Hybrid Speech/non-speech detector applied to Speaker Diarization of Meetings

Xavier Anguera[1,2], *Mateu Aguilo*[2], *Chuck Wooters*[1],
Climent Nadeu[2], *Javier Hernando*[2]

[1] International Computer Science Institute (ICSI)
1947 Center St., Suite 600, Berkeley, CA 94704, U.S.A.
[2] Technical University of Catalonia (UPC), TALP Research Group
Jordi Girona 1-3 D5, 08034 Barcelona, Spain
xanguera@icsi.berkeley.edu

Abstract

When performing speaker diarization, it is common practice to use an agglomerative clustering approach where the acoustic data is first split in small segments and then pairs of these segments are merged until a particular stopping point is reached. The diarization performance can be greatly improved by the use of a speech/non-speech detector. The use of a speech/non-speech detector helps the diarization system by preventing non-speech frames from "confusing" both the merging and the stopping processes. Over the years there has been extensive research on speech/non-speech detectors. Often times, speech/non-speech detectors require training data and their accuracy is strongly dependent on setting various thresholds correctly. In this work we present a hybrid speech/non-speech detector for use in our speaker diarization system within the meetings domain. Our proposed speech/non-speech system runs in two stages. The first stage performs an energy-based detection. The second stage performs a model-based decoding using the previous stage's data as a bootstrap for the acoustic models, thus avoiding the need for any outside training data. We show an improvement of 14% and 10% relative on a development and test set.

1. Introduction

The goal of speaker diarization is to segment an audio recording into speaker-homogeneous regions [1]. Typically, this segmentation must be performed with little knowledge of the characteristics of the recording or of the talkers in the recording. For example, we may know the source and date of the audio recording (e.g. CNN Nightly News or a NIST meeting), but we typically do not know how many speakers occur in the recording, how many speakers are male vs. female, whether there are commercials, music, or other noises, etc. The more that the system can automatically detect, the better the result will be.

It has been shown in [2] that the speaker diarization performance can be improved by the use of a speech/non-speech detector as a first step to the agglomerative clustering process. The speech/non-speech system used in this previous work was based on acoustic models that needed to be trained on data as similar as possible to test data. This poses a robustness problem when we intend to use the diarization system on ”unseen” data, and slows down the portability of the system to new environments, where new training data needs to be labelled/located and new speech/non-speech models need to be trained. Among the systems that do not use acoustic models for speech/non-speech detection, the most widely used always include energy as a feature. The performance of such systems is dependent on setting appropriate thresholds and these thresholds are typically tuned using some development data.

In this paper we present a novel system to perform speech/non-speech detection, and its application to speaker diarization in the meetings environment. Such system takes advantage of the fact that most non-speech in meetings is silence. It first performs an energy-based detection of the silence portions in the input data using energy derivative filtering based on [3]. This system only needs a coarse setting of a threshold, which is then iteratively modified until obtaining a reasonable amount of silence data. The second stage of the system models speech and silence given the output from the first stage, and creates a final speech/non-speech segmentation to be used in the diarization system. By running this two-stage system, we avoid using any external training data to obtain an initial set of acoustic models. From these initial models we can iterate between segmenting the data and retraining the models to obtain the final speech/non-speech segmentation.

In section 2 we present the speaker diarization system used when evaluating the speech/non-speech system. In section 3 the energy-based decoder is introduced, and in 4 the model-based decoding and the hybrid system are explained. In section 6 we show the experiments performed and finally we present the conclusions of this work.

2. Agglomerative Speaker Diarization

The speaker clustering system used in this paper is based on [2] and [4]; it follows an agglomerative clustering approach. The data is initially split into K clusters (where $K >$ estimated number of speakers which is computed automatically), and then iteratively merges the clusters (according to a merge metric based on ΔBIC) until a stopping criterion is met. The acoustic data is modelled using an ergodic hidden Markov model (HMM), where the initial number of states is equal to the initial number of clusters (K). Upon completion of the algorithm's execution, each remaining state is taken to represent a different speaker. Each state contains a set of M_D sub-states, imposing a minimum duration on the model (we use $M_D = 3$ seconds). Within the state, each one of the sub-states share a probability density function (PDF) modelled via a Gaussian mixture model (GMM).

Our clustering algorithm for the meetings domain, consists of the following steps:

1. Run speech/non-speech detection on the input data to be evaluated.

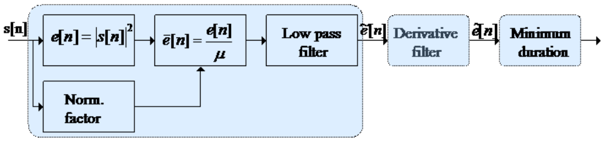

Figure 1: *Energy-based detector blocks diagram*

2. Extract acoustic features from the data and remove non-speech frames from the diarization processing.
3. Create models for K initial clusters via linear initialization.
4. Perform iterative merging using the following steps:
   (a) Run a Viterbi decode to resegment the data.
   (b) Retrain the models via an Expectation-Maximization (EM) algorithm using the segmentation from step (a).
   (c) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.0.
   (d) If no such pair of clusters is found, stop and output the current clustering.
   (e) Merge the pair of clusters found in step (c). The models for the individual clusters in the pair are replaced by a single, combined model.
   (f) Go to step (a).

For the merging and clustering stopping criteria, we use a variation of the commonly used Bayesian Information Criterion (BIC) [5]. The ΔBIC compares two possible models: two clusters belonging to the same speaker or to different speakers. The variation used was introduced by Ajmera et al. [4], [6], and consists of the elimination of the tunable parameter λ by ensuring that, for any given ΔBIC comparison, the difference between the number of free parameters in both models is zero.

The use of a speech/non-speech detector is an important part of the system. The inclusion of non-speech frames into the clustering process makes it difficult to correctly differentiate between two models because the non-speech frames tend to make the two models appear more similar than they really are. Even though there has been extensive use of model-based speech/non-speech detectors in the literature (including [2], [7]) and of energy-based detectors, we present an alternative hybrid system that attempts to solve some of the problems from both approaches. On one hand, we avoid accurate tuning of the energy threshold by using an iterative search of a rough speech/non-speech segmentation used to initialize the model-based decoder. On the other hand, by using such initialization on the model-based decoder, we avoid having to train its models with pre-labelled data, resulting in a system that is free of trianing data.

In the following sections we present the energy-based detector and the model-based decoder that have been used, and how they are combined to obtain the hybrid system.

3. Energy Detection

The first stage of the process consists on an energy-based speech/non-speech detector which can be divided into three major blocks as seen in figure 1. Each of these blocks are explained below. First of all, the data is preprocessed using common engineering techniques (see 3.1 below) with the purpose of increasing the quality of the speech signal. Then a derivative filter is applied over the energy signal (see 3.2 below). Finally we use a thresholding method together with a minimum duration enforcement via a Finite State Machine (FSM) to detect silences (see 3.3 below.)

3.1. Data Preprocessing

A Wiener filtering is applied over the waveforms in order to reduce noise effects and reverberations. This is done using a noise reduction algorithm developed for the Aurora 2 front-end proposed by ICSI, OGI, and Qualcomm [8]. The algorithm performs Wiener filtering with typical engineering modifications, such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor.

Due to the different sources and recording setups the average amplitude of the signal to be processed can vary over a large range. Therefore it needs to be normalized to be able to bring consistency in the follow-on processing. A standard energy average over all the recording would not be plausible due to the existence of extended silence regions and of sudden noise bursts. We chose a way to compute the normalization constant μ which is more robust to these effects as shown in equation 1.

$$\mu = \frac{1}{P}\sum_{p=1}^{P} max(s \cdot [pTF_s], \cdots, s[(p+1) \cdot TF_s]) \qquad (1)$$

where P is the total amount of non overlapped blocks of duration TF_s (with F_s being the sampling rate in samples/second, and T in seconds) in the recording. Each block of samples ranges from $p \cdot TF_s$ to $(p+1) \cdot TF_s$

Finally a low-pass Butterworth filter deletes all high band noises leaving only information of the signal below $4kHz$. This is done because the major part of the energy of the signal is contained in this band and we need no information but energy in this part of the detection. This Butterworth filter has been implemented using its IIR form.

3.2. Derivative Filtering

Given the normalized and filtered energy signal ($\tilde{e}[n]$) we use a derivative filter in order to enhance the speech/non-speech change-points. This processing helps prevent degradation due to low signal-to-noise ratios or nonstationary environments and was first introduced by [3]. Such filter is defined via the following impulse response,

$$h[n] = \{-f[-W \leq n \leq 0], f[1 \leq n \leq W]\} \qquad (2)$$

Where,

$$\begin{aligned} f[n] &= e^{An}[K_1\sin(An) + K_2\cos(An)] \\ &\quad +e^{-An}[K_3\sin(An) + K_4\cos(An)] \\ &\quad +K_5 + K_6e^{sn} \qquad (3) \end{aligned}$$

And,

$$\begin{aligned} A &= 0.41s \\ s &= \frac{7}{W} \\ W &= \text{Half of the window length.} \qquad (4) \end{aligned}$$

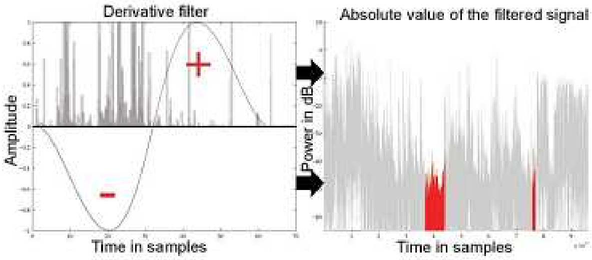

Figure 2: Left, filter over $\tilde{e}[n]$. Decision of silence in red after the thresholding.

And the values of the coefficients $[K_1 \ldots K_6] = [1.583, 1.468, -0.078, -0.036, -0.872, -0.56]$, for a chosen window length $W = 31$. The selection of an appropriate value for the W parameter is important as it sets the temporal resolution of the detector.

As shown in fig. 2 the result of the convolution of $\tilde{e}[n]$ and $h[n]$, $\hat{e}[n]$ is thresholded and labelled, each sample, as *speech* or *non-speech*.

3.3. Time Constraints on Speech/non-Speech

After the energy is filtered the third time we need to impose some time constraints to avoid changing too quickly between *speech* and *non-speech*. A finite state machine(FSM) has been implemented for this purpose. In this FSM described in figure 3 the time constraints are forced through *enter times* and *leave times* according to the values of $\hat{e}[n]$ using two thresholds (*enter thrld*, Θ_{enter} and *leave thrld*, Θ_{leave}) on each sample. The selection of the right thresholds is crucial to the correctness of the detector and, although the energies have been initially normalized, might differ from meeting to meeting. We define the *enter thrld* to be an order of magnitude bigger than *leave thrld* and its value is iteratively defined by the hybrid system described below. As for the appropriate minimum time of either speech or non-speech states we estimate it using the development data.

Inside the FSM, the conditions to go from *non-speech* to *speech* are the same to go from *speech* to *non speech*. This way to go from *speech* to *non-speech*, $\hat{e}[n]$ has to be higher than *enter thrld*, and vice versa:

$$\begin{aligned} \hat{e}[n_1] \geq \Theta_{enter} \;\&\; State_t = NSP \rightarrow &\quad State_{t+1} = SP \\ \hat{e}[n_2] \leq \Theta_{exit} \;\&\; State_t = SP \rightarrow &\quad State_{t+1} = NSP \end{aligned} \tag{5}$$

where NSP is a non-speech state and SP is a speech state.

4. Model-based Speech/Non-Speech decoder

The second stage of the process consists of a model-based speech/non-speech detector which obtains an initial segmentation (used for training its models) from the output of the energy-based detector. It then produces the speech/non-speech labels that are used for the speaker diarization task. By training the models from the output of the energy-based detector, we avoid the need for any external training data (or pretrained models).

The model-based decoder is composed of a two states ergodic HMM (following the same architecture as the speaker Diarization system), where one state models silence using a single gaussian model, and the speech state uses a GMM with M mixtures (M¿1). In each state we impose a minimum duration MD. We use EM-ML to train the models and Viterbi to decode the acoustic data. We iteratively segment and train both models until the overall meeting likelihood stops increasing, then we output the speech/non-speech labels.

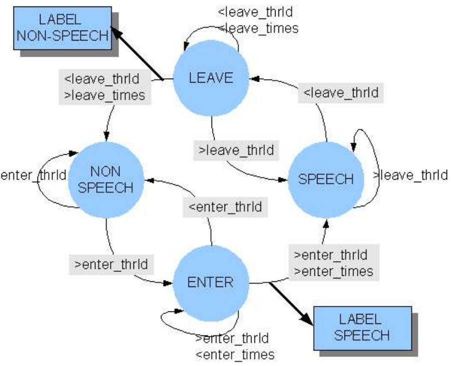

Figure 3: State machine used to apply time constraints.

In order for the speech and silence models to represent well the acoustic information, there needs to be enough frames of data in the input segmentation for each model. As seen in [9] the silence data can be modelled with a single gaussian with a very narrow variance. On the other hand, the speech information is much “broader” and dependant on the speakers present in the meeting. It is therefore important for the data used in training the silence model to contain as little speech data as possible. This translates into a very small “missed speech” rate in the energy based detector.

5. Hybrid Speech/non-Speech Detection

The hybrid Speech/non-Speech detector introduced here is composed of a 2 step process, as seen in figure 4, combining the energy-based detector and the model-based decoder. The output of the energy detector is used exclusively to initialize the model-based decoder, whose output is used as the speaker diarization speech/non-speech input.

As described above, the functioning of the energy detector depends on setting a threshold value properly. In an exclusively energy-based system such threshold has to be defined using a development set as close as possible to the test set to obtain optimum results. By using a model-based decoding as a second step we can relax the need for a perfectly tuned threshold since the aim now is to obtain a rough distinction between speech and non-speech. The Energy detector is initially run with a very low threshold pair (1e-5/1e-6). While the number of non-speech segments found (N_{sil}) is smaller than 10 we raise both thresholds by an order of magnitude and rerun the system. This is done iteratively until $N_{sil} > 10$). At that point, if $N_{sil} > 100$ we consider that there are too many silence segments and we go back to a smaller threshold step size until obtaining between 10 and 100 non-speech segments. The selection of the range (10,100) is defined *a grosso modo* in order to obtain a sufficient amount of silence frames to train the silence model in the model-based decoder with a low percentage of speech labelled as silence.

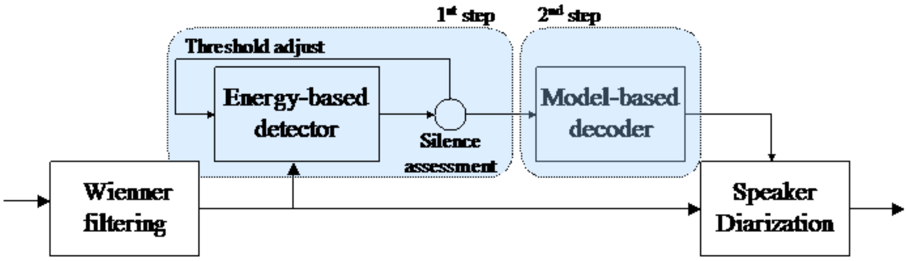

Figure 4: *Hybrid Speech/non-speech detector blocks diagram*

Such speech/non-speech segments are used to train the two models in the model-based decoder, which performs iterative viterbi decodings and EM-ML training on the data until reaching likelihood convergence.

The use of two well known speech/non-speech detection techniques back-to-back allows for the creation of a more robust system than using either of them alone. On one hand, on a system totally energy-based we will encounter that the optimum thresholds defining the speech and non-speech segments are different from one recording type to another (as it depends on the room, microphones used, distance of the people to them, etc.) and therefore they need to be optimized using data from the same source, becoming very dependent on it. On the other hand, in a totally cluster-based system, we need pre-labelled data in order to train the models (or somehow generated initial models), which is also very dependent on the type of recording. By using both systems we can process any type of data we obtain on its own, without the burden of similar data collection or annotation.

There are three main parameters that need to be determined in this hybrid system in order to obtain optimum results. These are the minimum duration of the speech/non-speech in the energy-based detector, the number of gaussian mixtures assigned to speech in the model-based decoder and the minimum duration of speech and non-speech in such decoder.

6. Experiments and Results

Both speech/non-speech and speaker diarization experiments were conducted using the acoustic data distributed for the NIST Rich Transcription 2004 and 2005 Spring Meeting Recognition Evaluation, RT04s and RT05s ([10]). This consists of excerpts from multi-party meetings in English collected at six different sites on various time periods. From each meeting only an excerpt of 10 to 12 minutes is evaluated. Although a number of distant microphones is available for each meeting, only the most centrally located microphone (as defined by NIST as the SDM channel) was used to test the algorithms presented here. We merged the RT04s development and evaluation data to create a development set (to a total of 16 meeting excerpts) to adjust the parameters, and the evaluation data from the NIST RT05s evaluation is used as an evaluation set to validate the chosen parameters.

To evaluate the system we used a total of three metrics. To evaluate the speech/non-speech accuracy we used two metrics which correspond to the possible errors found. They are the Missed Speech (MISS), which accounts of the percentage of the total evaluated time that the reference accounts for speech and the system labels it as silence; and the False Alarm (FA) speech, which is the percentage of silence that is labelled as speech. The sum of both error types is the total speech/non-speech error.

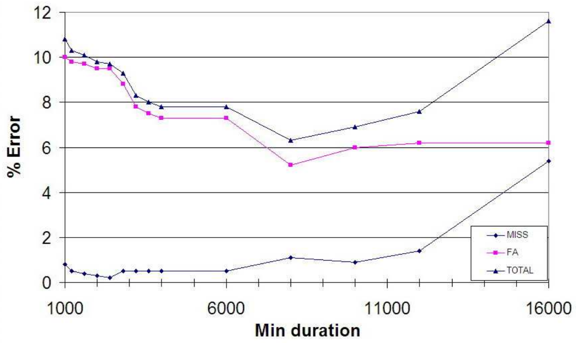

Figure 5: *Energy-based system errors depending on its segment minimum duration*

To evaluate the speaker diarization algorithm, the metric used is the same as is used in the NIST RT evaluations and is called Diarization Error Rate (DER). It is computed by first finding an optimal one-to-one mapping of reference speaker ID to system output ID and then obtaining the error as the percentage of time that the system assigns the wrong speaker label.

In order to compute the error rates from the system output files, ground-truth speaker diarization references have been generated via forced alignment using the ICSI-SRI speech-to-text (STT) system presented for the RT05s NIST evaluation (see [11]). For each meeting all speakers present in the meeting wear an individual headphone microphone (IHM). The data from the IHM channel was hand transcribed, and the STT system was used to align the reference text to the acoustic data in each channel. Then a single file was created by merging all of the alignments. When presenting the metrics we don't evaluate any speech in in the input signal that is overlapped (more than one speaker talking at the same time). One show in the evaluation set had to be descarded due to the lack of official transcripts for a speaker who had called in to the meeting on a speaker-phone.

For all metrics, the overall results given below are the time weighted averages among all meetings in the development or evaluation set.

6.1. Speech/Non-Speech Experiments

We used the development set to estimate the minimum duration of the speech and non-speech segments in the energy-based detector. In figure 5 we can see the MISS and FA scores for various durations (in # frames). While for a final speech/non-speech system we would choose the value that gives the minimum total error, in this case our goal is to obtain enough non-speech data to train the non-speech models in the second step. It is very

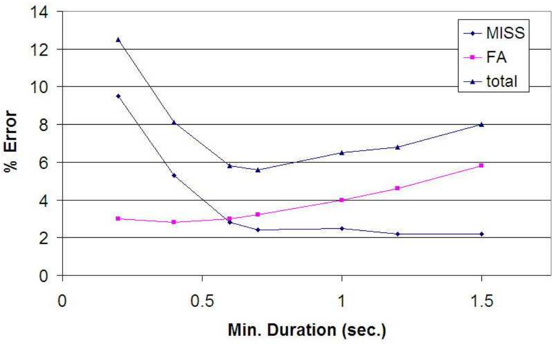

Figure 6: *Model-based system errors depending on its segment minimum duration*

important to choose the value with smaller MISS so that the non-speech model is as pure as possible (the speech model contains more Gaussian mixtures, therefore a bigger FA rate does not influence it as much). We observe how in the range between duration 1000 and 8000 the MISS rate remains quite flat, which indicates that even though, when applying the selected value to a different set of data, if this new set doesn't contain a minimum value for the MISS rate at the same value are our development set, it will most probably still be a very plausible solution. We choose a duration = 2400 (150ms duration) with MISS = 0.3% and FA=9.5% (total 9.7%).

The same procedure is followed to select the minimum duration for the speech and non-speech segments decoded using the model-based decoder, using the minimum duration determined by the previous analysis of the energy-based detector. In figure 6 we can see the FA and MISS error rates for different minimum segment sizes (the same for speech and non speech); such curve is almost identical when using different # mixtures for the speech model, we choose a complexity of 2 gaussian mixtures for the speech model. In contrast to the energy-based system, this second step does output a final result, therefore we are interested in finding the value that minimizes the total percent error. We choose the minimum value of 5.6% error using a minimum duration of 0.7 seconds. If in the energy-based detector we had chosen the parameters that minimize the overall speech/non-speech error (which is at 8000 frames, 0.5 seconds) we would have obtained a minimum error of 6.0% after the cluster-based decoder step.

sp/nsp system	Development data			evaluation data		
	MISS	FA	total	MISS	FA	total
All-speech system	0.0%	11.4%	11.4%	0.0%	13.2%	13.2%
Pre-trained models	1.9%	3.2%	**5.1**%	1.9%	4.6%	6.5%
hybrid (1st part)	0.4%	9.7%	10.1%	0.1%	10.4%	10.5%
hybrid system(all)	2.4%	3.2%	5.6%	2.8%	2.1%	**4.9**%

Table 1: Speech/non-speech errors on development and test data

In table 1 we present the results for the development and evaluation sets using the selected parameters. The ”all-speech” system shows the total percentage of data labelled as non-speech in the reference (ground truth) files. After obtaining the forced alignment from the STT system, there existed many non-speech segments with a very small duration. In the NIST RT evaluations a silence segment is only considered within two segments belonging to the same speaker when it is longer than 0.3 seconds. A postprocessing of the segments was done to conform to this rule. The second row shows the speech/non-speech results using SRI speech/non-speech system [11] which has been trained using training data coming from various meeting sources and its parameters optimized using the development data presented here and the forced alignment reference files. If tuned using the hand annotated reference files provided by NIST for each data set, it obtains a much bigger FA rate, possibly due to the fact it is more complicated in hand annotated data to follow the 0.3s silence rule. The third and forth rows belong to the results for the presented algorithm. The third row shows the errors in the intermediate stage of the algorithm, after the energy-based decoding. These are not comparable with the other systems as the optimization in here is done regarding the MISS error, and not the TOTAL error. The forth row shows the result of the final output from both systems together.

Although the speech/non-speech error rate obtained for the development set is worse than what is obtained using the pre-trained system, it is almost a 25% relative better in the evaluation set. As we will see in the next section, in both cases the new proposed speech/non-speech output helps reduce the DER error in the speaker diarization task.

6.2. Speaker Diarization Experiments

In order to test the usability of the speech/non-speech output for the speaker diarization of meetings data we have run the system explained in section 2 on the development and test data. In table 2 we present the Diarization Error Rates (DER) for the speaker diarization system presented in section 2 using different speech/non-speech system outputs.

sp/nsp system	Development	evaluation
All-speech	27.50%	25.17%
Pre-trained models	19.24%	15.53%
hybrid system	**16.51**%	**13.97**%

Table 2: DER using different speech/non-speech systems

The use of any speech/non-speech detection algorithm improves the performance of the speaker diarization system. Both systems perform much better than just using the diarization system alone. This is due to the agglomerative clustering technique, which starts with a large amount of speaker clusters and tries to converge to an optimum number of clusters via cluster-pair comparisons. As non-speech data is distributed among all clusters, the more non-speech they contain, the less discriminative the comparison is, leading to more errors.

In both the development and evaluation sets the final DER of the proposed speech/non-speech system outperforms by a 14% relative (development) and a 10% relative (evaluation) the system using pre-trained models. We can see how the DER on the development set is better, even though the proposed system has a worse speech/non-speech error. This indicates that the proposed system obtains a set of speech/non-speech segments that are more tightly coupled with the diarization system.

7. Conclusions and future work

In this paper we present a new hybrid speech/non-speech detector and we use it for the task of speaker diarization of meeting data. The proposed system first performs an energy-based detection with an automatic threshold setting to obtain a rough speech/non-speech segmentation, then a second part uses a model-based system where HMM acoustic models are trained with the data from the first step for speech and non-speech and outputs the final segmentation. This speech/non-speech segmentation is used on the diarization system, improving its performance. We show an improvement on Diarization Error Rate (DER) of 14% and 10% relative on the development and testing sets.

In the meetings environment the major source of non-speech is silence, which is what we focus on detecting with this system. In other environments, like broadcast news, silence is normally reduced to the minimum, and what mostly appears are other noises and music. The proposed system could be adapted to be able to process such recordings by exchanging the energy-based speech/non-speech detector by a music detector or a detector able to label other non-speech events present in the recording.

8. References

[1] D. Reynolds and P. Torres-Carrasquillo, “Approaches and applications of audio diarization,” in *ICASSP’05*, Philadelphia, PA, March 2005, pp. 953–956.

[2] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system,” in *Fall 2004 Rich Transcription Workshop (RT04)*, Palisades, NY, November 2004.

[3] Q. Li, J. Zheng, A. Tsai, , and Q. Zhou, “Robust endpoint detection and energy normalization for real-time speech and speaker recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 10(3), 2002.

[4] J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *ASRU’03*, US Virgin Islands, USA, Dec. 2003.

[5] S. Shaobing Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.

[6] J. Ajmera, I. McCowan, and H. Bourlard, “Robust speaker change detection,” *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 649–651, 2004.

[7] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system,” in *Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.

[8] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, “Qualcomm-icsi-ogi features for asr,” in *ICSLP’02*, 2002.

[9] X. Anguera, C. Wooters, and J. Hernando, “Purity algorithms for speaker diarization of meetings data,” in *ICASSP’06*, Toulouse, France, May 2006.

[10] NIST rich transcription evaluations, website: http://www.nist.gov/speech/tests/rt.

[11] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, “Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system,” in *Rich Transcription 2005 Spring Meeting Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.