MULTISPEAKER SPEECH ACTIVITY DETECTION FOR THE ICSI MEETING RECORDER

Thilo Pfau[1], Daniel P.W. Ellis[2], and Andreas Stolcke[1,3]

[1]International Computer Science Institute, Berkeley, CA, [2]Department of Electrical Engineering, Columbia University, New York, NY
[3]Speech Technology and Research Laboratory, SRI International, Menlo Park, CA

tpfau@icsi.berkeley.edu, dpwe@ee.columbia.edu, stolcke@speech.sri.com

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. Therefore, we have developed a more sophisticated approach for multichannel speech activity detection using 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. Feature normalization and crosscorrelation processing are used to increase the channel independence and to detect crosstalk. The use of both energy normalization and crosscorrelation based postprocessing results in a 35% relative reduction of the frame error rate.

Speech recognition experiments show that it is beneficial in this multispeaker setting to use the output of the speech activity detector for presegmenting the recognizer input, achieving word error rates within 10% of those achieved with manual turn labeling.

1. INTRODUCTION

The Meeting Recorder 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, 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 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 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 transcription.

The obvious approach to this problem, energy thresholding on each close-mic'd channel, turns out to give poor results. 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 microphone-wearing person on 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 noises, or other contact noises, at a level as strong or stronger than the voice.
- *Channel variation*: The range of microphones and microphone techniques between and within meetings means that the absolute speech level, and the relative level of background noise, vary widely over the corpus.

The multiparty spontaneous speech recorded on multiple separate microphones for this project is not represented in any standard task or database, and many of these problems have attracted little or no previous attention. For these reasons, we 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 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 crosscorrelation. Section 4 presents experimental results with the new S/NS detector, Section 5 presents a discussion, and Section 6 gives conclusions.

2. HMM-BASED S/NS DETECTION

2.1 Baseline architecture

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" and "nonspeech" – and a number of intermediate state pairs to impose time constraints on transitions between the two main states. Both main and intermediate states use the same multivariate Gaussian density, i.e., one Gaussian for "speech" and one Gaussian for "nonspeech".

2.2 Modifications for the Meeting Recorder 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 "nonspeech" 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.

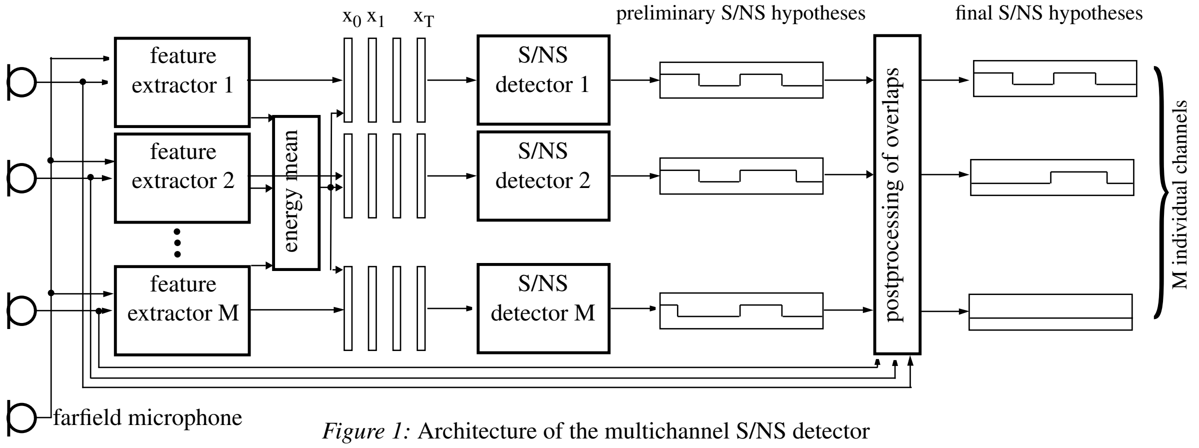

Figure 1: Architecture of the multichannel S/NS detector

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.

Zero crossing rate is independent of signal scaling, but 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 by:

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

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. The minimum frame energy is used as an estimate of the "noise floor" 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 detection modules are used to create preliminary S/NS hypotheses for each of the M channels (see Figure 1), which are then fed into a postprocessing module which focuses on correcting overlap regions (i.e., regions where several hypotheses show activity) as described below.

3. CROSSCORRELATION ANALYSIS

The peak normalized short-time crosscorrelation,

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

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).

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%)

Table 1: Frame counts of erroneous overlap labeling

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

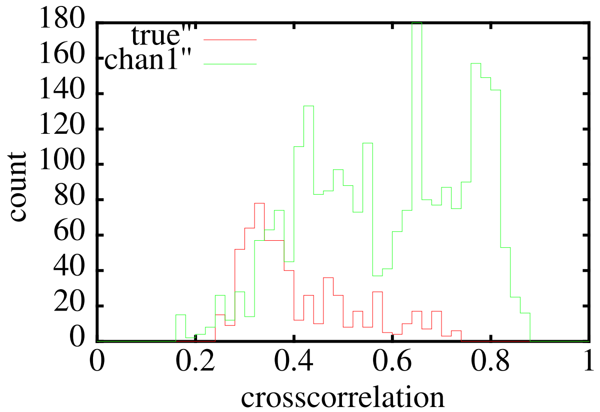

Figure 2: Smoothed maximum correlation for true overlaps and frames mislabeled overlap where channel 1 was active.

surprising, since the lapel microphone of channel 0 will pick up more speech from other speakers than the headset microphone of channel 1.

Figure 2 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 is active, while excluding few or none of the truly overlapped frames.

4. EXPERIMENTS AND RESULTS

4.1 S/NS detection

4.1.1 Training and test data

The training data consists of the first 20 minutes of conversational speech of a four-speaker meeting with three males and one female, wearing 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 manually created using the Transcriber tool [5].

The test data consists of conversational speech from four different multispeaker meetings. Five consecutive minutes were chosen from each channel of these meetings totalling 135 minutes (27 channels, 5 minutes each). The chosen regions involved several speakers and showed frequent speaker changes and/or speaker overlaps. These regions were manually marked with the categories “speech” and “nonspeech”.

4.1.2 Results

S/NS detection is evaluated using the frame error rate for the two-class problem of classification into “speech” and “nonspeech”, as well as the percentage of false alarms and false rejections.

Table 2 shows that the average frame error rate is 18.0% without energy normalization, 13.7% with energy normalization but without postprocessing, and 12.0% when both energy normalization and postprocessing are applied. This is a relative improvement of 35% which is caused by a decrease in the number of false alarms, whereas the number of false rejections is slightly increased.

energy norm.	post-process.	FER (%)	FRJ (%)	FAL (%)
no	no	18.6	1.7	16.9
yes	no	13.7	1.8	11.9
yes	yes	12.0	2.2	9.8

Table 2: S/NS detection results with/without energy normalization and with/without postprocessing.
FER: frame error rate, FRJ: false rejections, FAL: false alarms

4.2 Automatic Speech Recognition

4.2.1 Recognition system

The recognizer was a stripped-down version of the large-vocabulary conversational speech recognition system fielded by SRI in the March 2000 Hub-5 evaluation [6]. The system performs vocal-tract length normalization, feature normalization, and speaker adaptation using all the speech collected on each channel. The acoustic model consisted of gender-dependent, bottom-up clustered (genonic) Gaussian mixtures. The Gaussian means are adapted by two linear transform so as to maximize the likelihood of a phone-loop model, an approach that is fast and does not require recognition prior to adaptation. The adapted models are combined with a bigram language model for decoding. As an expedient we omitted more elaborate adaptation, cross-word triphone modeling, and higher-order language and duration models from the full SRI recognition system (which yield about a 20% relative error rate reduction on Hub-5 data). Note that both the acoustic models and the language model of the recognizer were identical to those used in the Hub-5 system, i.e., did not include any meeting training data

4.2.2 Test data

Six different meetings were used as test data for the recognition experiments (see Table 3). Only native American speakers with a sufficient word count were included in the ASR test set and the digits reading portions of the meetings were excluded (see [1]).

Meeting	Manual	Automatic	Unsegmented
all	41.6	45.8	73.2
no lapel	41.4	45.4	59.1

Table 3: Word error rate in percent for different segmentations
all: weighted average of all channels
no lapel: weighted average of all channels except lapel channel

4.2.3 Experiments and results

To evaluate the influence of the S/NS detection on ASR performance three types of experiments were conducted using different segmentations of the test data:

1. **Manual** segmentation: the test data of each channel was segmented according to the transcript. Only portions containing speech from the foreground speaker were fed to the recognizer.

2. **Automatic** segmentation: the output of the S/NS detector was used. Only “speech” portions were given to the recognizer.

3. **“Unsegmented”** data: each channel was continuously divided into segments covering the complete signal. The chunking into segments was necessary to feed the recognizer with signals of tractable length. As an expedient we used the feature normalizations and speaker-adapted models from the “manual” condition; the results are therefore an optimistic estimate of recognition error in this condition.

Whereas the manual segmentation provides an upper bound for recognition accuracy, the unsegmented data is expected to show how the recognizer itself can handle the multispeaker situation. Table 3 shows the recognition results for the three different segmentation types for each meeting of the test data. The automatic S/NS segmentation achieves word error rates within 10% relative of the ideal manual labelings.

5. DISCUSSION

In the preliminary labeling, error analysis revealed a signficant difference between the case of channel 1 alone and the case of channel 1 active simultaneous with other channels. In the latter case the peak correlation is well below the mode of the histogram in Figure 2. The smaller peak value indicates that sources such as activity in another 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 crosscorrelation exceeds 0.5 with one of the other channels.

The use of the "normalized energy" of Equation 1 reduces the error rate by 26.4% relative (rows 1 and 2 of Table 2), mainly by a decrease in false alarms at the cost of some added false rejections. A closer analysis of the results shows that, without energy normalization, the error rates of the lapel microphones are particularly high, which makes us believe that the normalization is essential to cope with the channel variations found on this data.

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

The combined use of energy normalization and the postprocessing reduces the accuracy of the system in detecting true speech segments, but the number of falsely detected speech segments is reduced by a much greater amount. The relative cost in transcription of these two kinds of error is not known: Transcribers must take care to detect speech segments which were missed by the system, but the number of distracting "empty" segments is reduced.

Crosscorrelation analysis suggests a different approach to the problem of crosstalk, namely, estimating the coupling between different channels and using the estimates to cancel the crosstalk signals. We are investigating such an approach based on the Block Least Squares algorithm described in [7]. 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 few centimeters, these movements are highly detrimental to this approach.

The ASR experiments show that in the framework of a multispeaker setting, it is crucial to provide reliable information about speeh activity. As is true for most recognizers, the ASR system was not designed to distinguish between foreground and background speech, and the "unsegmented" test condition shows that indeed it fails to do so even with headset microphones, resulting in higher insertion rates. This is consistent with earlier results where elevated insertion rates were found even on hand-segmented meeting data when segment boundaries were not always tightly fitted around the foreground speech [8]. On the other hand, the automatic S/NS detector tends to miss some speech segments, thus reducing recognition accuracy due to an increased number of deletion errors.

A possible direction for future research in this area could be a combination of S/NS detection and speaker verification methods to distinguish between foreground and background speech.

6. CONCLUSIONS

We have presented an HMM based approach to speech activity detection which utilizes feature normalization and crosscorrelation postprocessing. The method was applied to presegment speech data of multispeaker meetings in the framework of the ICSI Meeting Recorder project [1].

To meet the requirements of the multispeaker setting and improve the channel independence of the system, normalized features are used. The proposed energy normalization method leads to reductions in frame error rate. In addition, the experiments show that a crosscorrelation threshold is appropriate for detecting crosstalk. Both approaches have been combined successfully.

The S/NS detection results show that our system is able to capture most of the speech segments in the different channels. Since the S/NS detection produces output for each channel separately, the system is able to detect regions of speaker overlap.

Recognition results indicate that automatic segmentation leads to error rates about 10% higher than using a manual segmentation, but to considerably better performance when compared to speech recognition on unsegmented data.

ACKNOWLEDGMENTS

The authors wish to thank Jane Edwards for providing the reference transcripts used for evaluation, David Gelbart and Adam Janin for helping to create speech/nonspeech segmentations for training, David Gelbart for adapting the 'Transcriber' tool, Don Baron for providing the tools for chunking the signal files, Liz Shriberg and Chuck Wooters for help with data quality control, and Nelson Morgan and the Meeting Recorder team for discussions about the ongoing work.

This work has been funded under the DARPA Communicator project (in a subcontract from the University of Washington), supplemented by an award from IBM. Thilo Pfau is funded by a DAAD scholarship.

REFERENCES

[1] Morgan, N. et al., "The Meeting Project at ICSI", Proc. Human Language Tech. Conf., 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. et al., "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/Projects/Transcriber

[6] Stolcke, A. et al., "The SRI March 2000 Hub-5 Conversational Speech Transcription System", Proc. NIST Speech Transcription Workshop, College Park, MD, May 2000

[7] Woudenberg, E., Soong, F., and Juang, B., "A Block Least Squares Approach to Acoustic Echo Cancellation", *Proc. ICASSP-99*, Phoenix, vol. 2 pp. 869-872.

[8] Shriberg, E., Stolcke, A., and Baron, D., "Observations on overlap: Findings and implications for automatic processing of multi-party conversation", Proc. Eurospeech-2001, Aalborg.