Crosscorrelation-based Multispeaker Speech Activity Detection

KORNEL LASKOWSKI, QIN JIN AND TANJA SCHULTZ

Carnegie Mellon

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

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 Motivation

Automatic transcription of meetings and the study of multispeaker meeting audio has recently become very popular. A first processing step in almost all tasks in this field is utterance segmentation.

This problem is hard for two main reasons:

1. In uninstrumented rooms, speakers are recorded using personal microphones. Both lapel microphones and headset microphones exhibit a considerable amount of **crosstalk**, or backchanneling.
2. In unstructured meetings, multiple speakers may speak simultaneously, leading to a considerable amount of **speech overlap**.

Addressing both problems successfully is therefore a functional prerequisite for subsequent analysis. In particular, speaker adaptation techniques for speech recognition call for clean, single-speaker audio segments.

Additional problems arising during segmentation are due to unbalanced calibration of all microphone channels, possibly changing meeting topology, and speaker breathing and (head) motion, resulting in low frequency noise.

Work described here contributes to the overall effort at the Interactive Systems Labs in the NIST Rich Transcription 2004 Spring Meeting Recognition Evaluation (RT-04S).

2 Data

All experiments reported here 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), as well as room microphones placed on the conference table. In this work we focus on the task of automatic segmentation of all personal microphone channels, that is, the discovery of portions where a participant is speaking in his/her personal microphone channel. The algorithm we propose does not require knowledge of the microphone type.

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 delayed and attenuated. 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

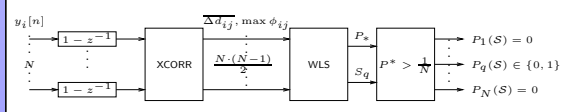
$$y_i[n] = \sum_{j=1}^{N} \alpha_{ij} x_j[n - \Delta n_{ij}] + \eta_i \tag{2}$$

where η_i is a noise term.

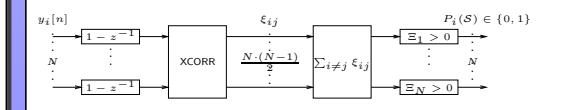
In the general case, all α_{ij} are positive, ie. all microphones pick up all speakers to some extent.

3.2 Baseline

A strightforward approach to this problem is to use **energy thresholding**, independently on each personal channel. We implement this in our baseline. 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 very poor performance.



IMTD segmentation



JMXC segmentation

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. 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.

If only 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 be (almost) 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 (we assume radially symmetric microphone responses). 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.

The microphone whose abscissa is smallest is hypothesized as being worn the speaker.

3.4 Joint Maximum Crosscorrelation

In a second algorithm, we consider the peak magnitude of the crosscorrelation.

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

$$\xi_{ij} = \log_{10} \frac{\max \phi_{ij}}{\phi_{jj}} \tag{5}$$

where the ϕ_{jj} is the power of $y_j[n]$ in the current analysis frame.

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 $y_j[n]$.

For every speaker S_i, we compute the sum

$$\Xi_i = \sum_{i \neq j} \xi_{ij} = \sum_{i \neq j} \log_{10} \frac{\max \phi_{ij}}{\phi_{jj}} \tag{6}$$

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 microphone signal $y_j[n]$ contains more power.

4 Experimental Results

4.1 Segmentation

In the following table we show the segmentation results on the development set, using miss rate (MS) and false alarm rate (FA), for all three algorithms, with and without smoothing.

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**

We also split the data into lapel microphone and headset microphone channels, and show JMXC segmentation performance separately for both below:

Mic Type	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

These results demonstrate that for the RT04s data, the performance of the JMXC algorithm is relatively independent of microphone type.

4.2 Speech Recognition

In the following table we compare the first pass speech recognition performance on segments produced by the different segmentation systems. We also compute the performance gap in word error rate relative to the ideal.

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

5 Discussion and Conclusions

The **baseline** algorithm suffers from a high false alarm rate due to crosstalk and speech overlap.

The **IMTD** algorithm reduces the false alarm rate, but at the expense of an increase in the miss rate. This is due to its inability to postulate simulataneous speakers. Also, where there is very little crosstalk, the algorithm suffers because there are no clear peaks in the crosscorrelation.

JMXC significantly reduces both types of error. In contrast to IMTD, it can posit simultaneous speakers; furthermore, the peak crosscorrelation value is a more robust feature than the sample lag at which it occurs.

Recognition accuracy using the JMXC algorithm for segmentation is fairly close to that obtained using human segmentation, for a wide variety of meetings and different microphone types.