

Hidden Markov Model Based Speech Activity Detection for the ICSI Meeting Project

Thilo Pfau¹ and Daniel P.W. Ellis²

¹International Computer Science Institute, Berkeley, CA

²Dept. of Electrical Engineering, Columbia University, New York, NY

Abstract

As part of a project into speech recognition in meeting environments, we have collected a corpus of multi-channel meeting recordings. We expected the identification of speaker activity to be straightforward given that the participants had individual microphones, but simple approaches yielded unacceptably erroneous labelings, mainly due to crosstalk between nearby speakers and wide variations in channel characteristics. We have therefore developed a more sophisticated approach for multichannel speaker activity detection based on a simple hidden Markov model (HMM).

A baseline HMM speech activity detector has been extended to use mixtures of Gaussians to achieve robustness for different speakers under different conditions. To further improve the channel independence, normalized features are used. The use of the proposed energy normalization yields a relative reduction in frame error rate by 26.4%. In a postprocessing step the crosscorrelation between different channels is used to detect crosstalk. Using this postprocessing step results in a further reduction of the frame error rate by 12.4%.

1. Introduction

The meeting project at ICSI aims at processing (transcription, query, search, and structural representation) of audio recorded from informal, natural, and even impromptu meetings. Details about the challenges to be met in this project, the data collection, and human and automatic transcription efforts undertaken in this project can be found in [1]. Each meeting in our corpus is recorded with close-talking microphones for each participant (a mix of headset and lapel mics), as well as several ambient (tabletop) mics.

In this paper we will focus on the task of automatically segmenting the individual participants' channels into portions where that participant is speaking or silent. We cast this as segmentation into "speech" (S) and "nonspeech" (NS) portions. Our interest in this preliminary labeling is threefold:

- Accurately pre-marking the speech segments greatly improves the speed of manual transcription, particularly when certain channels contain only a few words.
- Knowing the regions of active speech helps reduce errors and computation time for speech recognition experiments. For instance, speaker adaptation techniques assume segments contain data of one speaker only.
- Patterns of speech activity and overlap are valuable data for discourse analysis, and may not be extracted with the desired accuracy by manual transcribers.

The obvious approach to this problem, to use an energy threshold on each close-mic'd channel, turned out to work very poorly. Our investigation revealed the following problems:

- *Crosstalk*: In the meeting scenario, with participants sitting close together, it is common to get significant levels of voices other than that of the person wearing the micro-

phone in each channel. This is particularly true for the lapel mics, which pick up close neighbors almost as efficiently as the wearer (however, users prefer not to wear headsets).

- *Breath noise*: Meeting participants are often not experienced in microphone technique, and in many instances, the head-worn microphones pick up breath noise, or other contact noise at a level as strong or stronger than voice.
- *Channel variation*: The range of microphones and microphone technique between and within meetings means that the absolute speech level, and the relative level of background noise, varies enormously over the corpus.

For these reasons, we have found it necessary to develop a more sophisticated system to detect the activity of individual speakers.

The remainder of this paper is organized as follows. In Section 2 we will present both the architecture of the S/NS detector and the features used in the multichannel setting. Section 3 describes our approach to correcting crosstalk pickup via cross-correlation. Section 4 presents experimental results with the new S/NS detector, section 5 presents a brief discussion, and section 6 gives a conclusion.

2. HMM-based S/NS detection

2.1 Baseline architecture

2.1.1 Structure of the HMM

The S/NS detection module is based on a hidden Markov model (HMM) S/NS detector designed for automatic speech recognition on close talking microphone data of a single speaker [2]. The baseline detector is similar to the one used in [3], and consists of an ergodic HMM with two main states – "speech" (S) and "nonspeech" (NS) – and a number of intermediate state pairs (S' and NS') to impose time constraints on transitions between the two main states (see figure 1). Both main and intermediate states use the same multivariate Gaussian density, i.e. one Gaussian for "speech" (S and S') and one Gaussian for "nonspeech" (NS and NS').

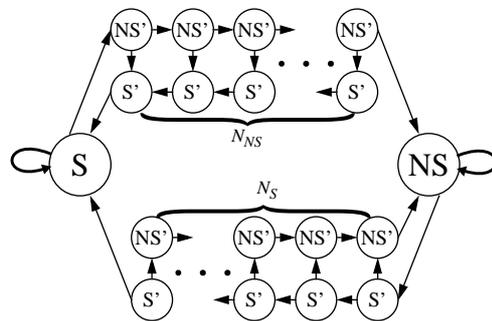


Figure 1: Structure of the HMM based S/NS detector.

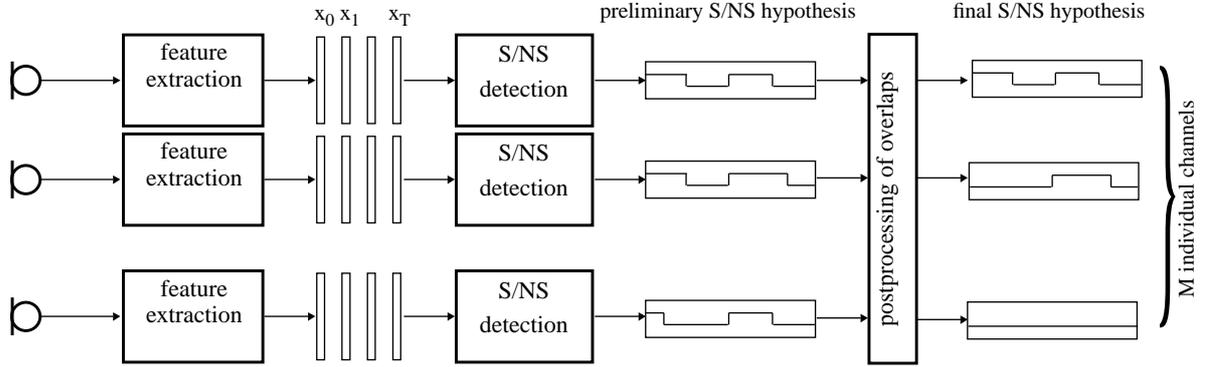


Figure 2: Architecture of the multichannel S/NS detector

2.1.2 S/NS detection process

The S/NS labeling is based on a Viterbi search with a simplified decision strategy, which does not require an absolute endpoint: Whenever one of the main states is found to have the maximum Viterbi path score, an immediate decision can be made. In all other cases (one of the intermediate states is assigned the maximum) the decision is delayed. The HMM structure ensures that the duration of this delay is limited to:

$$D = \max(2N_S, 2N_{NS}) \quad (1)$$

Based on the current decision of the detector, all dimensions of the current Gaussian are adapted to the mean and variance calculated over a sliding window. A threshold on the posterior probability is used to exclude time-frames for which the confidence in the decision is low. Experiments have proven a time constant of two seconds and a threshold for the posterior probability of 0.9 to be suitable.

2.2 Modifications for the meeting project

2.2.1 HMM with Gaussian mixtures

Crosstalk makes the distribution of features in the “nonspeech” state much more complex than in relatively static background noise. Therefore, a mixture of Gaussians are used for the “non-speech” state. A mixture is used also for the “speech” state, motivated by the fact that the S/NS detector for meeting data should be channel independent, i.e. cope with different speakers and different microphones without the need for retraining.

2.2.2 Features for S/NS detection

The wide variability of channel characteristics and signal level has considerable influence on the features used to model distributions within the HMM states. To avoid dependence on absolute level, we use a set of “normalized” features. The complete feature vector comprises 25 dimensions and is calculated over a 16 ms Hamming window with a frame shift of 10 ms. The feature vector contains loudness values of 20 critical bands up to 8kHz (distance between adjacent bands 1 bark), energy, total loudness, modified loudness [4], zerocrossing rate, and the difference between the channel specific energy and the mean of the farfield microphone energies.

Apart from the zero crossing rate, which is independent of a scaling of the signal, the other components of the feature vector are normalized as follows: Spectral loudness values are normalized to the sum over all critical bands. The total loudness and the modified loudness are normalized using the overall maximum within each channel.

The log-energy $E_j(n)$ of channel j at frame n is normalized according to equation 2:

$$E_{norm,j}(n) = E_j(n) - E_{min,j} - \frac{1}{M} \sum_k E_k(n) \quad (2)$$

First, the minimum frame energy $E_{min,j}$ of channel j is subtracted from the current energy value $E_j(i)$ to compensate for the different channel gains. Here the minimum frame energy is used as an estimate of the “noise floor” in each channel, to make this normalization mostly independent of the proportion of speech activity in that channel.

In the second step the mean (log) energy of all M channels is subtracted. This procedure is based on the idea that when a single signal appears in all the channels, the log energy in each channel will be the energy of that signal plus a constant term accounting for the linear gain coupling between that channel and the signal source. Subtracting the average of all channels should remove the variation due to the *absolute signal level*, leaving a normalized energy which reflects solely the *relative gain* of the source at channel j compared to the average across all channels. Signals that occur only in one channel, such as microphone contact and breath noise, should also be easy to distinguish by this measure, since in this case the relative gain will appear abnormally large for the local microphone.

2.2.3 Architecture of the multichannel S/NS detector

For a meeting with M individual channels, M independent S/NS detection modules are used to create a separate preliminary S/NS hypothesis for each of the M channels (see figure 2). The M preliminary hypotheses are then fed into a postprocessing module which focuses on correcting overlap regions (i.e. regions where several hypotheses show activity) as described in the next section.

3. Crosscorrelation analysis

The peak normalized short-time crosscorrelation,

$$\hat{\rho}_{ij} = \max_l \left\{ \frac{\sum_n (x_i[n] \cdot x_j[n+l])}{\sqrt{\sum_n x_i[n]^2 \cdot \sum_n x_j[n]^2}} \right\} \quad (3)$$

between the active channels i and j are used to estimate the similarity between the two signals. For “real” overlaps (two speakers speaking at the same time) the crosscorrelation is expected to be lower than for “false” overlaps (one speaker coupled into both microphones). For sound coming from a single source, the crosscorrelation shows a maximum at a time

skew corresponding to the difference in the arrival time of the signal at the two microphones.

The postprocessing module calculates the crosscorrelation function for time skews up to 250 samples (ca. 5m difference between the microphones) on 1024 point signal windows. The maximum is smoothed via median filtering over a 31 point window. When the smoothed maximum correlation exceeds a fixed threshold, the hypothesized “speech” region of the channel with the lower average energy or loudness is rejected. The threshold is chosen as described below.

We consider in particular the relation of a lapel microphone (channel 0) and a headset microphone (channel 1). Table 1 shows the counts of frames incorrectly labeled as overlapping (both channels active) in the preliminary analysis, broken down by the true state (according to hand labels).

Table 1: Frame counts of erroneous overlap labeling

true state	frame count
chan0 only	15 (0.6%)
chan0 and others	186 (7.0%)
chan1 only	1391 (52.4%)
chan1 and others	938 (35.3%)
other channels only	41 (1.5%)
no other channel	83 (3.1%)
total	2654 (100%)

As can be seen, the majority (88%) of erroneous overlap detections is found when channel 1 is active (rows “chan1 only” and “chan1 and others” of table 1), whereas activity in channel 0 is not combined with a large number of errors of this type. This is not surprising, since the lapel microphone of channel 0 will pick up more speech from other speakers than the headset microphone of channel 1.

Figure 3 shows the histograms of the smoothed maximum correlation between channel 0 and channel 1 for true overlap regions (according to the hand transcriptions) compared to error frames where channel 1 was active. It can be seen that choosing a threshold between 0.4 and 0.7 will successfully reject many of the cases when activity in channel 1 is causing the preliminary analysis to mistakenly label channel 0 as active, while excluding few or none of the truly overlapped frames.

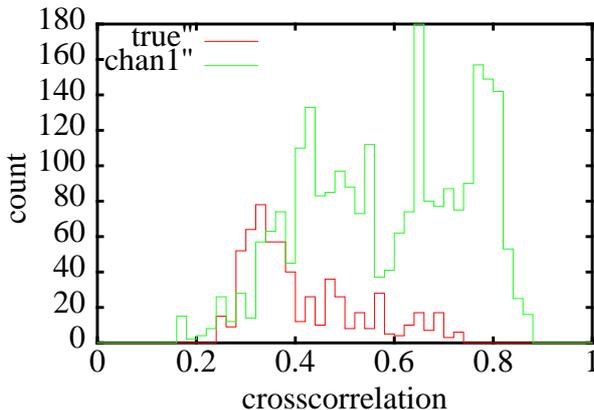


Figure 3: Smoothed maximum correlation for true overlaps and frames mislabeled overlap where channel 1 was active.

4. S/NS detection experiments

4.1 Training and test data

Training and test data for the experiments are chosen from five manually transcribed meetings of the ICSI meeting project.

The training data consists of the first 20 minutes of conversational speech of a four speaker meeting (MRM002) with three male and one female speakers, three wireless headset microphones and one wireless lapel microphone. For each channel a label file specifying four different S/NS categories (foreground speech, silence regions, background speech and breath noises) was created using the Transcriber tool [5]. Not all categories were found in all channels. In fact breath noises were marked only for one channel, where heavy breath sounds (due to a poor microphone setup) were audible. For the lapel channel, background speech was marked.

Table 2: Training data for S/NS detector

Meeting	Duration (minutes)	From - To (minute)	#Speakers (#Female, #Male)
MRM002	20	0 - 20	4 (1, 3)

Table 3: Test data for S/NS detector

Meeting	Duration (minutes)	From - To (minute)	#Speakers (#Female, #Male)
MRM003	5	17 - 22	6 (2, 4)
MRM004	5	29 - 34	8 (2, 6)
NSA003	5	12 - 17	6 (1, 5)
ROB004	5	20 - 25	7 (1, 6)

The test data consists of 20 minutes of conversational speech taken from four different meetings (MRM003, MRM004, NSA003 and ROB004). Five consecutive minutes were chosen from each of these meetings, which involved several different speakers and showed frequent speaker changes and/or overlapping between speakers.

4.2 Experimental results

In order to evaluate the quality of the S/NS detection module the frame error rate for the two class problem of classification “speech” and “nonspeech” is given. In addition to the frame error rate, two different types of errors, inserted speech frames (false alarms) and missed speech frames (false rejections), are given in table 4.

The frame error rates vary between 10.8% and 25.3% without energy normalization, between 10.2% and 17.3% with energy normalization but without postprocessing and between 8.3% and 16.9% when both energy normalization and postprocessing are applied. In this case the frame error rates are consistently lower than for the other two cases for all four meetings. The error rate reduction is caused by a decrease in false alarms, whereas the number of false rejections increases.

Whereas there is a great variation in the percentage of false alarms (“nonspeech” portions classified as “speech”), the percentage of false rejections (“speech” portions missed by the detector) is less variable. In general the S/NS detector tends to be too sensitive, since there are more false alarms than false rejections. A considerable amount of the false alarms is, however, of systematic nature resulting from heuristics to avoid truncating final consonants (see [2] for details).

Table 4: S/NS detection results with and without normalized energy and with and without crosscorrelation based postprocessing

energy normalization	post-processing	frame error rate (false rejections / false alarms) in %				
		MRM003	MRM004	NSA003	ROB004	average
no	no	22.5 (1.3/21.2)	25.3 (0.8/24.5)	15.8 (1.7/14.1)	10.8 (3.1/7.7)	18.6 (1.7/16.9)
yes	no	12.0 (1.4/10.6)	17.3 (0.8/16.5)	14.6 (2.0/12.6)	10.2 (3.0/7.2)	13.7 (1.8/11.9)
yes	yes	8.3 (2.4/5.9)	16.9 (1.3/15.6)	13.0 (2.6/11.4)	8.6 (3.1/5.5)	12.0 (2.2/9.8)

5. Discussion

Deeper analysis of the errors made by the preliminary labeling shows a significant difference between the case in which channel 1 is the only active channel, and the case of activity in more than one channel but including channel 1, in which case the peak correlation is well below the mode of the histogram of errors shown in figure 3. The smaller peak value indicates that other sources, such as activity in one of the other channels, might also contribute to the occurrence of this type of error. In fact, a high correlation between channel 0 and one of the remaining channels can be found in many of these cases; in 68% of these frames, the normalized cross-correlation exceeds 0.5 with one of the other channels.

The use of the “normalized energy” of equation 2 reduces the error rate by 26.4% relative (see rows 1 and 2 of table 4). The reduction is mainly caused by a decrease in false alarms. For some of the channels it leads to an increase in false rejections. A deeper analysis of the results shows, that without energy normalization, the error rates of the lapel microphones are especially bad. This makes us believe that the normalization is essential to cope with the channel variations found on the meeting recorder data.

A comparison of the S/NS detection results achieved with and without crosscorrelation based postprocessing (rows 2 and 3 of table 4) shows, that the use of a predefined threshold is an efficient way of reducing error rates. In average the frame error rate was reduced from 13.7% to 12.0%, which is a relative reduction of 12.4%. The reduction is again caused by a decrease in false alarms; however, it goes along with an increase in false rejections.

All in all, the combined use of both the energy normalization and the postprocessing reduces the accuracy of the system in detecting true speech segments. However, the number of falsely detected speech segments is reduced by a significantly greater amount. On the one hand, the transcribers therefore have to be more careful to detect speech segments which were missed by the system. On the other hand, the number of ‘empty’ segments distracting the transcribers is reduced.

Cross-correlation analysis suggests another approach to this problem - that of estimating the coupling between different channels and using the estimates to cancel the crosstalk signals. We are investigating an approach to this based on the Block Least Squares algorithm described in [6]. However, the situation is complicated by the very rapid changes in coupling that occur when speakers or listeners move their heads. Since the coupling filters are sensitive to changes of just a couple of centimeters, these movements are highly significant.

Our ultimate goal is to develop technologies to make a useful table-top meeting recorder unit. Such a device would likely have several microphones, and would again use cross-correlation analysis to support the detection of speaker activity and turns.

6. Conclusion

In this paper, an HMM based approach to speech activity detection was presented. It was used in the framework of the ICSI meeting project (see [1]) to provide additional information for the transcription process of the project.

A baseline HMM speech activity detector ([2], [3]) has been extended to use mixtures of Gaussians to achieve robustness for different speakers under different conditions. To further improve the channel independence, normalized features are used. It has been shown that the use of the proposed energy normalization method leads to reductions in the frame error rate. In a postprocessing step the crosscorrelation between channels is used to detect crosstalk. A predefined crosscorrelation threshold has been proven experimentally to be appropriate. Both approaches have been combined successfully.

The experimental results show, that the presented system is able to capture most of the speech segments in the different channels. Since the S/NS detection is performed for each channel separately, the system is able to detect regions, where more than one speaker is active at the same time.

However, the results also show, that still a considerable amount of false alarms is produced by the system. These often stem from human and nonhuman noise. Therefore future work will concentrate on finding noise robust features and/or improved noise modeling within the HMM.

Acknowledgements

The authors wish to thank Jane Edwards for providing the reference transcripts used for evaluation, Dave Gelbart and Adam Janin for helping to create speech/nonspeech segmentations for training, Dave Gelbart for adapting the ‘Transcriber’ tool, and Nelson Morgan and the meeting recorder team for discussions about the ongoing work.

References

- [1] Morgan, N. et al, “The Meeting Project at ICSI”, Proc. of the Human Language Technology Conference (in press), San Diego, CA, 2001.
- [2] Beham, M., Ruske, G., “Adaptiver stochastischer Sprache/ Pause-Detektor, Proc. of the DAGM-Symposium ‘Mustererkennung’, pp.60-67, Bielefeld, Germany, 1995.
- [3] Acero, A., Crespo, C., de la Torre, C., and Torrecilla, J.C., “Robust HMM-Based Endpoint Detector”, Proc. of Eurospeech 1993, pp. 1551-1554, Berlin, Germany.
- [4] Ruske, G., “Automatische Spracherkennung, Methoden der Klassifikation und Merkmalsextraktion”, second edition, Oldenbourg publ., München Wien, 1994.
- [5] Barras, C. et al., “Transcriber: A Tool for Segmenting, Labeling and Transcribing Speech”. <http://www.etca.fr/CTA/gip/Projets/Transcriber>
- [6] E. Woudenberg, E., Soong, F. & Juang, B., “A Block Least Squares approach to acoustic echo cancellation”, Proc. ICASSP-99, Phoenix, vol. 2 pp. 869-872.