Detection of overlapped speech using lapel microphones in meeting

Ryo Yokoyama[a,*], Yu Nasu[a], Koji Iwano[b], Koichi Shinoda[a]

[a] *Department of Computer Science, Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo 152–8552, Japan*
[b] *Faculty of Environmental and Information Studies, Tokyo City University, 3–3–1 Ushikubo, Tsuzuki-ku, Yokohama 224–8551, Japan*

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, 75.7%, was significantly better than that of the conventional method, 66.8%, which uses raw speech power and power spectral Pearson's correlation coefficient.

Keywords: Overlap speech detection, Spectral subtraction, Cosine distance

1. Introduction

In recent years, meeting speech recognition (Maganti et al., 2007; Nasu et al., 2011) and meeting speaker diarization (Boakye et al., 2008; Ben-Harush et al., 2009; Stolcke et al., 2010; Sun et al., 2010; Valente et al., 2010; Vijayasenan et al., 2010; Boakye et al., 2011; Stolcke, 2011; Valente et al., 2011; Yella et al., 2011; Vijayasenan et al., 2012; Zwyssig et al., 2012) have been effectively utilized 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 to this problem is to first 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 (Boakye et al., 2008; Ben-Harush et al., 2009; Boakye et al., 2011; Sun et al., 2011) and microphone arrays (Moore et al., 2003; Yamamoto et al., 2006; Rozgic et al., 2010) 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. In this study, we use lapel microphones 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 pocket or 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 one GMM for overlapped segments and one for non-overlapped segments, with a threshold set for their likelihood ratio. Then, the problem is what features we should use as input. The power summed over all the frequency bands has proven to be effective (e.g., Pfau et al., 2001; Wrigley et al., 2004; Xiao et al., 2011). Large powers in more than one microphone indicate overlapping. Some studies (Pfau et al., 2001; Wrigley et al., 2004; Ghosh et al., 2010; Xiao et al., 2011) focused on the similarity of the signals from different microphones. They tend to be similar with each other when only one speaker speaks, and be different when more than one speaker is speaking.

[*] Corresponding author. Tel.: +81 3 5734 3480; Fax: +81 3 5734 3480.
E-mail address: `yokoyama@ks.cs.titech.ac.jp` (R. Yokoyama).

Pfau et al. (2001) proposed an OSD method using both of these two features sequentially. It first detects the speech activity of signals from each microphone using the overall power. Then, if signals from more than one microphone are active, it calculates short-time cross-correlation (XC), which is a measure for signal similarity, between them. Those frames with short-time XC smaller than a certain threshold are detected as overlapping frames. In this method, the error in the first step cannot be recovered in the second step. Then, Wrigley et al. (2004) proposed an OSD method utilizing both of these features, the overall power and the short-time XC, at the same time as inputs to a detector and proved its effectiveness. Ghosh et al. (2010) found that XC with a long-term window yielded better results in cross-talk detection than that with a short-time window. Based on these studies (Wrigley et al., 2004; Ghosh et al., 2010), Xiao et al. (2011) proposed Pearson's correlation coefficient (PPC) as long-term XC, and reported that using the overall power and PPC as inputs is effective for OSD.

However, each of these features, the overall power and PPC, has a serious problem. The overall power may be contaminated by speech signals from the other speakers and detect overlapping segments incorrectly when only one speaker is speaking. 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 more than one speaker is speaking, and hence, may make the PPC large, leading to false detection of overlapped segments.

To solve the first problem, we focused on the source separation methods (e.g., Bell et al., 1995; Smaragdis, 1998; Aoki et al., 2001; Rickard et al., 2001) utilizing multiple channel signals from multiple microphones. Some of them used independent component analysis (ICA) (Bell et al., 1995; Smaragdis, 1998) and others employed binary masks in the spectrogram (Aoki et al., 2001; Rickard et al., 2001). The ICA-based methods, however, need high computational costs to calculate higher-order statistics. Signals separated by binary masking may include the other speaker's speech signals when the source signals are not sparse enough or when the estimation of masks is not reliable. Recently, cross-channel spectral subtraction (CCSS) was proposed (Nasu et al., 2011). It is based on spectral subtraction (Boll, 1979) and has less computational costs than ICA-based methods. It is also more robust than binary masking methods since it effectively estimates and suppresses interference components, while binary masking methods force all time-frequency components to be allocated to one channel.

In this paper, we propose two new features for OSD in meeting speech. One is CCSS power, which is an overall power obtained by CCSS. It is 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 small when more than one speaker is speaking. We use these two features as inputs to a GMM detector and examine their effectiveness using a meeting speech data with four speakers.

This paper is organized as follows. Section 2 review frequency analysis and spectral subtraction for speech signals. The two features used in this paper, CCSS power and ACC, are explained in Section 3 and Section 4, respectively. The experimental results are reported in Section 5. Finally, Section 6 concludes the paper.

2. Speech analysis

Let $x(t)$ be a recorded signal at time t. It can be written as:

$$x(t) = s(t) + n(t), \tag{1}$$

where $s(t)$ and $n(t)$ are the speech signal and the environmental additive noise, respectively.

It is generally assumed that speech is quasi-stationary, and hence, it can be analyzed in the frequency domain by the short-time Fourier transform (STFT). STFT transforms the recorded signal Eq. (1) to:

$$X(f,t) = \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau, \tag{2}$$

where $X(f,t)$ represents the STFT of the recorded signal in a frequency band f at time t, and $w(t)$ is a window function. In speech processing, the Hamming window function is typically used.

Since the recorded signal is usually discrete, discrete short-time Fourier transform (DSTFT) is often used for computation. It is calculated as:

$$X(f,t) = \sum_{\tau=-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau}, \tag{3}$$

and is denoted as STFT in this paper.

Spectral subtraction (Boll, 1979) is widely used for single channel speech enhancement. The power spectrum of the observed signal is approximated as:

$$|X(f,t)|^2 \approx |S(f,t)|^2 + |N(f,t)|^2, \tag{4}$$

where f and t are the frequency and frame indices, and $S(f,t)$ and $N(f,t)$ are spectra of speech and additive noise, respectively. With the estimated noise power spectrum $|\hat{N}(f,t)|^2$, the power spectrum of speech is estimated as:

$$|\hat{S}(f,t)|^2 = |X(f,t)|^2 - \alpha|\hat{N}(f,t)|^2, \tag{5}$$

where α is the subtraction factor.

3. CCSS power

3.1. Cross-channel spectral subtraction (CCSS)

In order to reduce the power from the other speakers, we introduce cross-channel spectral subtraction (CCSS) (Nasu et al., 2011), a source separation method based on spectral subtraction.

3.1.1. Algorithm

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:

$$X_i(f,t) = \sum_{j=1}^{N} G_{ij}(f,t)S_j(f,t), \tag{6}$$

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:

$$Y_j(f,t) = G_{jj}(f,t)S_j(f,t), \tag{7}$$

and substituting the transfer function with:

$$H_{ij}(f,t) = \frac{G_{ij}(f,t)}{G_{jj}(f,t)}, \tag{8}$$

the recorded signal can be written as:

$$X_i(f,t) = Y_i(f,t) + \sum_{j\neq i} H_{ij}(f,t)Y_j(f,t). \tag{9}$$

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

$$\begin{aligned}
&|X_i(f,t)|^2\\
&= \left| Y_i(f,t) + \sum_{j \neq i} H_{ij}(f,t) Y_j(f,t) \right|^2\\
&= |Y_i(f,t)|^2 + \sum_{j \neq i} |H_{ij}(f,t) Y_j(f,t)|^2\\
&+ \sum_{k=1}^{N} \sum_{j \neq i} |H_{ik}(f,t) Y_k(f,t) H_{ij}(f,t) Y_j(f,t)| \cos \theta_{kj,i},
\end{aligned} \tag{10}$$

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:

$$S_j(f,t) S_k(f,t) \simeq 0 \quad (j \neq k), \tag{11}$$

the third term of Eq. (10) becomes sufficiently small and can be ignored. Hence, the speech signal of the i-th speaker is represented as:

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \sum_{j \neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2, \tag{12}$$

in the same manner as in Eq. (5).

We will discuss the ways to estimate the transfer function $|\hat{H}_{ij}(f,t)|^2$ and the separated signal $|\hat{Y}_j(f,t)|^2$ in the next two subsections.

3.1.2. Estimation of transfer functions

We estimate the transfer functions $|\hat{H}_{ij}(f,t)|^2$ using frames in which only one speaker is speaking. It can be safely assumed that such frames exist in meetings. The signal channel recorded by the j-th microphone is expected to have the largest power when only the j-th speaker is speaking. Thus we calculate the target signal of the i-th channel as:

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j \neq i} |X_j(f,t)|^2, 0 \right), \tag{13}$$

and select the frames which suffice both of the following conditions:

$$\frac{1}{|F_1|} \sum_{f \in F_1} |\hat{Y}_j(f,t)|^2 > T_j(t), \tag{14}$$

$$\frac{1}{|F_1|} \sum_{f \in F_1} |\hat{Y}_k(f,t)|^2 < T_k(t), \quad \forall k \neq j \tag{15}$$

as the frames where only the j-th speaker is speaking, where F_1 is a set of frequency bands and $T_j(t), T_k(t)$ are predetermined thresholds. The thresholds were set as:

$$T_j(t) = \frac{2}{|F_1|} \sum_{f \in F_1} |\hat{N}_j(f,t)|^2, \tag{16}$$

$$T_k(t) = \frac{1}{|F_1|} \sum_{f \in F_1} |\hat{N}_k(f,t)|^2, \quad \forall k \neq j \tag{17}$$

where $|\hat{N}_j(f,t)|^2, |\hat{N}_k(f,t)|^2$ are the estimated noise power spectra of the j-th and k-th microphone. The noise power spectra are updated as:

$$|\hat{N}_i(f,t)|^2 = \rho_n|\hat{N}_i(f,t-1)|^2 + (1-\rho_n)|X_i(f,t)|^2, \tag{18}$$

using predetermined initial values $|\hat{N}_i(f,0)|^2$ and forgetting factor $\rho_n \in [0,1]$. This update is carried out during the noise frames in which nobody is speaking. When the signal at time t does not have a large power at any frequency band, it is estimated as a noise frame. We select the frame t which suffices the following condition:

$$\sum_{f\in F_1} |\hat{Y}_i(f,t)|^2 < \epsilon \sum_{f\in F_1} |X_i(f,t)|^2, \quad \forall i \tag{19}$$

where $\epsilon \in [0,1)$ and $|\hat{Y}_i(f,t)|^2$ is calculated in Eq. (13), as a noise frame. Since the noise power is usually large in lower frequency bands, we select F_1 from the low frequency region.

Let $|X_i(f,t)|^2$ be the power spectrum which includes the environmental noise. The clean power spectrum is calculated as:

$$|X'_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - |\hat{N}_i(f,t)|^2, 0\right), \tag{20}$$

using the noise estimated in Eq. (18). Since the transfer function is considered to be time-variable, we update it in some time intervals as:

$$|\hat{H}_{ij}(f,t)|^2 = \rho_h|\hat{H}_{ij}(f,t-1)|^2 + (1-\rho_h)\frac{|X'_i(f,t)|^2}{|X'_j(f,t)|^2}. \tag{21}$$

This update is carried out when only the j-th speaker is speaking, using the predetermined initial values $|\hat{H}_{ij}(f,0)|^2$ and forgetting factor $\rho_h \in [0,1]$.

3.1.3. Estimation of separated signals

The separated signals are estimated by an iterative process using the estimated transfer functions. We set the initial value as $|\hat{Y}_i^{(0)}(f,t)|^2 = |X'_i(f,t)|^2$ for $i = 1,\ldots,N$ and iteratively update it as:

$$|\hat{Y}_i^{(c)}(f,t)|^2 = \max\left(|X'_i(f,t)|^2 - \alpha^{(c)} \sum_{j\neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j^{(c-1)}(f,t)|^2, 0\right), \tag{22}$$

where c is the number of iterations and $\alpha^{(c)}$ is the subtraction factor of each iteration.

If we use Eq. (22) only once, some speech components of the target i-th speaker may be subtracted from the signal of the i-th microphone. This may cause large distortion in the target speaker's speech. We can obtain less distorted signals by improving the estimates of the second term of Eq. (22) with this iterative process.

Figs. 1 and 2 show examples of the raw signals recorded by 4 speakers' lapel microphones in meeting and the signals separated by CCSS. In Fig. 1, the 4th speaker is mainly speaking from 15 s to 65 s, and their attenuated signal appeared in the signals of the other microphones. When we use the power of raw signals for OSD, the segment where only one speaker is speaking may be detected incorrectly as a segment with overlapped speech. Whereas in Fig. 2, the signals of the other three microphones are much smaller than those in Fig. 1. Using the power of signals separated by CCSS is effective for achieving high OSD performance.

Nasu et al. (2011) evaluated the performance of CCSS using speech recognition. CCSS achieved significantly better recognition accuracies than the use of raw signals. They concluded that CCSS has high performance in source separation.

We evaluated the subjective overall quality of the signal separated by CCSS, using mean opinion score (MOS) that takes into account both the level of speech distortion by noise and the level of the target speech intelligibility. In this experiment, we selected two types of signals, the raw signal and the signal separated by CCSS, of 4 short time segments, making 8 signal samples in total. We asked 8 listeners to listen to them. Each listener chooses one of the five scores shown in Table 1 for each of the three questions about the signal quality. The results showed in Fig. 3 indicate that the overall signal quality of them is almost the same. Thus, it can be concluded that CCSS achieves high source separation performance without any significant degradation of the overall signal quality.

3.2. *Implementation to OSD*

The raw power (P) used as the feature for OSD in the previous methods (e.g., Xiao et al., 2011) is calculated as:

$$\mathrm{P}_i(t) = \sum_{f \in F_2} |X_i(f,t)|^2, \tag{23}$$

where F_2 is the set of all frequency bands in STFT.

Since signals of one microphone contain the speech from the other speakers, more than one microphone's power may be large even when only one speaker is speaking. In this case, false detection may occur.

To solve this problem, we use the power of the signal estimated in Eq. (22) by CCSS. It is defined as:

$$\mathrm{CCSS\_P}_i(t) = \sum_{f \in F_2} |\hat{Y}_i^{(c)}(f,t)|^2. \tag{24}$$

The CCSS power is not expected to include the power from the other speakers, and only one of microphone's CCSS power will be large when only one speaker is speaking. Thus, it can be used to distinguish between overlapped and non-overlapped segments.

4. Spectral similarity

4.1. *Difference of similarity*

Some studies have focused on the effect of cross-talk and use a spectral similarity between the signals from microphones to detect overlapped speech segments. In meetings, the similarity becomes high when only one speaker is speaking, and becomes low when more than one speaker is speaking. In the following explanation, the number of participants is fixed at two for simplicity.

Let the participants be i, j and only speaker i is speaking. The power spectrum of the recorded signals from microphone i, j is calculated as:

$$|X_i(f,t)|^2 = |\hat{Y}_i(f,t)|^2, \tag{25}$$
$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2 |\hat{Y}_i(f,t)|^2, \tag{26}$$

from Eq. (12).

Then, the power spectrum of the j-th microphone becomes:

$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2 |X_i(f,t)|^2, \tag{27}$$

and it becomes highly similar to $|X_i(f,t)|^2$, the spectrum of the i-th microphone.

On the other hand, consider the situation when both speakers i, j are speaking. The power spectrum of the recorded signals from microphone i, j is calculated as:

$$|X_i(f,t)|^2 = |\hat{Y}_i(f,t)|^2 + |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2, \tag{28}$$
$$|X_j(f,t)|^2 = |\hat{H}_{ji}(f,t)|^2 |\hat{Y}_i(f,t)|^2 + |\hat{Y}_j(f,t)|^2. \tag{29}$$

Since $|\hat{Y}_i(f,t)|^2$ and $|\hat{Y}_j(f,t)|^2$ are different, the similarity between $|X_i(f,t)|^2$ and $|X_j(f,t)|^2$ becomes low.

4.2. *Measurement of similarity*

4.2.1. *Power spectral Pearson's correlation coefficient*

In the previous method (Xiao et al., 2011), the power spectral Pearson's correlation coefficient (PPC) is employed to measure the similarity between the power spectra of the i-th microphone and the j-th microphone. It is defined as:

$$\mathrm{PPC}_{i,j}(t) = \frac{\left(\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\right) \cdot \left(\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\right)}{\left\|\boldsymbol{P}_i(t) - \bar{\boldsymbol{P}}_i(t)\right\| \left\|\boldsymbol{P}_j(t) - \bar{\boldsymbol{P}}_j(t)\right\|}, \tag{30}$$

where F_3 is a set of frequency bands and $\boldsymbol{P}_i(t)$ is the $|F_3| \times (2T+1)$ dimensional vector of power spectral components $|X_i(f,\tau)|^2$ for $f \in F_3, t-T \le \tau \le t+T$, and $\overline{\boldsymbol{P}}_i(t)$ is its mean over all the $|F_3|$ bands of all the $2T+1$ frames. Since most speech information is represented in the spectrum under 4 kHz, the set F_3 includes only the frequency bands under 4 kHz. 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 $\overline{\boldsymbol{P}}_i(t)$ in Eq. (30) is expected to be effective when there exists additional noise which should be subtracted. However, since speech signals are also normalized even when more than one speaker is speaking, PPC tends to become higher than without normalization. This may cause false detection of a single speaker when more than one speaker is speaking.

4.2.2. Amplitude spectral cosine correlation coefficient

Instead of PPC, we employ 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:

$$\mathrm{ACC}_{i,j}(t) = \frac{\boldsymbol{A}_i(t) \cdot \boldsymbol{A}_j(t)}{\left\|\boldsymbol{A}_i(t)\right\| \left\|\boldsymbol{A}_j(t)\right\|}, \tag{31}$$

where $\boldsymbol{A}_i(t)$ is the $|F_3| \times (2T+1)$ dimensional vector of amplitude spectral components $|X_i(f,\tau)|$ for $f \in F_3$, $t-T \le \tau \le t+T$. 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 in a meeting room.

5. Experiments

5.1. 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. We divided the meeting data into two in the middle, and the first half is used as a training set and the latter half is used as a test set. The speakers' positions are shown in Fig. 4. 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 25 ms). Their statistics are given in Table 2.

The recording was done at 16 kHz sampling frequency. STFT was performed using Hamming window with 50 ms width with 25 ms frame shift. We experimentally set the following factors which achieved the highest performance. In Eq. (14) and Eq. (15), the frequency bands are set to $F_1 = [60, 4000]$ Hz. For noise estimation in Eq. (18), the opening 500 ms segments are used as the initial value of $|\hat{N}_i(f,0)|^2$ with $\epsilon = 0.1$ and $\rho_n = 0.98$. The transfer functions in Eq. (21) were calculated on 10 divided frequency ranges of the overall frequency bands $|F_2| = 400$ to reduce the error and were updated with $\rho_h = 0.98$, $|\hat{H}_{ij}(f,0)|^2 = 0.20$. In the estimation of separated signals, we performed two iterations and set the subtraction factor at $\alpha^{(1)} = 1.0$, $\alpha^{(2)} = 4.0$ in Eq. (22). In the calculation of the spectral similarity, we set the window width $T = 10$ and used 200 bins in the lower half frequency of the overall frequency bands. The frequency range is $F_3 = [20, 4000]$ Hz. Each component in the GMM has a diagonal covariance matrix, and the number of Gaussian components in each GMM is 8.

The raw power (P) used in the previous method (Xiao et al., 2011), the CCSS power (CCSS_P) of the proposed method, the power spectral Pearson's correlation coefficient (PPC), and the amplitude spectral cosine correlation coefficient (ACC) are extracted from each frame. The dimensions 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 (Xiao et al., 2011) is denoted as P+PPC, and our proposal method is denoted as CCSS_P+ACC.

The log likelihood ratio of W_o to W_n is calculated as:

$$\begin{aligned}\Lambda(s) &= \ln \frac{P(s|W_\mathrm{o})}{P(s|W_\mathrm{n})} \\ &= \ln[P(s|W_\mathrm{o})] - \ln[P(s|W_\mathrm{n})],\end{aligned} \tag{32}$$

where $P(s|W_\mathrm{o})$ is the likelihood of s given W_o, and $P(s|W_\mathrm{n})$ is the likelihood of s given W_n. We decided the overlapped speech according to:

$$s = \begin{cases} \text{Overlapped speech,} & \Lambda(s) \geq T_\mathrm{border} \\ \text{Non-overlapped speech,} & \text{otherwise} \end{cases} \tag{33}$$

with the threshold T_border.

We used average precision (AP) (Zhu, 2004) as the measure of detection performance. We calculated $\Lambda(s)$ of all frames in Eq. (32), and listed them in descending order as $\Lambda_r(s)$, where r is the rank in this list. The AP is calculated as:

$$\mathrm{AP} = \sum_{r}^{M} P(r)\Big(R(r) - R(r-1)\Big) \tag{34}$$

where M is the total number of frames, $P(r)$ and $R(r)$ are the precision and the recall respectively when $T_\mathrm{border} = \Lambda_r(s)$ in Eq. (33), and let $R(0) = 0$. This AP corresponds to the area size inside the recall-precision curve. The closer the recall-precision curve is to the upper-right corner, the higher the detection performance is.

5.2. Results

The detection results are shown in Fig. 5. As can be seen, the proposed method, CCSS_P+ACC, achieves the highest performance. Its AP is 75.7% which is better than the 66.8% of P+PPC by 26.8% relative reduction of error.

We conducted experiments using part of the test dataset in which only two speakers participate to compare the performance of P and CCSS_P in Fig. 6. The overlapped region and the non-overlapped region in CCSS_P are more clearly separated than those in P. These results indicate that CCSS_P reduces the speech power from the other speakers, and thus, contributes to the high OSD performance of our method.

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 3. The highest AP of ACC is 50.6%. In Fig. 7, 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 when recall = 0.7 is shown in Fig. 8. CCSS_P tends to misdetect the frames whose power is relatively large. In these frames, CCSS_P's of more than one microphone become large because the power from the other speaker, who is speaking loudly, cannot be reduced enough by CCSS. For example, it can be seen in the upper-left region of Fig. 6(b) that CCSS_P$_1(t)$ in frames t, where only the 2nd speaker is speaking, tends to be larger when CCSS_P$_2(t)$ is relatively large and that there are some mixtures of the non-overlapped and the overlapped frames. On the other hand, ACC tends to misdetect the frames whose power is relatively small. In these frames, ACC between any microphones becomes small because, when only one speaker speaks quietly, the power from the other speaker is almost silence. In this way, CCSS_P and ACC make up for each other.

5.3. Evaluation with ICSI meeting corpus

5.3.1. Evaluation data

We evaluated our method using the first 20 minutes of the ICSI meeting data set (Bro027), which is in English with four male speakers (Janin et al., 2003). We divided the data into two in the middle, and used the first half as a training set and the latter half as a test set. The speech segments were hand-labeled, in which labels include laughter and coughing, and the label W_o or W_n was given to each frame (every 25 ms). Their statistics are given in Table 4. The other experimental conditions are the same as in Subsection 5.1.

5.3.2. Results

The detection results are shown in Fig. 9. As can be seen, the proposed method, CCSS_P+ACC, achieves the highest performance. Its AP is 73.3%, which is better than 64.4% of P+PPC. It reduced errors by 25.0%.

6. 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 proposed method is 75.7% which is better than 66.8% of the previous method by 26.8% relative reduction of error.

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 hand-labeled speech segments in this study. An overlapped detection method with unsupervised learning is required to reduce the annotation costs.

References

Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T., Kaneda, Y., 2001. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. Acoustical science and technology 22 (2), 149–157.

Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Computation 7 (6), 1129–1159.

Ben-Harush, O., Guterman, H., Lapidot, I., 2009. Frame level entropy based overlapped speech detection as a pre-processing stage for speaker diarization. IEEE International Workshop, Machine Learning Signal Prcess. MLSP, 1–6.

Boakye, K., Trueba-Hornero, B., Vinyals, O., Friedland, G., 2008. Overlapped speech detection for improved speaker diarization in multiparty meetings. Proc. ICASSP, 4353–4356.

Boakye, K., Vinyals, O., Friedland, G., 2011. Improved overlapped speech handling for speaker diarization. Proc. INTERSPEECH, 941–944.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27 (2), 113–120.

Ghosh, P.K., Tsiartas, A., Georgiou, P., Narayanan, S.S., 2010. Robust voice activity detection in stereo recording with crosstalk. Proc. INTERSPEECH, 3098–3101.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., Wooters, C., 2003. The ICSI meeting corpus. Proc. ICASSP, 364–367.

Maganti, H.K., Motlicek, P., Gatica-Perez, D., 2007. Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. Proc. ICASSP, 1037–1040.

Moore, D.C., McCowan, I.A., 2003. Microphone array speech recognition: experiments on overlapping speech in meetings. Proc. ICASSP, 497–500.

Nasu, Y., Shinoda, K., Furui, S., 2011. Cross-channel spectral subtraction for meeting speech recognition. Proc. ICASSP, 4812–4815.

Pfau, T., Ellis, D.P.W., Stolcke, A., 2001. Multispeaker speech activity detection for the ICSI meeting recorder. IEEE Workshop, Automatic Speech Recognition and Understanding. ASRU-01, 107–110.

Rickard, S., Balan, R., Rosca, J., 2001. Real-time time-frequency based blind source separation. Proc. ICA, 651–656.

Rozgic, V., Han, K.J., Georgiou, P.G., Narayanan, S.S., 2010. Multimodal speaker segmentation and identification in presence of overlapped speech segments. Journal of Multimedia 5 (4), 322–331.

Smaragdis, P., 1998. Blind separation of convolved mixtures in the frequency domain. Neurocomputing 22 (1)–(3), 21–34.

Sun, H., Ma, B., Khine, S.Z.K., Li, H., 2010. Speaker diarization system for RT07 and RT09 meeting room audio. Proc. ICASSP, 4982–4985.

Sun, H., Ma, B., 2011. Study of overlapped speech detection for NIST SRE summed channel speaker recognition. Proc. INTERSPEECH, 2345–2348.

Stolcke, A., Friedland, G., Imseng, D., 2010. Leveraging speaker diarization for meeting recognition from distant microphones. Proc. ICASSP, 4390–4393.

Stolcke, A., 2011. Making the most from multiple microphones in meeting recognition. Proc. ICASSP, 4992–4995.

Valente, F., Motlicek, P., Vijayasenan, D., 2010. Variational bayesian speaker diarization of meeting recordings. Proc. ICASSP, 4954–4957.

Valente, F., Vijayasenan, D., Motlicek, P., 2011. Speaker diarization of meetings based on speaker role n-gram models. Proc. ICASSP, 4416–4419.

Vijayasenan, D., Valente, F., Bourlard, H., 2010. Multistream speaker diarization beyond two acoustic feature streams. Proc. ICASSP, 4950–4953.

Vijayasenan, D., Valente, F., 2012. Speaker diarization of meetings based on large TDOA feature vectors. Proc. ICASSP, 4173–4176.

Wrigley, S.N., Brown, G.J., Wan, V., Renals, S., 2004. Speech and crosstalk detection in multichannel audio. IEEE Trans. Speech Audio Process., SAP-13 (1), 84–91.

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

Yamamoto, K., Asano, F., Yamada, T., Kitawaki, N., 2006. Detection of overlapping speech in meetings using support vector machines and support vector regression. IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, E89-A (8), 2158–2165.

Yella, S.H., Valente, F., 2011. Information bottleneck features for HMM/GMM speaker diarization of meetings recordings. Proc. INTERSPEECH, 953–956.

Zhu, M., 2004. Recall, Precision and Average Precision. Technical Report 09. Department of Statistics and Actuarial Science, University of Waterloo.

Zwyssig, E., Renals, S., Lincoln, M., 2012. On the effect of SNR and superdirective beamforming in speaker diarisation in meetings. Proc. ICASSP, 4177–4180.

Score	Distortion	Intelligibility	Total quality
5	Not distorted	Excellent	Excellent
4	Slightly distorted	Good	Good
3	Somewhat distorted	Fair	Fair
2	Fairly distorted	Poor	Poor
1	Very distorted	Bad	Bad

Table 1: Mean opinion score (MOS) evaluation criteria.

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

Table 2: Statistics of overlapped speech (W_o) and non-overlapped speech (W_n) in the training and test dataset of the self-recorded meeting.

	PPC	ACC	APC	PCC
AP	44.5	50.6	49.6	44.7

Table 3: AP (%) of the overlapped detection results using each spectral similarity features.

	Length	W_n	W_o
Train	10 min	75%	25%
Test	10 min	69%	31%

Table 4: Statistics of overlapped speech (W_o) and non-overlapped speech (W_n) in the training and test dataset of the ICSI meeting corpus.

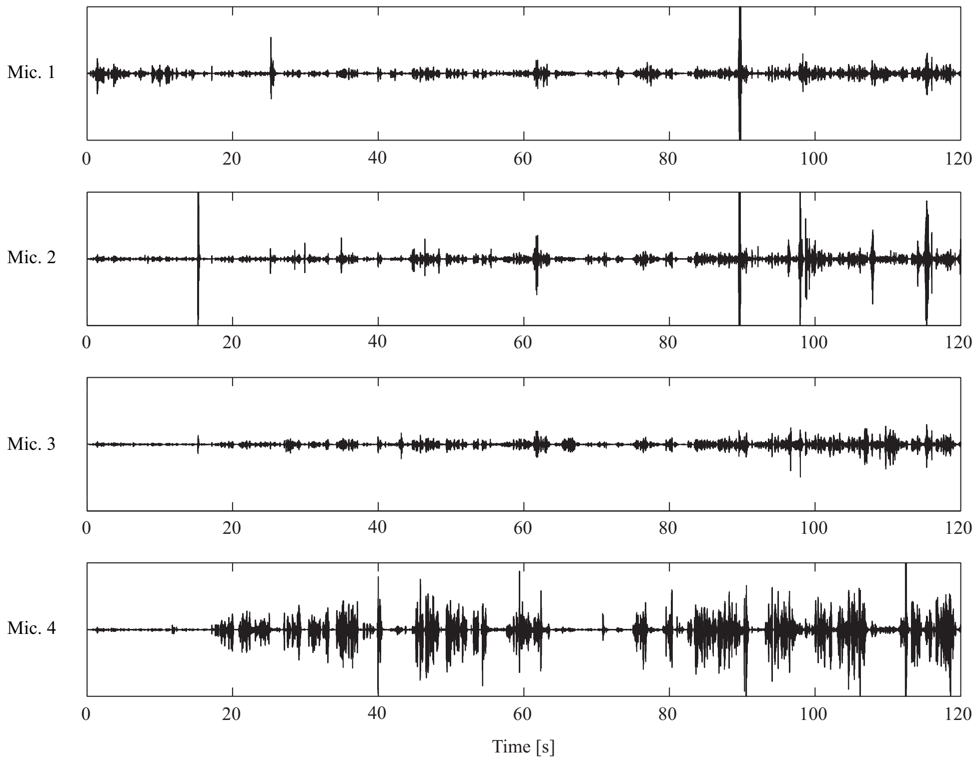

Fig. 1: Waveform of the raw signals.

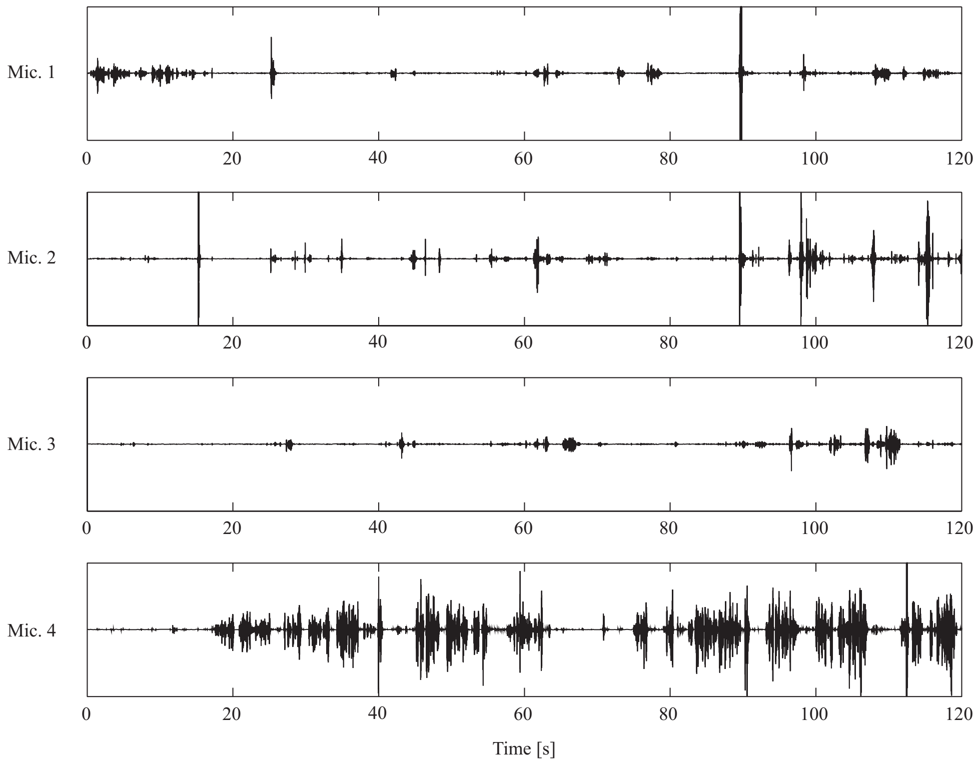

Fig. 2: Waveform of the signals separated by CCSS.

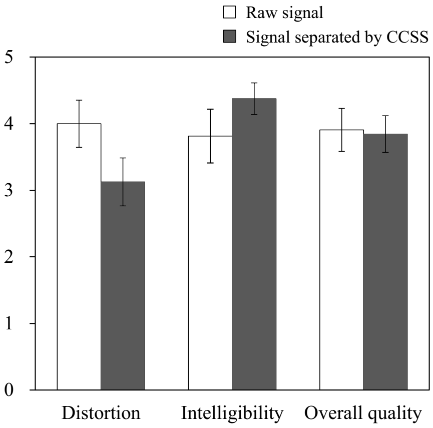

Fig. 3: Mean opinion score (MOS) of the raw signal and the signal separated by CCSS. The error bars indicate the 95% confidence intervals.

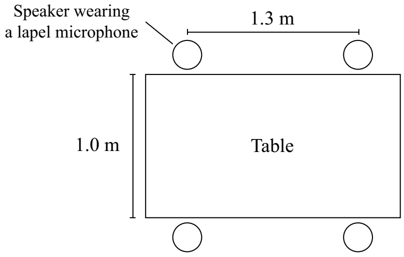

Fig. 4: Position of speakers in sit-down meeting.

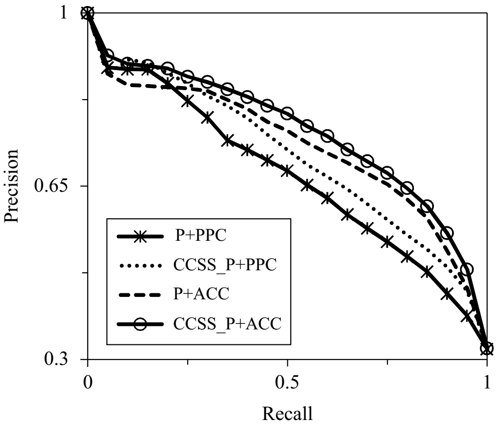

Fig. 5: Recall-precision curve of the OSD result in the test dataset of the self-recorded meeting.

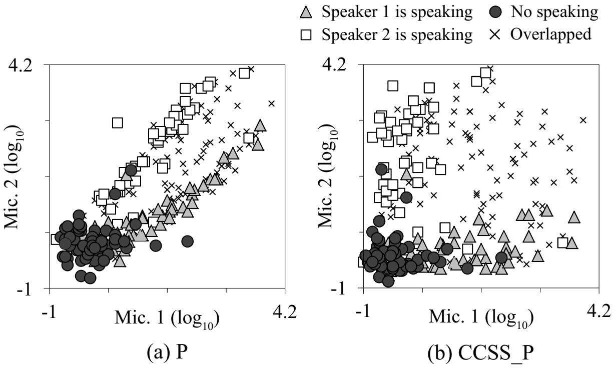

Fig. 6: Scatter diagram of frame labels obtained by OSD. The horizontal axis is the power obtained from the 1st microphone, and the vertical axis is the power obtained from the 2nd microphone.

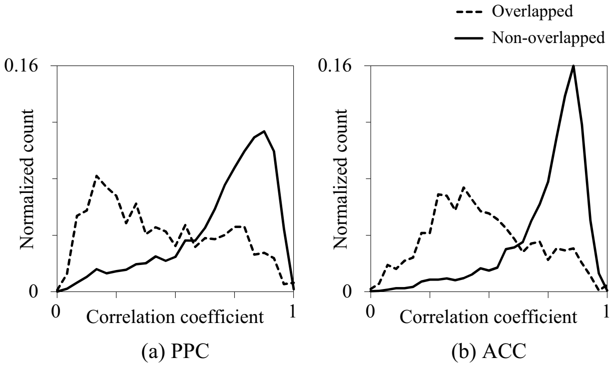

Fig. 7: Correlation coefficient histogram of the overlapped and non-overlapped frames. The horizontal axis is the value of PPC and ACC, and the vertical axis is the normalized count of overlapped and non-overlapped frames in each correlation coefficient value.

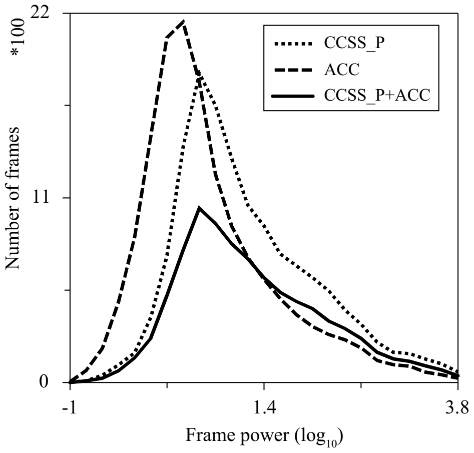

Fig. 8: Power histogram of the misdetected frames. The horizontal axis is the power of the non-overlapped frame misdetected as overlapped frame, and the vertical axis is the number of frames in each frame power level. The powers $\mathrm{P}_i(t)$ of the different channels ($i = 1, 2, 3, 4$) are separately counted and summed up in this histogram.

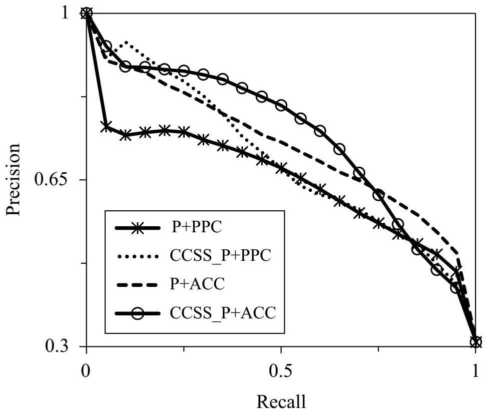

Fig. 9: Recall-precision curve of the OSD result in the test dataset of the ICSI meeting corpus.