

Statistical Speech Activity Detection based on Spatial Power Distribution for Analyses of Poster Presentations

Kentaro Ishizuka^{1,2}, Shoko Araki¹, and Tatsuya Kawahara²

¹ NTT Communication Science Laboratories, NTT Corporation, Japan

² Graduate School of Informatics, Kyoto University, Japan

{ishizuka, shoko}@cslab.kecl.ntt.co.jp, kawahara@i.kyoto-u.ac.jp

Abstract

This paper proposes a microphone array based statistical speech activity detection (SAD) method for analyses of poster presentations recorded in the presence of noise. Such poster presentations are a kind of multi-party conversation, where the number of speakers and speaker location are unrestricted, and directional noise sources affect the direction of arrival of the target speech signals. To detect speech activity in such cases without *a priori* knowledge about the speakers and noise environments, we applied a likelihood ratio test based SAD method to spatial power distributions. The proposed method can exploit the enhanced signals obtained from time-frequency masking, and work even in the presence of environmental noise by utilizing the *a priori* signal-to-noise ratios of the spatial power distributions. Experiments with recorded poster presentations confirmed that the proposed method significantly improves the SAD accuracies compared with those obtained with a frequency spectrum based statistical SAD method.

Index Terms: speech activity detection, microphone arrays, multi-party conversations, spatial power distribution

1. Introduction

Recently, multi-party conversations have been recorded in relation to, for example, Augmented Multi-party Interaction (AMI) [1], Computers in the Human Interaction Loop (CHIL) [2], and NIST Rich Transcription Meeting Recognition [3]. We have started a project on multi-modal recordings and analyses of poster sessions as theme-oriented multi-party conversations with static information resources [4].

The automatic indexing of such multi-party conversations can allow rapid retrieval from archives, automatic minute taking, and the automatic summarization of conversations. One essential piece of information for indexing speech data is the answer to the question, “Who spoke when?”, which can be determined with a speaker diarization method [5][6][7][8]. In particular, speech activity detection (SAD) methods play a crucial role in detecting the temporal period in which a speech signal exists from an observed signal that includes environmental noise.

Microphone array based SAD methods [9][10][11][12] have the advantage of utilizing spatial information. This contrasts with single-channel based methods [13][14][15], which can only utilize time-frequency information. However, such microphone array based SAD techniques sometimes require precise estimations of the directions of arrival (DOA) of speech signals in advance [9] or restrict speaker locations [10][11][12]. However, unlike meetings [3] in which the speakers usually sit around a table, we consider that the constraints imposed by the speakers’ locations are impractical in poster presentations because the participants can stand

anywhere and move freely. Since we cannot predict how many speakers will stand around the poster in a practical situation, parametric estimations of the number of speakers [16][17][18] should be avoided because the number is restricted by the number of microphones. In addition, there are directional noise sources when recording in real environments and this noise affects DOA estimations.

This paper proposes a SAD method without *a priori* knowledge about speaker location, number of speakers, and it is unaffected by environmental noise. The proposed method adopts the likelihood ratio test (LRT) approach [13] to spatial power distribution estimated by microphone arrays. Unlike the conventional SAD method [9], which utilizes *a priori* signal-to-noise ratios (SNRs) for the spectrum obtained from the estimated speaker location, the proposed method utilizes only *a priori* SNRs obtained from spatial power distributions. This method detects speech activities by utilizing the changes in power that come from a certain spatial range, and mitigate the effect of environmental noise incorporating stationary directional noise by utilizing the *a priori* SNRs obtained from spatial power distributions.

The remainder of this paper is organized as follows. Section 2 provides detailed explanations of the proposed SAD method. Section 3 describes evaluation experiments undertaken with recorded poster presentations to show the advantage of the proposed method in terms of diarization error rates (DERs) [3]. Section 4 concludes this study and describes future work.

2. Speech activity detection based on spatial power distribution

With the proposed method, the signals are observed by a microphone array consisting of M microphones, and then the observed signals are analyzed using a Hanning window with a certain amount of overlapping. This method works frame by frame. We consider the two following hypotheses for each analyzed frame assuming that the speech signal is uncorrelated with additive noise signals [13].

$$\begin{aligned} H_0: & \text{Speech absent; } \mathbf{X} = \mathbf{N} \\ H_1: & \text{Speech present; } \mathbf{X} = \mathbf{N} + \mathbf{S} \end{aligned}$$

where \mathbf{X} , \mathbf{N} , and \mathbf{S} represent the spatial power distributions of the observed signals, noise, and speech signals, respectively. Unlike conventional LRT based methods [9][13][14][15], the proposed method considers the hypotheses for the spatial power distributions instead of frequency spectra.

2.1. Spatial power distribution estimation

In this paper, the spatial power distributions of the observed signals are estimated by applying time-frequency masking [19][20] to the frequency spectra based on the estimated DOA for each time-frequency bin. To estimate DOAs, we first

obtain the instantaneous phase difference $\bar{\tau}_i(k)$ as follows.

$$\bar{\tau}_i(k) = [\tau_i^1(k), \tau_i^2(k), \dots, \tau_i^M(k)]^T \quad (1)$$

$$\tau_i^m(k) = \frac{\arg(X_i^m(k)/X_i^1(k))}{2 \cdot \pi \cdot f} \quad (2)$$

where $\tau_i^m(k)$ indicates the instantaneous phase difference between the signals observed by the m -th and first microphones at the k -th frequency bin of the i -th frame, $X_i^m(k)$ indicates a discrete Fourier representation, and f indicates the frequency corresponding to frequency bin k ($k = 1 \dots K$). Then, the microphone location vector D is introduced as follows.

$$D = [\bar{d}_1 - \bar{d}_1, \bar{d}_2 - \bar{d}_1, \dots, \bar{d}_M - \bar{d}_1]^T \quad (3)$$

where \bar{d}_m indicates the three dimensional location of the m -th microphone. Based on equations (1) and (3), we can obtain the angle $\theta_i(k)$ and elevation $\phi_i(k)$ by considering the following equation and assuming that the signal comes from a far field [21].

$$v_s \cdot D^{-1} \cdot \bar{\tau}_i(k) = [\cos \theta_i(k) \cos \phi_i(k), \sin \theta_i(k) \cos \phi_i(k), \sin \phi_i(k)]^T \quad (4)$$

where D^{-1} indicates the Moore-Penrose pseudoinverse of D , and v_s indicates the speed of sound. In this paper, we only consider the angle $\theta_i(k)$ as the estimated DOA of the observed signals. To obtain the spatial power distribution, time-frequency binary masks $Mask_i(k)$ are designed as follows based on the estimated DOA for each frequency bin.

$$Mask_i(k, \Theta_a) = \begin{cases} 0 & \text{if } \theta_i(k) \notin \Theta_a \\ 1 & \text{if } \theta_i(k) \in \Theta_a \end{cases} \quad (5)$$

where Θ_a indicates the discrete angle range for estimating the power of signals that comes from that range, a is the index of the range ($a = 1 \dots A$). The spatial power distribution $P_i(\Theta_a)$ is obtained by $Mask_i(k)$ as follows.

$$P_i(\Theta_a) = \frac{1}{K} \sum_{k=1}^K |Mask_i(k, \Theta_a) X_i(k)|^2 \quad (6)$$

This operation can be considered to constitute spatial power estimation using a kind of beamformer that only extracts the signal power from a narrow spatial range Θ_a . This spatial power distribution is calculated for all spatial ranges. Because the proposed method estimates the spatial power distribution based on the enhanced signals by utilizing time-frequency masking, it can exploit the differences between the spatial locations of the speakers and noise sources.

2.2. Statistical model-based speech activity detection

To model the spatial power distribution, we assume that the spatial power for each spatial range has a Gaussian distribution [13]. Based on this assumption, the probability density functions conditioned by H_0 and H_1 are as follows.

$$p_i(\mathbf{X}|H_0) = \prod_{a=1}^A \frac{1}{\pi \lambda_i^N(\Theta_a)} \exp\left(-\frac{P_i(\Theta_a)}{\lambda_i^N(\Theta_a)}\right) \quad (7)$$

$$p_i(\mathbf{X}|H_1) = \prod_{a=1}^A \frac{1}{\pi \{\lambda_i^N(\Theta_a) + \lambda_i^S(\Theta_a)\}} \exp\left(-\frac{P_i(\Theta_a)}{\lambda_i^N(\Theta_a) + \lambda_i^S(\Theta_a)}\right) \quad (8)$$

where $\lambda_i^N(\Theta_a)$ and $\lambda_i^S(\Theta_a)$, respectively, indicate the variances of \mathbf{N} and \mathbf{S} for the a -th spatial range. The likelihood ratio for the a -th spatial range can be calculated as follows.

$$\Lambda_i(\Theta_a) = \frac{p_i(\mathbf{X}|H_1)}{p_i(\mathbf{X}|H_0)} = \frac{1}{1 + \xi_i(\Theta_a)} \exp\left(\frac{\gamma_i(\Theta_a) \xi_i(\Theta_a)}{1 + \xi_i(\Theta_a)}\right) \quad (9)$$

where

$$\xi_i(\Theta_a) = \lambda_i^S(\Theta_a) / \lambda_i^N(\Theta_a) \quad (10)$$

$$\gamma_i(\Theta_a) = P_i(\Theta_a) / \lambda_i^N(\Theta_a) \quad (11)$$

Equations (10) and (11) can be considered *a priori* and *a posteriori* SNRs for the a -th spatial range. $\lambda_i^N(\Theta_a)$ is estimated from the frames which include only background noise incorporating stationary directional noise. In addition, instead of estimating $\lambda_i^S(\Theta_a)$, *a priori* SNRs are estimated in a decision-directed manner. The decision-directed *a priori* SNRs are estimated as follows.

$$\hat{\xi}_i(\Theta_a) = \alpha \frac{\hat{P}_{i-1}(\Theta_a)}{\lambda_{i-1}^N(\Theta_a)} + (1 - \alpha) \cdot u_0[\gamma_i(\Theta_a) - 1] \cdot (\gamma_i(\Theta_a) - 1) \quad (12)$$

where $u_0[\cdot]$ indicates the unit step function. $\hat{P}_{i-1}(\Theta_a)$ is estimated by applying a minimum mean squared error estimator [22] to the spatial power distribution. Equation (12) can reduce the bias of the likelihood ratios to H_1 . Based on equations (6), (9), (11), and (12), the likelihood ratio $\Lambda_i(\Theta_a)$ is estimated. Finally, the speech activity is decided by thresholding the average log likelihood ratios as follows.

$$\log \Lambda_i = \frac{1}{A} \sum_{a=1}^A \log \Lambda_i(\Theta_a) \quad (13)$$

If speech is present in the observed signals, this average log likelihood ratio becomes large. As regards the hangover procedure, we employ the hidden Markov model based hangover scheme [13] in the proposed method.

Our proposed method can be considered an extension of conventional LRT based SAD methods [9][13] because the proposed method considers all spatial directions to avoid the restrictions imposed by the number of speakers and the precise estimation of speaker locations. On the other hand, the proposed method can be thought of as a simplification of the conventional method, because it considers only the spatial power distributions regardless of the detailed structure of the frequency spectrum. However, this simplification certainly helps to relax the restrictions. In this respect, the proposed method can be used for multi-party conversations where such restrictions are difficult to apply.

3. Experiment

3.1. Recordings of poster presentations

Poster presentations were recorded with a microphone array consisting of 8 omni-directional microphones as shown in Fig. 1. The microphone array was mounted at the top of the poster presentation frame to record the presentations with various modalities [4] as shown in Fig. 2. For this recording, there was one presenter and two visitors for each poster presentation. The presenter and visitors could move freely.

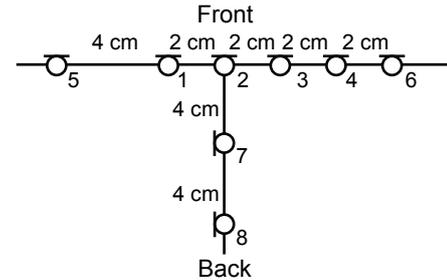


Figure 1: Microphone array for recording poster sessions. Numbers indicate microphone indices.



Figure 2: Poster presentation frame. Microphone arrays are set at the top of the frame.

Their initial recording locations are shown in Fig. 3. The microphones, amplifiers, and A/D converter were SONY ECM-77B, YAMAHA HA-8, and SDS DASBOX Model-500, respectively. The speech data were digitized at a sampling rate of 16 kHz with 24-bit quantization. Figure 4 shows an example of the recorded speech. The presence of environmental noise produced, for example, by computers in the room, people outside the room, and construction work under way outside the window, the SNRs for each microphone ranged from 0 to 6.5 dB.

Four recorded poster presentations were used in the following experiments. The speech onset/offset time stamps and speaker identifiers were hand labeled. The total data length for each presentation and total speaking time for each speaker are shown in Table 1. Owing to speech overlaps and non-speech periods, the total data length is not the same as the sum of all the speaking times for all the speakers.

3.2. Evaluation experiment

The performance of the proposed SAD method was evaluated by utilizing the recorded poster presentations shown in Table 1. The observed signals were analyzed by using 32 ms long Hanning windows with 16 ms overlaps. The space in front of the microphone array was split into 180 regions (i.e. $A = 180$) each 1 degree wide. The DOAs of the observed signals were estimated by using all the microphones, and time-frequency masks were applied to the spectrograms obtained at microphone 1 as shown in Fig. 1.

The DER, which is used for NIST Rich Transcription Meeting Recognition [3], is employed as a performance measure for this evaluation. The DER can completely account for the false-alarm speech time (FST), missed speech time

Table 1. Total data length and utterance length for each speaker. Speaker A is the poster presenter, and speakers B and C are visitors. The speakers are different for each session except for speaker A in sessions 1 and 2. Speech overlaps make that the total data length is not the same as the sum of the utterance lengths for each speaker.

	Speaker			Total
	A	B	C	
Session1	824.01	286.76	32.29	1036.89
Session2	788.98	129.45	129.01	913.42
Session3	1068.32	58.58	178.38	1149.86
Session4	1175.12	31.98	200.43	1290.66

(sec)

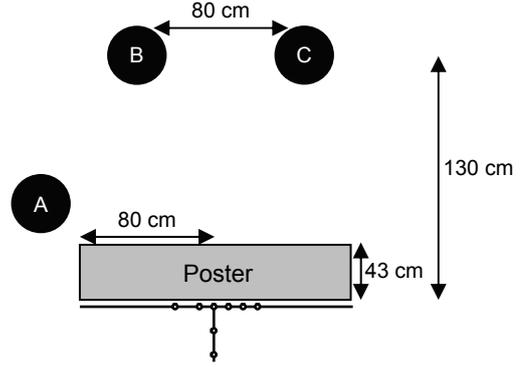


Figure 3: Approximate speaker positions. A, B, and C indicate the poster presenter and visitors. Participants were allowed to move freely.

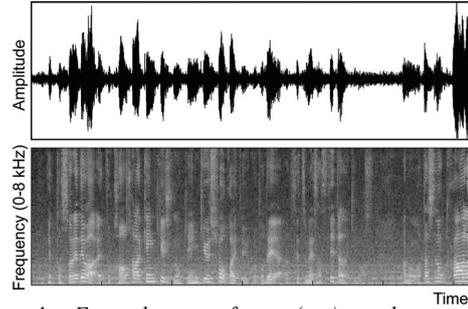


Figure 4: Example waveform (top) and spectrogram (bottom) for the recorded poster presentation.

(MST), and speaker error time (SET). When evaluating SAD, SET is not considered. The DER is calculated as follows.

$$DER = \frac{FST + MST}{Total\ length\ of\ data} \times 100\ (\%) \quad (14)$$

The evaluation criteria also follow those provided by NIST [3], that is, the speech segments were split with non-speech periods more than 300 ms in length, and vocalizations such as laughing and coughing are considered to be non-speech. The allowed tolerance for the difference between the system outputs and the correct labels was 250 ms.

To show the advantages of the proposed method in utilizing the spatial information, its performance was compared with that obtained using the frequency spectrum based statistical SAD method [13].

3.3. Results

Figure 5 shows examples of the spatial power distribution, *a priori* SNRs, and log likelihood ratios obtained with the proposed method. In this example, the speakers stand at around 60, 110, and 150 degrees. As shown in Fig. 5, the power of the environmental noise incorporating stationary directional noise was mitigated in the estimated *a priori* SNRs. In addition, because the log likelihood ratios are calculated by averaging for all spatial regions, they were unaffected by the speaker locations or the number of speakers, and exhibited large values when speech signals were present.

Table 2 shows the DERs obtained with the proposed method and the frequency spectrum based method [13]. For all sessions, the proposed method significantly outperformed the frequency spectrum based method in terms of DER. Because the noise power was widely distributed in the time-frequency region, the frequency spectrum based method was affected by the environmental noise. On the other hand, the proposed method could exploit the enhanced signals obtained

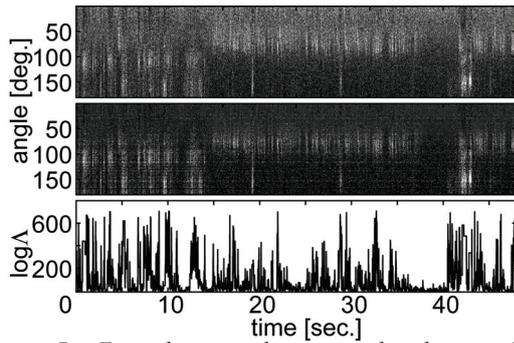


Figure 5: Example spatial power distribution (top), estimated *a priori* SNRs (middle), and log likelihood ratios (bottom) for the recorded poster presentation. White indicates the highest value and black the lowest value.

from time-frequency masking. This approach also works even in the presence of stationary directional noise sources by utilizing the *a priori* SNRs obtained from spatial power distributions.

4. Conclusion

We proposed an SAD method based on the LRT of the spatial power distribution. This method works in the presence of environmental noise by utilizing *a priori* SNRs of a certain spatial range. The spatial power distributions were estimated by applying time-frequency masking to the frequency spectrum. By considering the entire spatial range, the proposed method detects speech activity periods from observed signals regardless of the accuracy of the DOA estimation of the speaker locations. An experiment with recorded poster presentations showed that the proposed method outperformed the SAD method based on the LRT of the frequency spectrum in terms of DERs. This result confirmed the effectiveness of using spatial power distributions for SAD of such multi-party conversations as poster presentations.

It should be noted that this method does not consider whether or not the changes in power that come from a certain spatial range are caused by the presence of speech. Therefore, some mechanisms will be needed to cope with non-stationary directional noise sources. Moreover, the estimated log likelihoods for spatial power distributions can be utilized for the speaker diarization [5][6][7][8] of such multi-party conversations. The realization of such a speaker diarization method is also future work.

Acknowledgements

The authors thank Mr. Hisao Setoguchi and Mr. Katsuya Takanashi for their help in recording the poster presentations.

Table 2. Diarization error rate (DER), missed speech time (MST), and false alarm speech time (FST) obtained with the proposed and frequency spectrum based methods [13] for each session.

	Spatial power based SAD (proposed)			Frequency spectrum based SAD [13]		
	DER	MST	FST	DER	MST	FST
Session1	1.9	0.8	1.1	4.6	2.6	2.0
Session2	4.4	1.9	2.5	6.9	2.0	4.9
Session3	4.4	0.1	4.3	4.5	0.1	4.4
Session4	0.9	0.1	0.8	4.6	0.5	4.1

(%)

References

- [1] AMI Project: <http://corpus.amiproject.org/>
- [2] CHIL: <http://chil.server.de/>
- [3] NIST Rich Transcription Meeting Recognition: <http://nist.gov/speech/tests/rt/>
- [4] Setoguchi, H., Takanashi, K., and Kawahara, T., "Multi-modal conversational analysis of poster presentations using multiple sensors," Proc. ACM ICMI-2007 Workshop on Tagging, Mining and Retrieval of Human Related Activity Information, 44-47, 2007.
- [5] Tranter, S. E. and Reynolds, D. A., "An overview of automatic speaker diarization systems," IEEE Trans. on Audio, Speech, and Language Processing, 14, 1557-1565, 2006.
- [6] Anguera, X., Wooters, C., and Hernando, J., "Acoustic beamforming for speaker diarization of meetings," IEEE Trans. Audio, Speech, and Language Processing, 15, 2011-2022, 2007.
- [7] Macho, D., Padrell, J., Abad, A., Nadeu, C., Hernando, J., McDonough, J., Wölfel, M., Klee, U., Omologo, M., Brutti, A., Svaizer, P., Potamianos, G., and Chu, S., "Automatic speech activity detection, source localization, and speech recognition on the CHIL seminar corpus," Proc. ICME, 876-879, 2005.
- [8] Araki, S., Fujimoto, M., Ishizuka, K., Sawada, H., and Makino, S., "Speaker indexing and speech enhancement in real meetings / conversations," Proc. ICASSP, 93-96, 2008.
- [9] Potamitis, I. and Fishler, E., "Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays," J. Acoust. Soc. Am., 116, 2406-2415, 2004.
- [10] Davis, A., Low, S. Y., Nordholm, S., and Grbic, N., "A subband space constrained beamformer incorporating voice activity detection," Proc. ICASSP, 3, 65-68, 2005.
- [11] Hoffman, M. W., Li, Z., and Khataniar, D., "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," IEEE Trans. on Speech and Audio Processing, 9, 175-179, 2001.
- [12] Hioka, Y. and Hamada, N., "Voice activity detection with array signal processing in the wavelet domain," IEICE Trans. Fundamentals, E86-A, 2802-2811, 2003.
- [13] Sohn, J., Kim, N., and Sung, W., "A statistical model-based voice activity detector," IEEE Signal Processing Letters, 16, pp. 1-3, 1999.
- [14] Ramirez, J., Segura, J. C., Benitez, C., Garcia, L., and Rubio, A., "Statistical voice activity detection using a multiple observation likelihood ratio test," IEEE Signal Processing Letters, 12, 689-692, 2005.
- [15] Davis, A., Nordholm, S., and Togneri, R., "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," IEEE Trans. Audio, Speech, and Language Processing, 14, 412-424, 2006.
- [16] Capon, J., "High-resolution frequency-wavenumber spectrum analysis," Proc. IEEE, 57, 1408-1419, 1969.
- [17] Schmidt, R., "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas and Propagation, 34, 276-280, 1986.
- [18] Asano, F., Yamamoto, K., Ogata, J., Yamada, M., and Nakamura, M., "Detection and separation of speech events in meeting recordings using a microphone array," EURASIP Journal Audio, Speech, and Music Processing, 2007, Article ID 27616, 2007.
- [19] Yilmaz, Ö and Rickard, S., "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Processing, 52, 1830-1847, 2004.
- [20] Wang, D.-L. and Brown G. J. (eds.), *Computational Auditory Scene Analysis*, John Wiley and Sons, New Jersey, 2006.
- [21] Araki, S., Sawada, H., Mukai, R., and Makino, S., "DOA estimation for multiple sparse sources with normalized observation vector clustering," Proc. ICASSP, 5, 33-36, 2006.
- [22] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. on Acoustic, Speech, and Signal Processing, ASSP-32, 1109-1121, 1984.