Crosscorrelation-based Multispeaker Speech Activity Detection

Kornel Laskowski, Qin Jin, and Tanja Schultz

Interactive Systems Laboratories
Carnegie Mellon University, USA
{kornel|qjin|tanja}@cs.cmu.edu

Abstract

We propose an algorithm for segmenting multispeaker meeting audio, recorded with personal channel microphones, into speech and non-speech intervals for each microphone's wearer. An algorithm of this type turns out to be necessary prior to subsequent audio processing because, in spite of close-talking microphones, the channels exhibit a high degree of crosstalk due to unbalanced calibration and small inter-speaker distance. The proposed algorithm is based on the short-time crosscorrelation of all channel pairs. It requires no prior training and executes in one fifth real time on modern architectures. Using meeting audio collected at several sites, we present error rates for the segmentation task which do not appear correlated with microphone type or number of speakers. We also present the resulting improvement in speech recognition accuracy when segmentation is provided by this algorithm.

1. Introduction

The study of multispeaker meeting audio has recently seen a surge of activity at many levels of speech processing, as exemplified by the appearance of large meeting speech corpora from several groups and the ground-breaking evaluation paradigm launched by NIST, the Rich Transcription Evaluation on Meetings.

In this context, simultaneous speech/silence detection for all speakers becomes a functional prerequisite for subsequent analysis. In particular, speaker adaptation techniques for speech recognition call for clean, single-speaker audio segments.

This paper focuses on the automatic speech/silence segmentation of natural, multi-speaker data on each personal microphone channel. Microphones are of either headset or lapel type. Unexpectedly, even with close-talking microphones, due to unbalanced calibration and small inter-speaker distance, each participant's personal microphone picks up significant levels of voices from the other participants, making independent energy thresholding an unviable approach. The presence of extraneous speech activity in a given personal channel leads to a high word error rate due in large part to faulty insertion. Furthermore, portable microphones are subject to low frequency noise such as breathing and speaker (head) motion.

Work described in this paper contributes to the overall effort at the Interactive Systems Labs in the NIST Rich Transcription 2004 Spring Meeting Recognition Evaluation (RT-04S) [1]. There is growing interest in the meeting recognition task, and many important observations are available in the literature [2], [3]. To our knowledge, the only work which specifically addresses the simultaneous multispeaker segmentation problem is [4] at ICSI. While our conclusions are very similar to those in the ICSI study, the algorithm we propose is architecturally simpler. Specifically, it does not employ acoustic models for speech and non-speech states and thus requires no prior training.

Table 1: *Development dataset*

MeetingID	#Speakers	Mic Type
CMU_20020319-1400	6	lapel
CMU_20020320-1500	4	lapel
ICSI_20010208-1430	7	headset
ICSI_20010322-1450	7	headset
LDC_20011116-1400	3	lapel
LDC_20011116-1500	3	lapel
NIST_20020214-1148	6	headset
NIST_20020305-1007	7	headset

The remainder of this paper is organized as follows. In section 2 we briefly describe the data we used for the evaluation of our algorithm. In section 3 we outline several variants of our proposal, beginning with a baseline system which relies on energy thresholding alone. Section 4 presents our experimental results. Conclusions follow in section 5.

2. Data

All experiments throughout this paper were conducted on the RT-04S meeting data. Each meeting was recorded with personal microphones for each participant (a mix of headset and lapel microphones). The algorithm we propose does not require knowledge of the microphone type.

Both the development and the evaluation datasets from the NIST RT-04S evaluation were used. The data were collected at four different sites, including CMU [5], ICSI [6], LDC [7], and NIST [8]. The development dataset consists of 8 meetings, two per site. Ten minute excerpts of each meeting were transcribed. The evaluation dataset also consists of 8 meetings, two per site. Eleven minute excerpts of each meeting were selected for testing. All of the acoustic data used in this work is of 16kHz, 16-bit quality. Table 1 gives a detailed description of the RT-04S development dataset, on which we report detailed segmentation performance and speech recognition performance numbers. We also present final speech recognition performance on the RT-04S evaluation dataset.

3. Algorithms

3.1. Conceptual Framework

The audio for a single meeting consists of N time-aligned mono channels, where N is the number of speakers.

The response at microphone M_i, $y_i[n]$, is a combination of signals $x_j[n]$ from every acoustic source S_j in the room, both

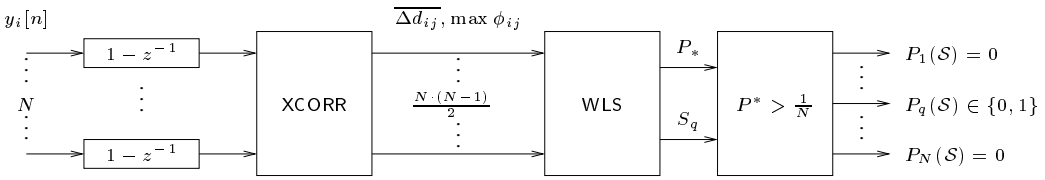

Figure 1: Architectural depiction of the IMTD algorithm

delayed and attenuated. We restrict our attention to exactly N possible sources, namely the vocal apparata of the N speakers wearing the microphones; we ignore the existence of other potential sound sources which we group at each microphone into a white noise term η_i. Furthermore we assume that the mouth-to-microphone distance for each speaker is negligible compared to the minimum inter-microphone distance; ie. $M_i \approx S_i$. This assumption is patently false but it allows for a simplified analysis involving the relative positions of only N points in a two-dimensional plane.

Each $x_j[n]$ is delayed and attenuated as a function of the distance d_{ij} between its source S_j and microphone M_i. The delay Δn_{ij}, measured in samples, is linearly proportional to the distance,

$$\Delta n_{ij} = \frac{f_s \, d_{ij}}{c} \tag{1}$$

where f_s is the sampling frequency and c is the speed of sound. For simplicity, we assume that $y_i[n]$ is a linear combination

$$y_i[n] = \sum_{j=1}^{N} \alpha_{ij} x_j[n - \Delta n_{ij}] + \eta_i \tag{2}$$

where η_i is a noise term.

In the general case, all α_{ij} are positive, ie. all microphones pick up all speakers to some extent.

3.2. Baseline

As already mentioned, we began with a baseline which relies on energy thresholding on each personal microphone channel. The energy threshold is equal to the average of the 200 lowest energies multiplied by a factor of 2. Any frame that has energy beyond the threshold will be considered as the participant's speech in that channel. As we will show in the experimental results section, the baseline system yields surprisingly poor performance.

3.3. Inter-microphone Time Differences (IMTD)

In our first experiment, we consider the use of inter-microphone time differences much as humans use interaural time differences to lateralize sources of sound [9]. In contrast to a single interaural lag in the latter, the meeting scenario offers an ensemble of $N \cdot (N-1)/2$ lags given N microphones/speakers, whose magnitudes are governed by much larger distances than head diameter as well as arbitrary seating arrangement.

Consider the general case with exactly one person S_q speaking during the current analysis frame. Then for each pair of microphone signals $\{y_i[n], y_j[n]\}$, $i \neq j$, the short-time crosscorrelation

$$\phi_{ij}[\Delta n] = \sum_{n} y_i[n] \cdot y_j[n + \Delta n] \tag{3}$$

exhibits a distinct peak at a lag corresponding to the difference in distance $\Delta d_{ij}^{(q)} = d_{iq} - d_{jq}$.

Given N points, we can compute $N \cdot (N-1)/2 > N$ distance differences. If the noise term, η, is both small and white, then this overdetermined system of equations will nevertheless be consistent, that is, for any three microphones $\{y_i[n], y_j[n], y_k[n]\}$,

$$\Delta d_{ik}^{(q)} = \Delta d_{ij}^{(q)} + \Delta d_{jk}^{(q)} \tag{4}$$

This defines an implicit transformation into polar coordinates, with speakers arranged radially around a single sound source, and in particular their projection onto the radial direction, spaced apart by the corresponding distance differences. After placing the origin arbitrarily in this single dimension, we solve for the positions of the listeners' microphones relative to that origin using a weighted least squares approximation, with the normalized peak crosscorrelation as the weight. The magnitude of the approximation error E indicates the degree to which the system of $N \cdot (N-1)/2 > N$ distance difference equations is consistent, and therefore the degree to which the hypothesis that a single speaker is speaking holds. We posit the probability that a single speaker is speaking (in a somewhat ad hoc fashion) as

$$P_* = e^{-E/\sqrt{N}} \tag{5}$$

which we can threshold as desired. Furthermore, the microphone whose abscissa is smallest is hypothesised as being worn by the speaker.

In situations where multiple speakers are speaking, maxima in the crosscorrelation spectra will not in general lead to a consistent system of distance difference equations; therefore E will be high. Likewise, during pauses, maxima in the spectra will occur at random lags since the microphone signals will be uncorrelated under the assumptions of our framework; likewise in this case, E will tend to be high.

The three main functional blocks of this algorithm: computation of all crosscorrelations, weighed least squares approximation and probability thresholding, are shown in Figure 1. In addition, we apply preemphasis to all channel signals, using a simple IIR filter ($1 - z^{-1}$), to reduce their low frequency contribution. Microphone motion and breathing both exhibit significant activity at low frequencies, and this method leads to significant reduction in the miss rate due to these phenomena on channels other than the foreground speaker's.

3.4. Joint Maximum Crosscorrelation (JMXC)

In a second competing algorithm, depicted in Figure 2, we employ the peak magnitude of the crosscorrelation between microphone signals as opposed to the lag at which it occurs.

After locating the peak in the crosscorrelation spectrum $\max \phi_{ij}$ between two microphone signals $\{y_i[n], y_j[n]\}$, we

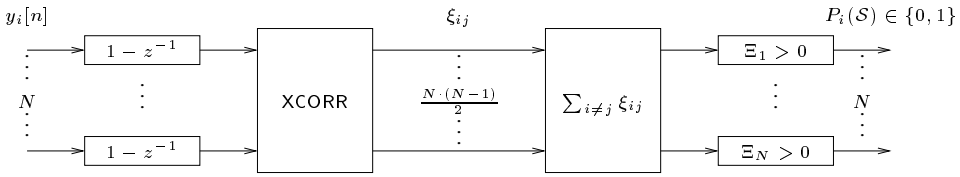

Figure 2: Architectural depiction of the JMXC algorithm

compute the quantity

$$\xi_{ij} = \log_{10} \frac{\max \phi_{ij}}{\phi_{jj}} \tag{6}$$

where the ϕ_{jj} is the power of $y_j[n]$ in the current analysis frame. ξ_{ij} attempts to measure to what extent speaker S_j is responsible for the peak cross-correlation $\max \phi_{ij}$, relative to speaker S_i. If speaker S_i is speaking and speaker S_j is silent, then ξ_{ij} will be positive, since $\max \phi_{ij}$ will be due to the power in $y_i[n]$, not the distant, attenuated copy $y_j[n]$. If both S_i and S_j are speaking, then their crosscorrelation spectrum will exhibit two peaks (symmetric about zero), but our search for a single peak will miss this bimodality and will only locate that which is higher. Under circumstances where the microphone gains are approximately equal, ξ_{ii} will be positive if S_i is the dominant speaker in the current analysis frame.

For every speaker S_i, we compute the sum

$$\Xi_i = \sum_{i \neq j} \xi_{ij} = \sum_{i \neq j} \log_{10} \frac{\max \phi_{ij}}{\phi_{jj}} \tag{7}$$

Per analysis frame, we hypothesize that S_i is speaking only if $\Xi_i > 0$. Otherwise, we assume that the power in $y_i[n]$ is due entirely to some other distant speaker(s) $S_{j \neq i}$, whose own microphone signal $y_j[n]$ contains more power.

3.5. Smoothing

The purpose of smoothing is to fill in the gaps between segments as we found that there is a high fraction of very short segments with short gaps between them. We perform merging in two steps. In the first step, we merge any two segments which have less than a 0.5s gap between them. Then we pad each segment with 0.5s at the start and end since it is hard to detect the exact beginning and ending points for each segment. In the second step, we check the new segments and merge any two segments which have less than a 0.3s gap between them. This two-step smoothing was found to give optimal segmentation accuracy in our experiments.

4. Experiments

In this section, we present our segmenation results and the speech recognition results based on segments provided by our algorithms. We use the miss rate (MS) and false alarm rate (FA) to measure segmentation performance. Given the hypothetical confusion matrix over segment durations for one channel M_i in Table 2, $\text{MS}_i = T_i^{(MS)}/(T_i^{(S)} + T_i^{(MS)})$ and $\text{FA}_i = T_i^{(FA)}/(T_i^{(S)} + T_i^{(FA)})$. Generally we seek systems which exhibit both a low miss rate and a low false alarm rate.

When reporting results for an entire meeting, we compute the overall miss rate

$$\text{MS} = \frac{\sum T_i^{(MS)}}{\sum T_i^{(S)} + \sum T_i^{(MS)}} \tag{8}$$

Table 2: *Hypothetical confusion matrix*

System Output	Reference	
	Speech	Non-speech
Speech	$T^{(S)}$	$T^{(FA)}$
Non-speech	$T^{(MS)}$	$T^{(N)}$

and the overall false alarm rate

$$\text{FA} = \frac{\sum T_i^{(FA)}}{\sum T_i^{(S)} + \sum T_i^{(FA)}} \tag{9}$$

The run-time performance for both algorithms is approximately 0.2 times real-time, as measured on a 2.8GHz Pentium 4 machine.

4.1. Segmentation Experiments

Segmentation results are shown in Table 3. As mentioned earlier, the performance of the baseline suffers from a high false alarm rate due to other speaker pickup. Our initial explorations were guided primarily by a desire to lower the false alarm rate.

IMTD with smoothing significantly reduces the false alarm rate, but at the expense of a large increase in the miss rate. This is due to the algorithm's inability to postulate simultaneous speakers, which is a frequent occurrence. In addition, meetings which exhibit very little channel crosstalk result in high errors because there are no clear peaks in the crosscorrelation.

JMXC significantly decreases both types of error relative to IMTD. This is due to its ability to postulate multiple speakers speaking simultaneosly. Also, the peak crosscorrelation value is a more robust feature than the sample lag at which it occurs.

Table 3: *Segmentation performance on devset meetings (in %)*

System	no smoothing		smoothing	
	MS	FA	MS	FA
baseline	7.2	66.2	—	—
IMTD	54.8	23.8	38.0	30.6
JMXC	33.2	4.2	**16.9**	**13.0**

In Table 4, we show the performance of the JMXC system on individual meetings. This data exhibits large variability, which appears uncorrelated with the microphone type and number of speakers. We think that this variability may be due to unquantified meeting characteristics such as overall degree of crosstalk, general meeting geometry including room acoustics, mean and standard deviation of signal-to-noise ratios and/or microphone variability within a meeting.

Table 4: *Individual JMXC segmentation performance (in %)*

Meeting ID	no smoothing		smoothing	
	MS	FA	MS	FA
CMU_20020319-1400	41.9	2.2	19.8	13.5
CMU_20020320-1500	28.8	5.7	11.8	17.4
ICSI_20010208-1430	22.3	4.8	11.1	16.1
ICSI_20010322-1450	22.1	8.7	9.0	17.2
LDC_20011116-1400	18.9	3.5	8.8	8.8
LDC_20011116-1500	36.1	3.1	23.1	13.3
NIST_20020214-1148	45.0	0.9	22.5	7.5
NIST_20020305-1007	47.0	3.2	25.5	9.1

We have tabulated the segmentation performance separately for lapel and headset microphone meetings in Table 5. The numbers suggest that the difference in performance is negligible if at all significant.

Table 5: *JMXC segmentation performance per mic type (in %)*

Meeting ID	no smoothing		smoothing	
	MS	FA	MS	FA
lapel	32.0	3.5	16.5	13.1
headset	34.4	4.9	17.2	12.9

We note that both of the explored algorithms actually perform non-silence detection; this includes speech as well as non-verbal sounds such as laughter. Other sources may also be picked up provided their acoustic distance to one microphone is much smaller than to any of the others. We expect that to some degree, non-verbal phenomena coming from the speaker may appear in the transcription and be useful to subsequent components of a meeting transcription system.

4.2. Application to Speech Recognition

Table 6 compares the first pass speech recognition performance based on different segmentation systems with the "ideal" segmentation using human labels. We also compute the performance gap in word error rate relative to the ideal.

Table 6: *Speech recognition performance on the RT-04S devset.*

System	Word Error Rate	Performance Gap
baseline	49.6%	25.3%
IMTD	68.6%	73.2%
JMXC	**43.6%**	**10.1%**
human	39.6%	—

JMXC was used to provide segmentation under the Individual Headset Microphone (IHM) condition for the ISL speech recognizer [10] in the NIST RT-04s evaluation. This system produced a 35.7% word error rate on the evaluation set in the final pass; refer to [10] for details.

5. Conclusions

We have presented a simple, fast algorithm, which requires no prior training, for detecting speech vs non-speech in multi-speaker meeting data. The experiments performed show that the algorithm is capable of providing useful segmentation of personal microphone audio in the presence of crosstalk on a wide range of meetings. It significantly improves the quality of audio usable for speaker adaptation in speech recognition; our results show only minor increase in word error rates relative to manually prepared segmentations. This algorithm was integrated into the ISL meeting transcription system used in the NIST RT-04S evaluation.

6. Acknowledgements

We would like to thank Christian Fügen and Florian Metze at Interactive Systems Labs in Karlsruhe for running the speech recognition experiments and for providing valuable feedback as to the performance of our segmentations during development.

7. References

[1] NIST, Rich Transcription 2004 Spring Meeting Recognition Evaluation, `http://www.itl.nist.gov/iad/894.01/tests/rt/rt2004/spring/`

[2] Burger, S., MacLaren, V., and Yu, H., "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style", Proc. ICSLP 2002, Denver, USA.

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

[4] Pfau, T., Ellis, D. P. W. and Stolcke, A., "Multispeaker Speech Activity Detection for the ICSI Meeting Recognizer", Proc. ASRU 2001, Madonna di Campiglio, Italy.

[5] Burger, S. and Sloane, Z. "The ISL Meeting Corpus: Categorical Features of Communicative Group Interactions", Proc. ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

[6] Janin, A., Ang, J., Bhagat, S., Dhillon, R., Edwards, J., Morgan, N., Peskin, B., Shriberg, E., Stolcke, A., Wooters, C., and Wrede, B. "The ICSI Meeting Project: Resources and Research", Proc. ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

[7] Strassel, S., and Glenn, M., "Shared Linguistic Resources for Human Language Technology in the Meeting Domain", Proc. ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

[8] Stanford, V., and Garofolo, J., "Beyond Close-talk - Issues in Distant Speech Acquisition, Conditioning Classification, and Recognition", Proc. ICASSP 2004 Meeting Recognition Workshop, Montreal, Canada.

[9] Moore, B. C. J., An Introduction to the Psychology of Hearing, Academic Press, 1997.

[10] Metze, F., Jin, Q., Fügen, C., Laskowski, K., Pan, Y., and Schultz, T., Issues in Meeting Transcription — The ISL Meeting Transcription System, submitted to Proc. ICSLP 2004, Jeju Island, Korea.