ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Overlapped Speech Detection in Meeting Using Cross-Channel Spectral Subtraction and Spectrum Similarity

Ryo Yokoyama[1], *Yu Nasu*[1], *Koichi Shinoda*[1], *and Koji Iwano*[2]

[1]Department of Computer Science, Tokyo Institute of Technology, 152–8552, Japan
[2]Faculty of Environmental and Information Studies, Tokyo City University, 224–8551, Japan
{yokoyama, nasu}@ks.cs.titech.ac.jp, shinoda@cs.titech.ac.jp, iwano@tcu.ac.jp

10.21437/Interspeech.2012-424

Abstract

We propose an overlapped speech detection method for speech recognition and speaker diarization of meetings, where each speaker wears a lapel microphone. Two novel features are utilized as inputs for a GMM-based detector. One is speech power after cross-channel spectral subtraction which reduces the power from the other speakers. The other is an amplitude spectral cosine correlation coefficient which effectively extracts the correlation of spectral components in a rather quiet condition. We evaluated our method using a meeting speech corpus of four speakers. The accuracy of our proposed method, 74.1%, was significantly better than that of the conventional method, 67.0%, which uses raw speech power and power spectral Pearson's correlation coefficient.

Index Terms: overlap speech detection, spectral subtraction, cosine distance

1. Introduction

In recent years, meeting speech recognition [1, 2] and meeting speaker diarization [2, 3, 4] have been effectively utilized in real applications in order to transcribe and browse meeting procedures. However, their performance is usually low at the overlapped speech segments where more than one speaker is speaking. One possible solution for this problem is first to detect the overlapped speech segments, and then to ignore them in the following process or to apply special techniques such as source separation to recover the signal from each speaker. We focus on overlapped speech detection (OSD) in this paper.

Several recording devices such as boundary microphones and microphone arrays have been employed for meeting speech processing. Boundary microphones are easy to use and inexpensive, but it is difficult to separate the speech signal of one speaker from those of the others. Microphone arrays can separate the speech signals better but are expensive and need to be calibrated carefully. Here we assume that each meeting participant wears a microphone. In this study, we use a lapel microphone for collecting meeting speech data. While the use of lapel microphones takes a little effort from each participant, it enables the identification of each participant's speech signal with relatively low costs. We can also use a smart phone in one's breast pockets or that in front of him/her on the meeting table as a microphone.

Most conventional OSD methods successfully use a GMM-based classifier, which consists of a GMM for overlapped segments and that for non-overlapped segments, and set a threshold for their likelihood ratio. Then, the problem is what features we should use as its input. The power summed over all the frequency bands has proven to be effective (e. g., [5]). Large powers in more than one microphone indicate overlapping. Some studies [5, 6] focused on the effect of cross-talk. The signals from microphones tend to be similar with each other when only one speaker speaks, and to be different when more than one speaker is speaking. For example, Xiao *et al.* [5] reported that Pearson's correlation coefficient between the power spectra of two microphones (PPC) is an effective input feature.

However, both of these features, the overall power and PPC, have a serious problem. First, the overall power may be contaminated by speech signals from other speakers and detect overlapping segments incorrectly when only one speaker is speaking. Second, PPC is normalized by the mean of the power spectral components over all the frequency bands of nearby frames. This normalization process is indeed effective when there exists stationary environmental noise which should be subtracted. However, it may also normalize speech signals when only one speaker is speaking, and hence, may not show good performance.

In this paper, we propose two new features for OSD in meeting speech. One is a CCSS power, which is an overall power obtained by cross-channel spectral subtraction (CCSS) [1]. CCSS is a source separation method and has proved to be very effective at excluding the speech signals from other speakers in meeting speech using lapel microphones. The other is an amplitude spectral cosine correlation coefficient (ACC) which does not include a feature normalization process and hence remains large when only one speaker is speaking. We use these two features as in-

puts to a GMM detector and examine their effectiveness using a meeting speech data with four speakers.

This paper is organized as follows. The two features used in this paper, CCSS power and ACC, are explained in Section 2 and Section 3, respectively. The experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2. CCSS power

2.1. Cross-channel spectral subtraction (CCSS) [1]

In order to reduce the power from the other speakers, we introduce cross-channel spectral subtraction (CCSS), a source separation method based on spectral subtraction [7].

Let the number of speakers be N. Consider that one lapel microphone is prepared for each speaker. We use the same suffix for one speaker and his/her microphone. Then, assuming the speech signals from multiple speakers are linearly mixed and ignoring noise, the signal recorded by the i-th microphone (of the i-th speaker) can be modeled as:

$$X_i(f,t) = \sum_{j=1}^{N} G_{ij}(f,t) S_j(f,t), \tag{1}$$

where $S_j(f,t)$ is the speech in a frequency band f at time t of the j-th speaker and $G_{ij}(f,t)$ is the transfer function from the j-th speaker to the i-th microphone. The transfer functions are time-variable, since they may change when speakers move around, while they are regarded as stationary in most conventional studies.

The target signal is the j-th speaker's speech recorded by the j-th microphone for each j. By defining it as:

$$Y_j(f,t) = G_{jj}(f,t) S_j(f,t), \tag{2}$$

and substituting the transfer function by:

$$H_{ij}(f,t) = \frac{G_{ij}(f,t)}{G_{jj}(f,t)}, \tag{3}$$

the recorded signal can be written as:

$$X_i(f,t) = Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t). \tag{4}$$

Then, the power spectrum of the recorded signal is calculated as:

$$\begin{aligned} &|X_i(f,t)|^2 \\ &= \left| Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \right|^2 \\ &= |Y_i(f,t)|^2 + \sum_{j \neq i} |H_{ij}(f,t) Y_j(f,t)|^2 \\ &+ \sum_{k=1}^{N} \sum_{j \neq i} |H_{ik}(f,t) Y_k(f,t) H_{ij}(f,t) Y_j(f,t)| \cos\theta_{kj,i}, \end{aligned} \tag{5}$$

where $\theta_{kj,i}$ is the phase difference between the speech of the k-th and j-th speakers observed with the i-th microphone.

Since the phases of different speakers are uncorrelated in each time-frequency bin, the expectation of $\cos\theta_{kj,i}$ is zero. Assuming that the sparseness of speech holds approximately, i. e., the following equation holds:

$$S_j(f,t) S_k(f,t) \simeq 0 \quad (j \neq k), \tag{6}$$

the third term of Eq. (5) becomes sufficiently small and can be ignored. Hence, the speech signal of the i-th speaker is estimated as:

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2. \tag{7}$$

2.2. Implementation to OSD

It can be safely assumed that, in the power obtained by the i-th microphone, the i-th speaker's voice is much larger than that of the other speakers when more than one speaker is speaking. Then $|\hat{Y}_j(f,t)|^2 = |X_j(f,t)|^2$ and $0 \leq |\hat{H}_{ij}(f,t)|^2 \leq 1$. In OSD, speech power is important, not distortion. In order to subtract most of the other speakers' speech power, letting $|\hat{H}_{ij}(f,t)|^2 = 1$, in the second term of Eq. (7), the target signal of the i-th channel is calculated as:

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j \neq i} |X_j(f,t)|^2, 0 \right), \tag{8}$$

and define a CCSS power as:

$$\text{CCSS\_P}_i(t) = \sum_{f \in F} |\hat{Y}_i(f,t)|^2, \tag{9}$$

where F is a set of frequency bands.

The previous method [5] used a raw power $|X_i(f,t)|^2$ instead of $|\hat{Y}_i(f,t)|^2$ in Eq. (9).

3. Spectral similarity

3.1. Power spectral Pearson's correlation coefficient

In the previous method [5], power spectral Pearson's correlation coefficient (PPC) is employed to measure similarity between the power spectra of the i-th microphone and the j-th microphone. It is defined as:

$$\text{PPC}_{i,j}(t) = \frac{\left(\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\right) \cdot \left(\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\right)}{\left\|\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\right\| \left\|\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\right\|}, \tag{10}$$

where $\boldsymbol{P}_i(t)$ is the $|F| \times (2T+1)$ dimensional vector of power spectral components $|X_i(f,\tau)|^2$ for $f \in F$, $t - T \leq \tau \leq t + T$, and $\bar{\boldsymbol{P}}_i(t)$ is its mean over all the $|F|$ bands of all the $2T+1$ frames.

Since PPC represents the similarity between two signals, it becomes large when only one speaker is speaking and becomes low when more than one speaker is speaking. Normalization using $\bar{\boldsymbol{P}}_i(t)$ is expected to be effective when there exists additional noise which should be subtracted. However, speech signals are normalized even when only one speaker is speaking too, and hence PPC becomes lower than that without normalization.

3.2. Amplitude spectral cosine correlation coefficient

Instead of PPC, we employed an amplitude spectral cosine correlation coefficient (ACC) to measure similarity between the amplitude spectra of the i-th microphone and the j-th microphone. It does not include the normalization process . It is defined as:

$$\mathrm{ACC}_{i,j}(t) = \frac{\boldsymbol{A}_i(t) \cdot \boldsymbol{A}_j(t)}{\|\boldsymbol{A}_i(t)\| \, \|\boldsymbol{A}_j(t)\|}, \tag{11}$$

where $\boldsymbol{A}_i(t)$ is the $|F| \times (2T+1)$ dimensional vector of amplitude spectral components $|X_i(f,\tau)|$ for $f \in F$, $t - T \leq \tau \leq t + T$. We use amplitude instead of power to keep the dynamic range of the coefficients small.

While ACC may not be better than PPC under noisy conditions, it is expected to be better in a rather quiet condition such as that of meeting speech data recorded by lapel microphones.

4. Experiments

4.1. Experimental conditions

We recorded a sit-down meeting of 19 minutes long, conducted in Japanese language by four speakers, one female and three male speakers. The speakers' positions are shown in Figure 1. The participants did not move from their seats, but they were allowed to change their posture as they desired. A lapel microphone was attached to the lapel of each speaker. The speech segments were hand-labeled, including laughter and coughing, and the label for overlapped speech (W_o) or that for non-overlapped speech (W_n) is given to each frame (every 10 ms). Their statistics are given in Table 1.

The recording was done at 16 kHz sampling frequency. STFT was performed using Hamming window with 20 ms width with 10 ms frame shift. Then, $|F| = 160$ in Eq. (9). We set the parameters in Eq. (11) $T = 25$ and used the lower half frequency 80 points, between 50 Hz and 4000 Hz. The number of Gaussian components in each GMM is 8.

A raw power (P) used in the previous method [5], a CCSS power (CCSS_P) of the proposed method, a power spectral Pearson's correlation coefficient (PPC), and an amplitude spectral cosine correlation coefficient (ACC) are extracted from each frame. The dimension of P and CCSS_P feature vectors are both 4, one from each of the four microphones, and those of PPC and ACC feature vectors are both 6, which is the number of pairs among the four speakers. The previous method [5] is denoted as P+PPC, and our proposal method is denoted as CCSS_P+ACC.

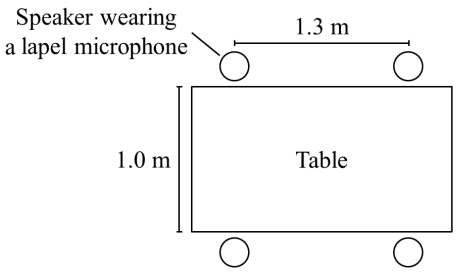

Figure 1: Position of speakers in sit-down meeting.

Table 1: Training and test dataset.

	Length	W_n	W_o
Train	9.7 min	70%	30%
Test	9.7 min	68%	32%

The log likelihood ratio of W_o to W_n as:

$$\begin{aligned} \Lambda(s) &= \ln \frac{P(s|W_\mathrm{o})}{P(s|W_\mathrm{n})} \\ &= \ln[P(s|W_\mathrm{o})] - \ln[P(s|W_\mathrm{n})], \end{aligned} \tag{12}$$

where $P(s|W_\mathrm{o})$ is the likelihood of s as W_o, and $P(s|W_\mathrm{n})$ is the likelihood of s as W_n. We use average precision (AP) as the measure of detection performance, since it includes both recall and precision.

4.2. Results

The detection results are shown in Figure 2. As can be seen, the proposal method, CCSS_P+ACC, achieves the highest performance. Its AP is 74.1% which is better than the 67.0% of P+PPC by 10.6% relative improvement.

We conducted the following experiments with a part of the test dataset in which only two speakers participate. We compare the performance of P and CCSS_P in Figure 3. The overlapped region and the non-overlapped region in CCSS_P are more clearly separated than those in P. We also compare the performance of PPC, ACC, an amplitude spectral Pearson's correlation coefficient (APC), and a power spectral cosine correlation coefficient (PCC) in Table 2. The AP of ACC achieves the highest score. And in Figure 4, it is shown that the ACC histogram of the overlapped frames and that of the non-overlapped frames are more clearly separated than those of PPC.

We analyze how CCSS_P and ACC complement each other in our proposed OSD. The power histogram of the non-overlapped frames misdetected as overlapped frames

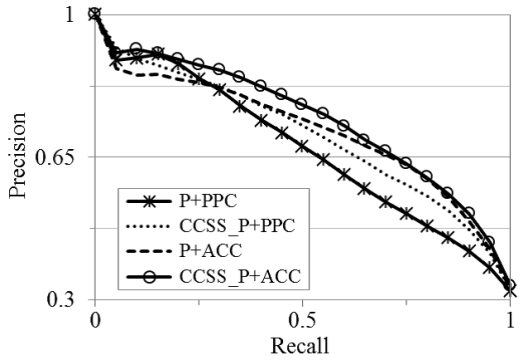

Figure 2: Recall-Precison curve of overlapped detection in meeting.

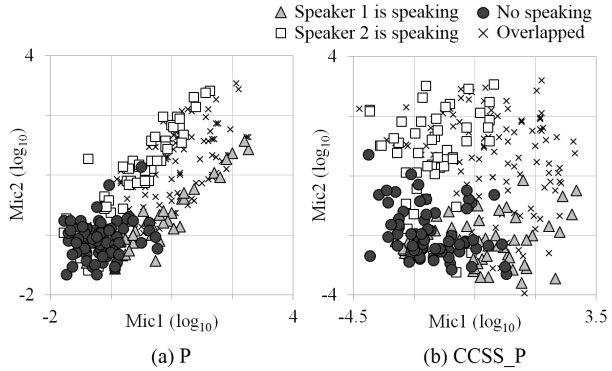

Figure 3: A scatter diagram of frame labels obtained by OSD. The horizontal axis is the power recorded by the 1st microphone, and the vertical axis is the power recorded by the 2nd microphone.

when Recall $= 0.5$ is shown in Figure 5. CCSS_P tends to misdetect the frames whose power is relatively large, and ACC tends to misdetect the frames whose power is relatively small. Thus, CCSS_P and ACC make up for each other.

5. Conclusion

In this study, we have proposed CCSS_P and ACC as the features for OSD in meeting speech. In our evaluation experiments, we compared our features with the previously proposed features, P and PPC. The AP of the proposal method is 74.1% which is better than 67.0% of the previous method by 10.6% relative improvement.

In spite of these improvements, misdetected frames still exist and more features are required to improve the OSD performance. One promising applicant would be entropy. In addition, we used the hand-labeled speech segments in this study. An overlapped detection method with unsupervised learning is required to reduce the annotation costs.

Table 2: AP (%) of each correlation coefficient.

	PPC	ACC	APC	PCC
AP	40.5	46.4	42.7	39.4

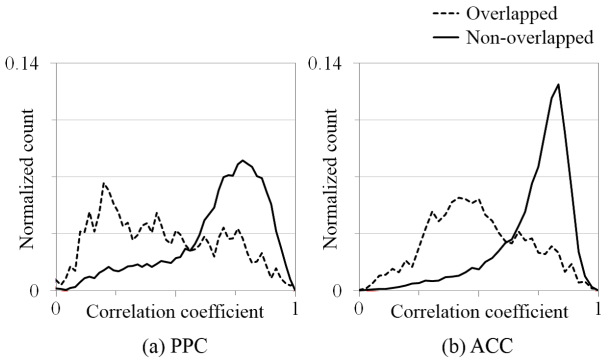

Figure 4: Correlation coefficient histogram of the overlapped and non-overlapped frames.

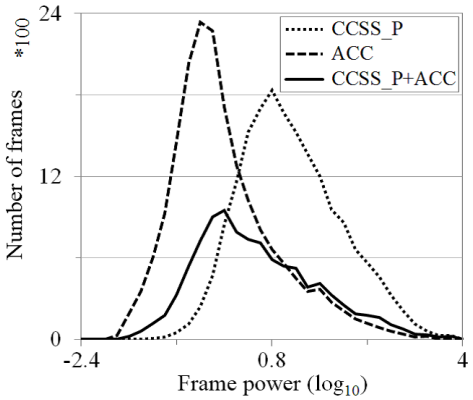

Figure 5: Power histogram of the misdetected frames.

6. References

[1] Y. Nasu, K. Shinoda, and S. Furui, "Cross-channel spectral subtraction for meeting speech recognition," in *Proc. ICASSP*, 2011, pp. 4812–4815.

[2] A. Stolcke, G. Friedland, and D. Imseng, "Leveraging speaker diarization for meeting recognition from distant microphones," in *Proc. ICASSP*, 2010, pp. 4390–4393.

[3] F. Valente, D. Vijayasenan, and P. Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *Proc. ICASSP*, 2011, pp. 4416–4419.

[4] O. Ben-Harush, H. Guterman, and I. Lapidot, "Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization," in *IEEE International Workshop on Machine Learning for Signal Prcessing*, 2009, pp. 1–6.

[5] B. Xiao, P.K. Ghosh, P. Georgiou, and S.S. Narayanan, "Overlapped speech detection using long-term spectro-temporal similarity in stereo recording," in *Proc. ICASSP*, 2011, pp. 5216–5219.

[6] S.N. Wrigley, G.J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," in *IEEE Transactions on Speech and Audio Processing*, 2004, vol. 13, no. 1, pp. 84–91

[7] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1979, vol. 27, no. 2, pp. 113–120