論文 / 著書情報
Article / Book Information

Title	Cross-channel spectral subtraction for meeting speech recognition
Authors	Yu Nasu, Koichi Shinoda, Sadaoki Furui
Citation	Proc. ICASSP2011, Vol. , No. , pp. 4812-4815
Pub. date	2011, 5
Copyright	(c) 2011 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	10.1109/ICASSP.2011.5947432
Note	This file is author (final) version.

CROSS-CHANNEL SPECTRAL SUBTRACTION FOR MEETING SPEECH RECOGNITION

Yu Nasu, Koichi Shinoda, Sadaoki Furui*

Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan

ABSTRACT

We propose Cross-Channel Spectral Subtraction (CCSS), a source separation method for recognizing meeting speech where one microphone is prepared for each speaker. The method quickly adapts to changes in transfer functions and uses spectral subtraction to suppress the speech of other speakers. Compared with conventional source separation methods based on independent component analysis (ICA) or that use binary masks, it requires less computational costs and the resulting speech signals have less distortion. In a recognition task of computer-simulated, partially-overlapped speech, CCSS improved the word accuracy from 66.5% to 77.7%. It also significantly improved the recognition accuracy of speech data in actual meetings.

Index Terms— Speech enhancement, sound source separation, spectral subtraction, meeting speech recognition

1. INTRODUCTION

Meeting speech recognition is useful for many purposes, such as in taking minutes and browsing meeting procedures. In natural conversations, speakers often give backchannels, such as an expression of agreement, and begin to speak before another speaker has finished. This produces overlapping speech by two or more speakers, a problem that makes meeting speech recognition difficult [1].

To solve this problem, sound source separation has been extensively studied (e.g., [2, 3, 4, 5]). Several methods utilizing multiple channel signals from multiple microphones have been proposed and proven to be effective. Some of them used independent component analysis (ICA) [2, 3] and others employed binary masks in the spectrogram [4, 5]. The ICA-based methods, however, need high computational costs to calculate higher-order statistics. Signals separated by binary masking are distorted when the source signals are not sparse enough or when the estimation of masks is not reliable.

Most of these studies used microphone arrays. They assume the positions of speakers are fixed during the meeting, which is not always true in actual meeting situations. In addition, a distant microphone array provides relatively low signal-to-noise ratio (SNR). Since the array is often placed on the meeting table, it may also be an obstacle when the participants want to share some materials. We can use headset microphones instead of a microphone array to obtain high SNR, but they have the disadvantage of each speaker having to wear a headset. Some speakers may feel uncomfortable wearing one, and furthermore a headset partially occludes one's face, thus making it difficult to use visual recognition techniques such as facial expression recognition. For these reasons, in this study we utilize lapel microphones for recording. While they need to be attached to the lapels of the speakers, they are easier to wear than headsets and do not influence facial image recognition techniques. Though the recognition accuracy of speech recorded with lapel microphones is degraded by overlapping speech [1], the degradation is significantly smaller than when using microphone arrays.

In noisy speech recognition, spectral subtraction [6], which reduces additive noise by subtraction of an estimated noise spectrum from the spectrum of the current frame, is one of the most popular methods for single channel speech enhancement. This method might also be an effective way to suppress other speakers' speech. Though it is effective and requires low computational costs, it assumes that the noise is stationary. Some methods adaptively estimate a non-stationary noise spectrum (e.g., [7]), but they cannot be directly applied to speech separation because they distinguish noise from speech and do not distinguish speech by different speakers.

This paper proposes a source separation method based on spectral subtraction for meeting speech recognition. With this method, which we call Cross-Channel Spectral Subtraction (CCSS), speech signals are recorded with lapel microphones. The method quickly adapts to changes in transfer characteristics by estimating coefficients successively and uses spectral subtraction to suppress the speech of other speakers. It has less computational costs than ICA-based methods and operates in real time. 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.

2. SPECTRAL SUBTRACTION

Spectral subtraction [6] is widely used for single channel speech enhancement. The power spectrum of the observed signal is approximated as:

$$|X(f,t)|^2 \approx |S(f,t)|^2 + |N(f,t)|^2, \tag{1}$$

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:

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

where α is the subtraction factor.

3. CROSS-CHANNEL SPECTRAL SUBTRACTION

3.1. Algorithm

Consider that one microphone is prepared for each speaker and let the number of speakers be N. Then, assuming the speech signals from multiple speakers are linearly mixed and ignoring noise, the signal recorded by the i-th microphone can be modeled as:

$$X_i(f,t) = \sum_{j=1}^{N} G_{ij}(f,t) S_j(f,t), \tag{3}$$

*The contact address is nasu@ks.cs.titech.ac.jp

where $S_j(f,t)$ is the speech 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:

$$Y_j(f,t) = G_{jj}(f,t)S_j(f,t), \tag{4}$$

and substituting the transfer function by:

$$H_{ij}(f,t) = \frac{G_{ij}(f,t)}{G_{jj}(f,t)}, \tag{5}$$

the recorded signal can be written as:

$$X_i(f,t) = Y_i(f,t) + \sum_{j\neq i} H_{ij}(f,t)Y_j(f,t). \tag{6}$$

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

$$\begin{aligned} &|X_i(f,t)|^2 \\ &= \left| Y_i(f,t) + \sum_{j\neq i} H_{ij}(f,t)Y_j(f,t) \right|^2 \\ &= |Y_i(f,t)|^2 + \sum_{j\neq i} |H_{ij}(f,t)Y_j(f,t)|^2 \\ &\quad + \sum_{k=1}^{N}\sum_{j\neq k} |H_{ik}(f,t)Y_k(f,t)H_{ij}(f,t)Y_j(f,t)| \cos\theta_{kj,i}, \end{aligned} \tag{7}$$

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:

$$S_j(f,t)S_k(f,t) \approx 0 \quad (j \neq k), \tag{8}$$

the third term of (7) becomes sufficiently small and it can be ignored. Hence, the speech signal of the i-th speaker is estimated as:

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \alpha \sum_{j\neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2, \tag{9}$$

in the same manner as in (2).

In practice, it is necessary to estimate unknown factors $|\hat{H}_{ij}(f,t)|^2$ and $|\hat{Y}_j(f,t)|^2$. We will discuss the ways to estimate them in the next two subsections.

3.2. Estimation of transfer functions

We estimate the transfer functions 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:

$$|\hat{Y}_i(f,t)|^2 = \max\left(|X_i(f,t)|^2 - \sum_{j\neq i} |X_j(f,t)|^2, 0 \right), \tag{10}$$

and select the frames which suffice the both conditions:

$$\frac{1}{|F|}\sum_{f\in F} |\hat{Y}_j(f,t)|^2 > T_{j1}(t), \tag{11}$$

$$\frac{1}{|F|}\sum_{f\in F} |\hat{Y}_k(f,t)|^2 < T_{k2}(t), \quad \forall k \neq j \tag{12}$$

as the frames where only the j-th speaker is speaking, with predetermined thresholds $T_{j1}(t)$ and $T_{k2}(t)$ where F is the frequency range to use.

Then, the power spectrum of the i-th recorded signal when only the j-th speaker is speaking is written by:

$$X_i(f,t) = \begin{cases} Y_j(f,t), & \text{if } i = j \\ H_{ij}(f,t)Y_j(f,t), & \text{otherwise} \end{cases} \tag{13}$$

and the transfer function $H_{ij}(f,t)$ can be estimated as the quotient of $X_i(f,t)$ and $X_j(f,t)$ when only the j-th speaker is speaking.

Since the transfer function is considered to be time-variable, we update it in some time intervals as:

$$|\hat{H}_{ij}(f,t)|^2 = \rho_h |\hat{H}_{ij}(f,t-1)|^2 + (1-\rho_h)\frac{|X_i(f,t)|^2}{|X_j(f,t)|^2}. \tag{14}$$

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

3.3. Estimation of separated signals

The separated signals are estimated by an iterative process using the estimated transfer functions. We set the initial value as $|\hat{Y}_i^{(0)}(f,t)|^2 = |X_i(f,t)|^2$ and iteratively update it as:

$$|\hat{Y}_i^{(n)}(f,t)|^2 = |X_i(f,t)|^2 - \alpha_n \sum_{j\neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_i^{(n-1)}(f,t)|^2, \tag{15}$$

where α_n is the subtraction factor of each iteration.

In the first loop of the iteration, some speech components of the target speaker are subtracted and the signals are distorted. We can obtain less distorted signals by improving estimates of the second term of (15) with this iterative process.

4. EXPERIMENTS

4.1. Experimental conditions

We performed experiments to determine our method's effectiveness in recognizing meeting speech. The experiments were performed first by using computer-simulated speech with the Corpus of Spontaneous Japanese (CSJ) [8] and then by using actual meeting speech.

The thresholds defined in (11) and (12) were set as:

$$T_{j1}(t) = \frac{2}{|F|}\sum_{f\in F} |\hat{N}_j(f,t)|^2, \tag{16}$$

$$T_{k2}(t) = \frac{1}{|F|}\sum_{f\in F} |\hat{N}_k(f,t)|^2, \tag{17}$$

with the frequency range $F = [100, 1000]$ Hz.

The transfer functions were calculated on 64 divided frequency ranges and updated with $\rho_h = 0.98$ and $|\hat{H}_{ij}(f,0)|^2 = 0.10$, which were set experimentally.

In our preliminary investigations, we obtained significant improvements with two iterations and small changes in recognition accuracy occurred when the number of iterations was increased. Accordingly, we performed two iterations in the actual experiments. We experimentally set the subtraction factors at $\alpha_1 = 1.0$ and $\alpha_2 = 4.0$.

We compared the speech recognition results for speech separated with the proposed method (CCSS), speech separated with a conventional binary masking method [4] (Binary masking) that also operates in real time, and unprocessed speech as a baseline (Baseline). In the binary masking method [4], each time-frequency component was allocated to the channel with the largest power. Short-time Fourier transform was performed using a Hamming window of 64 ms frame with 32 ms frame shift. Separated signals were retransformed back to the time domain prior to acoustic feature extraction.

We used word correct (Corr.) and accuracy rates (Acc.) as performance metrics. The speech feature vector was 38 dimensional, comprising 12 MFCCs, their first and second derivatives, and the first and second derivatives of the power. It was extracted every 10 ms with a 25 ms frame and normalized using cepstral mean subtraction. The acoustic model was trained with Simulated Public Speech (SPS) data in the CSJ excluding its test set, for which no spectral subtraction was carried out. We used HTK [9] for the experiments. It was assumed that speech/non-speech segmentation was done correctly. Recognition was performed for all speech segments of every speaker.

4.2. Evaluation with computer-simulated speech

4.2.1. Evaluation data

Using SPS data in the CSJ test set, we synthesized four-channel speech data spoken by four male speakers. The data comprised 480 seconds of Japanese speech, i.e., 120 seconds spoken by each of the four speakers. The speech was split into utterance units of 1.2–7.6 seconds duration (2.7 seconds on average), and arranged at random intervals. The total number of morphemes was 1,593. The speech segments totaled 480 seconds in duration, and 58% of them were overlapped with another speaker's speech. The sampling rate was set to 16 kHz.

The evaluation data was generated by convoluting it with the speech and impulse responses and adding background noise to it. The impulse responses were measured in a meeting room which had a reverberation time (RT_{60}) of 0.4 seconds. Four unidirectional lapel microphones and four loudspeakers were arranged in the room as shown in Fig. 1. Background noise was recorded with the same microphone arrangement, and added to the speech where the SNR in speech segments was 25 dB on average.

4.2.2. Results

Table 1 shows the results obtained for unprocessed speech (Baseline), speech separated by the binary masking and CCSS methods, and non-overlapping, noise-free speech (Ideal). Although the speech segments were given, the recognition accuracy for unprocessed speech dropped because of the overlap. While the conventional binary masking method improved word accuracy by 5.0%, CCSS achieved a significantly better 11.2% improvement. Its accuracy was close to that of non-overlapping speech. We consider that the signals separated by CCSS are less distorted than those separated by the binary masking method (which may produce some distortion because of non-sparseness or mask estimation errors), and that they maintain the acoustic features that are necessary for speech recognition.

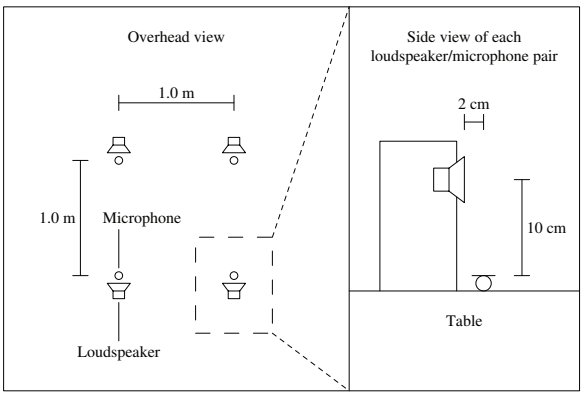

Fig. 1. Arrangement of loudspeakers and microphones. Four loudspeaker/microphone pairs were arranged on a table in a meeting room (left). A unidirectional lapel microphone was put in front of each loudspeaker (right).

Table 1. Recognition results for computer-simulated speech. Word correct (Corr.) and accuracy rates (Acc.) are shown for unprocessed speech (Baseline), speech separated with a conventional binary masking method [4] (Binary masking) and with the proposed method (CCSS), and non-overlapping and noise-free speech (Ideal).

	Baseline	Binary masking	CCSS	Ideal
Corr. [%]	75.3	78.3	82.6	84.4
Acc. [%]	66.5	71.5	77.7	79.0

4.3. Evaluation with sit-down meeting speech

4.3.1. Evaluation data

We recorded a sit-down meeting 20 minutes in duration, conducted in the Japanese language by four male speakers in the same meeting room described in the previous subsection. The speakers' positions are shown in Fig. 2. The participants could not move from their seats, but they were allowed to change their posture as they desired. The same unidirectional lapel microphones used in the previous experiment were attached to the lapels of the speakers. The speech segments and utterance transcriptions were hand-labeled. The total number of morphemes was 5,154. The speech segments totaled 1,496 seconds in duration, and 47% of them were overlapped by another speaker, the overlapping including laughter and coughing as well as speech.

4.3.2. Results

Table 2 shows the results obtained for unprocessed speech (Baseline) and speech separated by the binary masking and CCSS methods. Word accuracies were considerably worse than those obtained for computer-simulated speech. Since the speech was highly spontaneous and included many incomplete sentences and disfluencies caused by interruptions, recognizing it was a difficult task and would have been even if it had not had any overlap. However, CCSS improved word accuracy by 6.3% and, in this real-environment situation, again outperformed the conventional binary masking method.

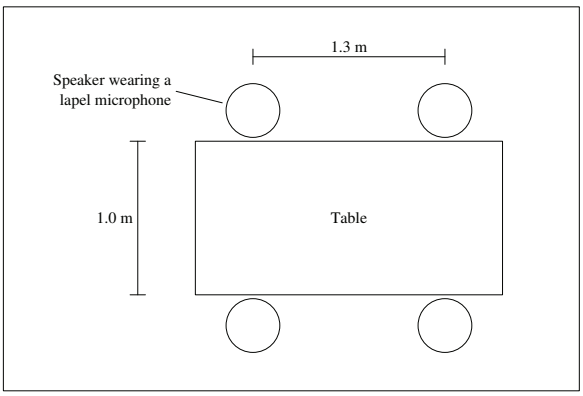

Fig. 2. Position of speakers in sit-down meeting

Table 2. Recognition results for sit-down meeting speech

	Baseline	Binary masking	CCSS
Corr. [%]	40.2	41.2	43.9
Acc. [%]	30.6	32.1	36.9

Table 3. Recognition results for stand-up meeting speech

	Baseline	Binary masking	CCSS
Corr. [%]	45.1	44.3	47.9
Acc. [%]	37.5	35.2	40.6

4.4. Evaluation with stand-up meeting speech

4.4.1. Evaluation data

We recorded a stand-up meeting 20 minutes in duration, conducted in Japanese language by four male speakers in the same meeting room described in the previous subsections. The speakers stood about 1.5 meters away from each other in front of a whiteboard on which one of them took notes. The same unidirectional lapel microphones as before were attached to the lapels of the speakers. The speech segments and utterance transcriptions were hand-labeled. The total number of morphemes was 3,980. The speech segments totaled 1,496 seconds in duration, and 28% of them were overlapped by another speaker, the overlapping including laughter and coughing as well as speech.

4.4.2. Results

Table 3 again shows the results obtained for unprocessed speech (Baseline) and speech separated by the binary masking and CCSS methods. Recognition accuracy was slightly diminished by the binary masking method. We assume that this is because the distortion of the separated signals that occurs with this method increases when the positions of speakers and microphones change. It appears that the decrease in accuracy due to the distortion outweighs the increase in accuracy due to separation. On the other hand, CCSS improved word accuracy by 3.1%, thus demonstrating that it is also an effective speech separation method for meetings in which the speakers are allowed to move around.

5. CONCLUSIONS

We have proposed a source separation method, CCSS, for meeting speech recognition, which separates target speech from concurrent speech signals. We confirmed its effectiveness through our experiments on computer-simulated speech and actual meeting speech. We found that CCSS was highly effective for computer-simulated speech and that it was also applicable to actual meeting speech. It performed significantly better than the conventional binary masking method. For computer-simulated speech it improved word accuracy from 66.5% to 77.7%, which was close to that obtained in the recognition of non-overlapping speech. It also operated well for stand-up meeting speech where the speakers were free to move around, while for this application the binary masking method could not improve the accuracy.

In the research reported in this paper, we estimated the transfer function between two microphones simply as the quotient of observed powers. There may be more effective ways to estimate the function and identifying them is a subject for future work.

While our method effectively reduced speech recognition error, its word accuracy obtained in experiments was only 36.9% and 40.6% for sit-down and stand-up meeting speech, respectively. These are not sufficiently high for applications such as the taking of minutes. Therefore, another subject for future work will be to achieve higher word accuracy with our speech recognition system.

6. ACKNOWLEDGEMENTS

This study was partly supported by a Grant-in-Aid for Scientific Research (B) 20300063 from the Japan Society for the Promotion of Science.

7. REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multiparty conversation," in *Proc. Eurospeech*, 2001, pp. 1359–1362.

[2] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[4] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda, "Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 149–157, 2001.

[5] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. ICA*, 2001, pp. 651–656.

[6] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. ASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. SAP*, vol. 9, no. 5, pp. 504–512, 2001.

[8] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000, vol. 2, pp. 947–952.

[9] "Hidden Markov Model Toolkit (HTK)," http://htk.eng.cam.ac.uk/.