2 Speaker clustering via the Bayesian Information Criterion

The most common approach for speaker clustering in speaker-diarization systems is agglomerative (bottom-up) clustering, where the Bayesian Information Criterion (BIC) is used as a similarity measure (Chen & Gopalakrishnan, 1999). The approach can be described in three main steps by following the agglomerative scheme presented in the previous section:

1. *initialization step:*
   - each segment C_i represents one cluster;
   - the initial clustering is $\vartheta_0 = \{C_i \mid i = 1, \ldots, N\}$
2. *merging procedure:*

   Repeat:
   - From among all the possible pairs of clusters (C_r, C_s) in ϑ_{t-1} find the one, say (C_i, C_j)), such that

$$\Delta_{BIC}(C_i, C_j) = \min_{C_r, C_s} \Delta_{BIC}(C_r, C_s) \tag{1}$$

   - Define $C_q = C_i \cup C_j$ and produce new clustering

$$\vartheta_t = (\vartheta_{t-1} - \{C_i, C_j\}) \cup \{C_q\} \tag{v}$$

3. *stopping criterion:*
   - The merging procedure is repeated until in ϑ_t there exists such pairs (C_r, C_s), for which

$$\Delta_{BIC}(C_r, C_s) < 0.0 \tag{3}$$

In the *merging procedure* the joining of clusters is performed by searching for the minimum Δ_{BIC} score among all the possible pair-wise combinations of clusters. The Δ_{BIC} measure is defined as:

$$\begin{aligned}\Delta_{BIC}(C_i, C_j) = \frac{1}{2}\Big(\big(K_{C_i} + K_{C_j}\big)\log|\Sigma_{C_i \cup C_j}| - K_{C_i}\log|\Sigma_{C_i}| - K_{C_j}\log|\Sigma_{C_j}|\Big) \\ -\frac{\lambda}{2}\Big(d + \frac{1}{2}d(d+1)\Big)\log\big(K_{C_i} + K_{C_j}\big),\end{aligned} \tag{4}$$

where the clusters C_i, C_j and $C_i \cup C_j$ are modeled by the full-covariance Gaussian distributions $N(\mu_{C_i}, \Sigma_{C_i})$, $N\left(\mu_{C_j}, \Sigma_{C_j}\right)$ and $N(\mu_{C_i \cup C_j}, \Sigma_{C_i \cup C_j})$, respectively. K_{C_i} and K_{C_j} are the number of sample vectors in the clusters C_i and C_j, respectively, and d is a vector dimension. λ is an open parameter, the default value of which is 1.0. Note that the term $\log|\Sigma|$ corresponds to the log of a determinant of a given full-covariance matrix Σ.

The $\Delta_{BIC}(C_i, C_j)$, defined in equation (4), operates as a model-selection criterion between two competing models, estimated from the data in clusters C_i and C_j. The first model is represented by a single Gaussian distribution, estimated from the data in $C_i \cup C_j$, while the second model is represented by two Gaussians, one estimated from the data in cluster C_i and the other from the data in cluster C_j. The first model assumes that all the data are derived from a single Gaussian process and therefore belong to one speaker, while the second model assumes that the data are drawn from two different Gaussian processes, and therefore belong to two different speakers. As such, the Δ_{BIC} represents the difference between the BIC scores of both models, where the first term in equation (4) corresponds to the difference in the quality of the match between the models and the data, while the second term is a penalty for the difference in the complexities of the models, with λ allowing the tuning of the balance between the two terms. Consequently, Δ_{BIC} scores above 0.0 correspond to better modeling with one Gaussian and thus favor one speaker, while Δ_{BIC}

scores below 0.0 favor the model with two separate Gaussians and thus support the hypothesis of two speakers.

While using the Δ_{BIC} measure in the merging process of speaker clustering, those clusters that produce the biggest negative difference in terms of Δ_{BIC} among all the pair-wise combinations of clusters are joined together. The merging process is *stopped* when the lowest BIC score from among all the combinations of clusters in the current clustering is higher than a specified threshold, which in our case was set to 0.0.

The most important role in such clustering is played by the penalty term in the BIC measure, which is weighted by the open parameter λ. In the original definition of BIC the parameter λ is set to 1.0 (Schwartz, 1976), but it was found that the speaker clustering performed much better if λ is considered as an open parameter that is tuned on the development data. The λ influences both the merging and the stopping criteria and needs to be chosen carefully to have the optimum effect. To avoid this, several modifications of the above approach have been proposed, but they all had only moderate success, since they either introduced a new set of open parameters (Ajmera & Wooters, 2003) or increased the computational cost of the speaker clustering (Zhu et al., 2005).

2.3 Speaker clustering with Gaussian Mixture Models

An alternative approach, which does not rely only on the BIC measure, was introduced in (Barras et al., 2006). The main idea was to improve the initial clustering with the BIC measure by introducing another stage of agglomerative clustering with Gaussian Mixture Models (GMMs).

This approach tends to stop the initial clustering stage (the BIC stage) early, and use the results to seed a second clustering stage with more initial data per cluster. As a result, the second stage can estimate more complex models for the speakers. In (Barras et al., 2006) they modeled clusters at this stage with GMMs by using methods from speaker-recognition tasks.

In this case the initial clustering is performed using the BIC method, described in the previous section, which is then continued by introducing the GMMs in the second stage of clustering. Before clustering, a Universal Background Model (UBM) with diagonal Gaussians is built on training data to represent the general speakers. In addition, some kind of feature normalization is applied to reduce the effects of the different acoustic environments. Next, the clustering is performed using the agglomerative clustering scheme presented in Section 2.2. The clusters are represented as GMMs and a cross log-likelihood ratio (Gauvain & Lee, 1994) is used as a similarity measure. The GMM for each cluster is obtained by a MAP adaptation (Reynolds et al., 2000) of the means of the pre-trained UBM. Explicitly, for each cluster C_i its model M_i is MAP adapted from the UBM B using the feature vectors x_i belonging to that cluster. Then, the cross log-likelihood ratio between the two clusters C_i and C_j is defined as (Barras et al., 2004):

$$CLR(C_i, C_j) = \log\frac{L(x_i|M_j)}{L(x_i|B)} + \log\frac{L(x_j|M_i)}{L(x_j|B)}, \quad (5)$$

Where the $L(x|M)$ in all four cases represents the average likelihood per frame of data x, given the model M. The pair of clusters with the highest CLR is merged and a new model is created. The process is repeated until the highest CLR is below a predefined threshold, chosen from the development data.

Several refinements can be made at all the stages of the presented speaker clustering. In order to reduce the effects of different acoustic environments, different types of feature-normalization techniques have been proposed. The most common is the feature-warping technique, which aims to reshape the histogram of the feature data, derived from the cluster segments, into a Gaussian distribution (Pelecanos & Sridharan, 2001). As far as the UBM is concerned, different UBMs can be trained and used, corresponding to the different gender and channel conditions that are expected in the audio data (Barras et al., 2004). Another method is to build a new UBM directly from the processing audio data prior to the data clustering (Moraru et al., 2005). Several improvements to the similarity measure have also been proposed. In the case where several UBMs are used in the speaker clustering, the GMMs are obtained through a MAP adaptation from the gender- and channel-matched UBM, and only these models (clusters) are then compared using the *CLR* measure (Barras et al., 2006). Alternative measures to the CLR have also been tested within this approach, like an upper-bound estimation of the Kullback-Leibler measure (Do, 2003; Ramos-Castro et al., 2005) or a penalized likelihood criterion, based on the BIC (Žibert, 2006b).

We implemented this approach by applying feature-warping normalization before the clustering, while just one general UBM was used for all the MAP adaptations of the GMMs. The UBM was trained directly from the processing audio data, and the derived GMMs were represented by diagonal-covariance Gaussians with 32 mixtures. We decided to use these rather small mixture-size GMMs (in the original approach (Barras et al., 2006) 128 mixtures were used), since we did not gain any improvement in the speaker clustering on the development data by increasing the number of mixtures in the GMMs. The second reason was that by using GMMs with a rather small number of parameters, we removed the need for running the initial stage of the BIC clustering in order to obtain more initial data per cluster.

3. Including prosodic information in the speaker clustering

Both the previously presented speaker-clustering approaches perform the clustering by measuring the similarity between the speaker data, based only on the acoustic representations. These selected acoustic representations perform reliably in most speaker-recognition systems, and they were an obvious choice in speaker-clustering approaches. Lately, however, several speaker-recognition systems have attempted to include prosodic information as well as acoustics for the representation of the speakers (Kajarekar et al., 2003; Reynolds et al., 2003; Shriberg et al., 2005; Baker et al., 2005). The fusing of both representations was an attempt to reduce the need for speaker modeling in various acoustic environments and to provide additional information about the speaker's speech characteristics.

We developed an approach to speaker clustering that included both acoustic and prosodic representations of the speakers. The main objectives were to derive the prosodic features from the speaker-cluster data and to integrate them into the basic acoustic representations of the speaker. In order to achieve this, we needed to adopt the presented agglomerative speaker-clustering approach to merge the cluster data by defining a new similarity measure that was able to fuse the similarity scores from both representations. In the following sections a derivation of the prosodic features for the speaker clustering and a new speaker-clustering approach, based on these features, are presented.

3.1 Prosodic features for the speaker clustering

The development of the prosodic features for the speaker clustering was inspired by a derivation of similar features for speaker recognition (Shriberg et al., 2005), where they focused on capturing the longer-range stylistic features of a person's speaking behavior. We followed this approach by producing three groups of prosodic features, which were related to pitch, energy and duration measurements in speech signals and were designed in such a way as to be suitable for speaker clustering.

A standard approach to extracting prosodic information from speech signals is to define the basic units of speech and then produce different features from the duration, pitch and energy measurements associated with these units (Noth et al., 2002). A key question is what kind of speech units should be applied and how much data is needed for a reliable estimation of the prosodic events? When prosodic information is modeled in combination with automatic speech-recognition systems, the usual way of producing prosodic features is to use recognized words as the basic speech units (Noth et al., 2002). In this case a large amount of training data should be available, which is not the case when modeling the prosodic information of the speakers from the speaker clusters. Consequently, the basic speech units should be defined on sub-word speech regions. In (Shriberg et al., 2005) the prosodic features were extracted from the syllable-based regions of speech, while we decided to use the voiced-unvoiced (VU) regions. Using the VU regions in speaker clustering has several advantages over the syllable-based representation. Both types of sub-word units operate at nearly the same speech-region levels and thus the same techniques for computing prosodic features can be applied, but the VU regions can be detected without the use of large-vocabulary speech-recognition systems and are language independent, which is not the case when the speech units are represented by syllables or words.

The procedure for computing the prosodic features from speech segments, using the VU regions as the basic speech units, was as follows. The energy and pitch measurements were made at the frame level, which in our case was set to 10 ms. The short-term energy was computed as the log of the signal energy, i.e., as the sum of the squared speech-signal amplitudes in the window-size range, which in our case was set to 32 ms. The energy was computed using the feature-extraction tool in the *HTK Toolkit* (Young et al., 2004). The pitch (*f0*) is estimated using the *get_f0* function in the *ESPS/Waves* toolkit (Talkin, 1995) and then post-processed using the median filtering. For the detection of voiced (V), unvoiced (U) and silent (S) regions in the speech, a generic phoneme-based speech recognizer was used. The recognizer was the same as the one presented in (Žibert et al., 2007), which was already applied in a speech-detection task, where it proved to be language independent. In addition, we aligned the voiced regions with the *f0* trajectory, where the voiced regions were either shortened or extended according to the *f0* values or some missing *f0* values were added in the cases of detected voiced regions. After the extraction and alignment of these measurements, we created three groups of prosodic features related to the energy, duration and pitch values in voiced-unvoiced-silent (V-U-S) regions:

Energy features:

1. *energy mean:* the estimated mean of the short-term energy frames in the speech segment;
2. *energy variance:* the estimated variance of the short-term energy frames in the speech segment;
3. *rising energy frame rate:* the number of rising short-term energy frames in the speech segment divided by the total number of energy frames;

4. *falling energy frame rate:* the number of falling short-term energy frames in the speech segment divided by the total number of energy frames.

Duration features:

5. *normalized VU speaking rate:* the number of changes of the V, U, S units in the speech segment divided by the speech-segment duration;
6. *normalized average VU duration rate:* the absolute difference between the average duration of the voiced parts and the average duration of the unvoiced parts, divided by the average duration of all the V, U units in the speech segment;

Pitch features:

7. *f0 mean:* the estimated mean of the *f0* frames computed only in the V regions of the speech segment;
8. *f0 variance:* the estimated variance of the *f0* frames computed only in the V regions of the speech segment;
9. *rising f0 frame rate:* the number of rising *f0* frames in the speech segment divided by the total number of *f0* frames;
10. *falling f0 frame rate:* the number of falling *f0* frames in the speech segment divided by the total number of *f0* frames.

All the above features were obtained from the individual speech segments associated within each cluster. The features were designed by following the approach for prosody modeling of speaker data (Shriberg et al., 2005) and the development of the prosodic features for word-boundary detection in automatically transcribed speech data (Gallwitz et al., 2002). Note that the features in 5 and 6 are the same as those used in speech detection based on phoneme-recognition features (Žibert et al., 2007). We decided to implement only those features that can be reliably estimated from relatively short speech segments and were suitable for prosody modeling in speaker clustering. A normalization of each feature was provided by averaging the selected measurements, either by segment duration or by the total number of counted frames in a segment.

The 10 presented features were carefully designed to capture the speaker-oriented prosodic patterns from relatively short speech segments; however, to obtain reliable prosodic information about a speaker there should be several segments present in a cluster. Therefore, the above prosodic features should be treated as a supplementary representation of the cluster data, which can provide a considerable improvement in speaker-clustering performance when larger amounts of cluster data are available.

3.2 Fusing of acoustic and prosodic information in speaker clustering

The development of prosodic features represented the first step of including prosodic information in speaker clustering. The next step was to provide an appropriate comparison of the different clusters represented by this set of features and to integrate the acoustic and prosodic information into a single, unified speaker-clustering approach. We decided to implement the same speaker-clustering approach as was used in the baseline BIC clustering, presented in Section 2.2, but we extended it by including both types of information in the merging process of clustering.

The main reason for integrating the prosodic features into the speaker clustering was to provide information in addition to the basic acoustic features in order to gain some improvement in the speaker clustering in the case of adverse acoustic conditions. Thus, a new clustering approach was designed, which enabled us to control the amount of each type

of information in the merging process of speaker clustering. To achieve this, two important issues had to be resolved:

1. an appropriate similarity measure for the comparison of the clusters represented by prosodic features had to be designed;
2. a fuzzy-based merging criterion had to be defined, which should appropriately combine the similarity scores of the acoustic and prosodic representations of the clusters.

In the baseline speaker-clustering approach the BIC was applied as the similarity measure between the clusters represented by the acoustic, MFCC features. In the merging stage of the baseline clustering approach two clusters were joined, providing their Δ_{BIC} score achieved the minimum among all the Δ_{BIC} scores. A similarity measure based on the prosodic features was needed to operate in the same manner: lower scores should correspond to more similar clusters and higher scores to less similar clusters. Both similarity measures were also required to be easily integrated into the fuzzy-based merging criterion of the speaker clustering. This could be ensured by enabling the normalization of the similarity scores of both measures and by the appropriate weighting of both similarities.

Taking all this into account, a new prosodic measure was proposed. The measure was defined on speaker clusters by computing the Mahalanobis distance between the principal components of the speaker segments represented by the prosodic feature vectors. This procedure involved the following steps:

1. Each segment s_i is represented by the vector $\boldsymbol{v}_{s_i}^{pros}$ constructed from 10 prosodic features, defined in Section 3.1.
2. A Principal Component Analysis (PCA; Theodoridis & Koutroumbas, 2003) is performed on all the processing segments s_i, represented by the vectors $\boldsymbol{v}_{s_i}^{pros}$. This involves computing the correlation matrix R^{pros} of the vectors $\boldsymbol{v}_{s_i}^{pros}$and decomposing the eigenvalue $R^{pros}= P\cdot\Lambda\cdot P^T$, where P represents the matrix of eigenvectors ordered by the eigenvalues, which are stored in the diagonal matrix Λ.
3. The Mahalanobis distance between the principal components of the speaker segments is computed:

$$d_{pros}(s_i, s_j) = \sum_{n=1}^{10} \frac{\left(w_{s_i}^n - w_{s_j}^n\right)^2}{\lambda_n}, \tag{6}$$

where $w_{s_i}^n$ is the principal component of $\boldsymbol{w}_{s_i} = \boldsymbol{P}^T \cdot \boldsymbol{v}_{s_i}^{pros}$ at the eigenvalue λ_n,n=1,…,10. $w_{s_i}^n$ is defined in a similar fashion.

4. The similarity measure between the speaker cluster C_i, composed of the speaker segments $\{ s_i \mid i= 1,\ldots, N_i\}$, and the speaker cluster C_j, composed of the speaker segments $\{ s_j \mid j= 1,\ldots, N_j\}$ is then defined as the average of the all the pair-wise combinations of segments from both clusters:

$$pros(C_i, C_j) = \frac{1}{N_i N_j} \sum_{s_i \in C_i} \sum_{s_j \in C_j} d_{pros}(s_i, s_j) \tag{7}$$

Lower scores in (7) correspond to a better similarity between the clusters represented by the corresponding prosodic features.

A development of the above prosodic measure was inspired by similar approaches of constructing distance measures on clusters with distances that are defined only on cluster samples (Theodoridis & Koutroumbas, 2003). In our case we used the Mahalanobis distance, which was computed for principal components of the prosodic features derived from the corresponding speaker segments. This was done so as to reduce the possible correlation effects of the selected prosodic features and to remove the influences of the different scalar ranges of the features on the distance computations. Note that the prosodic measure, defined in (7), operates in the same fashion as the BIC measure, defined in (4): lower scores correspond to a better similarity between clusters.

To integrate both the similarity scores into a merging criterion of speaker clustering some kind of score normalization needs to be applied to both similarity measures and the appropriate fusion scheme of joining both scores has to be defined. We decided to use the *min-max score normalization* of both similarity measures (Jain et al., 2005). The normalized version of the BIC measure from (4) was defined as:

$$norm_{\Delta_{BIC}}(C_r, C_s) = \frac{\Delta_{BIC}(C_r, C_s) - \min\limits_{C_i, C_j} \Delta_{BIC}(C_i, C_j)}{\max\limits_{C_i, C_j} \Delta_{BIC}(C_i, C_j) - \min\limits_{C_i, C_j} \Delta_{BIC}(C_i, C_j)}, \tag{8}$$

and a normalized version of the prosodic measure from (7) was defined as:

$$norm_{pros}(C_r, C_s) = \frac{pros(C_r, C_s) - \min\limits_{C_i, C_j} pros(C_i, C_j)}{\max\limits_{C_i, C_j} pros(C_i, C_j) - \min\limits_{C_i, C_j} pros(C_i, C_j)}. \tag{9}$$

The minimum and maximum values in equations (8) and (9) were computed from among all the pair-wise cluster combinations at the current step of merging. A controllable fusion of both representations of the speaker clusters in the merging criterion was obtained by producing a weighted sum of the normalized versions of both similarity measures:

$$fus(C_r, C_s) = \alpha \cdot norm_{\Delta_{BIC}}(C_r, C_s) + (1 - \alpha) \cdot norm_{pros}(C_r, C_s), \tag{10}$$

where α represents a weighting factor between the acoustic and prosodic representations of the speaker clusters. A merging of the clusters was then achieved by finding a minimum score among all the pair-wise combinations of clusters at the current step of clustering:

$$fus(C_i, C_j) = \min_{C_r, C_s} fus(C_r, C_s). \tag{11}$$

By using the above merging criterion the speaker clustering was performed by following the same clustering procedure as described in Section 2.2. The only difference was in step 2 of the procedure, where instead of a minimum of the Δ_{BIC} score in the merging step in equation (1), a minimum from among the fusion of scores from equation (11) was used. In this way we were able to include prosodic information in the baseline speaker-clustering approach.

4. Evaluation of the speaker-clustering approaches

An evaluation of all three presented clustering approaches was performed in two speaker-diarization tasks on broadcast news data. The evaluation experiments were conducted by

following the *NIST Rich Transcription Evaluation,* which has been the major evaluation technique for the speaker diarization of broadcast news data (Fiscus et al., 2004). A similar evaluation was also performed in the *ESTER Evaluation* using French radio broadcast news data (Galliano et al., 2005).
Our experiments were carried out on two broadcast news databases. The first includes 33 hours of BN shows in Slovene and is called the SiBN database (Žibert & Mihelič, 2004). The second was a multilingual speech database, COST278, which is composed of 30 hours of BN shows in nine European languages (Vandecatseye et al., 2004), and was already used for an evaluation of different language- and data-independent procedures in the processing of audio BN shows, (Žibert et al., 2005).

4.1 Evaluation measure

The speaker-clustering approaches were evaluated by measuring the speaker-diarization performance in terms of the diarization error rate (DER), (Fiscus et al., 2004). The DER measures the differences in the reference and hypothesized speaker segmentations. This is accomplished by finding a one-to-one mapping of the reference speaker segments to the hypothesis speaker labels so as to maximize the total overlap of the reference and the (corresponding) mapped hypothesis speakers. The speaker-diarization performance is then expressed in terms of the miss (speaker in reference but not in hypothesis), false-alarm (speaker in hypothesis but not in reference) and speaker-match (mapped reference speaker is not the same as the hypothesized one) error rates. While the miss and false-alarm error rates correspond to the speech/non-speech detection errors, the speaker-match error rate corresponds to the speaker-clustering errors. The overall diarization error (DER) is the sum of these three components.
We additionally modified the DER measure in order to more closely analyze the performance of the speaker-clustering approaches, regardless of the stopping criteria used in the clustering. We achieved this by computing the overall diarization-error-rate trajectory as the average of the DER trajectories of each processed audio file. The DER trajectory per each file was constituted from the DER values computed for a different number of speaker clusters. The number of speaker clusters was not defined in absolute figures, but as the relative difference compared to the actual number of speakers in each processed audio file. This enabled us to align the DER values of each file at the same evaluation points and produce the average trajectory as the final result. Such evaluations provided us with more valuable insights into how the different speaker-data representations could affect the speaker clustering and how well the merging process of clustering can be performed without any additional tuning of the proper stopping thresholds, since it is well known that an improper selection of the stopping thresholds can seriously degrade the speaker-clustering performance.

4.2 Experimental setup

We evaluated all three speaker-clustering approaches: a baseline system with the BIC, a GMM-based approach and a fusion-based approach, presented in Sections 2.2, 2.3, and Section 3, respectively.
Since we only wanted to assess the performance of the speaker-clustering approaches we used the same speech/non-speech-detection and audio-segmentation procedures in all the evaluation experiments. The speech/non-speech detection used the approach presented in (Žibert et al., 2007), while the audio segmentation used the approach presented in (Chen & Gopalakrishnan, 1999).

In all the tested speaker-clustering approaches we needed to set different open parameters. The parameters were chosen according to the optimal speaker-diarization performance of the corresponding clustering approaches on the development dataset, which was composed of 7 hours of BN audio data from the SiBN database. Detailed information of the experimental setup for each individual clustering approach is presented in the following list:

- **The baseline BIC approach:** (described in Section 2.2)
  The audio data were represented by *MFCC* features, which were composed of the first 12 cepstral coefficients (without the 0th coefficient) and a short-term energy with the addition of the Δ*MFCC* features. The Δ*MFCC* features were computed by estimating the first-order regression coefficients from the static *MFCC* features. The features were derived from audio signals every 10 ms by using 32-ms analysis windows, (Young et al., 2004). For the estimations of the Δ_{BIC} measure from equation (4) each cluster was modeled using full-covariance Gaussian distributions, and the penalty factor λ was set to 3.0, which was chosen according to the optimal clustering performance on the development dataset.
  This approach is referred to as the **clust_REF_BIC** approach in our experiments.
- **The UBM-MAP-CLR approach:** (described in Section 2.3)
  The audio data were represented by the same feature set as was used in the baseline BIC approach, but with the addition of feature warping (Pelecanos & Sridharan, 2001), which was performed on each segment separately. All the GMMs were constructed from 32 diagonal-covariance Gaussian mixtures. The UBM was estimated directly from the processing audio data by using the expectation-maximization algorithm (Theodoridis & Koutroumbas, 2003). No separate gender-derived models were trained. The MAP adaptation of (only) the UBM means was performed on each cluster to derive cluster-based GMMs. Next, the clusters where the highest CLR score in equation (5) was achieved were merged at each step of the merging process.
  This approach is referred to as the **clust_UBM_MAP_CLR** approach in our experiments.
- **The FUSION approach:** (described in Sections 3.1–3.2)
  The fusion of acoustic and prosodic representations is described by equation (10). The acoustic representation of the audio data was implemented by the same MFCC-based features as were used in the above approaches. The prosodic features were derived at every speaker segment and were not changed during the clustering. When combining the Δ_{BIC} measure from equation (8) and the prosodic measure from equation (9) into the weighted sum (10), the weighting parameter α needed to be set. This parameter was tuned on the development dataset and set to a value of 0.85. This was in accordance with our expectation that the main discriminative information for speaker clustering is stored in the acoustics, while the prosody provides only supplementary information. Note that we used the same penalty factor, λ=3.0, in the Δ_{BIC} measure as was used in the baseline BIC approach.
  This approach is referred to as the **clust_FUSION** approach in our experiments.

4.3 Evaluation results

An assessment of the selected clustering approaches was performed on the SiBN and the COST278 BN databases. The experiments were conducted in such a way as to evaluate the performance of the clustering approaches in various acoustic conditions. The SiBN database

consists of BN shows of one TV station, including the same set of speakers, and was collected in unchanged recording conditions. For this reason it was considered to represent relatively homogeneous data. On the other hand, the COST278 BN database consists of BN shows in different languages from several TV and radio stations, it includes a wide range of speakers and the data were collected under different recording conditions. As such it represented relatively inhomogeneous audio data in terms of different speakers and acoustic environments.

The speaker-diarization results, which were produced by running all three speaker-clustering approaches on the SiBN and COST278 BN databases, are shown in Figures 2 and 3, respectively. The DER results, plotted in Figures 2 and 3, should be interpreted as follows: the DER results at the evaluation point 0 correspond to the average of the DER across all the evaluated audio files, where the number of clusters is equal to the actual number of speakers in each file, the DER results at evaluation point +5 correspond to the average of the DER across all the evaluated audio files, where the number of clusters exceeds the actual number of speakers in each file by 5, and analogously, the DER results at evaluation point -5 correspond to the average of the DER across all the evaluated audio files, where the number of clusters is 5 clusters lower than the actual number of speakers in each file, and so on.

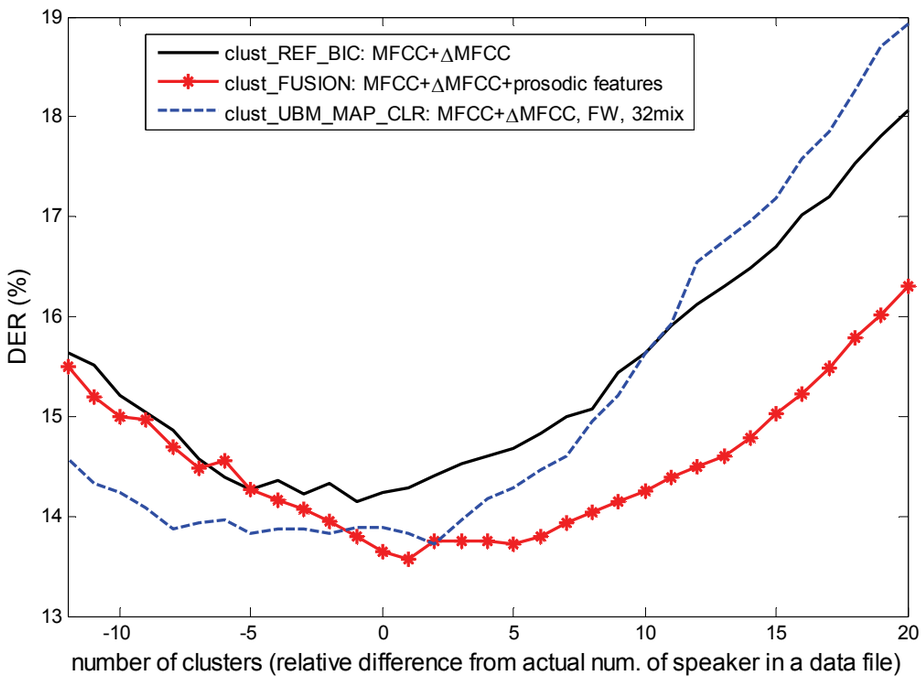

Fig. 2. Speaker-diarization results on the SiBN database when using different clustering procedures. The lower DER values correspond to better performance

The speaker-diarization results in Figure 2 correspond to the speaker-clustering performance of the tested approaches on the SiBN data. The overall performance of the speaker-clustering approaches varies between 13.5% and 16%, measured using the overall DER. The *clust_UBM_MAP_CLR* and *clust_FUSION* approaches perform slightly better than the baseline *clust_REF_BIC* approach across the whole range of evaluation points. When the

clust_FUSION and the *clust_REF_BIC* approaches are compared, it is clear that the SiBN results display significant differences in the speaker diarization performance of both approaches, which is in favor of the *clust_FUSION* approach. This indicates that adding the prosodic characteristics of speakers to the basic acoustic information could improve the speaker clustering. The same can be concluded from comparing the *clust_UBM_MAP_CLR* approach with the baseline BIC approach. The performance of the *clust_UBM_MAP_CLR* approach improved when enough clustering data were available for the GMM estimations, which resulted in lower DERs in comparison to the baseline BIC approach, when the number of clusters shrinks (the DER results display a better performance for the *clust_UBM_MAP_CLR* approach in the range below the evaluation point +10 in Figure 2).
It is also interesting to note that the DER trajectories of all the approaches achieved their minimum DER values around the evaluation point 0. This means that if all the clustering approaches were to be stopped when the number of clusters is equal to the number of actual speakers in the data, all the approaches would exhibit their optimum speaker-diarization performance. At that point the best clustering result was achieved with the *clust_FUSION* approach.

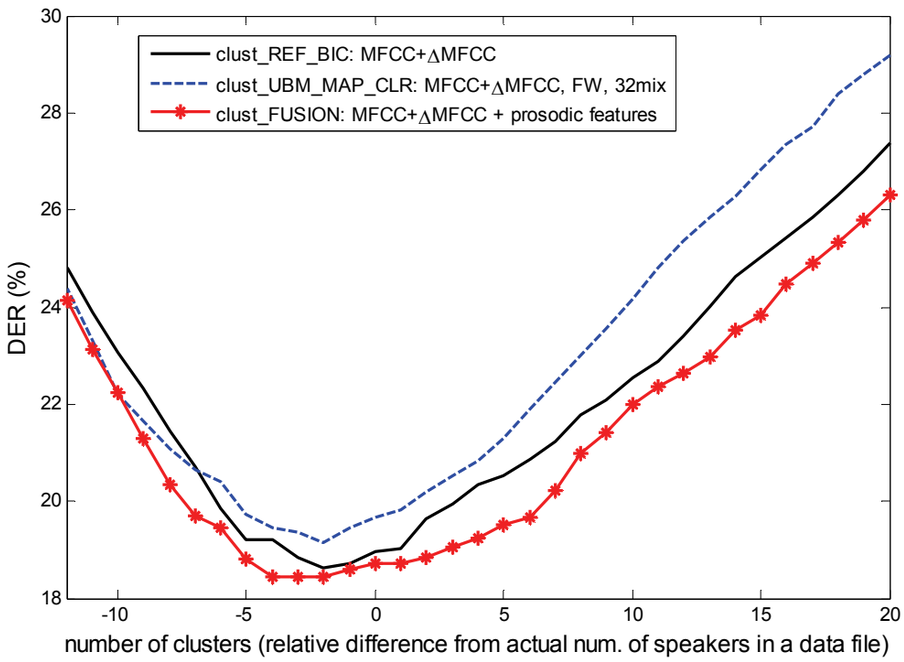

Fig. 3. Speaker-diarization results on the COST278 BN database when using different clustering procedures. The lower DER values correspond to a better performance.

Another interesting conclusion can be drawn from observing the flatness of the DER trajectories. Since the proposed evaluation measure aimed to compute the DER values at the relative numbers of clusters in each file, no stopping criteria needed to be applied; however, in practice the proper stopping of the clustering should be ensured. The optimum stopping criteria should end the merging process at the point with the lowest DER, which should coincide with the evaluation point 0, where the number of clusters is equal to the number of actual speakers in the data. Around this point it is better for the approaches to produce

relatively flat DER trajectories, which would result in a small loss of speaker-diarization performance, when the stopping criteria would not find the exact position for ending the merging process. In the case of the SiBN results, the DER trajectory, produced by the *clust_FUSION* approach, is flatter around the evaluation point 0 than the DER trajectories, produced by the *clust_REF_BIC* and *clust_UBM_MAP_CLR* approaches.

The speaker-diarization results in Figure 3 were produced by running the tested clustering approaches on the COST278 BN database. The results demonstrate the similar clustering performance of the approaches as in the case of the SiBN data, even though the overall DERs are higher than in the SiBN case. This was expected, since the COST278 BN data includes many more speakers in various acoustic environments than the SiBN data, and thus the clustering problem was more complex. In this situation the *clust_FUSION* approach produced the best overall speaker-diarization results, while the *clust_REF_BIC* approach performed slightly better than the *clust_UBM_MAP_CLR* approach. This means that in the case of adverse acoustic conditions it is better to model the cluster data by adding prosodic information to the cluster representations rather than modeling them just with acoustic representations (the *clust_REF_BIC* approach) or by a more precise acoustic modeling with the GMMs (the *clust_UBM_MAP_CLR* approach).

5. Discussion

In short, we have looked at three speaker-clustering approaches. The first was a standard approach using a bottom-up agglomerative clustering principle with the BIC as a merging criterion. In the second system an alternative approach was applied, also using bottom-up clustering, but the representations of the speaker clusters and the merging criteria are different. In this approach the speaker clusters were modeled by GMMs. In the clustering procedure during the merging process the universal background model was transformed into speaker-cluster GMMs using the MAP adaptation technique. The merging criterion in this case was a cross log-likelihood ratio (CLR). A totally new approach was developed within the fusion speaker-clustering system, where the speaker segments are modeled by acoustic and prosodic representations. The idea was to additionally model the speaker's prosodic characteristics and add them to the basic acoustic information. We constructed 10 basic prosodic features derived from the energy of the audio signals, the estimated the pitch contours, and the recognized voiced-unvoiced regions in the speech, which represented the basic speech units. By adding prosodic information to the basic acoustic features the baseline clustering procedure had to be changed to work with the fusion of both representations.

We performed two evaluation experiments where the overall diarization error rate was used as an assessment measure for the three tested clustering approaches. Experiments were performed on the SiBN and the COST278 BN databases. The evaluation results showed better performance for the tested systems in the SiBN case. This is due to the fact that the SiBN data included more homogeneous audio segments than the COST278 data, which resulted in an about 5% better performance for all of the clustering approaches. Furthermore, it was shown that speaker clustering, where the segments are modeled by speaker-oriented representations (speaker GMMs, prosodic features), were more stable and more reliable than the baseline system, where the segments are represented just by

acoustic information. The best overall results were achieved with the fusion system, where the clustering involved joining the acoustic and prosodic features. From this it can be concluded that the proposed fusion approach aimed at improving the speaker-diarization performance, especially in the case of processing BN data, where the speaker's speech characteristics across one BN show do not change significantly, but the speaker's clustering data can be biased due to different acoustic environments or background conditions.

6. Conclusion

Speaker clustering represents the last step in the speaker-diarization process. While the aim of the speech detection and speaker- and acoustic-segmentation procedures is to provide the proper segmentation of audio data streams, the purpose of the speaker clustering is to connect together segments that belong to the same speakers. In this chapter we solved this problem by applying agglomerative clustering methods. We concentrated on developing proper representations of the speaker segments for clustering and researched different similarity measures for joining the speaker segments that would result in a minimization of the overall diarization error for such systems. We realized three speaker-clustering systems, two of them operated on acoustic representations of speech, while the newly proposed one was designed to include prosodic information in addition to the basic acoustic representations. In this way we were able to impose higher-level information in the representations of the speaker segments, which led to improved clustering of the segments in the case of similar speaker acoustic characteristics in adverse acoustic conditions.

7. Acknowledgment

This work was supported by Slovenian Research Agency (ARRS), development project M2-0210 (C) entitled "AvID: Audiovisual speaker identification and emotion detection for secure communications."

8. References

Ajmera, J. & Wooters, C. (2003). A Robust Speaker Clustering Algorithm, *Proceedings of IEEE ASRU Workshop*, pp. 411-416, St. Thomas, U.S. Virgin Islands, November 2003.

Anastasakos, T.; McDonough, J.; Schwartz, R.; & Makhoul J. (1996) A Compact Model for Speaker-Adaptive Training, *Proceedings of International Conference on Spoken Language Processing (ICSLP1996),* pp. 1137-1140, Philadelphia, PA, USA, 1996.

Baker, B.; Vogt, R. & Sridharan, S. (2005). Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent Speaker Verification, *Proceedings of Interspeech 2005 - Eurospeech*, Lisbon, Portugal, September 2005.

Barras, C.; Zhu, X.; Meignier, S. & Gauvain, J.-L. (2004). Improving Speaker Diarization, *Proceedings of DARPA Rich Transcription Workshop 2004,* Palisades, NY, USA, November, 2004.

Barras, C.; Zhu, X.; Meignier, S. & Gauvain, J.-L. (2006). Multistage Speaker Diarization of Broadcast News. *IEEE Transactions on Speech, Audio and Language Processing, Special Issue on Rich Transcription*, Vol. 14, No. 5, (September 2006), pp. 1505-1512.

Beyerlein, P.; Aubert, X.; Haeb-Umbach, R.; Harris, M.; Klakow, D.; Wendemuth, A.; Molau, S.; Ney, H.; Pitz, M. & Sixtus, A. (2002). Large vocabulary continuous speech recognition of Broadcast News – The Philips/RWTH approach. *Speech Communication*, Vol. 37, No. 1-2, (May 2002), pp. 109-131.

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

Delacourt, P.; Bonastre, J.; Fredouille, C.; Merlin, T. & Wellekens, C. (2000). A Speaker Tracking System Based on Speaker Turn Detection for NIST Evaluation, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2000)*, Istanbul, Turkey, June, 2006.

Do, M. N. (2003). Fast Approximation of Kullback-Lebler Distance for Dependence Trees and Hidden Markov Models. *Signal Processing Letters*, Vol. 10, (2003), pp. 115-118.

Fiscus, J. G.; Garofolo, J. S.; Le, A.; Martin, A. F.; Pallett, D. S.; Przybocki M. A. & Sanders, G. (2004). Results of the Fall 2004 STT and MDE Evaluation, *Proceedings of the Fall 2004 Rich Transcription Workshop*, Palisades, NY, USA, November, 2004.

Galliano, S.; Geoffrois, E.; Mostefa, D.; Choukri, K.; Bonastre, J.-F. & Gravier, G. (2005). The ESTER phase II evaluation campaign of rich transcription of French broadcast news, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 1149-1152, Lisbon, Portugal, September, 2005.

Gallwitz, F.; Niemann, H.; Noth, E. & Warnke, V. (2002). Integrated recognition of words and prosodic phrase boundaries. *Speech Communication*, Vol. 36, No. 1-2, January 2002, pp. 81–95.

Gauvain, J. L.; & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech Audio Processing*, Vol. 2, No. 2, (April 1994), pp. 291-298.

Gauvain, J. L.; Lamel, L. & Adda, G. (2002). The LIMSI Broadcast News transcription system. *Speech Communication*, Vol. 37, No. 1-2, (May 2002), pp. 89-108.

Istrate, D.; Scheffer, N.; Fredouille, C. & Bonastre, J.-F. (2005). Broadcast News Speaker Tracking for ESTER 2005 Campaign, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 2445-2448, Lisbon, Portugal, September, 2005.

Jain, A.; Nandakumar, K. & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, Vol. 38, No. 12, (December 2005), pp. 2270-2285.

Kajarekar, S.; Ferrer, L.; Venkataraman, A.; Sonmez, K., Shriberg, E.; Stolcke, A. & Gadde, R.R. (2003). Speaker Recognition Using Prosodic and Lexical Features, *Proceedings of IEEE ASRU Workshop,* St. Thomas, U.S. Virgin Islands, November 2003.

Makhoul, J.; Kubala, F.; Leek, T.; Liu, D.; Nguyen, L.; Schwartz, R. & Srivastava, A. (2000). Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, Vol. 88, No. 8, (2000) pp. 1338-1353.

Matsoukas, S.; Schwartz, R.; Jin, H. & Nguyen, L. (1997). Practical Implementations of Speaker-Adaptive Training, *Proceedings of the 1997 DARPA Speech Recognition Workshop*, Chantilly VA, USA, February, 1997.

Meignier, S.; Bonastre, J.-F.; Fredouille, C. & Merlin T. (2000). Evolutive HMM for Multi-Speaker Tracking System, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000.

Moh, Y.; Nguyen, P. & Junqua, J.-C. (2003). Towards Domain Independent Clustering, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 85-88, Hong Kong, April 2003.

Moraru, D.; Ben, M. & Gravier, G. (2005). Experiments on speaker tracking and segmentation in radio broadcast news, *Proceedings of Interspeech 2005 - Eurospeech*, Lisbon, Portugal, September 2005.

Nedic, B.; Gravier, G.; Kharroubi, J.; Chollet, G.; Petrovska, D.; Durou, G.; Bimbot, F.; Blouet, R.; Seck, M.; Bonastre, J.-F.; Fredouille, C.; Merlin, T.; Magrin-Chagnolleau, I.; Pigeon, S.; Verlinde, P. & Cernocky J. (1999). The Elisa'99 Speaker Recognition and Tracking Systems, *Proceedings of IEEE Workshop on Automatic Advanced Technologies*, 1999.

Noth, E.; Batliner, A.; Warnke, V.; Haas, J.; Boros, M.; Buckow, J.; Huber, R.; Gallwitz, F.; Nutt, M. & Niemann, H. (2002). On the use of prosody in automatic dialogue understanding. *Speech Communication*, Vol. 36, No. 1-2, January 2002, pp. 45–62.

Pelecanos, J. & Sridharan, S. (2001). Feature warping for robust speaker verification. *Proceedings of Speaker Odyssey*, pp. 213–218, Crete, Greece, June 2001.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, Vol. 81, No. 9, (1993) pp. 1215-1247.

Ramos-Castro, D,; Garcia-Romero, D.; Lopez-Moreno, I. & Gonzalez-Rodriguez, J. (2005). Speaker verification using fast adaptive TNORM based Kullback-Leibler divergence, *Third COST 275 Workshop: Biometrics on the Internet*, University of Hertfordshire, Great Britain, October, 2005.

Reynolds, D. A.; Quatieri, T. F. & and R. B. Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, Vol. 10, No. 1, January 2000, pp. 19–41.

Reynolds, D. A.; Campbell, J. P.; Campbell, W. M.; Dunn, R. B.; Gleason, T. P.; Jones, D. A.; Quatieri, T. F.; Quillen, C.B.; Sturim, D. E. & Torres-Carrasquillo, P. A. (2003). Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition, *Proceedings of the Workshop on Multimodal User Authentication*, pp. 223-229, Santa Barbara, California, USA, December, 2003.

Reynolds, D. A. & Torres-Carrasquillo, P. (2004). The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations, Proceedings of the Fall 2004 Rich Transcription Workshop. Palisades, NY, USA, November, 2004.

Schwartz, G. (1976). Estimating the Dimension of a Model. *Annals of Statistics*, Vol. 6, pp. 461-464.

Shriberg, E.; Ferrer, L.; Kajarekar, S.; Venkataraman, A. & Stolcke, A. (2005). Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, Vol. 46, No. 3-4, (July 2005), pp. 455--472.

Sinha, R.; Tranter, S. E.; Gales, M. J. F. & Woodland, P. C. (2005). The Cambridge University March 2005 Speaker Diarisation System, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 2437-2440, Lisbon, Portugal, September, 2005.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In: *Speech Coding and Synthesis.* W. B. Kleijn & K. K. Paliwal, (Eds.), Elsevier Science, 1995.

Theodoridis, S. & Koutroumbas, K. (2003). *Pattern Recognition, second edition*. Academic Press, ISBN 0-12-685875-6, Elsevier, USA.

Tranter, S. & Reynolds, D. (2006). An Overview of Automatic Speaker Diarisation Systems. *IEEE Transactions on Speech, Audio and Language Processing, Special Issue on Rich Transcription*, Vol. 14, No. 5, (September 2006), pp. 1557-1565.

Tritschler, A. & Gopinath, R. (1999). Improved speaker segmentation and segments clustering using the Bayesian information criterion, *Proceedings of EUROSPEECH 99*, pp. 679-682, Budapest, Hungary, September, 1999.

Vandecatseye, A.; Martens, J.-P.; Neto J.; Meinedo, H.; Garcia-Mateo, C; Dieguez, J.; Zibert, J.; Mihelic, F.; Nouza, J.; David, P.; Pleva M.; Cizmar, A.; Papageorgiou, H.; Alexandris, C.; & Mihelič, F. (2004). The COST278 pan-European Broadcast News Database, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 873-876, Lisbon, Portugal, May 2004.

Woodland, P. C. (2002). The development of the HTK Broadcast News transcription system: An overview. *Speech Communication*, Vol. 37, No. 1-2, (May 2002), pp. 47--67.

Žibert, J. & Mihelič, F. (2004). Development of Slovenian Broadcast News Speech Database, *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*, pp. 2095-2098, Lisbon, Portugal, May 2004.

Žibert, J.; Mihelič, F.; Martens, J.-P.; Meinedo, H.; Neto, J.; Docio, L.; Garcia-Mateo, C.; David, P.; Zdansky, J.; Pleva, M.; Cizmar, A.; Žgank, A.; Kačič, Z.; Teleki, C. & Vicsi, K. (2005). The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation - Overview, Methodology, Systems, Results, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 629--632, Lisbon, Portugal, September, 2005.

Žibert, J.; Pavešić, N. & Mihelič, F. (2006a). Speech/Non-Speech Segmentation Based on Phoneme Recognition Features. *EURASIP Journal on Applied Signal Processing*, Vol. 2006, No. 6, Article ID 90495, pp. 1-13.

Žibert, J. (2006b).*Obdelava zvočnih posnetkov informativnih oddaj z uporabo govornih tehnologij, PhD thesis (in Slovenian language),* Faculty of Electrical Engineering, University of Ljubljana, Slovenia.

Žibert, J.; Vesnicer, B. & Mihelič, F. (2007). Novel Approaches to Speech Detection in the Processing of Continuous Audio Streams. In: *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, (Eds.), 23-48, I-Tech Education and Publishing, ISBN 978-3-902613-08-0, Croatia.

Zhou, B. & Hansen, J. (2000). Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion, *Proceedings of International Conference on Spoken Language Processing (ICSLP 2000)*, pp. 714-717, Beijing, China, October, 2000.

Zhu, X.; Barras, C.; Meignier, S. & Gauvain, J.-L. (2005). Combining Speaker Identification and BIC for Speaker Diarization, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 2441-2444, Lisbon, Portugal, September, 2005.

Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V. & Woodland, P. C. (2004). *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department, Cambridge, United Kingdom.

Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen