2

Novel Approaches to Speech Detection in the Processing of Continuous Audio Streams

Janez Žibert, Boštjan Vesnicer, France Mihelič
Faculty of Electrical Engineering, University of Ljubljana
Slovenia

1. Introduction

With the increasing amount of information stored in various audio-data documents there is a growing need for the efficient and effective processing, archiving and accessing of this information. One of the largest sources of such information is spoken audio documents, including broadcast-news (BN) shows, voice mails, recorded meetings, telephone conversations, etc. In these documents the information is mainly relayed through speech, which needs to be appropriately processed and analysed by applying automatic speech and language technologies.

Spoken audio documents are produced by a wide range of people in a variety of situations, and are derived from various multimedia applications. They are usually collected as continuous audio streams and consist of multiple audio sources. These audio sources may be different speakers, music segments, types of noise, etc. For example, a BN show typically consists of speech from different speakers as well as music segments, commercials and various types of noises that are present in the background of the reports. In order to efficiently process or extract the required information from such documents the appropriate audio data need to be selected and properly prepared for further processing. In the case of speech-processing applications this means detecting just the speech parts in the audio data and delivering them as inputs in a suitable format for further speech processing. The detection of such speech segments in continuous audio streams and the segmentation of audio streams into either detected speech or non-speech data is known as the speech/non-speech (SNS) segmentation problem. In this chapter we present an overview of the existing approaches to SNS segmentation in continuous audio streams and propose a new representation of audio signals that is more suitable for robust speech detection in SNS-segmentation systems. Since speech detection is usually applied as a pre-processing step in various speech-processing applications we have also explored the impact of different SNS-segmentation approaches on a speaker-diarisation task in BN data.

This chapter is organized as follows: In Section 2 a new high-level representation of audio signals based on phoneme-recognition features is introduced. First of all we give a short overview of the existing audio representations used for speech detection and provide the basic ideas and motivations for introducing a new representation of audio signals for SNS segmentation. In the remainder of the section we define four features based on consonant-vowel pairs and the voiced-unvoiced regions of signals, which are automatically detected by

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

a generic phoneme recognizer. We also propose the fusion of different selected representations in order to improve the speech-detection results. Section 3 describes the two SNS-segmentation approaches used in our evaluations, one of which was specially designed for the proposed feature representation. In the evaluation section we present results from a wide range of experiments on a BN audio database using different speech-processing applications. We try to assess the performance of the proposed representation using a comparison with existing approaches for two different tasks. In the first task the performance of different representations of the audio signals is assessed directly by comparing the evaluation results of speech and non-speech detection on BN audio data. The second group of experiments tries to determine the impact of SNS segmentation on the subsequent processing of the audio data. We then measure the impact of different SNS-segmentation systems when they are applied in a pre-processing step of an evaluated speaker-diarisation system that is used as a speaker-tracking tool for BN audio data.

2. Phoneme-Recognition Features

2.1 An Overview of Audio Representations for Speech Detection

As briefly mentioned in the introduction, SNS segmentation is the task of partitioning audio streams into speech and non-speech segments. While speech segments can be easily defined as regions in audio signals where somebody is speaking, non-speech segments represent everything that is not speech, and as such consist of data from various acoustical sources, e.g., music, human noises, silences, machine noises, etc.

Earlier work on the separation of audio data into speech and non-speech mainly addressed the problem of classifying known homogeneous segments as either speech or music, and not as non-speech in general. The research was focused more on developing and evaluating characteristic features for classification, and the systems were designed to work on already-segmented data.

Saunders (Saunders, 1996) designed one such system using features pointed out by (Greenberg, 1995) to successfully discriminate between speech and music in radio broadcasting. For this he used time-domain features, mostly derived from zero crossing rates. In (Samouelian et al., 1998) time-domain features, combined with two frequency measures, were also used. The features for speech/music discrimination that are closely related to the nature of human speech were investigated in (Scheirer & Slaney, 1997). The proposed measures, i.e., the spectral centroid, the spectral flux, the zero-crossing rate, the *4-Hz* modulation energy (related to the syllable rate of speech), and the percentage of low-energy frames were explored in an attempt to discriminate between speech and various types of music. The most commonly used features for discriminating between speech, music and other sound sources are the cepstrum coefficients. The mel-frequency cepstral coefficients (MFCCs) (Picone, 1993) and the perceptual linear prediction (PLP) cepstral coefficients (Hermansky, 1990) are extensively used in speaker- and speech-recognition tasks. Although these signal representations were originally designed to model the short-term spectral information of speech events, they were also successfully applied in SNS-discrimination systems (Hain et al., 1998; Beyerlein et al., 2002; Ajmera, 2004; Barras et al., 2006; Tranter & Reynolds, 2006) in combination with Gaussian mixture models (GMMs) or hidden Markov models (HMMs) for separating different audio sources and channel conditions (broadband speech, telephone speech, music, noise, silence, etc.). The use of these representations is a natural choice in speech-processing applications based on automatic

speech recognition since the same feature set can be used later on for the speech recognition. An interesting approach was proposed in (Parris et al., 1999), where a combination of different feature representations of audio signals in a GMM-based fusion system was made to discriminate between speech, music and noise. They investigated energy, cepstral and pitch features.
These representations and approaches focused mainly on the acoustic properties of data that are manifested in either the time and frequency or the spectral (cepstral) domains. All the representations tend to characterize speech in comparison to other non-speech sources (mainly music). Another perspective on the speech produced and recognized by humans is to treat it as a sequence of recognizable units. Speech production can thus be considered as a state machine, where the states are phoneme classes (Ajmera et al., 2003). Since other non-speech sources do not possess such properties, features based on these characteristics can be usefully applied in an SNS classification. The first attempt in this direction was made by Greenberg (Greenberg, 1995), who proposed features based on the spectral shapes associated with the expected syllable rate in speech. Karnebäck (Karnebäck, 2002) produced low-frequency modulation features in the same way and showed that in combination with the MFCC features they constitute a robust representation for speech/music discrimination tasks. A different approach based on this idea was presented in (Williams & Ellis, 1999). They built a phoneme speech recognizer and studied its behaviour with different speech and music signals. From the behaviour of the recognizer they proposed posterior-probability-based features, i.e., *entropy and dynamism*, and used them for classifying the speech and music samples.

2.2 Basic Concepts and Motivations

The basic SNS-classification systems typically include statistical models representing speech data, music, silence, noise, etc. They are usually derived from training material, and then a partitioning method detects the speech and non-speech segments according to these models. The main problem with such systems is the non-speech data, which are produced by various acoustic sources and therefore possess different acoustic characteristics. Thus, for each type of such audio signals one needs to build a separate class (typically represented as a model) and include it in a system. This represents a serious drawback with SNS-segmentation systems, which need to be data independent and robust to different types of speech and non-speech audio sources.
On the other hand, the SNS-segmentation systems are meant to detect speech in audio data and should discard non-speech parts, regardless of their different acoustic properties. Such systems can be interpreted as two-class classifiers, where the first class represents speech samples and the second class represents everything else. In this case the speech class defines the non-speech class. Following on from this basic concept one should find and use those characteristics or features of audio signals that better emphasize and characterize speech and exhibit the expected behaviour with all other non-speech audio data.
While the most commonly used acoustic features (MFCCs, PLPs, etc.) perform well when discriminating between different speech and non-speech signals, (Logan, 2000), they still only operate on an acoustic level. Hence, the data produced by the various audio sources with different acoustic properties needs to be modelled by several different classes and represented in the training process of such systems. To avoid this, we decided to design an

audio representation that would better determine the speech and perform significantly differently on all other non-speech data.
One possible way to achieve this is to see speech as a sequence of basic speech units that convey some meaning. This rather broad definition of speech led us to examine the behaviour of a simple phoneme recognizer and analyze its performance on speech and non-speech data. In that respect we followed the idea of Williams & Ellis, (Williams & Ellis, 1999), but rather than examine the functioning of phoneme recognizers, as they did, we analyzed the output transcriptions of such recognizers in various speech and non-speech situations.

2.3 Features Derivation

Williams & Ellis, (Williams & Ellis, 1999), proposed a novel method for discriminating between speech and music. They proposed measuring the posterior probability of observations in the states of neural networks that were designed to recognise basic speech units. From the analysis of the posterior probabilities they extracted features such as the mean per-frame entropy, the average probability dynamism, the background-label ratio and the phone distribution match. The entropy and dynamism features were later successfully applied to the speech/music segmentation of audio data (Ajmera et al., 2003). In both cases they used these features for speech/music classification, but the idea could be easily extended to the detection of speech and non-speech signals, in general. The basic motivation in both cases was to obtain and use features that were more robust to different kinds of music data and at the same time perform well on speech data.
In the same manner we decided to measure the performance of a speech recognizer by inspecting the output phoneme-recognition transcriptions, when recognizing speech and non-speech samples (Žibert et al., 2006a). In this way we also examined the behaviour of a phoneme recognizer, but the functioning of the recognizer was measured at the output of the recognizer rather than in the inner states of such a recognition engine.
Typically, the input of a phoneme recognizer consists of feature vectors based on the acoustic parameterization of speech signals, and the corresponding output is the most likely sequence of pre-defined speech units and time boundaries, together with the probabilities or likelihoods of each unit in a sequence. Therefore, the output information from a recognizer can also be interpreted as a representation of a given signal. Since the phoneme recognizer is designed for recognizing speech signals it is to be expected that it will exhibit characteristic behaviour when speech signals are passed through it, and all other signals will result in uncharacteristic behaviour. This suggests that it should be possible to distinguish between speech and non-speech signals just by examining the outputs of phoneme recognizers.
In general, the output from speech recognizers depends on the language and the models included in the recognizer. To reduce these influences the output speech units should be chosen from among broader groups of phonemes that are typical for the majority of languages. Also, the corresponding speech representation should not be heavily dependent on the correct transcription produced by the recognizer. Because of these limitations and the fact that human speech can be described as concatenated syllables, we decided to examine the functioning of recognizers in terms of the consonant-vowel (CV) level (Žibert et al., 2006a) and by inspecting the voiced and unvoiced regions (VU) of recognized audio signals (Mihelič & Žibert, 2006).

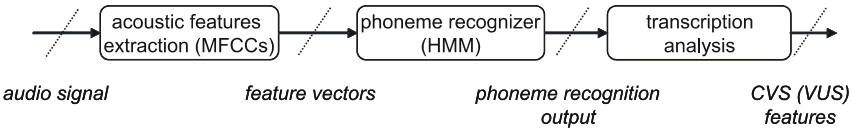


Figure 1. A block diagram showing the derivation of the phoneme-recognition features

The procedure for extracting phoneme-recognition features is shown in Figure 1. First, the acoustic representation of a given signal is produced and passed through a simple phoneme recognizer. Then, the transcription output is translated to specified phoneme classes, in the first case to the consonant (C), vowel (V) and silence (S) classes, and in the second case to the voiced (V), unvoiced (U) and silence (S) regions. At this point the output transcription is analysed, and those features that resemble the discriminative properties of speech and non-speech signals and are relatively independent of specific recognizer properties and errors are extracted. In our investigations we examined just those characteristics of the recognized outputs that are based on the *duration* and the *changing rate* of the basic units produced by the recognizer.

After a careful analysis of the functioning of several different phoneme recognizers for different speech and non-speech data conditions, we decided to extract the following features (Žibert et al., 2006a):

- **Normalized CV (VU) duration rate** of consonant-vowel (CV) or voiced-unvoiced (VU) pairs, defined in CV case as:

$$\frac{|t_C - t_V|}{t_{CVS}} + \alpha \cdot \frac{t_S}{t_{CVS}}, \tag{1}$$

where t_C is the overall time duration of all the consonants recognized in a signal window of time duration t_{CVS}, and t_V is the time duration of all the vowels in a window of duration t_{CVS}. The second term denotes the proportion of silence units (term t_S) represented in a recognized signal measured in time units. α serves as a weighting factor to emphasize the number of silence regions in a signal and has to be $0 \le \alpha \le 1$. In the VU case the above formula stays the same, whereas unvoiced phonemes replace consonants, voiced substitute vowels and silences are the same.

It is well known that speech is constructed from CV (VU) units in combination with S parts; however, we observed that speech signals exhibit relatively equal durations of C (U) and V (V) units and a rather small proportion of silences (S), which yielded small values (around 0.0) in Equation (1), measured on fixed-width speech segments. On the other hand, non-speech data were almost never recognized as a proper combination of CV or VU pairs, which is reflected in the different rates of C (U) and V (V) units, and hence the values of Equation (1) tend to be more like 1.0. In addition, when non-speech signals are recognized as silences, the values in the second term of Equation (1) follow the same trend as in the previous case.

Note that in Equation (1) we used the absolute difference between the durations, $|t_C - t_V|$, rather than the duration ratios, $\frac{t_C}{t_V}$ or $\frac{t_V}{t_C}$. This was done to reduce the effect of labelling, and not to emphasize one unit over another. The latter would result in the poor performance of this feature when using different speech recognizers.

- **Normalized average CV (VU) duration rate**, defined in the CV case as

$$\frac{\left|\overline{t_C} - \overline{t_V}\right|}{\overline{t_{CV}}}, \tag{2}$$

where $\overline{t_C}$ and $\overline{t_V}$ represent the average time durations of the C and V units in a given segment of a recognized signal, while $\overline{t_{CV}}$ is the average duration of all the recognized (C,V) units in the same segment. In the same way the normalized, average VU duration rate can be defined.
This feature was constructed to measure the difference between the average duration of the consonants (unvoiced parts) and the average duration of the vowels (voiced parts). It is well known that in speech the vowels (voiced parts) are in general longer than the consonants (unvoiced parts), and as a result this should be reflected in recognized speech. On the other hand, it was observed that non-speech signals do not exhibit such properties. Therefore, we found this feature to be sufficiently discriminative to distinguish between speech and non-speech data.
This feature correlates with the normalized time-duration rate defined in Equation (1). Note that in both cases the differences were used, instead of the ratios between the C (U) and V (V) units. This is for the same reason as in the previous case.

- **Normalized CV (VU) speaking rate**, defined in the CV case as

$$\frac{n_C + n_V}{t_{CVS}}, \tag{3}$$

where n_C and n_V are the number of C and V units recognized in the signal for the time duration t_{CVS}. The normalized VU speaking rate can be defined in the same manner. In both cases the silence units are not taken into account.
Since phoneme recognizers are trained on speech data they should detect changes when normal speech moves between phones every few tens of milliseconds. Of course, speaking rate in general depends heavily on the speaker and the speaking style. Actually, this feature is often used in systems for speaker recognition (Reynolds et al., 2003). To reduce the effect of speaking style, particularly spontaneous speech, we decided not to count the S units.
Even though the CV (VU) speaking rate in Equation (3) changes with different speakers and speaking styles, it varies less than for non-speech data. In the analyzed signals speech tended to change (in terms of the phoneme recognizer) much less frequently, but the signals varied greatly among different non-speech data types.

- **Normalized CVS (VUS) changes**, defined in the CV case as

$$\frac{c(C,V,S)}{t_{CVS}}, \tag{4}$$

where $c(C,V,S)$ counts how many times the C, V and S units exchange in the signal in the window of duration t_{CVS}. The same definition with V, U and S units can be produced in the VU case.
This feature is related to the CV (VU) speaking rate, but with one significant difference. Here, just the changes between the units that emphasize the pairs and not just the single units are taken into account. As speech consists of such CV (VU) combinations one should expect higher values when speech signals are decoded and lower values in the case of non-speech data.
This approach could be extended even further to observe higher-order combinations of the C, V, and S units to construct n-gram CVS (VUS) models (like in statistical language modelling), which could be additionally estimated from the speech and non-speech data.

As can be seen from the above definitions, all the proposed features measure the properties of recognized data on the pre-defined or automatically obtained segments of a processing signal. The segments should be large enough to provide reliable estimations of the proposed measurements. They depend on the size of the proportions of speech and non-speech data that were expected in the processing signals. We tested both possibilities of the segment sizes in our experiments. The typical segment sizes varied between 2.0 and 5.0 seconds in the fixed-segment size case. In the case of automatically derived segments the minimum duration of the segments was set to 1.5 seconds.
Another issue was how to calculate the features to be time aligned. In order to make a decision as to which proportion of the signal belongs to one or other class the time stamps between the estimation of consecutive features should be as small as possible. The natural choice would be to compute the features on moving segments between successive recognized units, but in our experiments we decided to keep a fixed frame skip, since we also used them in combination with the cepstral features.

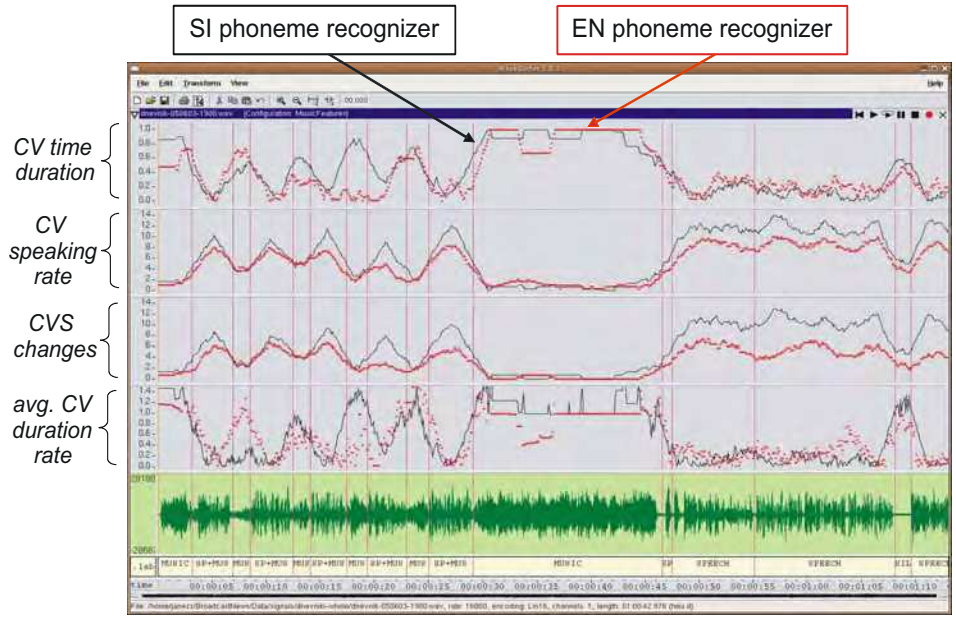


Figure 2. Estimation of the phoneme-recognition CVS features in a portion of a single broadcast news show in Slovene. The top four panes show the estimated CVS features from an audio signal that is shown in the fifth pane. The bottom pane displays the audio signal with the corresponding manual transcription. The top four panes consist of two lines. The black (darker) line represents the features obtained from a phoneme-based speech recognizer built to recognise Slovene speech, while the red (brighter) line displays the features obtained from the phoneme recognizer for English. All the data plots were produced using the *wavesurfer* tool, available at http://www.speech.kth.se/wavesurfer/.

Figure 2 shows phoneme-recognition features in action. In this example the CVS features were produced by phoneme recognizers based on two languages. One was built for Slovene (darker line in Figure 2), the other was trained on the TIMIT database (Garofolo et al., 1993) (brighter line), and was therefore used for recognizing English speech data. This example was extracted from a Slovenian BN show. The data in Figure 2 consist of different portions of speech and non-speech. The speech segments are built from clean speech, produced by different speakers in a combination with music, while the non-speech is represented by music and silent parts. As can be seen from Figure 2, each of these features has a reasonable ability to discriminate between the speech and non-speech data, which was later confirmed by our experiments. Furthermore, the features computed from the English speech recognizer, and thus in this case used on a foreign language, exhibit nearly the same behaviour as the features produced by the Slovenian phoneme decoder. This is a very positive result in terms of our objective to design features that should be language and model independent.

3. Speech Detection in Continuous Audio Streams

While there has been a lot of research done on producing appropriate representations of audio signals suitable for discriminating between speech and non-speech on already-segmented audio data, there have not been so many experiments conducted for speech detection in continuous audio streams, where the speech and non-speech parts are interleaving randomly. Such kinds of data are to be expected in most practical applications of automatic speech processing.

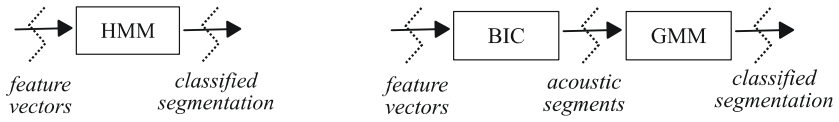
Most recent research in this field addresses this problem as part of large-vocabulary continuous-speech-recognition systems (LVCSRs), like BN transcription systems (Woodland, 2002; Gauvain et al., 2002; Beyerlein et al., 2002) or speaker-diarisation and speaker-tracking systems in BN data (Zhu et al., 2005; Sinha et al., 2005; Žibert et al., 2005; Istrate et al., 2005; Moraru et al., 2005; Barras et al., 2006; Tranter & Reynolds, 2006). In most of these investigations, energy and/or cepstral coefficients (mainly MFCCs) are used for the segmenting, and GMMs or HMMs are used for classifying the segments into speech and different non-speech classes. An alternative approach was investigated in (Lu et al., 2002), where the audio classification and segmentation were made by using support-vector machines. Another approach was presented in (Ajmera et al., 2003), where speech/music segmentation was achieved by incorporating GMMs into the HMM classification framework. This approach is also followed in our work and together with MFCC features it serves as a baseline SNS segmentation-classification method in our experiments.

In addition to our proposed representations, we also developed a method based on the acoustic segmentation of continuous audio streams obtained with the Bayesian information criterion (BIC) (Chen & Gopalakrishnan, 1998) and followed by the SNS classification.

In the following sections both segmentation-classification frameworks are described and compared using different audio-data representations.

3.1 Speech/Non-Speech-Segmentation Procedures

Block diagrams of the evaluated SNS-segmentation systems are shown in Figure 3.



(a) (b)

Figure 3. Block diagram of the two approaches used in the SNS segmentation. In (a) the segmentation and classification are performed simultaneously using HMM Viterbi decoding. In the second approach (b), firstly, the audio segmentation based on acoustic changes is performed by using the BIC segmentation procedure, followed by the GMM speech/non-speech classification.

The basic building blocks of both systems are GMMs. These models were trained with the EM algorithm in a supervised way (Young et al., 2004). In the first case, see Figure 3 (a), we followed the approach presented in (Ajmera, 2004), which was primarily designed for speech/music segmentation. Here, the segmentation and classification were performed simultaneously, by integrating already-trained GMMs into the HMM classification framework. We built a fully connected network consisting of N HMMs, as shown in Figure

4, where *N* represents the number of GMMs used in the speech/non-speech classification. Each HMM was constructed by simply concatenating the internal states associated with the same probability density function represented by one GMM. The number of states was fixed (*M* states in Figure 4) and set in such a way as to impose a minimum duration on each HMM. All the transitions inside each model were set manually, while the transitions between different HMMs were additionally trained on the evaluation data. In the segmentation process Viterbi decoding was used to find the best possible state sequence corresponding to speech and non-speech classes that could have produced the input-features sequence.

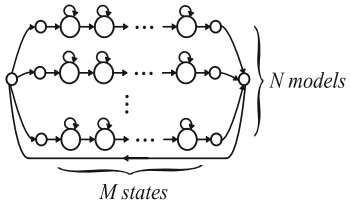


Figure 4. Topology of the HMM classification network used in the first procedure of the SNS segmentation.

In the second approach, see Figure 3 (b), the segmentation and classification were performed sequentially. The audio was segmented by applying the BIC measure to detect the acoustic-change points in the audio signals (Chen & Gopalakrishnan ,1998; Tritschler & Gopinath, 1999). Hence, in the first step of this procedure the segments based on acoustic changes were obtained, i.e., speaker, channel, background changes, different types of audio signals (music, speech), etc. In the next step these segments were classified as speech or non-speech. This classification was based on the same GMM set, which was also incorporated in the HMM classifier from the previous approach. In this way we could compare both methods using the same models. This approach is suited to the proposed CVS (VUS) features, which operate better on larger segments of signals than on smaller signal windows on a frame-by-frame basis.

In addition to both approaches, we also explored the fusion of different audio representations for SNS segmentation. The fusion of different representations was achieved at the score level in GMMs. We experimented with the fusion of the MFCC and the proposed CVS features, where each of the audio-signal representations form a separate feature stream. For each stream, separate GMMs were trained using the EM method. For SNS-segmentation purposes a similar HMM classification network to that of the non-fusion cases was built, see Figure 3 (a) and Figure 4, where in each state the fusion was made by computing the product of the weighted observation likelihoods produced by the GMMs from each stream. The product-stream weights were set empirically to optimize the performance on the evaluation dataset. In this way a fusion of the MFCC and CVS feature representations was performed by using a state-synchronous two-stream HMM (Potamianos et al., 2004).

3.3 Evaluation of Speech/Non-Speech Segmentation

We experimented with different approaches and representations of audio signals in order to find the best possible solution for the SNS discrimination in continuous audio streams.

We tested three main groups of features: acoustic features, represented by MFCCs, and the proposed CVS and VUS phoneme recognition, defined in Section 2. In addition, we combined both types of representations in one fusion SNS-segmentation system. All the representations were compared by using two SNS-segmentation approaches, presented in Section 3.1.

There were two evaluation databases: the development database, which was used to tune all the parameters of the audio representations and the SNS-segmentation systems, and the test dataset, which was composed of 12 hours of BN shows. The development dataset consisted of 6 hours of television entertainment and BN shows in different languages. A total of 4 hours of audio data were used to train the GMMs for SNS classification, the rest were used for setting the open parameters of the SNS-segmentation procedures to optimize their performances. The test database was used to compare the different audio representations and approaches in the SNS-segmentation task. This database is part of the audio database of BN shows in Slovene, which is presented in (Žibert & Mihelič, 2004).

3.3.1 Evaluation measures

The SNS-segmentation results were obtained in terms of the percentage of frame-level accuracy. We calculated three different statistics in each case: the percentage of true speech frames identified as speech, the percentage of true non-speech frames identified as non-speech, and the overall percentage of speech and non-speech frames identified correctly (the overall accuracy).

Note that in cases where one class dominates in the data (i.e., in the test-data case) the overall accuracy depends heavily on the accuracy of that class, and in such a case it cannot provide enough information on the performance of such a classification by itself. Therefore, in order to correctly assess classification methods one should provide all three statistics.

3.3.2 Evaluated SNS-segmentation systems

As a baseline system for the SNS classification we chose the MFCC features' representation in combination with the HMM classifier. We decided to use 12 MFCC features together with the normalized energy and first-order derivatives as a base representation, since no improvement was gained by introducing second-order derivatives. In that case 128-mixture GMMs for the modelling of several different speech and non-speech classes were trained. This baseline audio representation together with the HMM-based SNS segmentation is referred to as the *HMM-GMM: MFCC-E-D-26* system throughout the evaluation sections.

The above-described system was compared with different SNS approaches where phoneme-recognition features were used on their own and with the fusion system, where a combination of the MFCC and the CVS features were applied. The CVS and VUS features were obtained from two phoneme recognizers. One was built on Slovenian data, trained from three speech databases: GOPOLIS, VNTV and K211d, (Mihelič et al., 2003). It is referred as the *SI-phones* recognizer throughout the evaluation sections. The second was built from the TIMIT database, and thus was used for recognizing English speech. It is referred to as the *EN-phones* recognizer in all our experiments. Both phoneme recognizers were constructed from the HMMs of monophone units joined in a fully connected network.

Each HMM state was modelled by 32 diagonal-covariance Gaussian mixtures, built in a standard way, i.e., using 39 MFCCs, including the energy, and the first- and second-order derivatives, and setting all of the HMM parameters using the Baum-Welch re-estimation (Young et al., 2004). The phoneme sets of each language were different. In the *SI-phones* recognizer, 38 monophone base units were used, while in the TIMIT case, the base units were reduced to 48 monophones, according to (Lee & Hon, 1989). In both recognizers we used bigram phoneme language models in the recognition process. The recognizers were also tested on parts of the training databases. The *SI-phones* recognizer achieved a phoneme-recognition accuracy of about 70% on the GOPOLIS database, while the *EN-phones* recognizer had a phoneme-recognition accuracy of around 61% in a test part of the TIMIT database. Since our CVS (VUS) features were based on transcriptions of these recognizers, we also tested both recognizers on CVS recognition tasks. The *SI-phones* recognizer reached a CVS recognition accuracy of 88% on the GOPOLIS database, while for the *EN-phones* recognizer the CVS accuracy on the TIMIT database was around 75%. The same performance was achieved when recognizing the VUS units.

The CVS (VUS) features were calculated from phoneme-recognition transcriptions on the evaluation databases produced by both the *SI-phones* and *EN-phones* recognizers using the formulas defined in Section 2. The CVS (VUS) representations of the audio signal obtained from the *SI-phones* recognizer are named *SI-phones CVS (SI-phones VUS)* In the same manner, the CVS and VUS representations obtained from the *EN-phones* recognizer are marked as *EN-phones CVS (EN-phones VUS)*. The models used for classifying the speech and non-speech data were 2-mixture GMMs.

In the CVS (VUS) features case we tested both segmentation procedures, which were already described in Section 3.1. The segmentation performed by the HMM classifiers, based on trained speech/non-speech GMMs is referred to as the *HMM-GMM* and the segmentation based on the BIC measure, followed by the GMM classification, is referred to as the *BICseg-GMM*. In the *HMM-GMM* case the CVS (VUS) feature vectors were produced on a frame-by-frame basis. Hence, a fixed window length of 3.0 s with a frame rate of 100 ms was used in all the experiments. In Equation (1), α was set to 0.5. In the second approach the BIC segmentation produced acoustic segments computed from 12 MFCC features, together with the energy. The BIC measure was applied by using full-covariance matrices and a lambda threshold set according to the development dataset. These segments were then classified as speech or non-speech, according to the maximum log-likelihood criteria applied to the GMMs modelled by the CVS (VUS) features.

The fusion SNS-segmentation system was designed to join the MFCC and CVS feature representations into a two-stream HMMs classification framework. The GMMs from the *MFCC-E-D-26* and *SI-phones CVS* representations were merged into HMMs, and such an SNS-segmentation system is called a *HMM-GMM: fusion MFCC+CVS* system.

In the *HMM-GMM*-segmentation case the number of states used to impose the minimum duration constraint in the HMMs was fixed. This was done according to (Ajmera et al., 2003). Since in our evaluation-data experiments speech or non-speech segments shorter than 1.4 s were not found, we set the minimum duration constraint to 1.4 s, which corresponded to a different number of states with different types of representations. All the transition probabilities (including self-loop transitions) inside the HMMs were fixed to 0.5.

The HMM classification based on the Viterbi algorithm was made with the *HTKToolkit* (Young et al., 2004), while we provided our own tools for the BIC segmentation and the GMM classification and training.

3.3.3 Development Data Evaluations

The development dataset was primarily designed to serve for determining the models and for the tuning of other open parameters of the evaluated SNS-segmentation systems. Hence, this dataset was divided into the training part (4 hours) and the evaluation part (2 hours). In this subsection experiments on the evaluation data are outlined.

The evaluation data were intended mainly for tuning the threshold probability weights to favour the speech and non-speech models in the classification systems in order to optimize the overall performance of the SNS-segmentation procedures. Such optimal models were then used in the SNS-segmentation systems on the test data.

When plotting the overall accuracy of the SNS segmentation of the evaluation data against different choices of threshold probability weights, we were able to examine the performances of the evaluated approaches in optimal and non-optimal cases. In this way the constant overall accuracy of an SNS segmentation under different choices of probability weights could indicate the more stable performance of such an SNS-segmentation system in adverse acoustic or other audio conditions. The results of such experiments are shown in Figures 5 and 6.

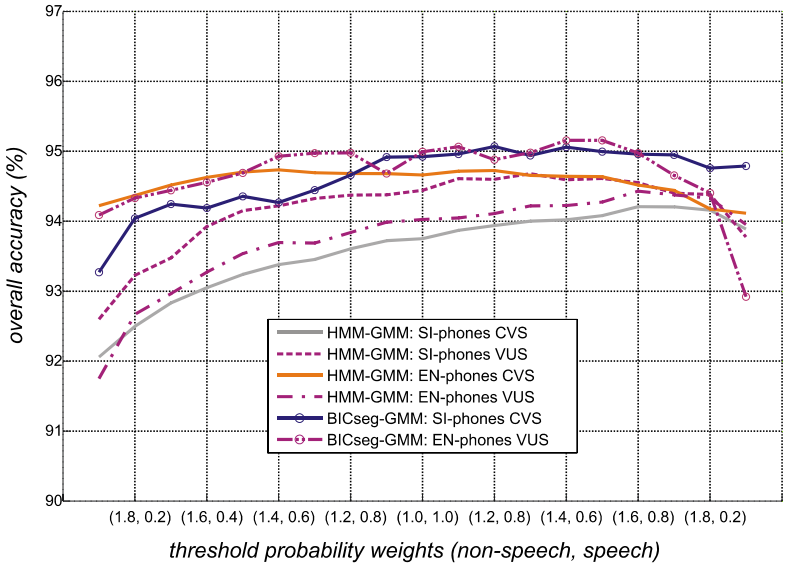


Figure 5. Determining the optimal threshold probability weights of the speech and non-speech models to maximize the overall accuracy of the CVS and VUS feature representations with different SNS-segmentation procedures.

Figure 5 shows a comparison of the different types of phoneme-recognition features with different SNS-segmentation procedures. As can be seen from Figure 5, all the segmentation methods based on both types of phoneme-recognition features are stable across the whole range of operating points of the threshold probability weights. The overall accuracy ranges between 92% and 95%. No important differences in the performance among the approaches based on the HMM classification and the BIC segmentation can be observed, even though the *BICseg-GMM* systems operated, on average, slightly better than their HMM-based counterparts. The same can be concluded when comparing CVS and VUS features computed from different phoneme recognizers. There is no significant difference in the performances when using *SI-phones* and *EN-phones* recognizers, even though the audio data in this development set are in Slovene. This proves that the phoneme-recognition features performed equally well, regardless of the spoken language that appeared in the audio data. When comparing the CVS and VUS feature types, no single conclusion can be made: the VUS features performed better than the CVS features when the *SI-phones* recognizers were used, but the opposite was the case when the CVS and VUS features derived from the *EN-phones* recognizers were applied.

In summary, the CVS and VUS features were stable and performed equally well across the whole range of threshold probability weights. They are also language independent and perform slightly better when they are derived from larger segments of data, like in the case of the *BICseg-GMM*-segmentation procedures. Therefore, we decided to use just the *SI-phones CVS* features in all the following evaluation experiments.

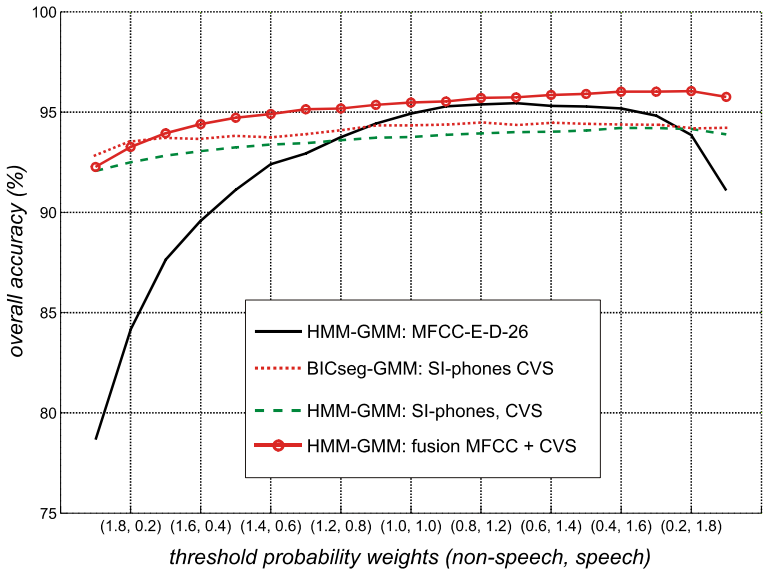


Figure 6. Determining the optimal threshold probability weights of the speech and non-speech models to maximize the overall accuracy of the different audio representations and SNS-segmentation procedures.

Figure 6 shows a comparison of the phoneme-recognition and acoustic features. The MFCC representation achieved the maximum accuracy, slightly above 95%, at the operating point (0.8,1.2). Around this point it performed better than the CVS-based segmentations, but in general the segmentation with just the MFCC features is more sensitive to the different operating points of the probability weights. The best overall performance was achieved by the fusion of both representations. The accuracy was increased to 96% (maximum values) around those operating points where the corresponding base representations achieved their own maximum performances. The fusion representation is also stable across the whole range of threshold probability weights due to the base CVS representation.
In general, it can be concluded that the CVS and VUS phoneme-recognition features were more stable than the acoustic MFCC features across the whole range of optimal and non-optimal cases. Therefore, it can be expected that they would also perform better in situations when the training and working conditions are not the same. In addition, a fusion of the phoneme and acoustic feature representations yielded the results with the highest overall accuracy.

3.3.4 Test Data Evaluations

In order to properly assess the proposed methods we performed an evaluation of the SNS-segmentation systems on 12 hours of audio data from BN shows. The results are shown in Table 1.

Classification & Features Type	*Speech Recognition (%)*	*Non-Speech Recognition (%)*	*Overall Accuracy (%)*
HMM-GMM: MFCC	97.9	58.8	95.4
HMM-GMM: *SI-phones* recognition, CVS units	98.2	91.3	97.8
BICseg-GMM: *EN-phones* recognition, CVS units	98.3	90.9	97.9
HMM-GMM: Fusion: MFCC + CVS	**99.3**	**86.4**	**98.5**

Table 1. Speech- and non-speech-segmentation results on 12 hours of audio data from BN shows.

The results in Table 1 were obtained when the optimum set of parameters was applied in all the evaluated SNS-segmentation procedures. The results on the test data reveal the same performance for the different methods as was the case in the development experiments. The results on the test set show that the proposed CVS representations of the audio signals performed better than just the acoustic MFCC representations. The advantage of using the proposed phoneme-recognition features becomes even more evident when they are compared in terms of speech and non-speech accuracies. In general, there exists a huge

difference between the CVS and the MFCC representations in correctly identifying non-speech data with a relatively small loss of accuracy when correctly identifying speech data. One of the reasons for this is the stability issue discussed in the previous subsection. In all cases of the CVS features (regardless of the segmentation method) this resulted in an increased overall accuracy in comparison to the MFCC features.
When comparing the results of just the CVS representations no substantial differences in the classifications can be found. The results from the *SI-phones* and the *EN-phones* recognizers confirm that the proposed features really are independent of the phoneme recognizers trained on speech from different languages. They also suggest that there are almost no differences when using different segmentation methods, even though in the case of the BIC segmentation and the GMM classification we got slightly better results.
In the case of fusing the MFCC and CVS features we obtained the highest scores in terms of overall accuracy, and the fusion of both representations performed better than their stand-alone counterparts.
In general, the results in Table 1 and in Figures 5 and 6 speak in favour of the proposed phoneme-recognition features. This can be explained by the fact that our features were designed to discriminate between speech and non-speech, while the MFCC features were, in general, developed for speech-processing applications. Another issue concerns stability, and thus the robustness of the evaluated approaches. For the MFCC features the performance of the segmentation depends heavily on the training data and the training conditions, while the classification with the CVS features in combination with the GMMs performed reliably on the development and test datasets. Our experiments with fusion models also showed that probably the most appropriate representation for the SNS classification is a combination of acoustic- and recognition-based features.
In next section the impact of the evaluated speech-detection approaches on speech-processing applications is discussed.

4. The Impact of Speech Detection on Speech-Processing Applications

In the introduction we explained that a good segmentation of continuous audio streams into speech and non-speech has many practical applications. Such a segmentation is usually applied as a pre-processing step in real-world systems for automatic speech processing: in automatic speech recognition (Shafran & Rose, 2003), like a broadcast-news transcription (Gauvain et al., 2002; Woodland, 2002; Beyerlein et al., 2002), in automatic audio indexing and summarization (Makhoul et al., 2000; Magrin-Chagnolleau & Parlangeau-Valles, 2002), in audio and speaker diarisation (Tranter & Reynolds, 2006; Barras et al., 2006; Sinha et al., 2005; Istrate et al., 2005; Moraru et al., 2005), in speaker identification and tracking (Martin et al., 2000), and in all other applications where efficient speech detection helps to greatly reduce the computational complexity and generate more understandable and accurate outputs. Accordingly, an SNS segmentation has to be easily integrated into such systems and should not increase the overall computational load.
Therefore, we additionally explored our SNS-segmentation procedures in a speaker-diarisation application of broadcast-news audio data. We focused mainly on the impact of different SNS-segmentation approaches to the final speech (speaker) processing results. The importance of accurate speech detection in each task of the speaker diarisation is evaluated and discussed in the following section.

4.1 Evaluation of the Impact of Speech Detection in a Speaker-Diarisation System

4.1.1 Speaker Diarisation-System Framework

Speaker diarisation is the process of partitioning input audio data into homogeneous segments according to the speaker's identity. The aim of speaker diarisation 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 LVCSR 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 diarisation system could also be used in speaker-identification or speaker-tracking systems, (Delacourt et al., 2000; Nedic et al., 1999).

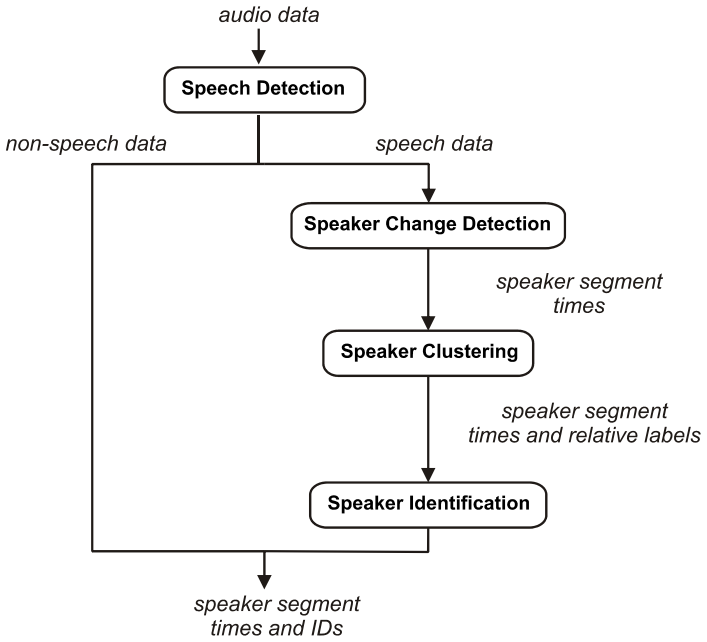


Figure 6. The main building blocks of a typical speaker-diarisation system. Most systems have components to perform speech detection, speaker or acoustic segmentation and speaker clustering, which may include components for gender detection and speaker identification.

Most speaker-diarisation systems have a similar general architecture to that shown in Figure 6. 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 further processing. The speech data are then chopped into homogeneous segments. The segment boundaries are located by finding acoustic changes in the signal, and each segment is expected to contain speech from only one speaker. The resulting segments are then clustered so that each cluster corresponds to just one speaker. At this stage, each cluster is labelled 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 a speaker diarisation 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-diarisation tasks can be found in (Tranter & Reynolds , 2006).

Our speaker-diarisation system, which was used for the current evaluation of speech-detection procedures, serves for speaker tracking in BN shows (Žibert, 2006b). All the components of the system were designed in such a way as to include the standard approaches from similar state-of-the-art systems. While the component for speech detection was derived from one of the SNS-segmentation procedures in each evaluation experiment, the audio segmentation, the speaker clustering and the speaker-identification procedures were the same in all experiments. The segmentation of the audio data was made 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 neighbouring 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 of the 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 or connect together segments that belong to the same speakers. In our system we realized this by applying a standard procedure using a bottom-up agglomerative clustering principle with the BIC as a merging criterion (Tranter & Reynolds, 2006). A speaker-identification component was adopted from a speaker-verification system, which was originally designed for the detection of speakers in conversational telephone speech (Martin et al., 2000). The speaker-verification system was based on a state-of-the-art Gaussian Mixture Model – Universal Background model (GMM-UBM) approach (Reynolds et al., 2000). The system made use of 26-dimensional feature vectors, composed of 12 MFCCs together with a log energy and their delta coefficients, computed every 10 ms and subjected to feature warping using a 3-s-long sliding window (Pelecanos & Sridharan, 2001). The log-likelihood scores produced by the system were normalized using the ZT-norm normalization technique (Auckenthaler et al., 2000).

All the open parameters and all the models used in each task of our speaker-diarisation system were estimated from the training data in such a way as to maximise the overall performance of the system.

4.1.2 Evaluation of the Impact

Since our speaker-diarisation system was constructed from four basic building blocks, we performed the evaluation of our speech-detection procedures after each processing block. Hence, the impact of the speech-detection procedures was measured, when using them as a pre-processing step of an audio-segmentation task, when using them together with an audio segmentation for speaker clustering, and in the final step, when measuring the overall speaker-tracking performance. The BN audio data used in the evaluation were the same as in the case of the evaluation of speech-detection procedures only, in Section 3.3.4. The audio-segmentation results, when using different speech-detection procedures, are shown in Table 2 and the final speaker-clustering and speaker-tracking results are shown in Figures 7 and 8, respectively.

Segmentation: baseline BIC method SNS segmentation:	*Recall (%)*	*Precision (%)*	*F-measure (%)*
Manual SNS segmentation	78.1	78.2	78.1
HMM-GMM: MFCC	60.0	80.7	68.8
HMM-GMM: *SI-phones* recognition, CVS	72.2	77.0	74.5
BICseg-GMM: *EN-phones* recognition, CVS	75.2	76.1	75.6
HMM-GMM: Fusion: MFCC + CVS	**75.7**	**76.8**	**76.3**

Table 2. Audio-segmentation results on BN audio data, when using different SNS-segmentation procedures.

The audio-segmentation performance in Table 2 was measured using three standard measures (Kemp et al., 2000): recall, precision and the F-measure. The recall is defined as the rate of correctly detected boundaries divided by the total number of boundaries, while the precision corresponds to the rate of correctly detected boundaries divided by the total number of hypothesized boundaries. Both measures are closely related to the well-known false-acceptance and false-rejection rates. The *F-measure* joins the recall and the precision in a single overall measure.

The overall segmentation results in Table 2 speak in favour of the proposed phoneme-recognition features (CVS), when using them as a representation of audio signals in speech-detection procedures. As has already been shown in the evaluation of speech-detection procedures alone (see Table 1), a baseline approach *HMM-GMM: MFCC* performed poorly in the detection of non-speech data. The non-speech segments were not detected, and consequently too many non-speech boundaries were not found. Therefore, the recall was too

low, and regardless of the relatively high precision the overall audio-segmentation results were not as good as in the other cases. We achieved relatively good results with both CVS representations in comparison to the manual SNS segmentation (in the first row of Table 2). The best overall results were achieved with the fusion representation of the MFCC and CVS features. The corresponding audio-segmentation results are just approximately 2% worse (measured by all three measures) than in the manual SNS-segmentation case. This proves that proper speech detection is an important part of an audio-segmentation system and that a good SNS segmentation can greatly improve the overall audio-segmentation results. This fact becomes even more obvious when different speech-detection procedures were compared in a speaker-clustering task, shown in Figure 7.

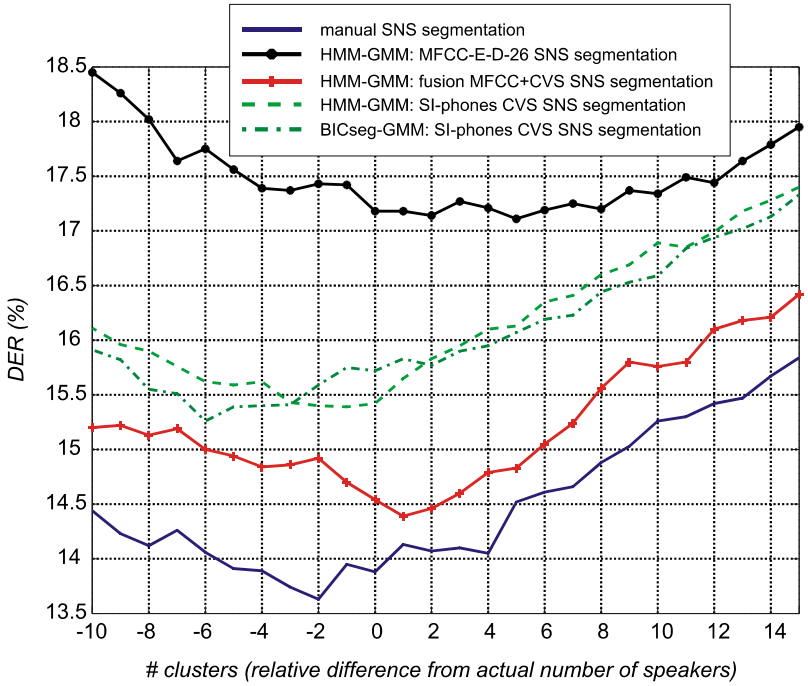


Figure 7. Speaker-clustering results when using different SNS-segmentation procedures. The lower DER values correspond to better performance.

Figure 7 shows a comparison of the four speech-detection procedures when using them together with the audio segmentation in the speaker-clustering task. The speaker clustering was evaluated by measuring the speaker-diarisation performance in terms of the diarisation error rate (DER), (Fiscus et al., 2004). The comparison was made when no stopping criteria were used in the speaker-clustering procedure. Hence, the impact of different speech-detection approaches was compared across the whole range of possible numbers of speaker clusters.

In Figure 7, the overall performance of the speaker clustering when using different SNS-segmentation procedures varies between 13.5% and 18.5%, measured using the DER. The speaker-clustering system, where the manual SNS segmentation was applied, was the best performing of all the evaluated procedures. In second place was the SNS segmentation with the fusion of the CVS and MFCC features. The DER results show on average an approximately 1% loss of performance with such speaker clustering. Speaker-clustering approaches show comparable performance, where just CVS representations of the audio signals were used in combination with different SNS-segmentation systems. A baseline speaker-clustering approach with MFCC features performed, on average, 3% worse (in absolute figures) than the best-evaluated approaches. These results also indicate the importance of speech detection in speaker-clustering procedures.

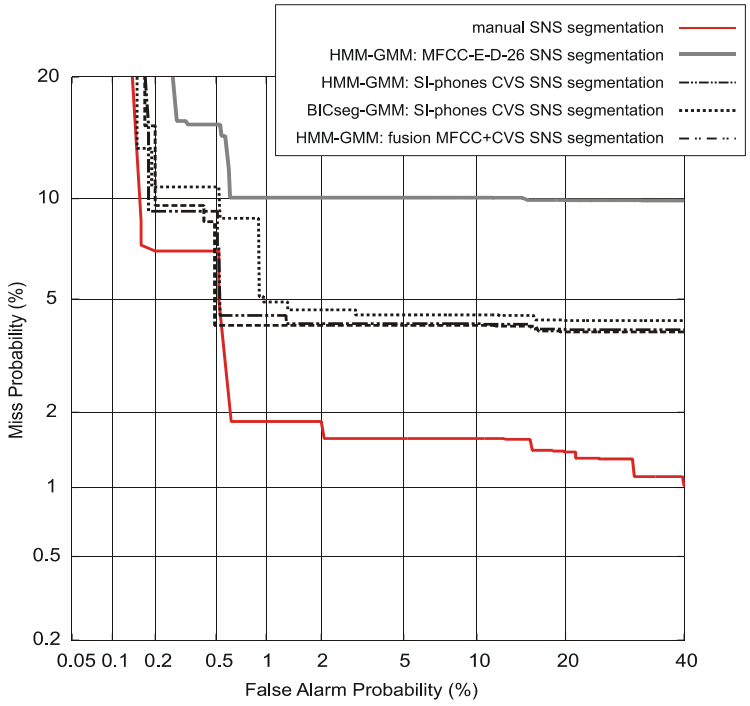


Figure 8. Overall speaker-tracking results plotted with DET curves. Lower DET values correspond to better performance.

The overall performance of the evaluated speaker-diarisation (SD) system is depicted in Figure 8, where the overall speaker-tracking results are shown. The results are presented in terms of the false-acceptance (FA) and false-rejection (FR) rates, measured at different operating points in the form of detection-error trade-off (DET) curves (Martin et al., 2000). In our case, the evaluated speaker-tracking system was capable of detecting 41 target speakers from the audio data, which included 551 different speakers. The target speakers were

enrolled in the system beforehand from the training part of the evaluation BN database. The performances of the evaluated speaker-tracking systems were therefore assessed by including all 41 target speakers, with the addition of the non-speech segments, and the results were produced from the FA and FR rates measured at the time (frame) level.
Figure 8 presents the evaluation results from four tested speaker-tracking systems. In all the evaluated systems the components for the audio segmentation, the speaker clustering and the speaker identification were the same, while the speech-detection procedures were different. The overall speaker-tracking results from Figure 8 reveal the same performance for the evaluated systems as for the speaker-clustering case. The best performance was achieved when using manual SNS segmentation, while the speaker-tracking system with the baseline SNS segmentation (*HMM-GMM:MFCC*) performed worse than all the other tested systems. Our proposed SNS-segmentation approaches with CVS features produce nearly the same overall evaluation results, which are in general 3% worse (in absolute figures) than the speaker-tracking results obtained using manual SNS segmentation. Note that the fusion of the MFCC and CVS features did not improve the evaluation results in comparison to the systems when just the CVS features were used, as was the case in previous evaluations.
We can conclude that, in general, the impact of SNS segmentation on speaker-diarisation and speaker-tracking systems is direct and indirect. As shown in the evaluation of an audio segmentation, good speech detection in continuous audio streams is a necessary pre-processing step if we want to achieve good segmentation results. And since audio segmentation serves as a front-end processing component for speaker clustering and speaker tracking, an erroneous audio segmentation influences the speaker-clustering performance. Speech detection alone has a direct impact on the performance of the speaker-diarisation performance. Since speaker-clustering performance (measured using the DER) and speaker-tracking performance (measured using the DET) are expressed in terms of a miss (speaker in reference but not in hypothesis), a false alarm (speaker in hypothesis but not in reference) and speaker error (mapped reference speaker is not the same reference as the hypothesized speaker), the errors in the speech detection produce a miss, a false alarm and false rejection errors in the overall speaker-diarisation results assessed by both evaluation measures. All types of errors are consequently integrated in the DER and DET plots in Figures 7 and 8.

5. Conclusion

This chapter addresses the problem of speech detection in continuous audio streams and explores the impact of speech/non-speech segmentation on speech-processing applications. We proposed a novel approach for deriving speech-detection features based on phoneme transcriptions from generic speech-recognition systems. The proposed phoneme-recognition features were designed to be recognizer and language independent and could be applied in different speech/non-speech segmentation-classification frameworks. In our evaluation experiments two segmentation-classification frameworks were tested, one based on the Viterbi decoding of hidden Markov models, where speech/non-speech segmentation and detection were performed simultaneously, and the other framework, where segments were initially produced on the basis of acoustic information by using the Bayesian information criterion and then speech/non-speech classification was performed by applying Gaussian mixture models.

All the proposed feature representations and segmentation methods were tested and compared in the different tasks of a speaker-diarisation system, which served for speaker tracking in audio broadcast-news shows. The impact of the speech detection was measured in four different tasks of a speaker-diarisation system. The evaluation results of the audio segmentation, the speaker clustering and the speaker tracking demonstrate the importance of a good speech-detection procedure in such systems. In all tasks, our proposed phoneme-recognition features proved to be a suitable and robust representation of audio data for speech detection and were capable of reducing the error rates of the evaluated speaker-diarisation systems. At the same time the evaluation experiments showed that the speech/non-speech segmentation with the fusion of the acoustic and the phoneme features performed the best among all the systems, and was even comparable to the manual speech/non-speech segmentation systems. This confirmed our expectations that probably the most suitable representation of audio signals for the speech/non-speech segmentation of continuous audio streams is a combination of acoustic- and recognition-based features.

6. Acknowledgment

This work was supported by the Slovenian Research Agency (ARRS), development project L2-6277 (C) entitled "Broadcast news processing system based on speech technologies."

7. References

Ajmera, J.; McCowan, I. & Bourlard, H. (2003). Speech/music segmentation using entropy and dynamism features in HMM classification framework. *Speech Communication*, Vol. 40, No. 3, (May 2003), pp. 351-363.

Ajmera, J. (2004). *Robust audio segmentation, PhD thesis,* EPFL Lausanne.

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.

Auckenthaler, R.; Carey, M. & Lloyd-Thomas, H. (2000). Score normalization for text-independent speaker verification system. *Digital Signal Processing*, Vol. 10, No. 1, January 2000, pp. 42-54.

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.

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.

Garofolo, J. S.; Lamel, L. F.; Fisher, W. M.; Fiscus, J. G.; Pallett, D. S. & Dahlgren, N. L. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus, *U.S. Dept. of Commerce, NIST*, Gaithersburg, MD, USA, February, 1993.

Greenberg, S. (1995). The ears have it: The auditory basis of speech perceptions, *Proceedings of International Congress of Phonetic Sciences (ICPhS 95)*, pp. 34-41, Stockholm, Sweden, August , 1995.

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.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, Vol. 87, No. 4, (1990), pp. 1738-1752.

Hain, T.; Johnson, S. E.; Tuerk, A.; Woodland, P. C. & Young, S. J. (1998). Segment Generation and Clustering in the HTK Broadcast News Transcription System, *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133-137, Lansdowne, VA, USA, February, 1998.

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.

Karnebäck, S. (2002). Expanded examinations of a low frequency modulation feature for speech/music discrimination, *Proceedings of International Conference on Spoken Language Processing (ICSLP2002 -Interspeech 2002)*, pp. 2009-2012, Denver, Colorado, USA, September, 2002.

Kemp, T.; Schmidt, M.; Westphal, M. & Waibel, A. (2000). Strategies for Automatic Segmentation of Audio Data, *Proceedings of the IEEE International Conference on Acoustic Signal and Speech Processing (ICASSP 2000)*, pp.1423-1426, Istanbul, Turkey, June, 2000.

Lee, K. F. & Hon, H. W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Transactions on Acoustic Speech and Signal Processing*, Vol. 37, No. 11, (1989), pp. 1641-1648.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling, *Proceedings of the International Symposium on Music Information Retrieval*, Plymouth, Massachusetts, USA, October, 2000.

Lu, L.; Zhang, H.-J. & Li, S. Z. (2003). Content-based audio classification and segmentation by using support vector machines. *ACM Multimedia Systems Journal*, Vol. 8, No. 6, (March 2003) pp. 482-492.

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.

Magrin-Chagnolleau, I. & Parlangeau-Valles, N. (2002). Audio indexing: what has been accomplished and the road ahead, *Proceedings of Joint Conference on Information Sciences, (JCIS 2002)*, pp. 911-914, Durham, North Carolina, USA, March, 2002.

Martin, A.; Przybocki, M.; Doddington, G. & Reynolds, D. (2000). The NIST speaker recognition evaluation - overview, methodology, systems, results, perspectives. *Speech Communications,* Vol. 31, No. 2-3, June 2000, pp. 225-254.

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.

Mihelič, F.; Žibert, J. (2006). Robust speech detection based on phoneme recognition features, *Proceedings of Text, speech and dialogue (TSD 2006)*, pp. 455-462, Brno, Czech Republic, September, 2006.

Mihelič, F.; Gros, J.; Dobrišek, S.; Žibert, J. & Pavešić, N. (2003). Spoken language resources at LUKS of the University of Ljubljana. *International Journal of Speech Technology*, Vol. 6, No. 3, (July 2003) pp. 221-232.

Moraru, D.; Ben, M. & Gravier, G. (2005). Experiments on speaker tracking and segmentation in radio broadcast news, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 3049-3052, 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.

Parris, E. S.; Carey, M. J. & Lloyd-Thomas, H. (1999). Feature fusion for music detection, *Proceedings of EUROSPEECH 99*, pp. 2191-2194, Budapest, Hungary, September 1999.

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.

Potamianos, G.; Neti, C.; Luettin, J. & Matthews, I. (2004). Audio-Visual Automatic Speech Recognition: An Overview. In: *Issues in Visual and Audio-Visual Speech Processing*, Bailly, G.; Vatikiotis-Bateson; E. & Perrier, P.(Eds.), MIT Press, Cambridge.

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.; 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. & 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.

Samouelian, A.; Robert-Ribes, J. & Plumpe, M. (1998). Speech, silence, music and noise classification of TV broadcast material, *Proceedings of International Conference on Spoken Language Processing (ICSLP1998)*, pp. 1099-1102, Sydney Australia, November-December, 1998.

Saunders, J. (1996). Real-time discrimination of broadcast speech/music, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP1996)*, pp. 993-996, Atlanta, USA, 1996.

Scheirer, E. & Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP1997)*, pp. 1331-1334, Munich, Germany, April, 1997.

Shafran, I. & Rose, R. (2003). Robust speech detection and segmentation for real-time ASR applications, *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003)*, pp. 432-435, Hong Kong, Hong Kong, April, 2003.

Sinha, R.; Tranter, S. E.; Gales, M. J. F. & Woodland, P. C. (1999). The Cambridge University March 2005 Speaker Diarisation System, *Proceedings of Interspeech 2005 - Eurospeech*, pp. 2437-2440, Lisbon, Portugal, September, 2005.

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.

Williams, G. & Ellis, D. P. W. (1999). Speech/music discrimination based on posterior probabilities, *Proceedings of EUROSPEECH 99*, pp. 687-690, Budapest, Hungary, September, 1999.

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.

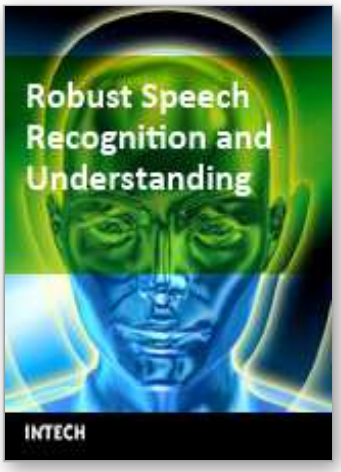
Žibert, J.; Pavešić, N. & Mihelič, F. (2006a). Speech/Non-Speech Segmentation Based on Phoneme Recognition Features. *EURASIP Journal on Applied Signal Processing*, Vol. 2006, No. 6, Article ID 90495, pp. 1-13.

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.

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.

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.

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.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

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