INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Factorial Modeling for Effective Suppression of Directional Noise

Osamu Ichikawa[1], Takashi Fukuda[1], Gakuto Kurata [1], Steven J. Rennie[2]

[1]Watson Multimodal, IBM, Tokyo 103-8510, Japan
[2]Watson Multimodal, IBM, Yorktown Heights, NY 10598, USA
[1]{ichikaw, fukuda1, gakuto}@jp.ibm.com, [2]sjrennie@us.ibm.com

Abstract

The assumed scenario is transcription of a face-to-face conversation, such as in the financial industry when an agent and a customer talk over a desk with microphones placed between the speakers and then it is transcribed. From the automatic speech recognition (ASR) perspective, one of the speakers is the target speaker, and the other speaker is a directional noise source. When the number of microphones is small, we often accept microphone intervals that are larger than the spatial aliasing limit because the performance of the beamformer is better. Unfortunately, such a configuration results in significant leakage of directional noise in certain frequency bands because the spatial aliasing makes the beamformer and post-filter inaccurate there. Thus, we introduce a factorial model to compensate only the degraded bands with information from the reliable bands in a probabilistic framework integrating our proposed metrics and speech model. In our experiments, the proposed method reduced the errors from 29.8% to 24.9 %.

Index Terms: microphone array, post-filtering, beamformer, speech recognition, factorial model

1. Introduction

Recently, the accuracy of automatic speech recognition (ASR) has been much improved by using deep neural networks. However, it is still insufficient to transcribe casual conversations, especially those in cocktail parties. On the other hand, ASR may find more business opportunities in less severe conditions, such as conversations between two people on limited topics, for example, conversations at a teller counter in a bank, in a medical exam room, at a pharmacy counter, and in a municipal office.

In these situations, the speech from each person must be transcribed separately. This diarization is difficult with a single microphone; however, it is relatively easy with multiple microphones because the direction of arrival (DOA) can provide cues. We can place microphones in parallel with two speakers such that, for example, the agent is on the left and the customer is on the right.

If the left speaker is the target of transcription, the right speaker is considered to be an interfering speaker or a directional noise source. To suppress noise, various beamforming technologies have been studied. A delay and sum beamformer (DSBF) focuses in the target direction by synchronizing channels. As for an adaptive beamformer, a minimum variance distortionless response beamformer (MVDR) [1] was proposed to minimize ambient noise while maintaining a constant gain in the target direction. Blind signal separation (BSS) based on independent component analysis (ICA) was also proposed; however, its performance is similar to that of MVDR [2] when the sound sources are correctly localized. For extra noise reduction, subtraction of estimated noise with a sub-beamformer was explored [3][4]. For decomposition, non-negative matrix factorization (NMF) [5] was enhanced as multi-channel NMF [6].

These speech separation methods have various pros and cons. One shared aspect is that they may have a spatial aliasing problem when the microphone interval is too large. In this case, the sounds of the directional noise cannot be distinguished from the sounds from the target direction in terms of the phase difference between the microphones. Because of this, the methods cannot completely eliminate sounds from non-target directions when using a small number of microphones and leave residual speech as by-products. Unfortunately, ASR is often more sensitive to such residual speech than humans are. Therefore, a practical approach is to combine beamformer technology with other techniques to enhance the suppression performance.

Binary selection is one such technique. It includes two types of masking based on inter-channel amplitude difference [7] or inter-channel phase difference [7][8]. The “amplitude” approach requires unidirectional microphones. The problem is that they detect a flipped phase for a sound coming from the opposite side. Therefore, they cannot be used as an adaptive beamformer in our configuration. On the other hand, the “phase” approach can use omnidirectional microphones; however, the problem with this method is that the microphone interval is theoretically limited.

Another technique is post-filtering. This technique essentially uses the correlation between channels. The basic form is the Zelinski post-filter [9], which assumes that noise is uncorrelated between channels. Therefore, it was not designed for directional noise cases. The post-filtering technique has been further enhanced [10][11][12] and is also known as a multi-channel Wiener filter. To work with spatially correlated noise components, a post-filter was combined with a generalized sidelobe canceller (GSC) [13] and a second post-filter [14]. However, there have been only a few attempts [15] that explicitly coped with spatial aliasing between the target speaker and the directional noise source.

In this paper, we assume that only a few microphones are available because a large-scale microphone array with many elements is expensive. Also, if the software processing is performed on a cloud system, transmission of too many channels may overload the intranet infrastructure.

Our previous study [15] showed that an adaptive beamformer recovers the original speech from the mixed speech signal better when the microphone interval is set to a larger distance, regardless of the spatial aliasing limit. In fact, many retail small-scale microphone arrays have larger

http://dx.doi.org/10.21437/Interspeech.2017-852

microphone intervals with a risk of spatial aliasing. In such a configuration, the adaptive beamformer and post-filter (Zelinski type) may have residual components from the directional noise source in the frequency bands where the spatial aliasing occurs. Our previous study proposed an aliasing-aware post-filter mainly to suppress noise in alternate speech segments; however, it did little for mixed speech segments. This is the motivation for this study.

Because the degradation only occurs in certain frequency bands, it is reasonable to compensate the degraded bands with information from reliable bands. This approach has been widely explored in previous studies on missing features [16][17][18]. In Section 2.2, we introduce a unique metric to account for spatial aliasing. It is integrated with a factorial model as a confidence metric as described in section 2.4.

2. Proposed Method

2.1. System configuration

The MVDR-based adaptive beamformer and Zelinski post-filter are introduced to reduce the directional noise from the interfering speaker. They are followed by a novel part of factorial modeling. For this model, the signal and metrics should be converted with Mel-filter banks because the model is designed in the log-Mel domain.

This system uses two DOA indices: one for the target speaker (which is to be searched for on the left side of the microphones) and another for the interfering speaker on the right side. The microphones are placed in parallel with the speakers. To detect the DOAs, cross-spectrum phase (CSP) analysis (also known as GCC-PHAT) is performed for the two channel inputs of all the microphone pairs.

2.2. Aliasing metric

Figure 1 shows an example of Zelinski post-filter values in a frame when only the interfering speaker speaks. Because the target speaker does not speak, all the values are expected to be zeros, yet we observe several bands that are not close to zero. Therefore, the output of a Zelinski post-filter cannot completely eliminate the interfering speech. This problem also occurs with MVDR.

It is caused by spatial aliasing. If we know the DOAs for both the target speaker and the interfering speaker (i.e. the directional noise source), we can estimate which bands are susceptible to interference. For this purpose, we introduce a new metric $E_{m,n}$ for the microphone pair of *m* and *n* as

$$E_{m,n}(q,T)=\cos\left(2\pi\cdot q\cdot\left(\hat{i}_{m,n}(T)-\hat{j}_{m,n}(T)\right)/N\right), \qquad (1)$$

where $\hat{i}_{m,n}$ is the DOA index for the target speaker for the microphone pair of *m* and *n*, as determined by CSP analysis. *N* is the discrete Fourier transform (DFT) size, and $\hat{j}$ is the DOA index for the interfering speaker. As shown in Figure 2, if $E(q)$ is close to one, the observed phase at bin *q* of the sound from the target direction is hardly distinguishable from the one from the interfering speaker direction. That means the output of the post-filtering (and also that of MVDR) is unreliable for that bin. In Figure 1, an example of the aliasing metric is shown by the dashed line. In this example, H' often takes non-zero values when the aliasing metric takes higher values.

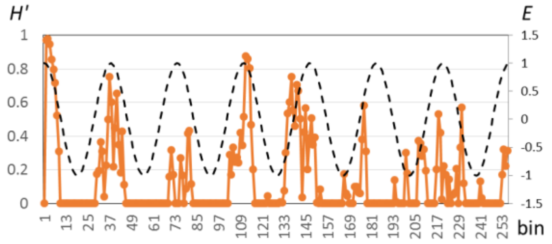

Figure 1: *Example of Zelinski post-filter values H' (non-negative part) when interfering speaker talked and target speaker was silent. Aliasing metric (E) is also plotted (dashed line). Two-microphone system was used for this example.*

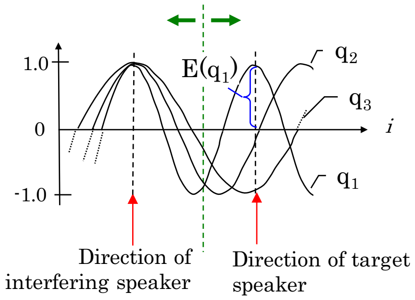

Figure 2: *Aliasing metric.*

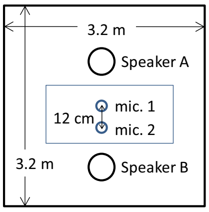

Figure 3: *Recording environment.*

When there are more than two microphones, $E_{m,n}$ should be aggregated for all the microphone pairs as

$$E(q,T)=\frac{2}{M(M-1)}\sum_{m=1}^{M-1}\sum_{n=m+1}^{M}E_{m,n}(q,T), \qquad (2)$$

for the *q*-th bin. This is further converted into the Mel domain as

$$e_d=\left[\sum_{q}\left\{\max(0,E(q))B_{d,q}\right\}\right]\Big/\sum_{q'}B_{d,q'}, \qquad (3)$$

where $B_{d,q}$ is the weight of the *d*-th Mel-filter bank for the *q*-th bin and for the *d*-th band. Then, e is referred to as the aliasing metric.

2.3. Spectral density metric

We also obtain a Mel-filtered version of the Zelinski post-filter transfer function H as

$$v_d=\max\left(0,\ \sum_{q}H(q)\cdot B_{d,q}\Big/\sum_{q'}B_{d,q'}\right), \qquad (4)$$

where v is regarded as a spectral density metric.

It has been suggested that a large value indicates that the band in the frame must be dominated by the target speech signal [17]. However, this is not always true when spatial aliasing occurs. Therefore, this metric is combined with the aliasing metric in the factorial model.

2.4. Factorial model

The factorial model is designed as the product of Gaussians. We use two-factor formulation to integrate the metrics. It is similar to that used to capture harmonic information [19]. The first factor represents a compensation vector referring to a clean speech GMM (Gaussian mixture model). The second factor is designed as a Gaussian whose mean is zero that may make the compensation vector null. Thus, depending on the balance of the two factors, the extent of the compensation can be controlled. It is implemented in a probabilistic framework based on the confidence.

Our goal is to compensate the Zelinski post-filtered output y as

$$x = y - g, \tag{5}$$

where x is the estimated clean speech of the target speaker. g accounts for the residual noise component. It is given by VTS (vector Taylor series) formulation for each band d using the mismatch function G as

$$g_d = \mathrm{G}(x,n)_d = \log\left(1+\exp\left(n_d - x_d\right)\right). \tag{6}$$

where n is the residual noise of the interfering speaker. For this purpose, we first calculate the probability distribution $p(g|y,n,e)$. Then, we obtain the estimated compensation vector $\hat{g}$ by using the minimum mean square error (MMSE).

First, we approximate the probability by the following factorial model in the log-Mel spectrum domain,

$$p(g|y,n,e,v) \propto p(g|y,n) \cdot p(g|e,v). \tag{7}$$

Here, the frame index T is omitted.

The first model $p(g|y,n)$ is designed to output the VTS compensation. It is a GMM whose k-th component has a higher probability around $G(\mu_{x,k},\mu_n)$. It refers to a clean speech GMM with mean $\mu_{x,k}$, variance $\Sigma_{x,k}$, and prior probability γ_k for the k-th Gaussian. We introduce a wild assumption that the residual noise of the interfering speaker is characterized as a single Gaussian with mean μ_n and variance Σ_n, which are measured in noise segments.

If Equation (7) does not have the second model $p(g|e,v)$, it works as full compensation of VTS. Because this paper approximates residual speech with a single Gaussian, its entire application may hurt the output. It should only be applied to the band at the frame in which the observation is not reliable. In such a background, the second model is introduced. It is designed to have a peak around zero. Zero means no compensation. The strength of the second model is controlled by manipulating the model variance.

The first model $p(g|y,n)$ is written by using VTS as

$$p(g|y,n) = \sum_k^K \rho(y)_k \mathrm{N}\left(g;\mu_{g,k},\Sigma_{g,k}\right), \tag{8}$$

$$\mu_{g,k} = G\left(\mu_{x,k},\mu_n\right), \tag{9}$$

$$\Sigma_{g,k} \cong F\left(\mu_{x,k},\mu_n\right)^2 \cdot \left(\Sigma_n - \Sigma_{x,k}\right) \tag{10}$$

where ρ_k is the posterior probability. F is an auxiliary function defined for each band as

$$F(x,n)_d = \left(1+\exp\left(x_d - n_d\right)\right)^{-1}. \tag{11}$$

As for the second model $p(g|e,v)$, our approach designs it as a Gaussian associated with the k-th component:

$$p(g|e,v)_k = \mathrm{N}\left(g;0,\psi(e,v)_k\right). \tag{12}$$

Therefore, the product in Equation (7) is actually taken for each component of the GMM.

The variance ψ should have a small value when e is small and v is large. When the variance becomes small, the probability distribution has higher probability around the mean value, namely zero. We calculate the variance ψ by scaling the variance of the k-th Gaussian at the d-th band in the speech model. Firstly, the geometric average of v and (1-e) are taken and processed with sigmoid function as

$$\theta_d = 1/\left(1+\exp\left(-\alpha\left(\sqrt{v_d(1-e_d)}-\beta\right)\right)\right), \tag{13}$$

where α and β are constants for the sigmoid and are set to 5.0 and 0.5, respectively. If we only use the aliasing metric, this can be written as

$$\theta_d = 1/\left(1+\exp\left(-\alpha\left((1-e_d)-\beta\right)\right)\right), \tag{14}$$

Then, the variance of the second model is obtained as

$$\psi_{k,d} = \Sigma_{x,k,d} \cdot \left(\theta_d^{-1} - \lambda\right), \tag{15}$$

where ε is a very small constant and λ is a constant (set to 0.5 in our experiments). These values are derived from our previous study [19]. Then, $p(g|y,n,e,v)$ can be written as a single GMM:

$$\begin{aligned} p(g|y,n,e,v) &= \sum_k^K Z_k^{-1}\rho(y)_k \cdot\mathrm{N}\left(g;\mu_{g,k},\Sigma_{g,k}\right)\cdot\mathrm{N}\left(g;0,\psi(e,v)_k\right) \\ &= \sum_k^K Z_k^{-1}\rho(y)_k \cdot\mathrm{N}\left(g;\mu''_{g,k},\Sigma''_{g,k}\right). \end{aligned} \tag{16}$$

Here, Z is a normalizing constant. The variances and means for z are given respectively by

$$\Sigma''_{g,k} = \left(\Sigma_{g,k}{}^{-1} + \psi_k{}^{-1}\right)^{-1}, \tag{17}$$

$$\mu''_{g,k} = \Sigma''_{g,k}\left(\Sigma_{g,k}{}^{-1}\mu_{g,k} + \psi_k{}^{-1}0\right). \tag{18}$$

We used a diagonal approximation in the clean speech model. The posterior probabilities are then given by

$$\rho(y)_k = \frac{\gamma_k \cdot \mathrm{N}\left(y;\mu''_{y,k},\Sigma''_{y,k}\right)}{\sum_{k'}\gamma_{k'}\cdot\mathrm{N}\left(y;\mu''_{y,k'},\Sigma''_{y,k'}\right)}, \tag{19}$$

where

$$\Sigma''_{y,k} = \left(\Sigma_{y,k}{}^{-1} + \psi_k{}^{-1}\right)^{-1}, \tag{20}$$

$$\mu''_{y,k} = \left(\Sigma_{y,k}{}^{-1}\cdot\mu_{y,k} + \psi_k{}^{-1}\cdot\mu_{x,k}\right)\cdot\Sigma''_{y,k}, \tag{21}$$

$$\mu_{y,k} \cong \mu_{x,k} + G\left(\mu_{x,k},\mu_n\right), \quad \text{and} \tag{22}$$

$$\Sigma_{y,k} \cong \left\{1 - F\left(\mu_{x,k},\mu_n\right)\right\}^2 \cdot \Sigma_{x,k} + F\left(\mu_{x,k},\mu_n\right)^2 \cdot \Sigma_n. \tag{23}$$

The compensation term is estimated with the MMSE as

$$\hat{g} = \int g \cdot p(g|y,n,e,v)dg \cong \sum_k^K \rho(y)_k \cdot \mu''_{g,k}. \tag{24}$$

According to (20), the variance used for the posterior probability $\Sigma''_{y,k,d}$ becomes smaller than the original variance $\Sigma_{y,k,d}$ for the d-th band when the band is reliable, where aliasing metric e_d is close to zero, and spectral density metric v_d is large. This makes the d-th band Gaussian more sensitive, thus contributing more to the posterior estimation. On the other hand, a band with low confidence has a larger variance, thus making less contribution to the posterior estimation in Equation] (19). Also, for such a degraded band, the MMSE output has more compensated value via the interpolated mean of Equation (18), where $\psi_{k,d}$ becomes large.

3. Experiments

In this paper, an evaluation was performed with two microphones in a small quiet meeting room, as shown in Figure 3. Two omnidirectional microphones were placed on a table between two speakers acting as an agent and a customer. The distance between the microphones was 12 cm. Three sessions were recorded with four different male speakers. They made simulated financial conversation in Japanese consisting of 134 utterances in total. The beamformer operated at a 16-kHz sampling frequency. Then, we simulated mixed speech data using this recording. For the agent speech testing, some parts of the customer speech were selected and then added over the entire recording in stereo. We performed the same task for the customer data. Table 1 shows the SNRs (signal-to-noise ratios) for the recordings and the simulated mixed speech data.

Because we focus on mixed speech, segments in which only the interfering speaker was active were not decoded by ASR. This corresponds to the use of DOA-based VAD (voice activity detection). In this paper, the metric to evaluate the performance of the proposed method is the ASR error rate. The acoustic model is a CNN-based quin-phone model trained with several hundred hours of speech data recorded in clean and noisy environments. The number of context-dependent phones is 7000. The input feature consists of 40-dimension log-Mel spectra, their delta, and delta-delta. The language model is a general purpose tri-gram model with a few hundreds of thousand vocabularies.

For evaluation of the factorial model, a clean speech GMM with 256 prototypes was trained in a 40-dimensional log-Mel domain with the clean speech data in CENSREC-3 (condition 4) [20], which were recorded with a far-field microphone in a stationary car. Because the channel characteristic is not compatible with our environment, the means of the Gaussian were shifted with channel bias estimated with the clean recordings of each session.

The DOAs' accuracy is important for better system performance. In this experiment, the two DOAs were tracked through the session. They were updated only when a CSP peak was observed on their respective sides.

Table 2 summarizes the experimental results. Case 0 is for reference (without overlapping). This result can be considered as an upper bound on the system performance. Case 1 is the baseline using the single microphone nearest the speaker. The error rate was quite high due to the interfering speech. Case 2 uses the simple MVDR system. It showed much improvement, but the error rate was still high. In Case 3, the Zelinski post-filter was applied to the MVDR output. We obtained additional gain with the post-filter. Case 4 performed model-based compensation for the output of Case 3. This is equivalent to implementing only the first model of Equation (7). Because this scheme performs VTS compensation for the entire output while the compensation involves approximation, it showed a small degradation compared to Case 3. Case 5 is a sub-set of the proposed system performing factorial modeling for the output of Case 3, but it uses only the aliasing metric with Equation (14). It reduced the errors by 11.7 % compared to Case 3 in favor of the factorial modeling. Case 6 is the full proposed system using both metrics. It further reduced the errors by 16.4 % compared to Case 3. Because Case 5 and Case 6 performed VTS compensation only for the necessary part in time-frequency space using information from reliable bands, they outperformed Case 4.

Table 1. *SNR in dB of recordings (clean) and simulated mixed speech (overlapped). It is averaged for left and right channels. 'A', 'B', 'C' and 'D' are speaker IDs.*

	Session 1		Session 2		Session 3	
	A	B	B	A	C	D
Clean	12.1	14.4	12.2	10.5	14.9	15.1
Overlapped	-2.1	2.1	2.1	-2.1	-0.6	0.6

Table 2. *Experimental results consisting of character error rate (CER) [%]. S, I, and D indicate counts of substitution, insertion, and deletion errors, respectively. T indicates the total number of reference characters. Because of ambiguity of Japanese word boundaries, CER was used instead of word error rate. FM denotes factorial modeling.*

System	S / I / D / T	CER
Case 0) Oracle	115 /57 /77 /2161	12.4
Case 1) Baseline	1125 /355 /247 /2161	87.4
Case 2) MVDR	365 /172 /211 /2161	39.6
Case 3) MVDR + Zelinski	323 /79 /169 /2161	29.8
Case 4) Case 3 + model	320 /44 /280 /2161	34.4
Case 5) Case 3 + FM (e)	263 /79 /159 /2161	26.3
Case 6) Case 3 + FM (e, v)	247 /67 /166 /2161	24.9

4. Conclusion

We have proposed a factorial modeling method to suppress directional noise efficiently in situations where spatial aliasing occurs. Although an adaptive beamformer with a small number of microphones performs better at a wider microphone interval than the aliasing limit, it leaves residual noise in particular bands. The proposed factorial model compensates the degraded bands by using information obtained from reliable bands indicated by the proposed metrics in a probabilistic framework. We believe we can improve our system more by introducing more Gaussians to represent residual speech while our current noise model uses only one.

5. Acknowledgement

This study was conducted using the CENSREC-3 database developed by the IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group.

6. References

[1] M. Brandstein and D. Ward, Eds., “Microphone Arrays: Signal Processing Techniques and Applications,” *Springer-Verlag*, 2001.

[2] H. Saruwatari and Y. Takahashi, "Blind Source Separation for Speech Application Under Real Acoustic Environment," *Independent Component Analysis for Audio and Biosignal Applications, InTech*, Chapter 3, pp. 41-66, 2012.

[3] L.J. Griffiths and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transaction Antennas & Propagation*, vol. AP-30, no. 1, pp. 27–34, 1982.

[4] Y. Ohashi, T. Nishikawa, H. Saruwatari, A. Lee, and K. Shikano, "Noise-robust hands-free speech recognition based on spatial subtraction array and known noise superimposition," *Proceedings of Intelligent Robots and Systems*, pp. 2328–2332, 2005.

[5] P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 177–180, 2003.

[6] A. Ozerov, C. Févotte, R. Blouet, and J.L. Durrieu, “Multichannel non-negative tensor factorization with structured constraints for user-guided audio source separation,” *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 257–260, 2011.

[7] 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*, 22 (2), 149-157, 2001.

[8] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 117–120, 2005.

[9] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms,” *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2578–2581, 1988.

[10] I.A. McCowan, C. Marro, and L. Mauuary, "Robust Speech Recognition using near-field superdirective beamforming with post-filtering," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1723–1726, 2000.

[11] C. Marro, Y. Mahieux, and K.U. Simmer, "Analysis of noise reduction and dereverberation techiniques based on microphone arrays with postfiltering," *IEEE Transactions on speech and audio processing*, 6(3), pp. 240–259, 1998.

[12] N. Ito, N. Ono, and S. Sagayama, "A blind noise decorrelation approach with crystal arrays on designing post-filters for diffuse noise suppression," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 317–320, 2008.

[13] Israel Cohen, "Analysis of Two-Channel Generalized Sidelobe Canceller (GSC) With Post-Filtering," *IEEE Transactions on speech and audio processing*, vol. 11, no. 6, 2003.

[14] J. Bitzer, K. U. Simmer, and K.-D. Kammeyer, "Multi-microphone noise reduction by post-filter and superdirective beamformer," *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 1999, pp. 100–103.

[15] O. Ichikawa, T. Fukuda, and R. Tachibana, “Effective speech suppression using a two-channel microphone array for privacy protection in face-to-face sales monitoring,” *Acoustical Science and Technology (Special Issue on Applied System)*, pp. 507-515, 2015.

[16] B. J. Borgström and A. Alwan, “A Statistical Approach to Mel-Domain Mask Estimation for Missing-Feature ASR,” *IEEE Signal Processing Letters*, vol. 17, pp. 941-944, 2010.

[17] S. Yamamoto, JM. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, “Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory,” *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 1477 – 1482, 2005.

[18] S. Badiezadegan and R. C. Rose, “Mask Estimation in Non-stationary Noise Environments for Missing Feature Based Robust Speech Recognition”, *Proceedings of the International Speech Communication Association (Interspeech)*, pp. 2062-2065, 2010.

[19] O. Ichikawa, S. J. Rennie, T. Fukuda, and M. Nishimura, "Model-based noise reduction leveraging frequency-wise confidence metric for in-car speech recognition," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4921-4924, 2012.

[20] M. Fujimoto, S. Nakamura, K. Takeda, S. Kuroiwa, T. Yamada, N. Kitaoka, K. Yamamoto, M. Mizumachi, T. Nishiura, A. Sasou, C. Miyajima, and T. Endo, “CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework”, *Proceedings of International Workshop on Real world Multimedia Corpora in Mobile Environment (RWCinME)*, pp. 53–60, 2005.