Non-parametric Speaker Turn Segmentation of Meeting Data

Petr Motlíček, Lukáš Burget, Jan Černocký

Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66 CZ
{motlicek,burget,cernocky}@fit.vutbr.cz

Abstract

An extension of conventional speaker segmentation framework is presented for a scenario in which a number of microphones record the activity of speakers present at a meeting (one microphone per speaker). Although each microphone can receive speech from both the participant wearing the microphone (local speech) and other participants (cross-talk), the recorded audio can be broadly classified in three ways: local speech, cross-talk, and silence. This paper proposes a technique which takes into account cross-correlations, values of its maxima, and energy differences as features to identify and segment speaker turns. In particular, we have used classical cross-correlation functions, time smoothing and in part temporal constraints to sharpen and disambiguate timing differences between microphone channels that may be dominated by noise and reverberation. Experimental results show that proposed technique can be successively used for speaker segmentation of data collected from a number of different setups.

1. Introduction

Meetings are a fundamental activity, in which speech provides sharing and developing information between a group of people. Due to this fact, meetings present an important application domain for speech processing technologies. In recent years, the study of multispeaker meeting audio has become very intensive at many levels of speech processing [1], [3], as exemplified by the appearance of large speech meeting corpora from several groups.

Generally, speaker segmentation and clustering consists of identifying who is speaking and when, in a long meeting conversation. In ideal case, a speaker turn segmentation and clustering system will discover how many people are involved in the meeting and output clusters corresponding to each speaker.

This paper presents results achieved on challenging data and takes advantage of AMI (Augmented Multiparty Interaction) project [2]. One of the tasks is to record data, comprising real meetings, with a range of equipment and configurations. Such data is going to be orthographically transcribed. The challenging problem is its automatic segmentation according to a speaker turns as well as Voice Activity Detection (VAD). Furthermore, as it was identified in [3], around 10-15% of words in meeting or telephone conversation contain some degree of overlapping speech. These overlapped speech segments are problematic for speech recognition. Patterns of speaker activity can provide valuable information regarding the structure of the meeting.

Meetings take place in a room equipped with multimodal sensors. Each meeting recording, we have used in our experiments, is composed of at least four speakers sitting around a table in a room. Audio information is acquired from lapel mounted microphones. Since, in all recorded meetings participants do not move, single sources stay most of the time on the same place.

The paper describes a system for automatic speaker segmentation and clustering of meetings based on multiple close-talking microphones. It might seem that for these personal microphones the task will degrade to a speech/silence detection. However, even for close-talking microphones, due to unbalanced calibration and small inter-speaker distances, each participant's personal microphone is highly influenced by activity from the other participants. Therefore, simple independent energy thresholding would lead to an unviable approach. The presence of extraneous speech activity in a given personal channel causes an increase of word error-rate mainly due to faulty insertions. Furthermore, portable microphones are subject to low frequency noise such as breathing and speaker motion.

Our proposed algorithm is based on the cross-correlation of all channel pairs. However, meeting data used for testing vary a lot especially in the level of cross-talk, which is largely caused by properties of microphones used for recordings (headsets, lapels, . . .). In order to have the real applicable and robust speaker segmentation algorithm, we also exploit the energy-level information as well as values of maxima of the cross-correlations. This has shown to be helpful especially when segmenting data from headset microphones.

Comparable works have primarily concentrated on speech/silence identification task [4], or on classification of cross-talk [5]. Also most of the previous work on speaker segmentation has focused on a single channel data such as mono recordings of broadcast audio, and has mainly relied on changes in the statistical properties of the speech spectrum to detect change points. These works belong to the parametric approaches, where the data is modeled usually by GMMs and then optimized using Maximum-Likelihood criterion [6], [7]. However, due to maximum applicability, we focus on an algorithm that makes as few assumptions as possible about the nature of the multiple audio channels, without any manual annotation. We merely assume that the sound field has been sampled at several different points, resulting in several acquired recordings. Then, we take into account only differences between signals received by each sensor. Practically, the points of microphone locations in the field are fixed, thus we do not employ any information about their actual locations. The speaker segmentation system has access to several of these synchronized recordings, but has no control over, or information about, where the sensors are placed.

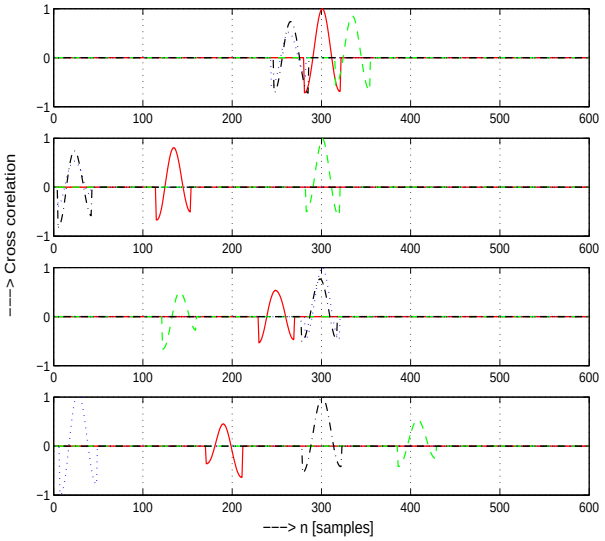

Figure 1: CCFs for particular channels; short rectangular window is applied around the global maximum of each CCF. Each panel plots CCFs of one channel with all others. Obviously, one CCF which represents auto-correlation sequence has always global maximum in the center of the plot.

2. Speaker turn segmentation

Our speaker turn segmentation system is based on timing difference cues. All possible comparisons among two microphone channels are made to obtain time-difference vectors. Our recently developed system is able to process four microphone channels, but the approach can be easily extended for arbitrary number of channels.

The different distances between each microphone and a particular speaker result in time differences that can be simply estimated by finding the *lag* l that maximizes short-time cross-correlation functions (CCFs) Θ:

$$\Theta(l,i,j) = \sum_{n=-N/2}^{N/2} s(n,i)s(n+l,j), \qquad (1)$$

where $s(n,i)$ is a sample sequence from microphone i. l can theoretically vary in interval $l \in (-\infty,\infty)$. However, CCFs are computed over an N-point window. The window must be sufficiently long so that time differences between particular microphones will become evident. In our experiments, where input data used was recorded at 16kHz, $N = 300$. Therefore, the length of short-time speech segments is approximately equal to 19ms, which is sufficient to capture sounds from sources up to 6 meters distant from the receivers. In order to eliminate the influence of decaying sides of Θ and to simply increase the length of speech segments, $s(n,i)$ sequence is padded on both right and left sides to create 620 samples long sequence. Then, 321 long resulting CCF is returned. Estimation of Θ is executed 100 times per second. The chosen lag for each microphone pair $\{i,j\}$ at time m is:

$$l(m,i,j) = arg\max_{l}\ \Theta(l,m,i,j). \qquad (2)$$

For the case of four microphone channels, we compute short-time cross-correlations of each channel with the other three signal sources, as well as the auto-correlation of this channel. Finally, the highest value of lag is found and the corresponding channel position is returned:

$$k(m,i) = arg\max_{j}\ l(m,i,j),\ 0 \le j \le 3. \qquad (3)$$

k represents the hypothesis of the source localized channel.

Since the algorithm is repeated for each source channel, it should eliminate possible noise integrated into one channel. Therefore, it results in four hypothesis of active channel. Examples of CCFs for particular channels are given in Fig. 1.

2.1. Time-smoothing of cross-correlations

The use of raw short-time CCFs for searching lags carrying information about localized source is practically impossible. Some other pre-processing operations need to be done in order to improve correct estimation of lags due to corrupting source channels mostly by additive noise. In our experiments, it has shown to be useful to smooth matrix of CCFs within some temporal window defined by length W_1, where rows of the matrix are CCFs computed for each 10ms time-step. The matrix captures temporal-evolution of CCFs for each i,j, and is graphically given in Fig. 3.

The smoothing operation is done independently for each i, j (independently also for each column of the matrix), thus, we need to have $i \times j$ buffers of length W_1 to store time-evolutions of all CCFs. Mathematically, smoothed CCF at time m is obtained:

$$\tilde{\Theta}(l,m,i,j) = \frac{\sum_{n=m-W_1/2}^{m+W_1/2}\Theta(l,n,i,j)}{W_1+1}, \forall\ l,i,j. \qquad (4)$$

For 10ms sampling rate of input speech segments, we use $W_1 = 100$.

2.2. Temporal constraints

Additional pre-processing block able to improve source localization hypothesis is the definition of some temporal constraints to remove outliers from the sequence of time estimates. We search for "optimal" path through multidimensional (time – cross-correlation) space. For each CCF at time m, all local maxima bigger than $0.8 \times l_M(m)$ are detected, where $l_M(m)$ is the global maximum observed at the CCF. Lags corresponding to these local maxima $l_x(m)$ are stored, where $x = 1,2,3,\ldots$. This process is executed for all CCFs across some temporal window W_2. Finally, the most appropriate lag $l_y(m,i,j)$ is found, where:

$$y = arg\min_{x}\left(\left|\frac{\sum_{n=m-W_2/2}^{m+W_2/2} l_M(n,i,j)}{W_2+1} - l_x(m,i,j)\right|\right) \qquad (5)$$

Obviously, the algorithm is applied independently for different combinations i,j. The length of temporal window is $W_2 = 30$.

Time-evolution scheme of CCFs over 5 seconds window is given in Fig. 3. Upper plot shows matrix of rough CCFs, whereas lower plot is obtained after application of time-smoothing operation. When additional temporal constraints are applied, the final path through the matrix is found (plotted as a bold line in Fig. 3). As we can see, temporal constraints would also be able to detect the silence parts in the speech. However, this information is not very accurate and so far not exploited in our experiments.

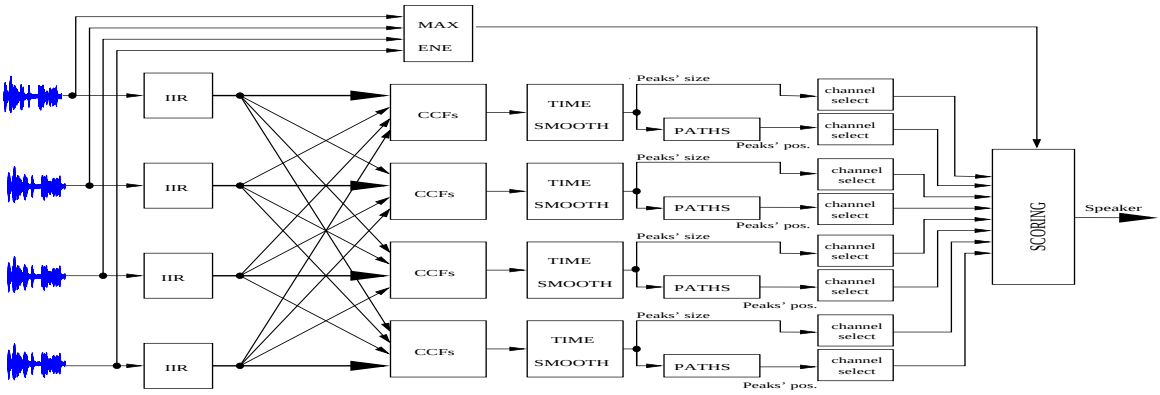

Figure 2: Speaker turn segmentation system.

2.3. Energy information

Another used hypothesis about speaker turns is derived from energy levels of the individual source channels. In our experiments, the source channels are globally normalized over time to have the same energy levels. Simply, a channel corresponding to the maximum energy is taken as the possible hypothesis.

Intuitively, the global maxima of CCFs carry information about the active channel. Therefore, value of each CCF corresponding to the l_y, from $\tilde{\Theta}(l_y, m, i, j)$, is found. Then, we do simple comparisons of these values for each source channel i, and the source channel corresponding to the highest value is detected:

$$k(m,i) = arg\max_{j} \tilde{\Theta}(l_y, m, i, j),\ 0 \leq j \leq 3. \qquad (6)$$

Such operation is repeated for all i. Result derived from majority voting of theses four hypothesis provide a last input to the scoring system.

2.4. Scoring

Majority voting based scoring approach is applied to decide about the speaker turns. The reliability of different hypothesis can vary for different meeting data. For example, the energy levels of source channels are much more reliable in case of low cross-talks between channels and vice-versa. Let us assume that four audio channels (each for one speaker) are available for each meeting. Then, inputs for the scoring system are four hypothesis estimated from time delays (no weights), the channel's position related to maximum energy level (weighted by W_e constant) and the channel's position derived from determined values $\tilde{\Theta}(l_y, m, i, j)$ (weighted by W_p constant).

CCF	ENE	PEAKS	T_SM	PATHS	FA
X					25%
X	X				38%
X	X	X			54%
X	X		X	X	71%
X	X	X	X		79%
X			X	X	82%
X	X	X	X	X	87%

Table 1: Overall results achieved on Brno meeting data ($W_e = W_p = 2$.)

The final decision is made for each speech segment (thus every 10ms) and is given by majority voting of the described weighted inputs.

3. Experimental data

The experiments throughout this paper were conducted on three meeting data sets, which are available within AMI project [2]. Sampling frequency of data is 16kHz:

- Brno meeting data [9] contains recordings from four lapel microphones. The recordings are heavily cross-talked due to physical properties of microphones. Approximately, over 2 hours of English-spoken meetings were used. Word-level orthographic transcriptions are available for this data.
- A sub-set of ICSI meeting data [8] were also used with total length almost 12 hours. Among others, four channels recorded by close-talking head-worn microphones were used. This data is lightly affected by cross-talk. Thus, we could not heavily rely on time differences between channels so that weights W_e and W_p are bigger. Since the data was force-aligned, speech/silence boundaries are available for each speaker channel.
- A first series of AMI-pilot meetings following the AMI Remote Control Design scenario have been made in the IDIAP room [10]. The total length of pilot meetings is almost 2 hours. Lapel microphones were used for speaker turn segmentation.

4. Experimental results

The best results were achieved when input signals were filtered by band pass filter (500Hz - 6kHz, 12 order Butterworth IIR filter), providing 60dB rejection outside the pass band.

Brno meeting data [9] was used as initial data set for evaluation of particular processing steps of the described speaker localization system which is graphically given in Fig. 2. To assess the system performance, we measured frame accuracies (FA) which take into account number of correctly labeled frames to the total number of frames (in percents). We did not evaluate speech segments belonging to the silence part or to the part with overlapping speakers.

The overall results for particular steps are given in Tab. 1. As can be seen, the use of rough time difference information based hypothesis returns low performance (*CCF column*). The

	TIME DELAYS	ENE	PEAKS	OVERALL
BRNO	82.74%	71.13%	80.15%	**86.72%**
ICSI	33.37%	94.03%	94.30%	**94.30%**
AMI	87.83%	78.45%	83.33%	**90.24%**

Table 2: FA when only one particular input is provided for speaker turn segmentation. Last column shows the final FA for all available inputs used in the scoring system. The best results were obtained for these constants: Brno data $W_e = W_p = 2$, ICSI data $W_e = W_p = 3$, AMI data $W_e = W_p = 2$.

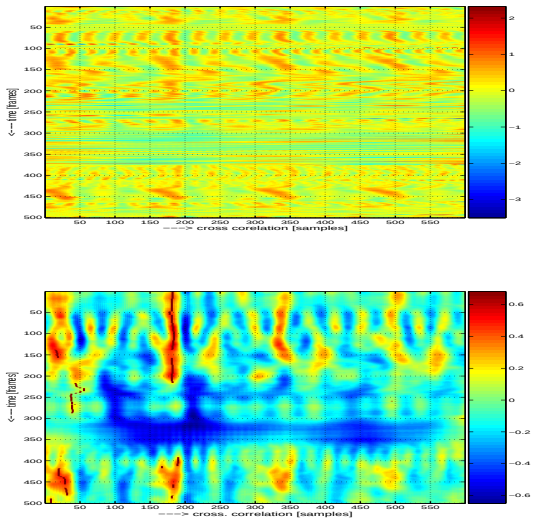

Figure 3: Time-evolution scheme of CCFs within window of 5 seconds: Upper panel – CCFs. Lower panel – time-smoothed CCFs with application of temporal constraints as described in Sect. 2.2.

time smoothing (*T_SM column*) of CCFs applied with temporal constraints (*PATHS column*) significantly improves achieved results. If speech segment energies (*ENE column*) and values of maxima of CCFs (*PEAKS column*) are taken into account, another substantial improvement is obtained.

Tab. 2 highlights particular inputs to the scoring system for all the data sets.

5. On-line speaker turn segmentation

For demonstration purposes, we have developed a speaker segmentation system that is able to detect speaker turns in real time. The system has been proposed together with acoustic based keyword spotter (KWS). Furthermore, on-line pre-processing of visual input from the camera, scanning the whole scene using the hyperbolic mirror has been used. Four channel audio input was recorded similarly to Brno meeting data.

Speaker turn segmentation together with KWS is capable of running on standard PC (1.4GHz Pentium, 256MB RAM) in real-time.

6. Conclusions

Achieved results show that the between-channel timing information revealed by CCFs brings sufficient information about speaker turns, especially in case of segmenting heavily crosstalked data. Obviously, this first effort leaves much room for additional improvement.

Achieved results on these two data sets show the different properties of the databases. The best segmentation of Brno and AMI-pilot meeting data is obtained from timing information captured by CCFs. Additional speaker turn hypothesis based on energy levels and maxima of CCFs further improves final decision. On the other hand, time differences do not help (confuse scoring system) when ICSI meeting data is segmented. Here, speaker-turn decision based on maxima of CCFs is the most reliable.

Recently, the presented speaker-turn segmentation system has been practically used for initial pre-processing of newly recorded meeting data. Due to, we also created robust VAD system based on classical MFCCs classified using Neural Network trained on ICSI training data set. Such rough speaker-turn and speech/silence segmentation are exploited by annotators to create word-level orthographic transcriptions of new AMI meeting data.

7. Acknowledgments

This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811, Grant Agency of Czech Republic under project No. 102/05/0278, and by CESNET project No. 119/2004. Jan Cernocky was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

8. References

[1] S. Burger, V. MacLaren, H. Yu. “The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style”, in proc. of ICSLP, Denver, USA, 2002.

[2] Official site of AMI project, “http://www.amiproject.org”.

[3] E. Shriberg, A. Stolcke, D. Baron. “Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation”, in proc. of Eurospeech, Aalborg, 2001.

[4] T. Pfau, D.P.W. Ellis, A. Stolcke. “Multispeaker speech activity detection for the ICSI meeting recorder”, in proc. of ASRU Workshop, Italy, December 2001.

[5] S. N. Wrigley, G. J. Brown, V. Wan, S. Renals. “Feature Selection for the Classification of Crosstalk in Multi-Channel Audio”, in proc. of Eurospeech, Geneva, 2003.

[6] J. Ajmera, H. Bourlard, I. Lapidot, I. McCowan. “Unknown-multiple speaker clustering using HMM”, in proc. of ICSLP, Denver, USA, 2002.

[7] T. Kemp, M. Schmidt, M. Westphal, A. Waibel. “Strategies for automatic segmentation of audio data”, in proc. of the IEEE ICASSP, Istanbul, Turkey, 2000.

[8] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters. “The ICSI Meeting Corpus”, in proc. of the IEEE ICASSP, Hong Kong, China, 2003.

[9] Brno meeting data, “http://www.fit.vutbr.cz/research/grants/ami/meetings.htm”.

[10] D. Moore. “The IDIAP smart meeting room”, IDIAP Communication 02-07, 2002.