ENSEMBLE INTEGRATION OF CALIBRATED SPEAKER LOCALIZATION AND STATISTICAL SPEECH DETECTION IN DOMESTIC ENVIRONMENTS

Yuuki Tachioka, Tomohiro Narita

Information Technology R&D Center
Mitsubishi Electric Corporation
Kamakura, 247–8501, Japan

Shinji Watanabe, Jonathan Le Roux

Mitsubishi Electric Research Laboratories
Cambridge, MA 02139, USA

ABSTRACT

This paper describes speaker localization and speech detection techniques for domestic environments. In real environments, it is hard to localize speakers because reverberation causes discrepancy from the simple spherical wave assumption. We propose a template-based method that calibrates the localization errors included in conventional methods. In addition, we use statistical speech detection methods to deal with noises. However, in this challenge, there are five rooms and leaked utterances from other rooms must be rejected. This kind of rejection is hard to perform by only using speech detection results. To address this problem, we also propose a method that integrates speech localization and speech detection using a minimum cost criterion or a classifier-based strategy. The proposed method achieved an accuracy of 0.712 for speaker localization and an F value of 0.743 for speech detection on the development set compared with the baseline 0.559 and 0.570, and 0.666 and 0.706 on the test set compared with the baseline 0.517 and 0.602.

Index Terms— Speaker localization, speech detection, calibration, rejection

1. INTRODUCTION

Speaker localization and speech detection are important and effective techniques for distant applications. One such application is automatic speech recognition using distant microphones, e.g., in home devices. Under such conditions, it is necessary to enhance the target speech. Although there are many 'blind' speech enhancement methods solely exploiting speech characteristics [1], the additional use of speakers' positions has been shown to improve robustness and effectiveness [2, 3] over blind approaches. For example, speaker localization techniques can effectively suppress directive noise.

The Distant-speech Interaction for Robust Home Applications (DIRHA) project [4] tackles the problem of distant speech interaction in home environments using multiple microphones. A challenge was derived from this project, comprising two major tasks: speaker localization and speech detection.

For speaker localization, speakers must be localized in 2D or 3D. It is fairly easy to determine the speaker direction only (1D). For example, the Cross Spectrum Phase (CSP) method [5] with prior distributions is shown in [6] to be effective even under noisy environments. However, 2D speaker localization is much harder than direction estimation, because it is susceptible to errors, but it is also more attractive. Recently, some 2D localization techniques have been proposed. Among them, the 2D-CSP method [7] is simple and effective. This method compares the observed time difference of arrivals (TDOAs) to the theoretical TDOAs for candidate points and picks up the point that achieves the smallest difference between them, but its performance degrades under reverberant environments because, due to reverberation, observed TDOAs do not match their theoretical TDOAs. To reduce the effect of these errors, some passive calibrations are needed [8]. We propose a template-based method that replaces the reference (theoretical) TDOAs by observed TDOAs for correct points to compensate the effect of discrepancy.

For speech detection, statistical methods [9, 10] have achieved great success. These methods are robust to noise. However, one difficulty of this challenge is that there are five rooms and the utterances from other rooms must be rejected. Speech detectors can discriminate speech from noise but cannot easily discriminate between speech from the target room and speech from other rooms. To address this problem, integration of speaker localization and speech detection is needed. We propose to utilize speaker localization results for speech detection through the use of either a minimum cost criterion or a classifier-based strategy.

This paper first describes the conventional 2D-CSP source localization method [7] and proposes a template-based method that calibrates errors in Section 3. Next, statistical speech detection methods [9, 10] are described in Section 4, and finally, an integration method of speaker localization and speech detection is described in Section 5. Experiments show that the proposed template-based method improves the localization performance and that our classifier-based strategy improves speech detection performance in Section 6.

2. SYSTEM OVERVIEW

Figure 1 shows a schematic diagram of the proposed system, which consists of a speaker localization part and a speech detection part. For the speaker localization part, M input pairs are selected from N microphone inputs and the corresponding M TDOAs τ are calculated by the CSP method. Comparing these TDOAs with the theoretical TDOAs, the 2D-CSP method outputs localized coordinates $\mathbf{s}$ with costs $P(\mathbf{s})$, and the template-based method compensates for errors using reference TDOAs. For the speech detection part, likelihood ratio approaches are adopted. Here, Sohn's method [9] and a switching Kalman filter based method [10] are used. Detections are done per microphone input, and the N detection results are combined using majority voting. In the real data, there are system replies between utterances. These replies are detected separately, and the corresponding utterances are deleted if they exist in the above detection results. Finally, the detection results are modified using a minimum cost criterion or a classifier-based strategy which combine costs P and average powers in each room.

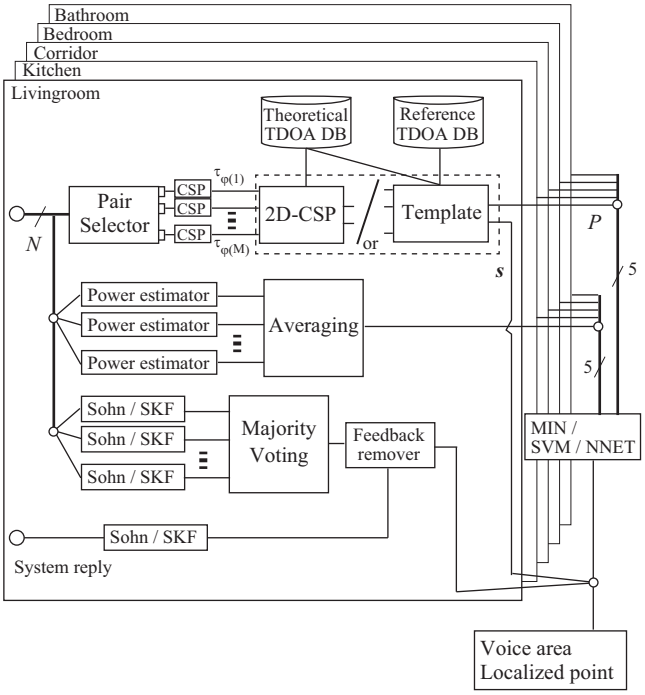

Fig. 1: Schematic diagram of the proposed system for the "Livingroom" localization and detection. (CSP: cross spectrum phase analysis, TDOA: time difference of arrival, Sohn: Sohn's speech detection, SKF: switching Kalman filter based speech detection, MIN: minimum cost criterion, SVM: support vector machine, NNET: neural network)

3. LOCALIZATION METHODS

3.1. 2D-CSP method

The original CSP method [5] only estimates the direction of arrival under the plane wave assumption. Under the condition that the microphones are distributed over a broad area, the source locations can be estimated using triangulation. On the other hand, the 2D-CSP method localizes speakers under the spherical wave assumption [7]. When the speaker position is $\mathbf{s}$ and the i^{th} microphone position among N microphones is $\mathbf{r}_i$, the theoretical TDOA τ_{ij}^{theo} between microphones i, j $(1 \leq i, j \leq N)$ is

$$\tau_{ij}^{theo}(\mathbf{s}) = \frac{|\mathbf{r}_i - \mathbf{s}| - |\mathbf{r}_j - \mathbf{s}|}{c}, \tag{1}$$

where c is the speed of sound. The CSP method estimates the TDOAs from the cross spectra of observed short-time Fourier transform coefficients $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ [5]. TDOA τ_{ij}^{csp} is obtained as the optimal solution of the problem:

$$\tau_{ij}^{csp} = \arg\max_{\tau} \left(\mathcal{F}^{-1} \left(\frac{\boldsymbol{X}_i \odot \boldsymbol{X}_j^*}{|\boldsymbol{X}_i||\boldsymbol{X}_j|} \right) \right), \tag{2}$$

where $\mathcal{F}$ is a short-time Fourier transform, and * and $\odot$ respectively denote the complex conjugate and the element-wise multiplication of two vectors.

For each candidate point $\mathbf{s}$ for a speaker, the cost function $P(\mathbf{s})$ is calculated by adding the difference between observed TDOAs τ^{csp} of M arbitrary pairs of microphones $(2 \leq M \leq {}_N C_2)$ and the corresponding theoretical ones τ^{theo}. If τ^{theo} is near to τ^{csp}, the cost function P will be small. The speaker position $\mathbf{s}$ is selected from the candidate points $\mathbf{S}$ as that which minimizes $P(\mathbf{s})$:

$$\arg\min_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}) = \arg\min_{\mathbf{s}\in\mathbf{S}} \sum_{m=1}^{M} \left| \tau_{\varphi(m)}^{theo}(\mathbf{s}) - \tau_{\varphi(m)}^{csp} \right|^2, \tag{3}$$

where $\varphi(m)$ is the m^{th} microphone pair. Because one microphone pair can only indicate that the sound source is located on a hyperbola, two and more different microphone pairs (i.e., three and more microphones) are needed to estimate the location.

3.2. Template-based method

In real situations, the theoretical TDOA and the observed TDOA for the correct position can differ, due for example to reverberation or measurement errors. The cost function P in Eq. (3) can be generalized, resulting in the following optimization problem:

$$\arg\min_{\mathbf{s}\in\mathbf{S}} P(\mathbf{s}) = \arg\min_{\mathbf{s}\in\mathbf{S}} \sum_{m=1}^{M} \left| \tau_{\varphi(m)}^{ref}(\mathbf{s}) - \tau_{\varphi(m)}^{csp} \right|^2, \tag{4}$$

where $\tau^{ref}(\mathbf{s})$ is a reference TDOA for position $\mathbf{s}$. In the 2D-CSP method, the theoretical TDOA is used as a reference, but observations generally contain errors ϵ, such that

$$\tau_{\varphi(m)}^{theo}(\mathbf{s}) \approx \tau_{\varphi(m)}^{csp} - \epsilon_{\varphi(m)}(\mathbf{s}). \tag{5}$$

To reduce the influence of errors, we propose a template-based method that modifies the reference TDOA $\tau_{\varphi(m)}^{ref}$ as in Eq. (6). These errors ϵ are calculated for every point $\mathbf{s} \in \mathbf{S}$ on the development set.

$$\tau_{\varphi(m)}^{ref}(\mathbf{s}) \approx \tau_{\varphi(m)}^{theo}(\mathbf{s}) + \epsilon_{\varphi(m)}(\mathbf{s}). \tag{6}$$

These modified references are expected to cancel out the errors.

4. SPEECH DETECTION METHODS

4.1. Conventional likelihood ratio test (Sohn's method)

One of the simplest and most effective conventional likelihood ratio test methods [9] is described here. Let $\boldsymbol{X} = \{X_k\}_{k=1}^{K_X}$ be the observed K_X-dimensional spectra. The power spectra $|X_k|^2$ are assumed to be independent conditionally on the noisy speech model λ^S in noisy speech frames (H_S) and on the noise model λ^N in non-speech frames (H_N):

$$\begin{aligned} p(\boldsymbol{X}|\lambda^S, H_S) &= \prod_{k=1}^{K_X} \frac{1}{\pi[v_k^S + v_k^N]} e^{-\frac{|X_k|^2}{v_k^S + v_k^N}}, \\ p(\boldsymbol{X}|\lambda^N, H_N) &= \prod_{k=1}^{K_X} \frac{1}{\pi v_k^N} e^{-\frac{|X_k|^2}{v_k^N}}, \end{aligned} \tag{7}$$

where v_k^S and v_k^N are the variance of speech and noise spectra, respectively. The log-likelihood ratio of speech and noise at the k^{th} dimension is then given by

$$\Lambda_k(X_k|\lambda^S, \lambda^N) = \ln \frac{p(X_k|\lambda^S, H_S)}{p(X_k|\lambda^N, H_N)}. \tag{8}$$

The geometric mean of the likelihood ratios is used to determine whether individual frames are speech or noise, as

$$\Lambda(\boldsymbol{X}|\lambda^S, \lambda^N) = \frac{1}{K_X} \sum_{k=1}^{K_X} \Lambda_k(X_k|\lambda^S, \lambda^N) \underset{H_N}{\overset{H_S}{\gtrless}} \eta, \tag{9}$$

where if $\Lambda(\boldsymbol{X}|\lambda^S, \lambda^N)$ is greater than some threshold η, the frame is considered to be in a (noisy) speech state, and otherwise in a noise state. The noise model is estimated in advance using observed noise, and the speech model is estimated by maximum likelihood estimation, i.e., $\partial\Lambda_k(X_k)/\partial\lambda_k^S = 0$, which results in the relationship $v_k^S = |X_k|^2 - v_k^N$. This shows that the speech model λ_k^S is estimated assuming that the speech and noise powers are additive.

4.2. Switching Kalman filter based method

The state-of-the-art switching Kalman filter based speech detection method [10] builds the noisy speech model frame by frame, from a prepared clean speech model and a noise model which is estimated online. The features considered are the K_Y-dimensional log-Mel spectra $\boldsymbol{Y} = \{Y_k\}_{k=1}^{K_Y}$. In the log-Mel domain, the observed features of speech can be represented as a logarithmic summation of those of clean speech and noise. The likelihoods under the noisy speech and the noise models are each given through a Gaussian mixture model (GMM) whose components are updated by switching Kalman filters. The likelihood ratio calculation is performed in the same way as in Eqs (8) and (9), replacing the Gaussians on X_k by the GMMs on Y_k.

5. INTEGRATION OF LOCALIZATION AND SPEECH DETECTION

In this challenge, the utterances from other rooms must be rejected. We propose to use localization results in the other rooms to do so.

5.1. Minimum cost criterion

Our first approach is to compare the localization cost P in the target room P_{in} with those in the other rooms P_{out}. If a speaker is localized in multiple rooms, selecting the speaker location which results in the minimum cost across rooms appears to be the most reasonable. However, simple comparisons lead to many false rejections, because the cost features are dependent on the room shape and microphone settings and thus cannot be simply compared. We thus introduce a tolerance parameter η', and for each frame, set a flag f indicating whether the frame's cost is close to being the smallest among all rooms:

$$f = \begin{cases} \text{true} & \forall P_{out}, P_{in} < \eta' P_{out} \\ \text{false} & \text{otherwise} \end{cases}$$

For each utterance, if the ratio of the number of true flags to the total number of frames is under some thresholds, the utterance is rejected.

5.2. Classifier-based strategy

In a second approach, we use a classifier $\mathcal{C}$ whose input is a concatenated vector of features from the target room $\mathbf{z}_{in}$ and features from the other rooms $\mathbf{z}_{out}$. After training the classifier on the development set, the classifier outputs are compared with a threshold η'' to estimate flags for utterance and each frame, as:

$$f = \begin{cases} \text{true} & \mathcal{C}([\mathbf{z}_{in}; \mathbf{z}_{out}]) > \eta'' \\ \text{false} & \text{otherwise} \end{cases}$$

These flags are then combined as in 5.1 to determine whether to reject the utterance.

6. EXPERIMENTAL SETUP

6.1. Database description

Synchronously recorded sound files (approximately 1-2 minutes) were provided by the DIRHA consortium. For simulating realistic environments, these databases were recorded in a real house, which consisted of five rooms: Kitchen, Livingroom, Corridor, Bathroom, and Bedroom. Localization and speech detection were limited to the Kitchen and Livingroom. For Kitchen and Livingroom, six-microphone circular microphone arrays were installed at the center of the room. Additionally, for all rooms, several two- or three-microphone arrays were installed on the walls encompassing the room. In total, 40 microphones were used. Microphone pairs were selected within each array, because microphones belonging to separate arrays, were far away and their correlations were too small.

A development set (**dev**) and a test set (**test**) were provided. According to the regulations, any parameter can be tuned on the **dev** set. Both sets consist of REAL and SIMULATIONS subsets. In the REAL set, for each task, there is only one speaker in one room, moving around the room. To simulate the dialog between speaker and system, system replies sometimes break in, but they are provided separately. In the SIMULATIONS set, there can be multiple speakers speaking in different rooms, but the speakers are still. System performance was evaluated using the provided evaluation tools.

6.2. Localization

Because height localization is less important than horizontal localization, we focused on the 2-D localization. (the `-2D` option was used for the evaluation tool.) The speech data were down-sampled from the original 48 kHz to 16 kHz for our experiments. The frame size was 960 and the frame shift was 800. We compared the performances of the 2D-CSP and the proposed template-based method with those of the multi-channel CSP method [11] and the SRP-PHAT[1] [12] with a long frame size (1 second). Fine errors were defined as localization errors less than 50 cm.

6.3. Speech detection

The speech detection performance was evaluated per utterance in terms of precision, recall, and F value. The frame size was 960 and the frame shift was 160 (with 16 kHz sampling). The maximum silence duration in utterances and minimum duration of utterances were set to 500 ms and 300 ms, respectively. For SKF, the number of Gaussian mixture components was 32, and 20-dimensional Mel-spectra were used. HMM hangover scheme [9] was used for both methods. After performing speech detection per file, majority voting was used to obtain the final speech detection results per room.

6.4. Integration

Localization costs P and segmental speech powers averaged over microphones in each room were used as the features $\mathbf{z}_{in}$ and $\mathbf{z}_{out}$. For the classifier-based strategy, we used SVM-light (v.6.02)[2] for support vector machine (SVM) based classification (linear SVM) and pyBrain (v.0.31)[3] for neural network (NNET) based classification, after normalizing the features to have unit variance. SVM and NNET were trained using binary outputs indicating whether the

[1] http://www.lems.brown.edu/array/tools/srplems.m

[2] http://svmlight.joachims.org/

[3] http://pybrain.org/

Table 1: Localization and speech detection results on the development set (**dev**). Methods are indicated for speech activity detection (SAD), source localization (LOC), and their integration (INT). Performance criteria for source localization are Fine Error (FE), Gross Error (GE), and Percentage of Correct localization (PCor). For speech detection, utterance-based criteria are used: Precision (P), Recall (Re), and F value.

Methods			REAL						SIMULATIONS						AVERAGE					
SAD	LOC	INT	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F
Oracle	2D-CSP	-	298	602	.685	-	-	-	309	925	.504	-	-	-	306	870	.540	-	-	-
	Template		303	592	.719	-	-	-	160	864	.643	-	-	-	200	817	.658	-	-	-
	M-CSP		347	1307	.177	-	-	-	348	1433	.208	-	-	-	348	1409	.202	-	-	-
	SRP-PHAT		*289*	826	.537	-	-	-	248	987	.509	-	-	-	257	957	.515	-	-	-
Sohn	2D-CSP	-	295	565	.709	.693	**.957**	.804	308	836	.525	.354	**.905**	.509	305	794	.559	.414	**.919**	.570
	Template	-	301	537	.746	.693	**.957**	.804	161	769	.657	.354	**.905**	.509	197	732	.673	.414	**.919**	.570
		MIN	301	537	.748	.744	**.957**	**.837**	161	769	.657	.354	**.905**	.509	197	732	.673	.419	**.919**	.575
		SVM	304	528	.757	.740	.826	.781	159	749	.681	.670	.836	.744	197	714	.695	.689	.833	.754
		NNET	299	**498**	**.779**	.797	.826	.811	151	732	.685	**.800**	.693	.743	193	692	.704	**.799**	.729	**.762**
SKF	2D-CSP	-	300	559	.699	.697	.812	.750	303	798	.548	.416	.894	.568	302	762	.574	.461	.872	.603
	Template	-	306	532	.744	.697	.812	.750	158	714	.678	.416	.894	.568	194	686	.689	.461	.872	.603
		MIN	306	528	.752	.699	.768	.732	158	709	.679	.414	.889	.565	194	682	.692	.457	.857	.596
		SVM	310	535	.741	.823	.783	.802	157	688	.699	.661	.841	.740	196	663	.707	.694	.826	.754
		NNET	**292**	503	.756	**.837**	.609	.705	**149**	**663**	**.704**	.733	.778	**.755**	**180**	**642**	**.712**	.753	.733	.743

Table 2: Localization and speech detection results on the test set (**test**).

Methods			REAL						SIMULATIONS						AVERAGE					
SAD	LOC	INT	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F	FE	GE	PCor	P	Re	F
Oracle	2D-CSP	-	301	622	.582	-	-	-	302	1076	.461	-	-	-	302	965	.497	-	-	-
	Template		297	584	.658	-	-	-	186	1094	.564	-	-	-	228	972	.592	-	-	-
Sohn	2D-CSP	-	298	585	.610	.868	.962	.913	303	1004	.479	.368	**.944**	.530	302	904	.517	.441	**.949**	.602
	Template	-	293	550	.673	.868	**.962**	.913	185	969	.590	.368	**.944**	.530	225	870	.613	.441	**.949**	.602
		MIN	293	545	.677	.882	**.962**	**.920**	186	970	.591	.365	.934	.525	225	868	.616	.441	.942	.600
		SVM	299	**505**	.678	**.917**	.316	.470	185	961	.592	.678	.939	**.788**	204	920	.602	.700	.762	**.730**
		NNET	287	542	.657	.900	.532	.668	**178**	969	.567	**.720**	.707	.714	211	889	.588	**.755**	.657	.703
SKF	2D-CSP	-	296	846	.624	.657	.937	.772	304	922	.526	.411	.859	.556	301	823	.557	.462	.881	.606
	Template	-	292	513	.683	.657	.937	.772	184	859	.637	.411	.859	.556	225	768	.651	.462	.881	.606
		MIN	292	512	**.684**	.651	.937	.768	184	857	.639	.411	.843	.553	225	766	.653	.461	.870	.602
		SVM	299	518	.668	.571	.367	.447	180	838	.644	.684	.813	.734	**203**	798	.647	.664	.686	.675
		NNET	**284**	507	.662	.692	.608	.647	187	**768**	**.667**	.712	.742	.727	215	**710**	**.666**	.707	.704	.706

source was in the target room or not. Parameters and thresholds for SVM and NNET were tuned using the **dev** set. For NNET, the number of hidden layers was two and the number of nodes in the hidden layers was 15 and 10 from the bottom. Finally, for REAL, the speech powers of the detected utterances in Livingroom and Kitchen were compared and only the highest one was used because there can be an active speaker only in one room.

7. RESULTS

7.1. Localization accuracy with oracle speech detection

To compare the localization accuracies among the above-mentioned methods, the first parts of Tables 1 and 2 show the results for oracle speech detection cases. The performance of the 2D-CSP method was higher than those of the multi-channel CSP and SRP-PHAT method. Moreover, the computational complexity was much smaller than those of the multi-channel CSP and SRP-PHAT method. We thus adopted the 2D-CSP method as a baseline. The performance of the template-based method was better than that of the 2D-CSP method significantly, proving effective for the localization in domestic environments.

7.2. Speech detection accuracy

The second and third parts of Tables 1 and 2 show the results with speech detection. The performance of SKF was slightly higher than that of Sohn's method. However, neither method by itself was very effective in rejecting noises or leaked utterances from the other rooms. Integration with localization proved effective, but only for the classifier-based strategy. As the classifiers are trained on **dev** data, we compare the results on the **test** set. The performance of the minimum cost criterion was equivalent to that of the baseline. SVM significantly improved the F value, especially with Sohn's method, while NNET improved the F value more consistently with Sohn's method and SKF.

8. CONCLUSIONS

We have introduced an effective template-based method that can compensate the discrepancy between the simple spherical wave assumption and the observations, and showed its effectiveness for real domestic environments. In addition, to reject utterances that cannot be easily rejected only by speech detection, we proposed to integrate speaker localization and speech detection. Doing so using classifiers such as SVMs and neural networks improved the speech detection performance.

9. REFERENCES

[1] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*, pp. 509–519. Springer, 2008.

[2] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba, Y. Kubo, M. Souden, S.J. Hahm, and A. Nakamura, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech and Language*, vol. 27, pp. 851–873, 2013.

[3] Y. Tachioka, S. Watanabe, J. Le Roux, and J. Hershey, "Discriminative methods for noise robust speech recognition: A CHiME challenge benchmark," in *Proceedings of the 2nd International Workshop on Machine Listening in Multisource Environments*, 2013, pp. 19–24.

[4] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," in *Proceedings of LREC*, 5 2014.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 8 1976.

[6] Y. Tachioka, T. Narita, and T. Iwasaki, "Direction of arrival estimation by cross-power spectrum phase analysis using prior distributions and voice activity detection information," *Acoustical Science & Technology*, vol. 33, pp. 68–71, 1 2012.

[7] D.V. Rabinkin, R.J. Renomeron, A. Dahl, J.C. French, J.L. Flanagan, and M.H. Bianchi, "A DSP implementation of source location using microphone arrays," in *Proceedings of SPIE*, 1996, pp. 88–99.

[8] K. Ho and L. Yang, "On the use of a calibration emitter for source localization in the presence of sensor position uncertainty," *IEEE Transactions on Signal Processing*, vol. 56, pp. 5758–5772, 2008.

[9] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1 1999.

[10] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," *IEICE Transactions on Information and Systems*, vol. E91-D, pp. 467–477, 3 2008.

[11] K. Hayashida, M. Morise, and T. Nishiura, "Near field sound source localization based on cross-power spectrum phase analysis with multiple channel microphones," in *Proceedings of INTERSPEECH*, 9 2010, pp. 2758–2761.

[12] H. Do, H. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction(src) on a large-aperture microphone array," in *Proceedings of ICASSP*, 4 2007, vol. 1, pp. 121–124.