

Filtering the Unknown: Speech Activity Detection in Heterogeneous Video Collections

Marijn Huijbregts^{1,2}, Chuck Wooters², Roeland Ordelman¹

¹Department of Electrical Engineering, Mathematics and Computer Science,
University of Twente, Enschede, The Netherlands

²International Computer Science Institute,
1947 Center Street, Suite 600, Berkeley, CA 94704, USA

wooters@icsi.berkeley.edu, {huijbreg,ordelman}@ewi.utwente.nl

Abstract

In this paper we discuss the speech activity detection system that we used for detecting speech regions in the Dutch TRECVID video collection. The system is designed to filter non-speech like music or sound effects out of the signal without the use of predefined non-speech models. Because the system trains its models on-line, it is robust for handling out-of-domain data. The speech activity error rate on an out-of-domain test set, recordings of English conference meetings, was 4.4%. The overall error rate on twelve randomly selected five minute TRECVID fragments was 11.5%.

Index Terms: speech activity detection

1. Introduction

For the 2007 TREC Video Retrieval Workshop¹, The Netherlands Institute for Sound and Vision² has provided a collection of 400 hours of Dutch television broadcasts. This collection consists of a broad spectrum of video material. It contains for example broadcast news, children's programs, poetry recitals and documentaries on various topics (further referred to as Sound&Vision data). As part of the collection the TRECVID participants receive transcripts of the speech in the collection that we have automatically generated by means of automatic speech recognition (ASR). As it concerns a heterogeneous video collection, containing videos with a lot of sound effects, music, background noise and different graduations of 'near' silence, an important part of the process that is involved in such a speech recognition task is determining which parts of the videos actually contain speech. This task is typically referred to as speech activity detection (SAD). As a speech decoder will always try to map a sound segment to a sequence of words, processing non-speech portions of the videos (i) would be a waste of processor time, (ii) introduce noise in the transcripts due to assigning word labels to non-speech fragments, and (iii) reduce speech recognition accuracy in general when the output of the first recognition run is used for acoustic model adaptation purposes.

Various methods have been proposed to solve the Speech Activity Detection task. For audio that only contains speech and silence, SAD can be performed by simply determining the energy levels and discarding all segments under a certain threshold as silence. Unfortunately even for tasks with a clear distinction between speech and silence, like in broadcast news recordings,

it is difficult to determine the optimal threshold that discards all silence without removing speech fragments. To circumvent the need for threshold optimization, Gaussian Mixture Models (GMM) can be trained on speech and silence segments. The GMMs are used as probability density functions in a Hidden Markov Model (HMM). With a Viterbi search on the audio the speech-silence segmentation can be found. As long as there is a sufficient amount of training data available, the system can be extended to detect all kinds of audible events, such as music. Typically HMM systems do not only rely on energy as an input feature, but instead use features like Mel Frequency Cepstral Coefficients (MFCCs). Adding these kinds of features increases the SAD performance.

A disadvantage of SAD systems that use models trained on external data is that models need to be retrained when the task domain changes. A SAD system trained on broadcast news data for example, may perform very poorly on recordings of meetings. Given the large variation in audio conditions in the Sound&Vision data, selecting data for training silence and speech models that are more or less representative for the collection as a whole, is difficult. In addition, it is not straightforward to determine what kind of extra models are needed to filter out unknown audio fragments such as music or sound effects and to select data for training the models.

In this paper, we report on our approach for filtering non-speech out of the audio from the Sound&Vision TRECVID 2007 collection prior to the actual speech recognition task. Instead of entirely relying on pre-defined models, the system automatically trains dedicated models on the data under evaluation itself. This approach benefits from the main advantages of regular HMM systems, for example no thresholds need to be set, but it is more robust for changing audio conditions.

Our approach consists of two stages. First, bootstrap models are used to create a rough initial segmentation. The high confidence regions of this segmentation are used in the second stage to train a special silence, sound and speech model. In the next section the algorithm used to train these models is described. Section 3 describes the evaluation experiments that were performed on this system.

2. System description

The SAD system proposed by [1] at the NIST Spring 2006 Rich Transcription (RT06s) evaluation also consists of two stages. It first selects those regions in the audio with high and low energy levels. In the second stage it trains dedicated speech models on the high energy regions and silence models on the low energy

¹<http://trecvid.nist.gov>

²<http://portal.beeldengeluid.nl/>

levels. The major advantage of this approach is that no earlier trained models are needed making it robust for domain changes. The drawback of using energy however is that it is not possible to use this approach when the audio contains fragments with high energy levels that are not speech.

Instead of using energy as an initial confidence measure, we propose a system that uses the output of a broadcast news SAD system to determine regions of high confidence. In the second stage, new models are trained using segments with these high confidence scores only. In addition, a third model is trained along the way that (in general) will contain all segments with high energy that are not speech.

In the following subsections we describe the features, the broadcast news SAD system, and the algorithm we developed for training the three new models on only the data itself.

2.1. Feature extraction

As the audio may contain sounds with high energy levels that are not speech, energy is not used as a feature. Instead, the first twelve Mel Frequency Cepstral Coefficients (MFCC) coefficients are used together with zero-crossing and a factor that determines the number of frequencies active at a time frame. During MFCC calculation, 256 frequency bins are calculated (0-8KHz). When voiced speech is produced, only a certain number of these frequency bins will have a high energy level. Thus the fourteenth coefficient is simply the total number of frequency bins that exceed a threshold.

The features are calculated using 32ms Hamming windows that are shifted 10ms at a time. After calculating each feature vector, the delta's and delta-delta's are added resulting in a vector of 42 components.

2.2. Broadcast news SAD system

In the first stage, a broadcast news SAD system is used to create a first rough alignment. This system 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 figure 1). The minimum duration for silence is set to 30 states (300ms) and the minimum duration for speech is set to 75 states.

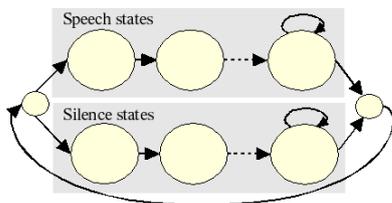


Figure 1: The Broadcast News HMM topology. The two strings of states represent speech and silence. All states in a string share one GMM trained on Dutch broadcast news data. All transition probabilities are set to one.

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) [2]. Three and a half hours of speech and half an hour of silence from 200 male and 200 female speakers were used. The models have been

initialized with a single Gaussian. The number of gaussians was increased iteratively until a mixture of 20 gaussians was reached.

The data was forced aligned to the reference transcription to ensure the correct placements of speech/silence boundaries. To make sure that only speech was used to train the speech model, all phones neighbouring silence were not used.

We evaluated the SAD system on a broadcast news show of half an hour. The total SAD error rate of the system on this show was 4.5%. This error rate was obtained by dividing the amount of misclassified data by the total amount of actual speech [3].

2.3. Algorithm

The broadcast news SAD system works well on clean broadcast news data that does not contain any other sounds than speech. In cases that the audio contains a few sounds such as anchor jingles these sounds will typically be classified as silence. Most non-speech sounds are not very well modelled by the speech model and will often fit the more general silence model best. Note that although the silence model is trained on silence, energy is not used as a feature and therefore the model can not distinguish low energy noise from high energy noise.

When data outside the broadcast news domain is used, the misclassification rate will certainly increase. However, by exploiting again the high confidence regions, new GMMs can be trained that better fit the out-of-domain data. Those new models are then improved iteratively by re-aligning the data and re-training the models a number of times.

In Figure 2 the algorithm is 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, chunking simplifies the tuning of the system parameters. In the final step of the algorithm, the chunks are concatenated. When two neighboring segments from different chunks are classified the same, the segments are merged.

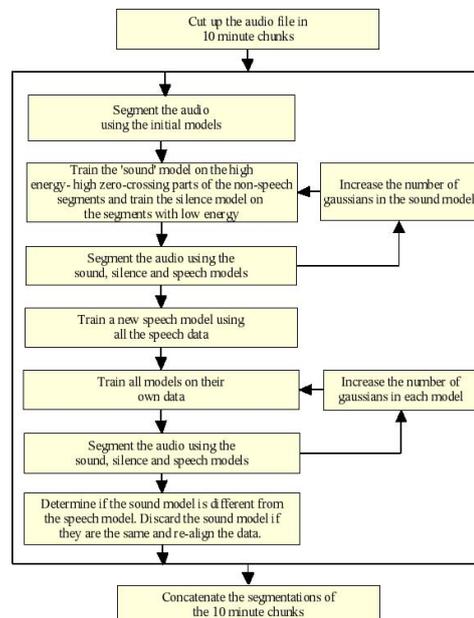


Figure 2: Overview of the speech activity detection system.

Each audio chunk is first segmented using the broadcast

news speech and silence models. At this point the majority of data classified as speech will actually be speech, but the silence segments may contain silence, sounds and also speech. Next, a silence and a sound model are created from part of the segments classified as silence to improve the initial segmentation. All data that is classified as silence is split into two data sets. A small amount of data with the lowest average energy is used to train a new silence model. Another small amount of data with high energy levels and high average zero crossing coefficients is used to train a sound model. The new models are trained using the same feature vectors described in section 2.1 and do not use energy as a feature.

Using the new silence and sound models and the old speech model, a new segmentation is created. This segmentation is used to train more refined silence and sound models. All data assigned to the sound and silence models by the new segmentation 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 as it is trained on the new data before the speech model and therefore may fit the data better. The remaining data is divided over the silence model and the sound model as before. The silence model receives data with low energy levels and the sound model receives data with high energy and zero crossing levels. This time though, the models will be assigned more data and consequently more gaussians will 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 and some sound effects appear to be modelled by the silence GMM very well. Because the goal of this system is to find all speech segments and discard everything else, this is not a problem.

After the silence and sound models are trained, a new speech model will be trained using all data that is classified as speech. By now, most non-speech will already be pulled out by the other models so that it is possible to train a good speech model on all data and not only on high confidence regions. Once the new speech model is created, all models are iteratively re-trained with increasing numbers of gaussians. At each training iteration the data is re-segmented.

This algorithm works for audio of various domains and with a range of non-speech sounds, but it is not well suited when the data only contains speech and silence. In that case the sound model will be trained solely on the speech that is misclassified at the first iteration (because the initial models are trained on a different domain, 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, we use the Bayesian Information Criterion (BIC) to check if the sound and speech model are the same.

For the BIC comparison, a new model θ is trained containing the sum of the number of gaussians in the two original models θ_a and θ_b . This *merged* model is trained using the training data of the original two models. If the sound and speech models are the same (if they both model speech) this merged model must be able to replace the two models without decreasing system performance. Because the total number of gaussians will not change if θ replaces θ_a and θ_b , the complexity of the system will not change. This makes it possible to use the following formula to calculate the BIC score for merging two models. D_a

file ID	BN SAD	% 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
EDL_20050216-1051	13.8	0.6	1.2	1.8
EDL_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

Table 1: SAD error rates for the RT06s conference meetings

is the data used to train model θ_a , D_b to train θ_b and D to train model θ .

$$BIC(\theta_a, \theta_b) = \log P(D|\theta) - \log P(D_a|\theta_a) - \log P(D_b|\theta_b) \quad (1)$$

If the BIC score is positive, the two models are replaced by the single new model θ .

3. Experiments

The SAD system is evaluated on three different benchmarks. To test its performance on out of domain data, the system was evaluated first on the RT06s conference meeting evaluation data. Not only are the topics of these meetings different from general broadcast news topics, also the audio conditions and the language do not match the Dutch broadcast news training data (section 3.1).

A speech/music test set is used to determine if the algorithm is able to classify music as non-speech (section 3.2), and finally, twelve fragments from the 2007 TRECVID collection were used for evaluating the system on the domain of focus (section 3.3). Again, for all evaluations the error percentage is obtained by dividing the amount of misclassified data by the total amount of data that actually contained speech [3].

3.1. Out-of-domain evaluation

In the first out-of-domain evaluation, the SAD system trained on Dutch broadcast news data was used to determine the initial segmentation. Table 1 contains the SAD results on the nine RT06s conference meetings. As a baseline, the error of the first segmentation, the Dutch Broadcast News system, is shown. 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 illustrates that the conference meeting data is indeed out-of-domain for our initial models. The overall error of the total system is only 4.4%. This is in line with the state-of-the-art at RT06s.³

3.2. The IDIAP speech/music evaluation

In the second evaluation, a speech/music test set described in [4] is 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.

³We have used this system to perform SAD for our RT07s speaker diarization submission.

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

Table 2: Classification results on the IDIAP speech/music test set. The scores are all percentages of correctly classified frames.

file ID	speech (sec)	BN SAD	% missed speech	% false alarm	% SAD 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	20.4	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	16.6	6.8	1.7	8.5
36058	179.62	34.8	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	18.3	7.5	2.9	10.4

Table 3: SAD error rates for the twelve TRECVID fragments. Each fragment is five minutes long. The third column contains the error of the first stage BN alignment.

The third file contains more speech than music while the fourth file contains more music. The performance was 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. Note that this is different from the measure used for the NIST evaluations described earlier, but because we did not have the exact speech alignment we decided to measure the performance on this set the same way as in [4].

In Table 3.2 the results of our system on the four files are listed. Our system did not perform as well as the best system in [4] (on average 95.2%), but considering the system was initialized with Dutch models, the average score of 92.1% can be regarded as satisfactory.

3.3. Dutch TRECVID collection evaluation

For the final TRECVID collection evaluation we randomly selected five minute fragments of twelve different documents from the fifty hours of video material from the 2007 TRECVID development set. We manually annotated these fragments and determined the speech regions by applying forced alignment on the Dutch speech. Table 3 lists the results of the system on these twelve fragments. The overall error is 11.5% of the total speech in the audio. Note that only 39 minutes of the in total one hour long test set is actual speech.

4. Discussion and conclusions

Filtering non-speech out of an acoustically heterogeneous video archive such as the Sound&Vision TRECVID collection is one of the many challenges when automatically annotating the archive. The size and variety of the collection makes it hard to train special sound models. Instead we have proposed a sys-

tem that will automatically train a sound model for all the audio files. This system is tested on three benchmarks and the results convinced us to use it in both the RT07s benchmark for our speaker diarization system and for filtering out non-speech for the TRECVID collection.

Overall on the TRECVID 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 these regions. On the other hand, using this SAD system only 3.2% non-speech will be processed. If we would not use this SAD step, this percentage would be 54% (21 minutes of the total test set is non-speech). Manual inspection of the missed speech showed that most missed speech is speech over various sources of non-speech. It is hard to perform correct ASR on this kind of speech and therefore we think that the gain of not processing 54% of non-speech is more important than missing 8.3% of the speech for further processing.

We are currently investigating other methods to obtain the initial segmentation for training the three models. Although the GMM based approach works reasonably well, we believe that we can make the system more robust if no models at all are needed for the initial segmentation.

A problem related to SAD that is not yet addressed by our current system is detecting and discarding foreign speech fragments. Similar to non-speech segments, feeding foreign speech into the ASR system will influence its performance negatively. For next years TRECVID evaluation we will address this problem.

5. Acknowledgements

The work reported here was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811) by the bsik-program MultimediaN which is funded by the Dutch government (<http://www.multimedien.nl>) and the EU projects MESH (IST-FP6-027685), and MediaCampaign (IST-FP6-027413). We would like to thank IDIAP for providing us with the speech/music benchmark files.

6. References

- [1] X. Anguera, C. Wooters, and J. Pardo, "Robust speaker diarization for meetings: Icsi rt06s evaluation system," in *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation, RT06s, Washington DC, USA*, ser. Lecture Notes in Computer Science, vol. 4299. Berlin: Springer Verlag, October 2007.
- [2] N. Oostdijk, "The spoken dutch corpus. overview and first evaluation." in *Second International Conference on Language Resources and Evaluation*, M. Gravididou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhaouer, Eds., vol. II, 2000, pp. 887–894.
- [3] J. Fiscus, J. Ajot, M. Michel, and J. Garofolo, "The rich transcription 2006 spring meeting recognition evaluation," in *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation, RT06s, Washington DC, USA*, ser. Lecture Notes in Computer Science, vol. 4299. Berlin: Springer Verlag, October 2007, pp. 309–322.
- [4] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.