# Improved Speech Activity Detection Using Cross-Channel Features for Recognition of Multiparty Meetings

*Kofi Boakye*[1] *and Andreas Stolcke*[1,2]

[1]International Computer Science Institute, Berkeley, CA, U.S.A.
[2]SRI International, Menlo Park, CA, U.S.A.
{kaboakye, stolcke}@icsi.berkeley.edu

## Abstract

We describe the development of a speech activity detection system using an HMM-based segmenter for automatic speech recognition on individual headset microphones in multispeaker meetings. We look at cross-channel features (energy and correlation based) to incorporate into the segmenter for the purpose of addressing errors related to cross-channel phenomena such as crosstalk. Results demonstrate that these features provide a marked improvement (18% relative) over a baseline system using single-channel features as well as an improvement (8% relative) over our previous solution of separate speech activity detection and cross-channel analysis. In addition, the simple cross-channel energy features are shown to be more robust—and consequently better performing—than the more common correlation-based features.

**Index Terms**: speech activity detection, multi-channel audio, crosstalk.

## 1. Introduction

The segmentation of an audio signal into regions of speech and nonspeech is a critical first step in the task of automatic speech recognition (ASR). This is especially the case within the context of multispeaker meetings with individual headset microphones (IHMs). In such meetings, the microphone channels often contain significant amounts of crosstalk—speech from speakers other than the wearer of the headset—which typically generates insertion errors if processed by the recognizer. In addition, breath or other contact noise can be present, particularly for inexperienced headset wearers with poor microphone technique, and produces similar results. Lapel microphones capture less extraneous mouth noise, but are even more prone to pick up crosstalk speech.

These phenomena present a significant challenge because they cannot be addressed using the energy-based methods developed from single-channel speech activity detection (SAD) systems. In such systems speech/nonspeech (S/NS) decisions are typically performed according to one or more (possibly adaptive) thresholds (as in [1], for example). Crosstalk and breath noise, however, often contain a substantial amount of energy, causing the thresholding methods to falsely trigger. Recent strong interest in the recognition and understanding of multispeaker meetings is demonstrated by initiatives such as the Interactive Multimodal Information Management (IM2), Augmented Multiparty Interaction (AMI), and the NIST Rich Transcription (RT) meeting recognition evaluations. Recent results in the NIST RT evaluations [2] show that errors in speech activity detection are one of the major sources of error in recognition from IHM recordings, providing us with the motivation for the work reported here.

This research was thus not directed at SAD per se, but at improving the ICSI-SRI meeting recognition system, with results measured in terms of word error rate (WER). Our previous approach to SAD for IHM recognition was to perform a time-based intersection of the output from two distinct segmenters:

- A segmenter based on hidden Markov models (HMMs) similar to that described in [3], but simpler in structure and utilizing standard cepstral features
- A local-energy detector that generates segments by zero-thresholding a "crosstalk-compensated" energy-like signal derived from the energy signals of all channels

The effectiveness of the approach lay in that the intersection procedure allowed the main weakness of each segmenter to largely cancel out that of the other: the false alarms arising from crosstalk in the HMM-based segmenter and those arising from breath noise in the local energy detector. Details of the system can be found in [2].

Though relatively well performing, having the cross-channel analysis be disjoint from the speech activity modeling was believed to be a suboptimal approach. This paper details the development of a modified system that addresses this issue by combining the two sources of information through the incorporation of cross-channel features in the HMM-based segmenter.

The remainder of the paper is organized as follows. Section 2 details the HMM-based segmenter, and the ASR system with which we measured segmentation performance is briefly described in section 3. The cross-channel energy modeling is detailed in Section 4. We present development experiments in Section 5, validation of the final system in Section 6, and discussion proceeds in Section 7. Conclusions are given in Section 8.

## 2. HMM-based S/NS Segmenter

### 2.1. HMM architecture

The S/NS segmenter is derived from an HMM-based speech recognition system. The system was modified and simplified to consist of only two classes – "speech" (S) and "nonspeech"' (NS) – each being represented with a three-state phone model. State emission probabilities are modeled using a multivariate Gaussian Mixture Model with 256 components and diagonal covariance matrices. Segmentation is carried out by decoding the full IHM channel waveform. The decoding is potentially performed multiple times, with decreasing transition penalty between the two classes, so as to generate segments that do not exceed 60 seconds in length.

### 2.2. Baseline features

The features used in the baseline system consist of 12th-order Mel-frequency cepstral coefficients (MFCCs), log-energy, along with their first and second differences. The features are computed over a window of 25ms advanced by 20ms and cepstral mean subtraction (CMS) is performed as a waveform-level normalization. Features such as these are common to many speech recognition systems and therefore provide an advantage over those used in [3]. In addition, the cepstral features, being largely independent of energy, provide information unavailable to energy-based systems, which could aid in distinguishing between local speech and other phenomena with similar energy levels (such as breaths and coughs).

### 2.3. Segmenter post-processing

To mitigate the effect of "clipped" segments (i.e., segments that cut off initial or final speech) that may be generated by the segmenter, a post-processing step is performed that pads the segment on both ends by a fixed amount (40ms). Similarly, a post-processing step that merges adjacent segments that have small separation (less than 0.4s) is also performed to "smooth" the segmentation. Segments are merged to a maximum of 60s. These time constraints had been optimized for best recognizer accuracy and a good tradeoff with recognizer runtime (long segments tend to use more decoding time), using our baseline segmentation models. They have not (yet) been reoptimized for the improved segmenter features presented here.

## 3. ASR System

For ASR we used the meeting recognition system fielded by ICSI-SRI in the NIST Spring 2005 Meeting Recognition evaluation (RT-05S), as described in detail in [2]. The recognizer uses multiple decoding passes and front ends for cross-adaptation between subsystems and successive refinement of hypotheses. It uses perceptual linear prediction (PLP) and MFCC acoustic features, the latter augmented with discriminative phone-posterior features estimated by multilayer perceptrons. Features are transformed with vocal tract length normalization and heteroscedastic linear discriminant analysis, as well as feature-level constrained maximum likelihood linear regression (CMLLR). Acoustic models are trained on about 2000 hours of telephone speech data, followed by maximum a posteriori (MAP) adaptation to about 100 hours of meeting data. The language model is a 4-gram estimated from a mix of telephone conversations, meeting transcripts, broadcast, and Web data. The system has two versions: one using two decoding passes for quick turnaround (the "fast" system), and one using an additional six decoding passes for best results (the "full" system).

## 4. Cross-Channel Modeling

For a given speaker and corresponding channel in the IHM condition, the primary complicating factor for speech activity detection is the presence of other speakers. Approaches that use information from the other channels (and thus about the speech activity of the other speakers) are best suited for this condition. Such a cross-channel approach was incorporated into the previous SAD system, as mentioned in Section 1, but in a way that kept it separate from the speech activity modeling. An alternative method explored here is the use of cross-channel features that are appended to the baseline feature vector. In this way cross-channel phenomena such as crosstalk can be better modeled, improving local speech activity modeling and detection. The features examined are given below.

**Log-energy differences (LEDs)** The log-energy difference represents the log of the ratio of short-time energy between two channels, and is computed between a given target IHM channel and each of the non-target channels. As with the baseline features, the short-time energy is computed over a window of 25ms with an advance of 20ms. This is a variation of the feature described in [4] with the simplifying removal of the sigmoid, because the raw values were considered more informative.

**Normalized log-energy differences (NLEDs)** In some cases differencing of the raw log-energy values may be suboptimal because of significant differences in microphone gains. To compensate for this, the normalization scheme described in [3] was adopted as a step prior to the energy differencing. This normalization consists of subtracting the minimum frame log-energy of a channel from all log-energy values in the channel. That is, for a channel $i$ at frame $n$

$$E_{norm}(n) = E_i(n) - E_{min,i} \quad (1)$$

where $E$ represents log-energy. This minimum frame log-energy serves as a noise floor estimate for the channel and has the advantage of being largely independent of the amount of speech activity in the channel.

**Normalized maximum cross-correlation (NMXC)** A more common cross-channel feature found in the literature [5, 6] is one based on short-time cross-correlation maxima between channels. The correlation between the channels serves as an indicator of crosstalk. We define the normalized maximum cross-correlation between a target channel $i$ and nontarget channel $j$ to be

$$\Gamma_{ij} = \frac{\max_\tau \phi_{ij}(\tau)}{\phi_{jj}(0)} \quad (2)$$

where $\phi_{ij}(\tau)$ represents the cross-correlation at lag $\tau$ and $\phi_{jj}(0)$ is the nontarget channel autocorrelation for lag 0 (i.e., its short-time energy). Cross-correlation and autocorrelation values are computed over a context window of 25ms using a Hamming window function with an advance of 20ms.

A key consideration in using cross-channel features is the potentially variable number of channels to be processed versus the requirement of a fixed feature vector size for the HMM-based segmenter. The solution adopted for this work was to use order statistics—specifically maximum and minimum—of the feature values generated on the different channels, as was done by Wrigley et al. in [5].

## 5. Development Experiments

Two development test sets were chosen for initial experiments to evaluate the performance of the cross-channel features described above, and to determine which methods to include in the final SAD system.

### 5.1. Results on AMI development set

The AMI development set consists of meetings contributed by the AMI program for the NIST RT-05S meeting recognition evaluation. These are scenario-based meetings, elicited as described in

[7], each involving four participants wearing headset microphones or head-mounted lapel microphones.

Because the meetings all contain the same number of channels, it is possible to create a feature vector of fixed length using values from all channels, rather than by using the maximum and minimum values only. This experiment was performed to determine the effect of the length standardization procedure.

For training of the HMM-based segmenter, the first 10 minutes from 35 of these meetings were utilized. Testing was performed on 12-minute excerpts from four additional meetings.

Table 1: *Performance comparisons for systems using AMI development data. Results obtained using "fast" ASR system.*

| System | Del | Subs | Ins | WER |
|---|---|---|---|---|
| baseline | 17.4 | 13.0 | 7.4 | 37.8 |
| base + LEDs (all) | 17.2 | 13.0 | 4.5 | 34.8 |
| base + LEDs (max & min) | 17.4 | 12.8 | 4.5 | 34.7 |
| base + NLEDs (max & min) | 17.1 | 12.0 | 4.4 | 33.5 |
| base + NMXC (all) | 17.2 | 12.8 | 4.3 | 34.3 |
| base + NMXC (max & min) | 17.4 | 12.1 | 4.5 | 34.1 |
| reference | 18.3 | 10.2 | 3.4 | 32.0 |

The results for the various systems are given in Table 1. "Reference" refers to a segmentation derived from the time marks in the reference for word error scoring. As these were preliminary experiments, the fast version of the ASR system was used. From these results we see that the systems with cross-channel features all represent a significant performance improvement from the baseline, and that this is largely due to the reduction of insertion errors. This suggests that these cross-channel features are indeed useful in distinguishing crosstalk from local speech, as crosstalk is a key source of insertion errors for the IHM condition.

Also of note is that using max and min feature values yields performance similar to using all cross-channel values. The tentative conclusion is that max and min are good representative values for the purposes of SAD, although one additional value was omitted. It should also help that the min and max features impose a consistent rank ordering on the available cross-channel values. Unfortunately, no substantial data sets are available to test these effects on a much larger number of channels.

A third observation is that the energy normalization technique produces about a 1% absolute improvement over the unnormalized case, thus establishing its effectiveness. In addition, these normalized log-energy difference features appear to be slightly better than the commonly used cross-correlation based features for this data set.

### 5.2. Results on RT-04S evaluation set

Having established the effectiveness of the features, we subsequently evaluated the cross-channel feature systems on the RT-04S evaluation set, this time using the full ASR system. This test set consists of 11-minute excerpts of meetings provided from each of the sources CMU, LDC, ICSI, and NIST. Each site contributed two meetings for a total of eight meetings. The meetings vary in style, number of participants, and room acoustics, potentially presenting a greater challenge than the AMI set. For this set the segmenter was trained using the first 10 minutes from each of 15 NIST meetings and 73 ICSI meetings.

Table 2 gives results on the RT-04S test set for the baseline HMM segmenter, the old intersection segmentation system used in [2] and briefly described in Section 1 (denoted by 'intersection'), various cross-channel feature systems, and the reference segmentation. Also note that the cross-channel features use only max and min values because of the variable number of speakers.

Table 2: *Performance comparisons for systems using RT-04S evaluation data. Results obtained using "full" ASR system.*

| System | WER | | | | |
|---|---|---|---|---|---|
| | ALL | CMU | ICSI | NIST | LDC |
| baseline | 29.6 | 33.1 | 23.4 | 20.0 | 38.7 |
| intersection | 27.9 | 32.5 | 21.4 | 20.2 | 34.9 |
| base + LEDs | 27.3 | 32.8 | 20.1 | 20.0 | 33.7 |
| base + NLEDs | 26.9 | 32.8 | 18.5 | 19.6 | 34.0 |
| base + NMXC | 28.1 | 31.7 | 24.9 | 19.0 | 33.8 |
| reference | 25.1 | 30.3 | 18.0 | 17.0 | 31.9 |

As with the AMI development set, these results reveal improved performance over the baseline system for the cross-channel feature systems, further confirming the effectiveness of these features. The cross-channel feature systems also represent an improvement over the intersection system, supporting the initial hypothesis that the disjoint cross-channel analysis and speech activity modeling was a suboptimal approach.

A comparison of the normalized cross-correlation and the normalized log-energy difference features produces somewhat different observations for this data set. For two of the four sources (CMU and NIST), the NMXC features produce substantially lower word error rates than the NLED ones. For the ICSI meetings, however, the WER with NMXC is *much* higher—about 30% relative. Further investigation reveals that the contributing factor for the higher WER is almost exclusively insertion errors (0.9 for the NLED features and 9.1 for the NMXC features), which suggests a poorer handling of crosstalk. This leads to a poorer overall performance for the NMXC features system (28.1% versus 26.9%) and indicates that the NLED features tend to be more robust than the NMXC ones. As a result, the NMXC features were removed from consideration for the final SAD system.

## 6. Final System Validation

The NIST RT-05S meeting recognition evaluation was selected as a test set for performing validation on the finalized system—the HMM-based segmenter with the baseline and NLED features. The test data is composed of 12-minute excerpts from 10 meetings. The meetings were contributed by five sites with two meetings per site: AMI, CMU, ICSI, NIST, and Virginia Tech (VT). Being drawn from a pool similar to the RT-04S data, these meetings also possess significant variation in style, number of participants, and room acoustics.

The segmenter was trained using the union of the AMI, ICSI, and NIST training meetings described earlier (see Section 5.1 and 5.2). We explored two options to train the segmenter: either to pool all training data to train a single model, or to train an AMI-only S/NS model for use on AMI test data, and a separate ICSI+NIST model for use on all other test meetings. Using results on separate development data to make the decision, we chose the two-model approach for the baseline and intersection methods, and the single-pooled-model approach for the new cross-channel features.

Table 3: *Performance comparisons for systems using RT-05S evaluation data. Results obtained using "full" ASR system.*

| System | WER | | | | | |
|---|---|---|---|---|---|---|
| | ALL | AMI | CMU | ICSI | NIST | VT |
| baseline | 29.3 | 22.1 | 23.3 | 20.5 | 45.8 | 35.8 |
| intersection | 25.9 | 23.3 | 23.3 | 24.5 | 34.5 | 23.6 |
| base + LEDs | 25.6 | 22.0 | 23.5 | 20.9 | 37.3 | 23.8 |
| + SDM | 24.7 | | | | 33.0 | |
| base + NLEDs | 23.9 | 21.9 | 23.1 | 20.6 | 30.9 | 22.9 |
| + SDM | 22.7 | | | | 25.2 | |
| reference | 19.5 | 19.2 | 19.9 | 16.8 | 21.4 | 20.6 |

Table 3 presents recognition performance results for the segmentation from the two log-energy difference systems (i.e., unnormalized and normalized) along with the key contrastive ones: the baseline segmenter, the old intersection system, and the reference segmentation. With regard to performance of the cross-channel features, the same trend can be seen as in the development experiments. The cross-channel system with NLEDs gives about an 18% relative WER reduction over the baseline and about an 8% relative reduction over the intersection approach.

One of the NIST meetings in this test set presented an unusual setup. The meeting had a speaker participating via a speakerphone, and, consequently, without corresponding IHM channel. In addition, two of the participants were silent during the entire meeting. This led to an inordinate amount of insertion errors triggered by crosstalk, as reflected in the very high baseline error rate in the NIST column. To better cope with unmiked speakers we experimented with a variant of our algorithm that included a single, centrally located, omnidirectional distant microphone (SDM) channel in the cross-channel feature computation. The intent was for this SDM to serve as a stand-in for any speakers without IHM.[1] The corresponding results are given in the table in the rows marked "+ SDM". As can be seen, the SDM approach worked very well for dealing with the particular situation of the NIST meeting, especially when coupled with the NLED features. The difference between that system and the reference channel on the NIST meetings was now comparable to that for the other meeting sources.

## 7. Discussion

The initial motivation for this work was the improvement of the ICSI-SRI ASR system for the IHM condition of the NIST meeting recognition evaluation, specifically the speech activity detection. The results presented here demonstrate that, to this end, we made substantial progress. Significant WER reductions (of up to 18% relative) were achieved using the log-energy difference features we have described. In the process, the technique of cross-channel modeling using these features was validated. Particularly notable is the extent to which performance gains can be made using such relatively simple features. Along these lines, the robustness of these features (as compared to the cross-correlation-based features) is also of note.

Last, there still exists a performance gap of about 2-3% absolute between our best automatic system and the reference segmentation. This suggests the possibility of further improvements. One way to achieve such improvements might be the inclusion of other cross-channel, as well as single-channel features (e.g., as in [5]).

[1] Recall again that in the IHM condition, the recognizer is not supposed to recognize speech spoken by speaker without personal microphones.

## 8. Conclusions

We have detailed the development of a speech activity detection system using an HMM-based segmenter with single-channel (cepstra, log-energy, and derivatives) and cross-channel (log-energy differences) features. Results show that the inclusion of these simple cross-channel features yields large reductions in ASR word error rate performance over both the case of no cross-channel analysis and that of cross-channel analysis independent of speech activity modeling. In addition, the benefit of the simple normalizing technique of minimum energy subtraction was demonstrated. Finally, the log-energy difference features were shown to exhibit greater robustness than the more prevalent cross-correlation-based features. The inclusion of a distant omnidirectional microphone in the cross-channel feature computation allows the suppression of crosstalk even from speakers without dedicated microphones.

## 9. Acknowledgments

This work was partly supported by the Swiss National Science Foundation through the research network IM2. This material is also based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DARPA or the Department of Interior-National Business Center (DOI-NBC).

## 10. References

[1] L.R. Rabiner and M.R. Sambu, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. on Acoustics, Speech, and Sig. Proc.*, vol. 25, no. 4, pp. 338–343, 1977.

[2] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, 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 *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, Steve Renals and Samy Bengio, Eds. 2005, vol. 3869 of *Lecture Notes in Computer Science*, pp. 463–475, Springer.

[3] T. Pfau, D.P.W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. IEEE ASRU Workshop*, 2001, pp. 107–110.

[4] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," in *Proc. IEEE ASRU Workshop*, 2003, pp. 333–338.

[5] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multi-channel audio," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[6] K. Laskowski, Q. Jin, and T. Schultz, "Crosscorrelation-based multispeaker speech activity detection," in *Proc. Interspeech 2004 - ICSLP*, 2004, Jeju Isand; Korea.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meetings corpus," in *Proc. of the Measuring Behavior 2005 Symposium on "Annotating and Measuring Meeting Behavior"*, 2005, AMI-108.