DOA ESITIMATION WITH HYBRID MODELING OF STCHASTIC DYNAMICS OF TARGET SOUND SOURCE AND DETERMINISTIC ENVIRONMENTAL CHARACTERISTICS FOR MOBILE APPLICATIONS

Mitsunori Mizumachi

Faculty of Engineering, Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan
Phone & Fax: (+81)-93-884-3245   E-mail: mizumach@ecs.kyutech.ac.jp

ABSTRACT

This paper proposes a robust direction-of-arrival (DOA) finder with two microphones on a mobile application to achieve accurate and stable estimation under noisy environments. DOA estimation is carried out by state estimation with a stochastic target source model and a deterministic noise model on a spatial space domain. Dynamics of the target source is modeled by the Markov model with a random walk process, and the environmental noise model represents the spectral characteristics of stationary noises. The noise model is employed to select the most dominant frequency of the target signal in time-variant noisy observation in the deterministic manner of the spectral subtraction. Once the frequency selection process is completed, state estimation is done using particle filters. The subband cross-correlation function at the dominant frequency is selectively used for updating the particle weights to estimate posterior distribution. The feasibility of the proposed method has been confirmed under stationary car interior noise and non-stationary factory floor noise environments in several SNR conditions.

1. INTRODUCTION

Recent high growth of computer science is going to bring up fruitful technologies for in-vehicle and mobile systems [1]. Multimedia human-machine interfaces are now available on various equipments. Browsing talk, which enables to talk while seeing the display as shown in Fig. 1, will be one of the most popular multimedia interfaces using cellular phones. In this paper, we proposed a robust direction-of-arrival (DOA) finder with two microphones for achieving attractive mobile interfaces.

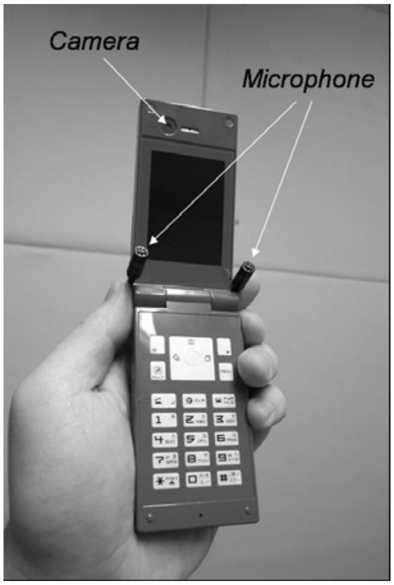

Figure 1: Scene of browsing talk using cellular phone with 2-ch microphone array.

DOA estimation has been an important issue in multi-channel acoustic signal processing, as DOA of an acoustic signal is useful for speech enhancement with beamforming, camera control for video conferencing systems, and so on [2]. Existing DOA estimators are roughly divided into three approaches: (1) time delay estimation approach relying on cross-correlation (CC) between arrival signals [3], (2) energy-distribution-based approach by delay-and sum beamforming, and (3) high-resolution spectral estimation approach such like the MUSIC [4]. To estimate DOA with high accuracy, however, the beamformer-based method requires a lot of spatially-separated microphones, and the parametric high-resolution spectral estimation method has to prepare a proper model in advance considering the number of sound sources carefully. The cross-correlation-based time delay estimation method has been widely used for real applications due to its flexibility with reasonable computational cost. In this paper, we deal with the problem to estimate the difference of arrival time between paired-microphone as DOA estimation.

Robustness against noise and reverberation should be discussed to make DOA finder available in the real world. Recently, advanced DOA estimators have been proposed to improve noise robustness with a prior knowledge on target signals and/or acoustic environments. Provided a target signal is speech, we can use well-known characteristics of speech signals: e.g. harmonicity caused by vibration of the vocal cord and spectral envelopes characterized by resonance in the vocal tract. In the spectral domain, a speech signal is locally dominant against interferences both at harmonic frequencies and around formant frequencies. Brandstein confirmed the importance of introducing speech modeling into multi-channel signal processing applications [5]. DOA finders using harmonicity of speech signals have already been proposed to improve the robustness under reverberant environments [6] [7]. Generally, it is difficult to estimate the fundamental frequency and the formant frequency accurately in noisy and reverberant conditions. The authors have proposed a DOA finder with frequency selectivity [8]. Subband DOA estimates at the dominant frequencies are expected to be more accurate, and are helpful for yielding the global DOA estimate with less influence of stationary background noises. In [8], dominant frequency components are selected by comparing the energy of the time-variant observed signal with that of the background noise in the same way to the spectral subtraction, which is originally aimed at noise reduction in the amplitude spectral domain [9].

Tracking moving sound sources is another interest in DOA estimation. Bayesian dynamic modeling of the dynamics of a sound source is considered to be reasonable, and a particle filter can be applied to estimate posterior distributions with relax assumptions on stochastic behaviors of observations [10]. As a pioneering work, Vermaak and Blake show that non-linear filtering is effective for speaker tracking under reverberant environments [11]. In this work, they model the dynamics of DOAs in the form of random model, and the DOA estimation itself is done by a rather simple method. Ward and Williamson proposed the particle filtering beamformer for sound source localization by using carefully-positioned four microphones [12]. Particle filters are also applied to tracking multiple sound sources and tracking time-varying numbers of speakers [13]. The authors proposed a DOA finder by two-step particle filtering with posterior distribution over bi-dimensional, spectro-spatial state space, which is specified by frequency and time lag [14]. The method regards rectified versions of both CC and whitened cross-correlation (WCC) functions as likelihoods. The likelihood is used to determine weights for particles in a two-step way: CC likelihood leads particles to a rough DOA region, and the second filtering with WCC likelihood aims at concentrating the particles near the true DOA on spectro-spatial space. The method regards kernel density in terms of time lag as a global correlation function, which gives a DOA estimate. It is too difficult to model the set of subband correlation functions on bi-dimensional spectro-spatial state space due to the heterogeneous dynamics between spectral change and source movement, although the proposed method succeeds in yielding accurate DOA estimates in the experimental condition which is carefully matched to the model.

In this paper, we aim at achieving accurate and stable DOA estimation by particle filtering with an environmental noise model. The subband correlation function at the dominant frequency is expected to provide much smooth DOA trajectory by particle filtering [8]. The proposed method carries out state estimation on spatial state space with the reliable likelihood, which is given as the rectified subband cross-correlation function at the dominant frequency [14]. Once the most dominant frequency of the target signal is extracted from noisy observations by using the noise model, the subband correlation function at the dominant frequency updates importance weights for particle filtering.

Rest of this paper is organized as follows. In Section 2, we review DOA estimation based on cross-correlation between two observed signals. In Section 3, we describe both a target source model as the system model on state estimation and an environmental noise model to prepare the reliable likelihood from noisy observations. In Section 4, we propose a DOA finder based on cross-correlation by particle filtering with the environmental noise model. In Section 5, we evaluate the performance of the proposed method in adverse conditions. Finally, conclusion is given in Section 6.

2. DOA ESTIMATION BASED ON CROSS-CORRELATION

2.1. Signal model

Let us assume that a target signal $s(t)$ is received by a pair of spatially-separated microphones. The observed signals, $x_1(t)$ and $x_2(t)$, which are received by the microphones M_1 and M_2, can be modeled as follows:

$$\begin{cases} x_1(t) = h_1(t) * s(t) + n_1(t), \\ x_2(t) = h_2(t) * s(t-\tau) + n_2(t), \end{cases} \qquad (1)$$

where $h_1(t)$ and $h_2(t)$ represent room impulse responses between the sound source and the microphones M_1 and

M_2 , respectively, $n_1(t)$ and $n_2(t)$ are channel-independent additive noises as the mixtures of measurement noises and signals delivered by non-target sound sources, and τ is the time lag of the signal arriving at the microphones M_1 and M_2 . For far-field sound sources, assuming that $h_1(t) \approx h_2(t)$, only difference in phase becomes a clue for estimating a DOA.

2.2. Cross-correlation-based DOA estimation

Concerning DOA estimation with cross-correlation, DOA is given straightforwardly from the time lag τ with the peak in a CC function. CC function has a blunt peak. On the other hand, WCC function as a kind of generalized cross-correlation (GCC) function has the sharp peak in a correlation function [3]. The GCC function $r_{x_1x_2}(\tau)$ is defined as follows.

$$r_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \Psi(f) R_{x_1x_2}(\tau, f) df , \qquad (2)$$

$$R_{x_1x_2}(\tau, f) = X_1(f) X_2^*(f) e^{j2\pi f\tau} , \qquad (3)$$

where $X_1(f)$ and $X_2(f)$ are the short-term Fourier transforms with Hanning windows of the observed signals $x_1(t)$ an $x_2(t)$, respectively, and $*$ denotes the complex conjugate. $R_{x_1x_2}(\tau)$ becomes a conventional CC function, when the generalizing operator $\Psi(f)$ takes a constant over the whole frequency. The WCC method employs the whitening operator as

$$\Psi(f) = \frac{1}{|X_1(f)||X_2(f)|}. \qquad (4)$$

Generally, WCC method accurately finds the local DOA of the signal having the highest energy in each frequency bin. Moreover, a matched filter, which extracts the target signal from the noisy observation, would be adopted as the optimum operator $\Psi(f)$, if target signal could be estimated exactly.

In this paper, CC function with the whitening operator, that is, WCC function, is employed for DOA estimation.

3. MODELING TARGET SOUND SOURCE AND ENVIRONMENTAL NOISE

3.1. Target sound source model

It is hard for DOA finders to give accurate DOA estimates under noisy environments, provided that DOA estimation is carried out in each short-term frame independently. Markov modeling of the temporal trajectory of DOA is a key to yielding accurate and stable DOA estimates, because human speakers moves continuously and smoothly in between short-term frames.

In this paper, we consider DOA estimation on mobile platforms including hand-held devices such as a cellular phone. In such situation, DOA estimation will be the problem to estimate the relative angle between the sound source and the microphone array. Temporal movements of both the source and the microphone array are too complicated to model the characteristics exactly in a deterministic approach, although both objects move smoothly in the short-term view. Therefore, random walk process is applied to represent stochastic speaker movement against a microphone array.

3.2. Environmental noise model

Acoustic environment as well as target sound sources is time-invariant and difficult to exactly predict its behavior. In this paper, assuming that the temporal change of the environmental noise signal is less dynamic than that of target speech signal, we describe the spectral characteristic of the acoustic environment roughly in the deterministic manner of the spectral subtraction [9]. In other words, deterministic noise model is prepared and updated before each utterance, while the model is fixed during each utterance.

We will have an advantage in estimating DOA, if the dominant frequency of the target signal is given against a noisy observation and the band-passed received signals are provided into a DOA finder. Once an enhanced observation, which is the noise-reduced observed signal, is obtained using the noise model, frequency selectivity is achieved by seeking the frequency with the global spectral peak of the enhanced observation.

4. DOA ESTIMATION BY PARFICLE FILTERING

Temporal trajectory of DOA is modeled by a state space model, and it is estimated through the state estimation procedure using particle filters. Observed signals pass through the band-passed filter, of which centre frequency is set at the most dominant frequency corresponding to the global spectral peak, and a WCC function is calculated using 2-ch enhanced observations $\mathbf{x}_k \equiv (x_{1,k}, x_{2,k})$ in the k-th frame. The normalized, subband WCC function is regarded as likelihood $p(\mathbf{x}_k \mid z_k)$, where z_k represents the true DOA at the k-th frame.

State estimation is formally done in a recursive form of the posterior distribution $p(z_{1:k} \mid \mathbf{x}_{1:k})$, where

$z_{1:k} = \{z_1, z_2, \cdots, z_k\}$ and $\mathbf{x}_{1:k} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k\}$ are the time evolution of true DOA and the sampled observations up to the k-th frame, respectively, as follows:

$$p(z_{1:k} \mid \mathbf{x}_{1:k}) \propto p(z_{1:k-1} \mid \mathbf{x}_{1:k-1}) p(\mathbf{x}_k \mid z_k) p(z_k \mid z_{k-1}). \qquad (5)$$

Actually, sequential state estimation is carried out by updating weighted particles according to Eq. (5). We employ a bootstrap filter which uses the system model as proposal (see [10] for more details). DOA estimation by particle filtering is performed as below.

STEP 0: (initial distribution)

Uniform distribution is adopted, and particles $\{z_0^{(l)}\}_{l=1}^{M}$ are drawn according to the distribution.

STEP 1: (filtering by noise robust WCC likelihood)

Particles at *k*-th frame are drawn from the system model with particles at the previous (*k*-1)-th frame, and the weights are updated for $l = \{1, 2, \cdots, M\}$ as $z_k^{(l)} \sim p(z_k \mid z_{k-1}^{(l)})$ and $w_k^{(l)} \sim p(\mathbf{x}_k \mid z_k^{(l)})$.

STEP 2: (resampling)

The particles $\{z_k^{(l)}\}_{l=1}^{M}$ are sampled with replacement in proportion to the weight $\{w_k^{(l)}\}_{l=1}^{M}$. The obtained particles form the proposal particle distribution $\{z_{k+1}^{(l)}\}_{l=1}^{M}$ in the next (*k*+1)-th frame.

STEP 3: (finding DOA)

Kernel density as a correlation function is prepared from the sampled particles in an ordinary way. Finding the peak of the kernel density, DOA is given from the time lag with the peak.

STEP 4: go to STEP 1

5. FERFIRMANCE EVALUATION

5.1. Experimental configuration

Performance of the proposed method is evaluated using target speech and noise sources in a less-reverberant, sound proofed room. Figure 2 illustrates the experimental configuration with two microphones and three sound sources. The target *Source A* was 1.0 m away at the front of the paired-microphone with the spacing of 5.0 cm and the target *Source B* was 1.0 m away at 25 degrees to the right. The *Noise source* was 2.0 m away at the front, but was set against the microphones so as to present non-directional background noises. The target speech and noise signals were prepared from the TI-digit speech database [15] and the NOISEX-92 database (car interior noise and factory floor noise) [16]. Note that car interior noise (Volvo 340 at 120 km/h, in 4th gear, on an asphalt road, in rainy conditions) is almost stationary and the factory floor noise contains not only the stationary component such as a machinery noise, but also some kinds of non-stationary, irregular, sudden noises.

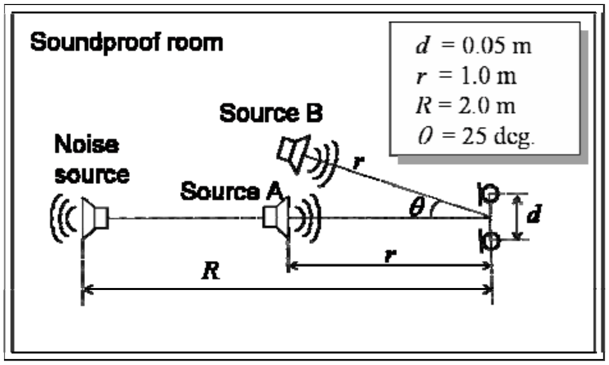

Figure 2: Arrangement of sound sources and microphones for performance evaluation

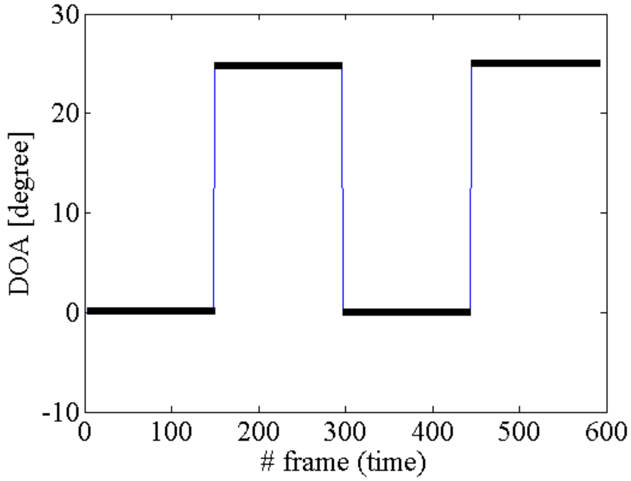

Figure 3: True DOA trajectory of target sound source

5.2. Experimental results and discussion

DOA estimation was carried out 600 trials in each SNR condition: 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. Either of the target sources produced the target speech signal by turns with the alternative DOAs at 0 and 25 degrees as shown in Fig. 3.

Figures 4 and 5 show the DOA estimates in the most severe 0 dB SNR conditions by the conventional WCC without frequency selection and filtering (***pink*** marks), the WCC with frequency selectivity and non-filtering (***sky blue*** marks), the filtering method with the normal full-band WCC likelihood (***red*** line), and the proposed filtering method with frequency selectivity (***blue*** line) under car interior noise and factory floor noise environments, respectively. Figures 6 and 7 summarize the mean among estimation errors by the filtering method with the normal full-band WCC likelihood (***red*** bar), and the proposed filtering method with frequency selectivity (***blue*** bar) in each SNR condition for each noise environment, respectively. Frequency selectivity with the noise model has obvious advantage over full-band use on noise robustness in DOA estimation, while filtering scheme with the target source model is indispensable to yield smooth DOA trajectory. Under the stationary car interior noise environment, the proposed method never deteriorates its performance even in the 0 dB SNR condition. It is found that noise modeling plays an important role for achieving robust DOA estimation, as the filtering method without noise model degrades its performance in inversely proportion to SNR. On the other hands, Figs. 5 and 7 suggest that the proposed method needs a further smart noise model in the presence of non-stationary noises.

6. CONCLUSION

In this paper, a noise robust DOA finder is presented for mobile applications. The proposed method improves noise robustness with a stochastic target sound source model and a deterministic environmental noise model. DOA estimation is carried out based on cross-correlation through state estimation on a spatial state space. The noise model contributes to determine the most dominant frequency, and the subband cross-correlation function at the dominant frequency is regarded as reliable likelihood. Sequential state estimation is done by particle filters with the robust likelihood. The proposed method succeeds in estimating DOAs with the error within a few degrees under car interior noise and the factory floor noise conditions. Future work includes the design of the robust noise model against non-stationary noises.

Acknowledgment

This work was partially supported by Grant-in-Aid for Scientific Research (No. 17760338) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

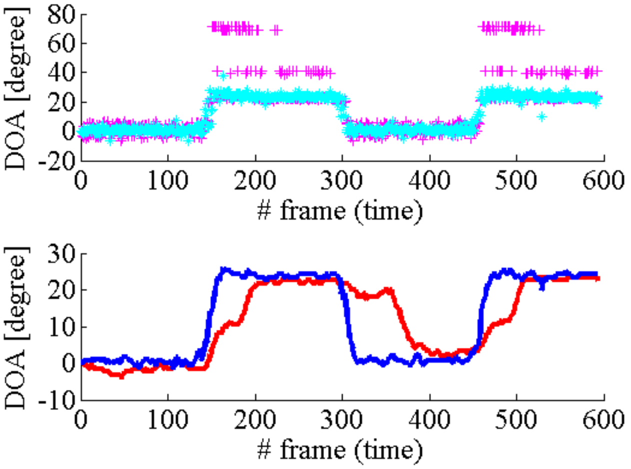

Figure 4: DOA estimates obtained by conventional WCC (**pink** marks), WCC with frequency selectivity (**sky blue** marks), filtering with normal full-band WCC likelihood (**red** line), and proposed method (**blue** line) under **car interior noise** in **0 dB** SNR condition.

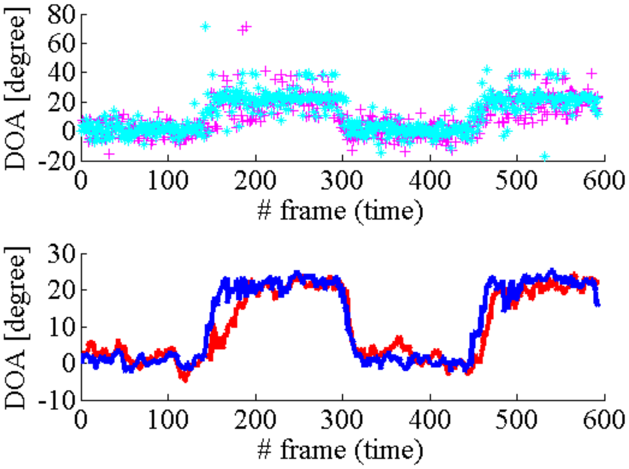

Figure 5: DOA estimates obtained by conventional WCC (**pink** marks), WCC with frequency selectivity (**sky blue** marks), filtering with normal full-band WCC likelihood (**red** line), and proposed method (**blue** line) under **factory floor noise** in **0 dB** SNR condition.

[1] H. Abut, J. H. L. Hansen, and K. Takeda (eds), *DSP for In-Vehicle and Mobile Systems*, Springer-Verlag, New York, 2005.

[2] M. S. Brandstein and D. B. Ward (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

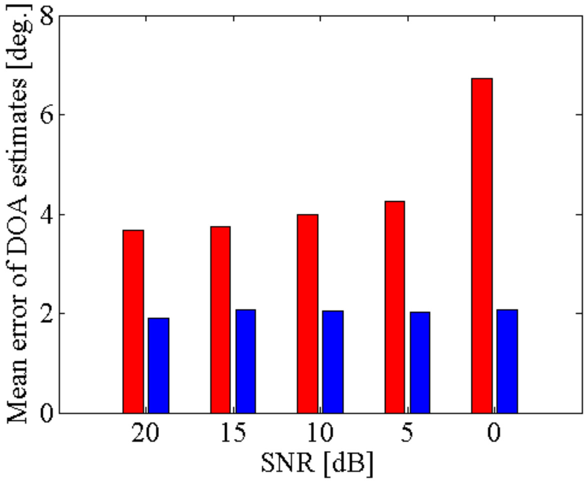

Figure 6: Mean error of DOA estimates obtained by filtering with normal full-band WCC likelihood (**red** line), and proposed method (**blue** line) under **car interior noise** in each SNR condition.

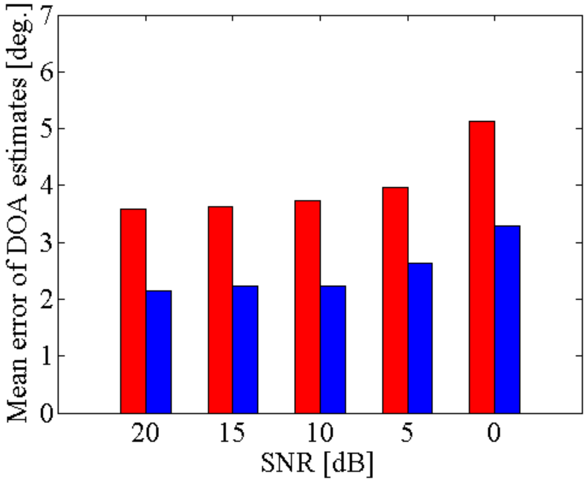

Figure 7: Mean error of DOA estimates obtained by filtering with normal full-band WCC likelihood (**red** line), and proposed method (**blue** line) under **factory floor noise** in each SNR condition.

[3] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,'” IEEE Trans. Acoust., Speech, Signal Process., vol. 24, pp. 320-327, 1976.

[4] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. Antennas Propagation, vol. AP-34, pp. 276-280, 1986.

[5] M. Brandstein, “On the Use of Explicit Speech Modeling in Microphone Array Applications,” Proc. ICASSP'98, pp. 613-616, 1998.

[6] M. Brandstein, “Time-Delay Estimation of Reverberated Speech Exploiting Harmonic Structure,” J. Acoust. Soc. Am., vol. 105, no. 5, pp. 2914-2919, 1999.

[7] Y. Hioka, Y. Koizumi, N. Hamada, “Improvement of DOA Estimation Method Using Virtually Generated Multichannel Data from Two-channel Microphone Array,” Proc. International Symposium on Information Theory and Its Applications, pp. 735-738, 2002,

[8] M. Mizumachi, M. Yuji, and K. Niyada, “DOA Estimation by Cross-correlation Approach with Frequency Selectivity,” Proc. RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP2006), CD-ROM, 2006.

[9] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust., Speech, and Signal Process., vol. 27, no. 2, pp. 113-120, 1979.

[10] A. Doucet, J. F. G. de Freitas, and N. J. Gordon (eds.), *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001.

[11] J. Vermaark and A. Blake, “Nonlinear Filtering for speaker tracking in noisy and reverberant environments,” Proc. ICASSP '01, vol. 5, pp. 3021-3024, 2001.

[12] D. B. Ward, E. A. Lehmann, and R. C. Williamson, “Particle filtering algorithms for tracking an acoustic source in a reverberant environment,” IEEE Trans. Speech and Audio Process., vol. 11, no. 6, pp. 826-836, 2003.

[13] B. Vo, S. Singh and A. Doucet, “Sequential Monte Carlo methods for Bayesian Multi-target filtering with Random Finite Sets,” IEEE Trans. Aerospace and Electronic Systems, vol. 41. no. 4, pp. 1224-1245, 2005.

[14] M. Mizumachi, N. Ikoma, and K. Niyada, “DOA estimation based on cross-correlation by two-step particle filtering,” Proc. 14th European Signal Processing Conference (EUSIPCO2006), CD-ROM, 2006.

[15] R. G. Leonard, “A database for speaker independent digit recognition,” Proc. ICASSP ’84, vol. 9, pp. 328-331, 1984.

[16] A. Varga, H.J.M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, no. 3, pp. 247-252, 1993.