ELSEVIER

Available online at www.sciencedirect.com
ScienceDirect

Speech Communication 53 (2011) 143–153

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Robust speech/non-speech classification in heterogeneous multimedia content

Marijn Huijbregts *, Franciska de Jong

University of Twente, Department of Computer Science, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Received 30 June 2009; received in revised form 3 June 2010; accepted 6 August 2010

Abstract

In this paper we present a speech/non-speech classification method that allows high quality classification without the need to know in advance what kinds of audible non-speech events are present in an audio recording and that does not require a single parameter to be tuned on in-domain data. Because no parameter tuning is needed and no training data is required to train models for specific sounds, the classifier is able to process a wide range of audio types with varying conditions and thereby contributes to the development of a more robust automatic speech recognition framework.

Our speech/non-speech classification system does not attempt to classify all audible non-speech in a single run. Instead, first a bootstrap speech/silence classification is obtained using a standard speech/non-speech classifier. Next, models for speech, silence and audible non-speech are trained on the target audio using the bootstrap classification. The experiments show that the performance of the proposed system is 83% and 44% (relative) better than that of a common broadcast news speech/non-speech classifier when applied to a collection of meetings recorded with table-top microphones and a collection of Dutch television broadcasts used for TRECVID 2007.

Keywords: Speech/non-speech classification; Rich transcription; SHoUT toolkit

1. Introduction

The quantity of digital audiovisual collections is growing every day but efficient methods of accessing and searching large audiovisual collections are lagging behind. Spoken word collections are sometimes made accessible and indexed using manually annotated metadata such as abstracts of the speech in each recording, but especially for large spoken word collections, the traditional method of manual annotation puts heavy demands on resources and due to financial constraints for some content repositories, to apply even the most basic form of archiving is hardly feasible. There is common agreement that automatic annotation of audiovisual collections based on the automatic transcription of the spoken words therein may boost the accessibility of these collections enormously (Byrne et al., 2004; Goldman et al., 2005; Garofolo et al., 2000).

For collections that have well-known characteristics and for which sufficient training data is available, such as archives of broadcast news or meeting recordings, it is possible to generate high quality speech transcriptions with automatic speech recognition (ASR) techniques. Unfortunately, for the majority of collections, the exact features of the content are unknown, training data is scarce and the content to be processed may be far more heterogeneous in nature than usually seen in the laboratory.

An example of such a heterogeneous collection that is currently being used as a video retrieval benchmark corpus in the 2007–2009 TRECVID[1] evaluations and 2008–2009

* Corresponding author.
E-mail address: marijn.huijbregts@let.ru.nl (M. Huijbregts).

[1] http://www-nlpir.nist.gov/projects/trecvid

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.008

VideoCLEF evaluations[2], is the Academia collection provided by The Netherlands Institute for Sound and Vision. This heterogeneous collection consists of Dutch news magazine, science news, news reports, documentaries, educational broadcasts and television shows for children. The Academia collection does not only contain speech, silence and broadcast news jingles, but also all kinds of sounds such as music, sound effects or background noise with high volume (traffic, cheering audience, etc). The Academia collection is not the only heterogeneous collection with a variety in recording conditions, various type of speech (spontaneous, prepared speech, etc) and a broad range of audible non-speech. Many other heterogeneous collections such as the Academia collection exist and a very robust speech recognition framework is needed in order to process them. The first step in such a framework in principle is to locate all speech fragments and separate them from any non-speech, but in the common approach for data sets such as the Academia collection, system tuning is needed for each audio condition and statistical models are required for each individual type of non-speech. In this paper we will discuss a novel method that makes it possible to distinguish speech from non-speech without the need to know what kind of sounds are present in the recording.

Speech/non-speech segmentation and classification, or Speech Activity Detection (SAD), is the task of detecting the fragments in an audio recording that contain speech. SAD is not only a classification task (distinguishing speech from silence or audible non-speech), but also a segmentation task: before a fragment can be classified as speech or non-speech, the start time and end time of that fragment need to be determined. For simplicity, in the remainder of this paper we will use the term speech/non-speech classification for the entire process of segmenting and classifying the audio into speech and non-speech classes.

Speech/non-speech classification is an important step in ASR not only because it is more practical to process small speech segments instead of an entire recording, but especially because by applying SAD, the performance of the ASR system can be enhanced. The added value is twofold. First, even though audible non-speech (such as sound effects, etc) does not contain any speech, if they are passed to an ASR system it will always output a hypothesis, leading to transcriptions with insertions of words not actually spoken. Second, because after determining which segments contain speech, it is possible to cluster the speech segments on a speaker basis and use these clusters for automatically tuning the ASR system (for example by applying speaker-specific vocal tract length normalization or performing unsupervised acoustic model adaptation). All non-speech presented to a speaker clustering system will contaminate the speaker models and this will decrease the clustering quality.

The algorithm of the SAD system that will be described in this paper is developed with the aim to have no parameters (including statistical models) that need tuning on a training set, so that when audio with unknown non-speech sounds needs to be processed, it is possible to perform high quality speech/non-speech classification directly without the need for in-domain training data.

The speech/non-speech classifier discussed in this paper is part of the large vocabulary continuous speech recognition system called SHoUT[3]. Before the SAD approach is introduced in Section 3, we will describe the three common approaches for speech/non-speech classification in Section 2. In Section 4, the algorithm used for the approach is described step-by-step. In Sections 5–8, we focus on four important aspects of the SAD system (feature extraction in Section 5, confidence measures in Section 6, bootstrapping in Section 7 and system parameter settings in Section 8). Finally, in Section 9 the evaluation of the SHoUT SAD system will be discussed.

2. Common classification methods

Before describing our own speech/non-speech classification approach in the following sections, in this section we will describe the three common approaches: *silence-based* classification, *model-based* classification and *metric based* classification. After describing these approaches we will define what we consider to be speech, in order to clarify which audio fragments the system should actually be able to label as speech.

2.1. Silence-based classification

Silence-based classification systems assume that audio only contains speech and silence. For example, Broadcast News (BN) recordings might contain some jingles, but the major part of the recording consists of speech and small pauses between utterances or topics (Huijbregts et al., 2001). Some systems make use of this pattern by segmenting on the basis of the silences in the audio. For detecting speaker changes, these systems assume that there is always a short silence between speakers. In case of BN recordings, this assumption is often valid. Unfortunately, for recordings with more spontaneous speech such as recordings of meetings, this assumption is often not valid at all.

There are two common methods of finding silences in an audio stream. The first method is calculating the energy of short (often overlapping) windows. The local minima of this energy series are considered silence. The second method is to run a fast ASR decoder (Wölfel et al., 2007). Most decoders contain a silence ‘phone’ that takes care of pauses between speech.

[2] http://www.clef-campaign.org

[3] SHoUT is a Dutch acronym freely translated as: ‘SpeecH recognition University Twente’.

In Pellom and Hacioglu (2003) the ASR acoustic models are used to create two special models: one for silence and one for speech. The speech model is created by combining the most dominant Gaussian mixtures of all phones into one Gaussian Mixture Model (GMM). A small Hidden Markov Model (HMM) is then created containing only two states. The first state uses the silence GMM for its Probability Distribution Function (PDF) and the second state uses the speech GMM. A Viterbi decoding run using this HMM will result in the speech/silence classification. Although in this approach the classification is silence-based, because of the use of the HMM it should also be considered a model-based approach.

2.2. Model-based classification

Model-based classification systems train one GMM for each class. These GMMs are used as PDF in a hidden Markov model where each state is connected to all other states. Performing a Viterbi decoding run using this HMM results in the segmentation and classification of an audio file. The advantage of this method is that it is very easy to add classes. The systems in Hain et al. (1998) and Gauvain et al. (1999) train a GMM model for silence, speech and music, but it is possible to create models for other classes such as sound effects or even known speakers (for example the anchor person in BN recordings).

Without taking special measures, HMMs with one state for each class tend to produce short segments, even when the transition probabilities from one class to the other are set low. In order to force minimum time constraints on segments, sometimes HMMs are created with a string of states per class that each share the same GMM. Each state in a string is connected to the next state and only the final state has a self-transition (see Fig. 1). The number of states in the string determine the minimum time of each segment. Another approach is to post-process the segmentation and join short speech segments or remove short silence segments.

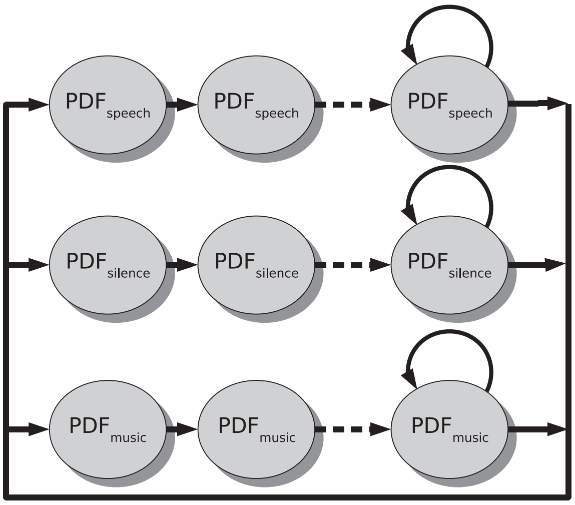

Fig. 1. An example HMM used in model-based segmentation. Each string of states represents one segmentation class and all states of a string share the same PDF.

The major disadvantage of model-based classification is that the GMMs need to be trained on some training set. If the acoustic characteristics of the audio to be processed are too different from the characteristics of the training data, the accuracy of the classification will be poor. Model-based classification has recently been used in various systems for finding speech and non-speech regions (Hain et al., 1998; Gauvain et al., 1999; Huang et al., 2007; Stolcke et al., 2007; van Leeuwen and Konečný, 2007).

2.3. Metric-based classification

Metric-based classification is one of the most common classification methods to date. In metric-based classification, a sliding window is used to investigate a short portion of the audio at each step. Typically, the window is cut into two evenly sized segments S_i and S_j and a distance metric is used to measure whether the two segments are similar and belong to the same class S, or if they are actually part of two separate segment classes. In the latter case, the point in time between the two segments is marked as segment border.

In the literature, a number of distance metrics have been proposed. Most of these metrics make use of models (often Gaussians or Gaussian mixtures) that are trained on S_i, S_j and S in order to calculate distances (Anguera, 2006). The most common distance metric is the *Bayesian Information Criterion* (BIC) (Schwartz, 1978). This metric uses some model M_i with $\#(M_i)$ parameters representing a segment of data S_i with N_i time frames (feature vectors) and it determines how well the model fits the data:

$$\mathrm{BIC}(M_i) = \log L(S_i, M_i) - \frac{1}{2}\lambda\#(M_i)\log N_i \tag{1}$$

λ is a free parameter that needs to be tuned on a training set. The value of this parameter influences when the BIC value is positive, meaning that the model fits the data, or negative, meaning that the model does not fit the data very well. Eq. (1) can be used to determine if the data of the two segments S_i and S_j fit M_i and M_j best or if the data of the two segments together ($S_i + S_j = S$) fit the model M trained on S the best:

$$\begin{aligned}\Delta\mathrm{BIC}(M_i,M_j) &= \mathrm{BIC}(M) - (\mathrm{BIC}(M_i) + \mathrm{BIC}(M_j))\\ &= \log L(S,M) - (\log L(S_i,M_i) + \log L(S_j,M_j))\\ &\quad - \lambda\Delta\#(M_i,M_j)\log N \end{aligned}\tag{2}$$

where $\Delta\#(M_i, M_j)$ is $\#(M) - (\#(M_i) + \#(M_j))$. If ΔBIC is negative, the model of the total segment S fits the data not as good as the two separate models and a segment border is placed between the two segments. ΔBIC was first used for segmentation and clustering in Chen and Gopalakrishnan (1998). In Anguera (2006) a mathematical proof of Eq. (2) is given. Note that when $\Delta\#(M_i, M_j)$ is zero, meaning that the number of free parameters in M equals the number of free parameters in M_i and M_j, the design parameter λ no longer influences the equation.

In combination with speaker clustering, the Bayesian Information Criterion has recently been used for speaker change detection in a number of systems (Cassidy, 2004; Istrate et al., 2006; van Leeuwen and Konečný, 2007; Rentzeperis et al., 2007).

2.4. What is considered speech?

Before we can develop a system that is able to separate speech from non-speech, we need to define what we actually consider to be speech. On the one hand we could define speech to be every word that a person produces, even if what he says is drowned out by loud noises or by other speech. For some applications it might be wanted that a speech activity system marks such corrupted speech as actual speech. On the other hand when SAD is used as a preprocessing step for ASR, corrupted speech that the ASR system is not able to process correctly anyway, might as well be marked as non-speech.

In this paper, audio fragments will be considered speech if the human transcriber is able to hear what is being said. The system is expected to process all types of speech as long as a person, the transcriber, is able to understand the content of this speech. Therefore the SAD system needs to be able to classify all speech fragments as actual speech, even if the fragments contain high levels of noise.

3. The SHoUT SAD approach

The algorithm implemented in the SHoUT SAD system is developed with the aim to be able to do without parameters that need tuning on an in-domain training set. If it is possible to create a system that is able to perform high quality speech/non-speech classification without the need of a training set for tuning parameters or training statistical models, such a system could be applied directly on any type of recording without the need of re-training the statistical models or fine-tuning its parameters on in-domain training data.

Our method is inspired by the model-based SAD approach with the distinction that the models are not trained on a training set, but during the classification process itself on the audio that is being processed. In order to train the models on the audio itself, a rough initial classification, the bootstrap classification, is needed. Anguera et al. (2007), who also use a bootstrap classification to generate statistical models, obtain the classification by applying silence-based classification, but when audible non-speech is expected to be present in the audio, a bootstrap classification based on silence will not be sufficient. A new solution is needed to solve this research problem. The SHoUT SAD system addresses the problem by applying a model-based classification component to create the bootstrap classification. After the initial classification step, three models are trained on the audio to be processed: a model trained on silence, a model trained on audible non-speech and a model trained on speech. Each of these models is trained on the data to be segmented. By applying the three models, the system is able to perform high quality SAD.

In the following section, the algorithm used for our approach will be described step-by-step.

4. The speech/non-speech classification algorithm

Our classification algorithm aims at training models for an HMM-based classification system on the audio to be processed instead of on a separate training set. The proposed system trains three models: a silence model, a model for audible non-speech and a model for speech.

In Fig. 2 the successive algorithm steps are shown. First the audio stream is cut-up in chunks of ten minutes. As the number of Gaussians needed in each GMM is dependent on the amount of data, using chunks simplifies the tuning of the system parameters. In the final algorithm step, the chunks are concatenated. When two neighboring segments from different chunks are assigned to the same class, the segments are merged.

For each chunk, first a bootstrap classification is created. This classification is used to train models for silence and audible non-speech (first light-gray box in Fig. 2). After training of these models, a model is created for all speech in the recording (second light-gray box). Once all three models are created, it is checked if the audible non-speech model is actually needed. If this is not the case, the non-speech model is discarded and two new models are trained for silence and speech (final light-gray box). In the following subsections, the three steps will be discussed further.

4.1. Step 1: bootstrapping

Each audio chunk is first segmented using a bootstrapping component which segments the data in speech and non-speech fragments. Although the performance of this bootstrapping component does not need to be optimal, it is important that the majority of the data classified as speech actually *is* speech. For the segments classified as non-speech, it is less of a problem when some speech segments are included, as long as most of the silence and sound segments are classified as non-speech.

For the bootstrap classification, various classification approaches can be taken. We decided to use the model-based approach as described in Section 2. In Section 7 we will describe the exact set-up of our bootstrapping component.

4.2. Step 2: training the models for non-speech

Next, a silence and a sound model are created from the part of the data classified as non-speech. Two measures are

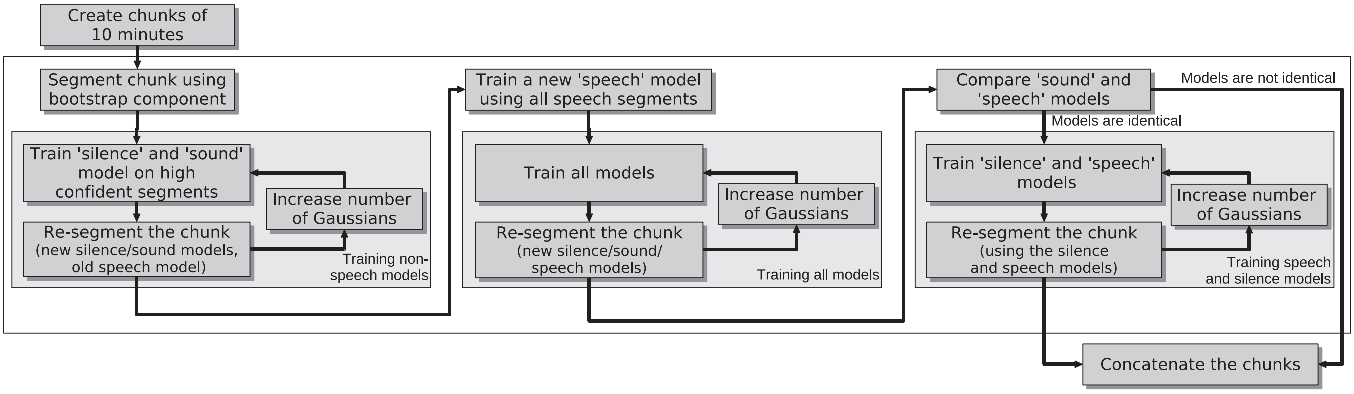
Create chunks of 10 minutes
Segment chunk using bootstrap component
Train 'silence' and 'sound' model on high confident segments
Increase number of Gaussians
Re-segment the chunk (new silence/sound models, old speech model)
Training non-speech models
Train a new 'speech' model using all speech segments
Train all models
Increase number of Gaussians
Re-segment the chunk (new silence/sound/ speech models)
Training all models
Compare 'sound' and 'speech' models
Models are not identical
Models are identical
Train 'silence' and 'speech' models
Increase number of Gaussians
Re-segment the chunk (using the silence and speech models)
Training speech and silence models
Concatenate the chunks

Fig. 2. The algorithm of the speech activity detection system. The audio recording is cut in chunks of 10 minute segments and the procedure within the outer box is performed for each chunk.

developed to calculate the confidence that a segment is actually silence or audible non-speech. In Section 6 these measures will be discussed. All non-speech segments are labeled with the two confidence scores and a small part of the non-speech data that is marked with the highest silence confidence score is used to train an initial silence model. A small amount of data that is labeled with high audible non-speech confidence scores is used to train the initial ‘sound’ model.

Using these silence and sound models and the primary speech model, a new classification is created. This classification is used to train silence and sound models that fit the audio very well simply because they are trained on it. All data assigned to the sound and silence models by the new classification are merged and any samples that were originally assigned to the speech model in the first iteration are subtracted from the set. This is done to avoid that the sound model starts pulling away all the data from the speech model. This risk is present because although the sound model is already trained on the data that is being processed, the speech model applied is still the old model trained on outside data. Therefore, the sound model may fit *all* of the data better (including speech segments) so that during the Viterbi alignment, speech segments may be assigned to the sound model.

The remaining data is divided over the silence model and the sound model as before. The silence model receives data with high silence confidence scores and the sound model receives data with high audible non-speech confidence scores. This time though, the confidence threshold is not set as high as the first time and consequently more data is available to train each model and therefore more Gaussians can be used to train each GMM. This procedure is repeated a number of times. Although the silence and sound models are initialized with silence and sound respectively, there is no guarantee that sound is never classified as silence. Energy is not used as a feature (see Section 5) and some sound effects appear to be modeled by the silence GMM very well. Because the goal is to find all speech segments and discard everything else, this is not considered a problem.

4.3. Step 3: training all three models

After the silence and sound models are trained, a new speech model will be trained using all data classified as speech. By now, non-speech will be modeled well by the sound and silence models so that a Viterbi alignment will not assign any non-speech to the speech model. This makes it possible to train the speech model on all data assigned to it and not only on high confidence regions. Once the new speech model is created, all models are iteratively retrained with increasing numbers of Gaussians. At each training iteration the data is re-segmented. Note that in this phase, all data is being used to train the models. During the earlier iterations, the data assigned to the speech class by the bootstrap classification component was not used to train the silence and sound models, but because now also the speech model is being retrained, it is less likely that using this data will cause the sound model to pull speech data away from the speech model.

4.4. Step 4: training speech and silence models

The algorithm works for audio of various domains and with a range of non-speech sounds, but it is not well suited for data that contains speech and silence only. In that case the sound model will be trained solely on the speech that is misclassified at the first iteration (because the initial models may be trained on data not matching the audio being processed, the amount of misclassified speech can be large). During the second training step the sound model will subtract more and more speech data from the speech model and finally instead of having a silence, sound and speech model, the system will contain two competing speech models. Therefore as a final check, the Bayesian Information Criterion (BIC) is used to check if the sound and speech model are the same (Cf. Eqs. (1) and (2)). As shown in Section 2, the design parameter in the BIC equation that needs tuning on matching data can be omitted when the number of Gaussians in the separate speech and sound models is the same as the number of Gaussians in the combined model. Therefore a new model is created from the data classified as speech *and* from the data classified as sound, with exactly as many Gaussians as the two separate models together. If the ΔBIC score is positive, both models are trained on speech data and the speech and sound models need to be replaced by a single speech model. Again, a number of alignment iterations is conducted to obtain the best silence and speech models.

4.5. Limitations and future extensions

In this section the algorithm of our SAD approach has been described. The main advantage of this algorithm is that it is possible to perform speech/non-speech classification on audio of unseen conditions for which no training data is available. It must be noted though that the algorithm has its limitations as well.

It is not possible to use the algorithm in a realtime, online application that requires instant classification of the incoming audio stream. Although our SAD system is able to run faster than realtime, because the entire recording is needed to train the models it is impossible to perform classification in an online fashion (instantly in only one processing pass of the audio). This limitation is no problem if the entire recording is available during processing (for example for a spoken document retrieval task), but for applications such as dialog systems, our algorithm can not be applied in its current state. Also, although the system runs faster than realtime, it will never be as fast as a straightforward single pass Viterbi HMM system, simply because it requires a number of Viterbi iterations. The evaluations given in Section 9 will show that the loss in speed is paid back with a significant increase in classification precision.

Because we use clean speech to train the bootstrap speech model, the bootstrap alignment may classify speech with a low speech-to-noise ratio incorrectly. In this paper we will only focus on separating audible non-speech from speech with reasonable (but varying) speech-to-noise ratios. Degraded speech will be a topic of future research, in which we envisage to incorporate the approach that can be summarized as follows.

A very pronounced example of the degraded speech problem is broadband recordings containing telephone speech. With the current implementation of the algorithm, telephone speech will be classified as audible non-speech. Because the speech model of the bootstrapping component is trained solely on broadband speech, during bootstrapping, all possible narrowband (telephone) speech will be classified as non-speech. The telephone speech will then be used for training the sound model and in the final ΔBIC step, the two classes will not be merged and all telephone speech will be classified as audible non-speech.

It is possible to solve the problem by adding a broadband/narrowband classification module, but it is preferable to adjust the classification system in such a way that it is possible to detect speech with various conditions that are all represented by bootstrap speech models. With the current system it is only possible to detect the broadband speech represented by the bootstrap BN model, but in future work we will add other models such as the CTS model or a model with highly degraded speech during the bootstrapping step to avoid such speech to be labeled as non-speech.

5. Feature extraction

Mel Frequency Cepstral Coefficients (MFCC) are frequently used as input feature vectors for speech activity detection. Also for the SHoUT SAD system, MFCC is chosen for feature extraction (twelve coefficients). It is common to add energy to the feature vector, but for the SHoUT SAD system, the energy feature is omitted because it will cause audible non-speech to be classified as speech. As will be discussed in Section 7, the bootstrapping classification component is trained on speech and silence but not on audible non-speech. If energy is used, it will play a dominant role in discriminating between the two classes. Because audible non-speech consist of high energy levels (compared to the low levels of silence), audible non-speech will most probably end up in the speech class. For the algorithm described in the previous section it is important that the majority of non-speech, also the audible non-speech, is actually labeled as such and therefore the energy feature is not used.

Although it is not known what kind of audible non-speech can be expected in the audio to be processed, it is reasonable to assume that a lot of these sounds will not be generated by a single human voice. In these cases, the *zero-crossing* feature might be a good addition to the MFCC features. The zero-crossing feature is calculated by counting the number of times that the amplitude crosses zero in one frame. It has been shown in Ito and Donaldson (1971) that for vowels pronounced by humans, the value of this coefficient is only varying within small boundaries while the value can be randomly high or low for other kinds of sounds. Zero-crossing is often used because it does not require complicated and time consuming calculations. In most work, zero-crossing is used in combination with the energy feature.

SHoUT uses the first twelve MFCC coefficients supplemented by the zero-crossing feature. From these thirteen features, the derivatives and the derivatives of these derivatives are calculated and added to the feature vector, creating 39 dimensional feature vectors. Each vector is calculated on a window of 32 ms audio and this window is shifted 10 ms in order to calculate the next vector.

6. Confidence measures

The SAD algorithm described in Section 4 needs two confidence measures: one for calculating the confidence that a certain fragment is silence and one to determine if a certain fragment is audible non-speech. For the confidence measures used, first all segments that are longer than one second will be cut-up in evenly sized shorter segments of one second each, so that all segments are comparable in length. The confidence measures will then return a certain amount of these *one-second-segments* that are most likely to be either silence or audible non-speech.

It is determined if a one-second-segment is silence by measuring the energy for each frame and calculating the mean energy of the segment. This calculation is performed for all candidate segments (all segments classified as non-speech by the bootstrap classification component) and the resulting values are placed in a histogram. Using the histogram it is possible to return a top amount of segments with the lowest mean energy. As described in Section 4, a very small amount is chosen for the first iteration and higher amounts are chosen for later iterations.

For determining an amount of one-second-segments that is most likely audible non-speech, first the same approach is taken as for silence segments: a top amount of segments is picked with the highest average energy. From these segments a top amount of segments is returned with the highest mean zero-crossing values. In other words, this algorithm returns the segments with the highest mean energy and zero-crossing values. Although audible non-speech segments will have high mean energy values, it is possible that speech segments even have higher average energy values. It is assumed that for these speech segments, the average zero-crossing values will be lower than for the audible non-speech.

7. The bootstrapping component: Dutch broadcast news SAD

The component that is used to create the initial speech/non-speech classification for the SHoUT SAD system is a standard model-based speech activity detection component, developed for finding speech segments in Broadcast News (BN) recordings. As BN shows do not contain a lot of audible non-speech, the component is not trained with any models for music, sound effects or other audible non-speech.

The component consists of an HMM with two strings of parallel states. The first string represents silence and the second string represents speech. The states in each string share one GMM as their probability density function. Using a string of states instead of single states ensures a minimum duration of each segment (see Fig. 1). The minimum duration for silence is set to 30 states (300 ms) and the minimum duration for speech is set to 75 states.

The speech and silence GMMs are trained on a small amount of Dutch broadcast news training data from the publicly available Spoken Dutch Corpus (CGN) (Oostdijk, 2000). Three and a half hours of speech and half an hour of silence from 200 male and 200 female speakers are used. The models are initialized with a single Gaussian. The number of Gaussians is increased iteratively until a mixture of 20 Gaussians is reached for both classes. The data is forced aligned to the reference transcription to ensure the correct placements of speech/silence boundaries. To make sure that only speech is used to train the speech model, all phones neighboring silence are not used.

The BN SAD component uses the feature extraction method described in Section 5. This means that frame energy is not used as a feature, but zero-crossing is. Because energy is not used, the discrimination between silence and speech will have to be made purely on MFCC features and zero-crossing. It is expected that speech will be well modeled using these features and that any audible non-speech encountered in the audio to be processed will fit the general silence model better than the speech model and will therefore be categorized as silence.

If most audible non-speech will be categorized as silence, it is possible to use the BN component as bootstrapping component for the SHoUT SAD system. In Section 9, the experiments will show that this is actually the case. It is even possible to use the BN component that is trained on Dutch speech as bootstrapping component for American English speech.

8. System parameters

The algorithm for the SHoUT SAD system is developed with the aim to have no parameters that need tuning on a training set. The only 'parameters' left in the system are the silence and speech models trained on broadcast news data. Fortunately, as will be shown in Section 9 it is possible to use these models for audio that does not match the training data at all and still get good end results.

The system makes use of other parameters such as the number of training iterations performed at each step of the algorithm or the number of Gaussians used to train the models. It is assumed though, that these parameters do not need tuning for specific audio conditions, and that the values of these parameters can be determined using a single development set. Therefore, a development set is created by adding sound effects and some musical fragments to a broadcast news recording. By trial and error the parameters were given their values. The number of Gaussians for each model in each phase of the algorithm are shown in Table 1. During all following experiments, the parameters were kept fixed at these values.

As can be seen in Table 1, the number of Gaussians for all models are low at first and increased after each iteration. The final number of Gaussians when the sound model is determined not to be the same as the speech model, will be 7 for the silence model, 16 for the speech model and 18 for the sound model. When the sound is determined to be the same as the speech model and new silence and speech models are being trained, the final number of Gaussians will be 5 for the silence model and 12 for the speech model.

Also the amount of data that is marked as high confidence silence or sound needs to be set. For both the silence model and for the sound model initially 20 s of data are used for each chunk. This 3.33% of the total chunk size is increased by another 20 s in the first three

Table 1
The system parameters that were given their values during development. The values of these parameters were kept fixed for all experiments. The parameters are listed here in the order that they were used in the algorithm.

Parameter	Value
Initial number of Gaussians for the silence model	2
Initial number of Gaussians for the sound model	2
Number of iterations for training the two models	5
Increase of Gaussians at each iteration for SIL	0
Increase of Gaussians at each iteration for SOUND	2
Stop increase of Gaussians at iteration	3
Initial number of Gaussians for the speech model	6
Number of training iterations for the three models	5
Increase of Gaussians at each iteration for SIL	1
Increase of Gaussians at each iteration for SOUND	2
Increase of Gaussians at each iteration for SPEECH	2
Number of training iterations for SIL/SPEECH when the SOUND model is discarded	7
Initial number of Gaussians for the silence model	2
Initial number of Gaussians for the speech model	2
Increase of Gaussians at each iteration for SIL	1
Increase of Gaussians at each iteration for SPEECH	2
Stop increasing Gaussians for SIL at iteration	3
Stop increasing Gaussians for SPEECH at iteration	5

iterations of training the two models. In the final two iterations of training the silence and sound models, simply all available data that is marked silence and sound, except for the data that is assigned to speech by the bootstrap classification, are used for retraining. In order to obtain the top amount of data with high zero-crossing values, a higher amount of data with high energy levels needs to be selected (see Fig. 3). Therefore, five times as much data with high energy values are selected as are selected for training the sound model. This means that initially, 100 s of data with high energy levels are selected and that at each iteration this is increased with another 100 s. Note that the settings for the amount of data used to train the silence and sound models are chosen independently of the actual amount of silence and sound in the recording. The settings were obtained by tuning the system on the development set and do not reflect the actual ratio of speech/silence/sound of typical recordings. To ensure that we did not over-tune this parameter-nor one of the other ones we evaluated the system on a number of data sets with varying characteristics. In the following section we will describe these experiments.

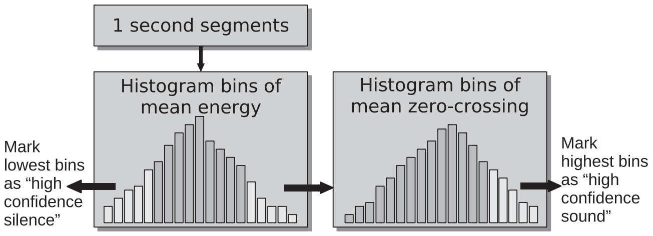
1 second segments
Histogram bins of mean energy
Histogram bins of mean zero-crossing
Mark lowest bins as "high confidence silence"
Mark highest bins as "high confidence sound"

Fig. 3. A top number of fragments with lowest energy is returned as being silence and a top number of the fragments with highest energy and highest zero-crossing is returned as being sound.

9. Evaluation

The SHoUT SAD system has been evaluated on four different benchmark collections. First it has been tested on broadcast news data. This experiment provides information about the performance of both the BN SAD component and the entire SAD system. Next, to test system performance on out-of-domain data, the system has been evaluated on data from the meeting recordings. Not only are the topics of the meetings in this evaluation set different from general broadcast news topics, also the audio conditions and the language do not match the Dutch broadcast news training data (Section 9.3). A speech/music test set has been used to determine if the algorithm is able to classify music as non-speech (Section 9.4), and finally, twelve fragments from the TRECVID 2007 collection have been used for evaluating the system on varying audio conditions (Section 9.5).

9.1. Evaluation metrics

Two measures are regularly used for assessing classification results. The first one is to measure for each class the percentage of time that the class was correctly assigned, or if one overall number is required, the percentage of time that all classes were correctly assigned:

$$\text{Score} = \frac{\sum_c C_c}{L} \cdot 100\% \qquad (3)$$

where C_c is the total time that class c was classified correctly and L is the total audio length.

The second method of assessing classification systems is to consider the result to be a special case of a speaker diarization system with all speakers mapped on one single class: the speech class (Fiscus et al., 2006). This was done for the Speech Activity Detection (SAD) task at benchmarks organized by the National Institute of Standards and Technology (NIST) in 2005 and 2006. The equation used to evaluate speaker diarization systems is simplified with the following measure as result:

$$\text{SAD error} = \frac{M + F}{S} \cdot 100\% \qquad (4)$$

where S is the total time of speech, M is the total time of speech that was not classified as speech (missed speech) and F is the total time of silence that was falsely classified as speech (false alarms). Note that this measurement results in an error percentage while the first measurement (Eq. (3)) results in a percentage of correctly assigned classes. Also, the SAD error measurement is a percentage of the total time of *speech* in the reference transcript, while using the first measurement for a SAD system would result in a percentage of the total time of the evaluation audio.

The music benchmark will be scored with the first metric defined in Eq. (3). This metric is used so that it is possible to compare the results to earlier work on this collection. The other experiments are scored using the SAD error rate defined by Eq. (4).

9.2. Baseline: broadcast news evaluation

A small evaluation set of half an hour of Dutch broadcast news recordings has been used to test the SAD system on the BN domain. The SAD error of the bootstrap component is 4.5% on this test set. The error rate of the SHoUT SAD system is also 4.5%. For each chunk, the comparison of the sound and speech models results in discarding the sound model.

The results of this initial test are as expected. Because the bootstrapping component is trained on Dutch BN, it was expected that it would perform well and that the re-training of the models has limited effect. Because there was no audible non-speech in this test set, it was also expected that the sound cluster would be discarded for all chunks.

9.3. Out-of-domain: English meeting evaluation

For a number of years, NIST organizes benchmarks for automatic rich transcription (speech-to-text and speaker diarization) on the conference meeting domain. In 2006, RT06s, the benchmark consisted of nine English spoken conference meetings. In contrast to the BN recordings, the meetings are not recorded with close-talking microphones but with far-field microphones placed on the meeting table.

The RT06s conference meeting test set was used to evaluate how well the SAD system performs on out-of-domain data, recorded with other conditions and with speech in another language than the training set. Table 2 contains the SAD error results on the nine RT06s conference meetings. As a baseline, the error of the bootstrap classification coming from the Dutch Broadcast News component, is shown. After the final alignment iteration, the overall error of the baseline on this test set is 26.9% whereas on in-domain Dutch broadcast news it was only 4.5%. This underlines that the conference meeting data is indeed out-of-domain for the bootstrapping models. The overall SAD error of the total system is only 4.4%. This is in line with the state-of-the-art at RT06s where the lowest SAD error reported was 4.3% (Fiscus et al., 2006). This experiment proves that it is not needed to use a high performance classification component as bootstrap component in order to achieve good results.

Table 2
SAD error rates for the RT06s conference meetings. The second column (Btstr) contains the%SAD error of the bootstrapping component (trained on Dutch BN data).

File ID	% Btstr	%Missed speech	%False alarm	%SAD error
CMU_20050912-0900	42.6	2.8	2.8	5.6
CMU_20050914-0900	41.5	2.3	3.5	5.8
EDI_20050216-1051	13.8	0.6	1.2	1.8
EDI_20050218-0900	16.4	0.8	2.1	2.9
NIST_20051024-0930	20.8	3.8	0.7	4.5
NIST_20051102-1323	17.0	0.8	1.5	2.3
TNO_20041103-1130	32.7	4.5	1.3	5.8
VT_20050623-1400	24.1	1.4	2.3	3.7
VT_20051027-1400	31.7	6.5	1.5	8.0
Overall error	26.9	2.50	1.90	4.40

9.4. The IDIAP speech/music evaluation

For this evaluation, the speech/music test set described in Ajmera et al. (2003) has been used. The data consists of four audio files that contain English broadcast news shows interleaved with various genres of music. The first file contains speech and music fragments of fifteen seconds each. The second file contains fragments of varying lengths but overall with the same amount of speech as music. The third file contains more speech than music while the fourth file contains more music. The performance is measured by (i) the percentage of true speech frames identified as speech, (ii) the percentage of true music frames identified as music and (iii) the overall percentage of speech and music frames identified correctly (see Eq. (3)).

The reference transcripts of this test set only consist of music and speech segments. Any pauses in speech (silence) are not annotated. Therefore, for this evaluation, if a silence segment is neighboring two speech segments, it is merged with these two segments. All other silence and sound segments are labeled as music. This means that silence between speech and music is always labeled as music although it might be the end or beginning of a speech segment.

In Table 3 the results of the SAD system on the four files are listed. The SHoUT SAD system does not perform as well as the best system in Ajmera et al. (2003) (on average 95.2%), but considering that it is initialized with Dutch models and that no tuning has been done on a training set similar to this data, the average score of 92.1% can be regarded as satisfactory.

9.5. Dutch TRECVID 2007 ASR evaluation

In this section we will provide evaluation results on part of the TRECVID 2007 collection, the collection of 400 hours of Dutch news magazine, science news, news reports,

Table 3
Classification results on the IDIAP speech/music test set. The scores are all percentages of correctly classified frames.

File ID	% Speech	% Music	% Overall
Set-1	90.2	95.7	92.9
Set-2	88.0	97.0	92.5
Set-3	85.1	99.9	92.5
Set-4	81.0	99.5	90.3

documentaries, educational broadcasts and television shows for children (see the introduction). The University of Twente provided the automatic transcriptions for the entire collection (Huijbregts et al., 2007) and the SAD system was used to obtain all speech segments and discard non-speech regions.

For evaluation, five minute fragments of twelve different documents have been randomly selected from the TRECVID 2007 collection. These fragments have been manually annotated and the speech regions are determined by applying forced alignment on the Dutch speech. Table 4 lists the results of the system on these twelve fragments. The overall error is 11.4% of the total speech in the audio. Note that only 39 min of the in total one hour long test set is actual speech. The bootstrapping BN SAD error is 20.3%. The most part of this error, 15.8%, is due to missed speech and 4.5% is due to false alarms.

9.6. Discussion

As can be seen from the evaluation in this section, with our SAD approach it is possible to perform high quality speech/non-speech classification on audio of unknown conditions for which no training data is available. Tables 2 and 4 show that the precision of the bootstrapping BN SAD component is significantly lower than that of the final classification. Note that for the BN experiment, where the audio conditions are similar to the training data of the bootstrapping component, the final classification is equally good as the bootstrapping classification.

On the TRECVID 2007 ASR evaluation set, 8.3% of the speech is classified as non-speech. This means that the ASR system will never be able to correctly recognize the words in this part of the audio. On the other hand, using this SAD system, only 3.2% non-speech will be processed by the ASR system. If speech activity detection is not used at all, the percentage of non-speech in the data would be 54% (21 min of the total test set is non-speech). Manual inspection of the missed speech showed that most missed segments consist of speech mixed with various sources of non-speech. It is hard to perform correct ASR on this kind of speech and therefore the loss of being able to process the missing 8.3% of the speech is considered less important than the gain of not needing to process the 54% of non-speech, which would have led to an increase of insertion errors.

Table 4
SAD error rates for the twelve fragments of the TRECVID 2007 ASR evaluation set. Each fragment is five minutes long, but the amount of speech in these fragments varies (second column). The third column contains the error of the bootstrapping component (trained on Dutch BN data).

File ID	Speech	BN	Missed	False	SAD
	(sec)	SAD	speech	alarm	error
15190	274.65	5.9	5.0	1.4	6.4
3273	156.86	38.6	3.9	9.0	12.9
34837	193.59	20.8	11.1	5.3	16.4
3484	196.91	37.2	18.1	0.2	18.2
34973	262.99	7.2	1.7	0.1	1.8
35202	168.71	15.8	4.4	3.0	7.4
35447	204.54	21.5	1.6	7.8	9.4
35757	215.79	16.5	6.8	1.7	8.5
36058	179.62	34.7	15.3	4.5	19.8
36366	73.32	37.4	6.0	15.9	21.9
36626	223.59	17.5	11.5	1.4	12.9
36641	176.06	20.6	15.7	0.1	15.7
Overall	2326.62	20.3	8.3	3.2	11.4

10. Conclusions and future work

Filtering non-speech out of an acoustically heterogeneous video collection such as the TRECVID 2007 collection is one of the many challenges for the task of automatically annotating a collection. The variety of such collections makes it hard to train task-specific audible non-speech models. To overcome this limitation, a SAD system was proposed that automatically trains a model for audible non-speech, a so called *sound* model, for each recording in the collection. The system has been tested on three benchmark collections with promising results.

The system, that is part of our automatic speech recognition toolkit SHoUT, comes with a number of system parameters (discussed in Section 8), but it does not contain any parameters that need tuning on a training set. This makes the algorithm robust for varying audio conditions. A high performing bootstrap classification component has been shown not to be needed in order to obtain good final results, and it is not a problem that the speech/silence models used in the bootstrap component are sometimes trained on data mismatching the audio to be processed. Having noted this, it would be interesting to investigate other methods for obtaining the initial classification that do not depend on statistical models. One method that might be able to replace the model-based approach is to initially segment on voiced speech fragments. Determining voiced speech regions can be done without the use of models and the majority of the audible non-speech of the resulting classification will actually be labeled as non-speech, making it possible to use the proposed algorithm.

The energy feature is not used in the system because it will decrease performance during bootstrapping. It would be interesting though to experiment with this feature in a later phase, for example during the final two iterations of silence/sound training. Also, although the energy feature is not useful during bootstrapping, the energy-delta and delta-delta features might actually be helpful in this early phase.

A problem related to SAD that is not yet addressed by the proposed system is the detection and discarding of foreign speech fragments. As with non-speech segments, feeding foreign speech into the ASR system will influence its performance negatively. Not surprisingly, as was shown by the experiments on the RT06s conference meeting data, our SAD system will classify speech from foreign languages as speech. A solution to this problem is to apply a language detection system directly after the SAD system. Speech from a language for which an ASR system is available can be passed to that system, while speech of other languages can be discarded.

In Section 4.5 it was noted that with the current system setup, highly degraded speech or narrowband speech will be classified as non-speech. This problem can be solved by extending the algorithm to cope with multiple speech classes. Instead of bootstrapping with one speech class, it should be possible to bootstrap with both a narrowband and a broadband speech model and to generate new speech models using the bootstrap classification. In future work we will extend our algorithm in such a way that the telephone speech will be classified correctly and that it is possible to apply special telephone models during ASR.

Acknowledgements

The work reported here was partly supported by the bsik-program MultimediaN which is funded by the Dutch government(http://www.multimedian.nl) and the EU projects AMI (IST-FP6-506811), MESH (IST-FP6-027685), and MediaCampaign (IST-PF6-027413). We would like to thank IDIAP for providing us with the speech/music benchmark files.

References

Ajmera, J., McCowan, I., Bourlard, H., 2003. Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Communication 40 (3), 351–363.

Anguera, X., 2006. Robust speaker diarization for meetings. Ph.D. thesis, Universitat Politecnica De Catalunya.

Anguera, X., Wooters, C., Pardo, J., 2007. Robust speaker diarization for meetings: ICSI RT06s evaluation system. Machine Learning for Multimodal Interaction (MLMI). In: Lecture Notes in Computer Science, vol. 4299. Springer Verlag, Berlin.

Byrne, W., Doermann, D., Franz, M., Gustman, S., Hajic, J., Oard, D., Picheny, M., Psutka, J., Ramabhadran, B., Soergel, D., Ward, T., Zhu, W.-J., 2004. Automatic recognition of spontaneous speech for access to multilingual oral history archives. IEEE Transactions in Speech and Audio Processing 12 (4), 420–435.

Cassidy, S., 2004. The macquarie speaker diarisation system for RT04S. proceedings of the NIST RT04s Evaluation Workshop. Montreal, Canada.

Chen, S.S., Gopalakrishnan, P., 1998. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop. Virginia, USA.

Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S., 2006. The Rich Transcription 2006 Spring Meeting Recognition Evaluation. In: Machine Learning for Multimodal Interaction, Third International Workshop, pp. 309–322.

Garofolo, J., Auzanne, G., Voorhees, E., 2000. The trec spoken document retrieval track: a success story. In: proceedings of the Recherche d'Informations Assiste par Ordinateur: ContentBased Multimedia Information Access Conference.

Gauvain, J.-L., Lamel, L., Adda, G., Jardino, M., 1999. The LIMSI 1998 hub-4e transcription system. In: Proceedings of the DARPA Broadcast News Workshop. pp. 99–104.

Goldman, J., Renals, S., Bird, S., de Jong, F.M.G., Federico, M., Fleischhauer, C., Kornbluh, M., Lamel, L., Oard, D.W., Stewart, C., Wright, R., 2005. Accessing the spoken word. International Journal on Digital Libraries 5 (4), 287–298.

Hain, T., Johnson, S., Tuerk, A., Woodland, P., Young, S., 1998. Segment generation and clustering in the HTK broadcast news transcription system. In: Proceedings DARPA Broadcast News Transcription and Understanding Workshop. Virginia, USA, pp. 133–137.

Huang, J., Marcheret, E., Visweswariah, K., Libal, V., Potamianos, G., 2007. The IBM rich transcription 2007 speech-to-text systems for lecture meetings. In: Proceedings of the NIST Rich Transcription 2007 Spring Meeting Recognition Evaluation, RT07s. Baltimore, USA.

Huijbregts, M., Ordelman, R., van Hessen, A., 2001. Prosody based boundary detection. Technical Report, University of Twente, Parlevink Group.

Huijbregts, M., Ordelman, R., de Jong, F., 2007. Annotation of heterogeneous multimedia content using automatic speech recognition. Proceedings of the Second International Conference on Semantics And digital Media Technologies (SAMT). Lecture Notes in Computer Science. Springer Verlag, Berlin.

Istrate, D., Fredouille, C., Meignier, S., Besacier, L., Bonastre, J.F., 2006. NIST RT05S evaluation: pre-processing techniques and speaker diarization on multiple microphone meetings. Machine Learning for Multimodal Interaction (MLMI). Lecture Notes in Computer Science. Springer Verlag, Berlin.

Ito, M., Donaldson, R., 1971. Zero-crossing measurements for analysis and recognition of speech sounds. Audio and Electroacoustics, IEEE Transactions on 19 (3), 235–242.

Oostdijk, N., 2000. The Spoken Dutch Corpus. Overview and first evaluation. In: Gravilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S., Stainhaouer, G. (Eds.), Second International Conference on Language Resources and Evaluation. Vol. II. pp. 887–894.

Pellom, B., Hacioglu, K., 2003. Recent Improvements in the CU Sonic ASR system for Noisy Speech: The SPINE Task. In: Proc. ICASSP.

Rentzeperis, E., Stergiou, A., Boukis, C., Pnevmatikakis, A., Polymenakos, L.C., 2007. The 2006 athens information technology speech activity detection and speaker diarization systems. Machine Learning for Multimodal Interaction (MLMI). Lecture Notes in Computer Science. Springer Verlag, Berlin.

Schwartz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Stolcke, A., Anguera, X., Boakye, K.,Çetin, O., Janin, A., Magimai-Doss, M., Wooters, C., Zheng, J., 2007. The SRI-ICSI spring 2007 meeting and lecture recognition system. In: Proceedings of the NIST Rich Transcription 2007 Spring Meeting Recognition Evaluation, RT07s. Baltimore, USA.

van Leeuwen, D., Konečný, M., 2007. Progress in the AMIDA speaker diarization system for meeting data. Multimodal Technologies for Perception of Humans. Lecture Notes in Computer Science. Springer Verlag, Berlin.

Wölfel, M., Stüker, S., Kraft, F., 2007. The ISL RT-07 speech-to-text system. In: Proceedings of the NIST Rich Transcription 2007 Spring Meeting Recognition Evaluation, RT07s. Baltimore, USA.