ESTIMATION OF ACTIVE SPEAKER'S DIRECTION USING PARTICLE FILTER

Mitsunori Mizumachi and Katsuyuki Niyada

Kyushu Institute of Technology
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka 804-8550, Japan
E-mail: mizumach@comp.kyutech.ac.jp

ABSTRACT

Building in-car human-machine interfaces, information on speaker's direction is helpful for speech enhancement and controling a video camera. Direction-Of-Arrival (DOA) estimation has been an essential problem in multi-channel acoustic signal processing. This paper proposes two-step particle filtering in a spectro-spatial domain for achieving robust DOA estimation under noisy environments such as in-car environments. The two-step filtering aims at combining the advantages of both traditional cross-correlation (CC) and generalized cross-correlation (GCC) methods. In multiple sound source conditions, proposal particle distribution given by DOA estimates, which are previously obtained, contributes to track the sudden change of an active sound source without latency. Experimental results show that the proposed method is superior both in accuracy and stability to conventional CC and GCC methods under noisy and slightly reverberant environments.

1. INTRODUCTION

Speech is ispensable means for achieving ideal human-machine communication in a car. Automatic speech recognition systems make attempts to enhance speech signals by beamforming techniques with signal directions under adverse environments [1]. To achieve robust human-machine speech communication, Direction-Of-Arrivals (DOAs) of speech signals have to be estimated accurately in noisy conditions.

Cross-Correlation (CC) between two received signals has been widely used for finding a DOA due to its simplicity and efficiency. Among cross-correlation-based approaches, generalized cross-correlation (GCC) method is the most popular technique [2]. The GCC method uses filtered signals, for example whitened signals, for calculating cross-correlation functions, because wide-band signals generally yield the sharp main-lobe at the true DOA in a correlation function. It is obvious that the sharp peak gives the accurate DOA and is robust against acoustic interferences. Provided a narrow-band target signal and a wide-band interference are observed simultaneously, however, the GCC method could not find the DOA of the target signal even if Signal-to-Noise Ratio (SNR) is high enough.

This paper proposes a robust cross-correlation-based DOA finder by particle filtering under noisy environments. Conventional Kalman filter can be applied to this problem, when the behaviors of both target and noise signals can be approximated by conventional statistical models. From the viewpoint of DOA estimation, occlusion, mutations, and rapid changes on signal directions should be considered, when an active sound source disappears or another sound source is born. Furthermore, spectra of speech signals dynamically change time by time in non-stationary noisy conditions. The particle filter can be applied to predict the behavior of the state space without any assumption on stochastic characteristics [3].

Target tracking is one of the suitable problems to be solved by a particle filter, because the voluntary movement of a target can be modeled by Markov process under non-Gaussian, non-stationary, practical noise conditions [4]. The particle filter has been already applied in target tracking for audio, video, radar applications [3] [5] [6] [7]. Vermaak *et al.* introduced a particle filter for tracking a moving sound source under reverberant environments [8] [9]. Ward and Williamson proposes the particle filter beamforming technique for sound source localization [10]. This beamformer-based scheme is attractive for speech enhancement, because it does not require intermediate time-delay estimates before beamforming as usual. Asoh *et al.* proposes audio and video information fusion by particle filtering for multiple speaker tracking [11]. To simplify the problem, observation model consists of two kinds of extracted features, that is, DOAs estimated by the MUSIC algorithm [12] and speakers' positions estimated by using video images [11]. However, the MUSIC algorithm requires a prior knowledge on the number of sound sources, and beamformer-based approaches require a number of microphones.

In this paper, observation model consits of non-parametric cross-correlation-based spectro-spatial functions, that is, a set of subband cross-correlation functions. Subband CC and GCC functions determine the weights for filtering particles in spectral and spatial domains, respectively. In other

words, likelihood of observation is given by half-wave rectified CC and GCC functions In the spectro-spatial domain, the two-step particle filtering is proposed to fuse the emphasized advantages of both traditional CC and GCC methods. The proposed method yields a smooth DOA trajectory, because a frame-based DOA estimate is modeled as a state in Markov process,

This paper is organized as follows. In Section 2, we formulate a signal model for DOA estimation, and review traditional CC and GCC methods. In Section 3, a new approach is proposed to fuse respective advantages of the CC and GCC methods by two-step particle filtering. In Section 4, we present experimental results to examine the performance of the proposed method compared with the conventional CC and GCC approaches. The experiments were performed by using target speech signals recorded in a real room with computer-generated, white Gaussian noises. Both single source and multiple source conditions are considered. Conclusion is given in Section 5.

2. DOA ESTIMATION BY CROSS-CORRELATION-BASED APPROACH

2.1. Signal model

Let us assume that a target signal $s(t)$ is received by two spatially-separated microphones M_i and M_j. The signals $x_i(t)$ and $x_j(t)$, which are received by the microphones M_i and M_j, respectively, can be simply modeled as follows.

$$x_i(t) = \alpha_i s(t) + n_i(t), \quad (1)$$
$$x_j(t) = \alpha_j s(t - \tau_{ij}) + n_j(t), \quad (2)$$

where α_i represents a decay factor in amplitude depending on the distance between the sound source and the i-th microphone, $n_i(t)$ is the additive noise including a measuring noise and acoustic signals coming from other sound sources, and τ_{ij} is the relative time difference when the signal $s(t)$ arrives at the microphones. To simplify the problem, signal distortion caused by reverberation is not considered. When the target sound source is located far from the microphones, we assume that $\alpha_i \approx \alpha_j$. Concerning the observed target signal, difference in phase is much dominant between received signals than that in amplitude.

2.2. Cross-correlation-based DOA estimation

We can estimate DOAs based on the time differences of arrivals. The most well-known approach to estimate the time difference is the cross-correlation-based method. Above all, the GCC method is widely used due to its simplicity and robustness against acoustic interferences [2]. The GCC function $R_{x_i x_j}(\tau)$ is defined as

$$R_{x_i x_j}(\tau) = \int_{-\infty}^{\infty} \Psi(f) X_i(f) X_j^*(f) e^{jf\tau} df, \quad (3)$$

where $X_i(f)$ and $X_j(f)$ are the short-term Fourier transforms of the windowed received signals $x_i(t)$ and $x_j(t)$, and $*$ denotes the complex conjugate. The weighting operator $\Psi(f)$ is introduced to achieve robust DOA estimation in adverse conditions. Provided the weighting operator takes a constant over the whole frequency, Eq. (3) gives a conventional CC function. In this paper, the weighting operator is provided as a whitening filter as follows.

$$\Psi(f) = \frac{1}{|X_i(f)||X_j(f)|} \quad (4)$$

DOA is given straightforwardly from the time lag τ_{ij} with the global maximum peak in the GCC (or CC) function.

2.3. Prospective characteristics of CC and GCC functions

The GCC function with the whitening filter has the sharper main-lobe compared with a conventional CC function. In most cases, the main-lobe appears at the correct time delay correspond to the DOA of the target signal. In case the band-width of the target signal is much narrower than that of an interference signal, however, the global peak in the GCC function may not locate at the true time delay even if SNR is high enough. In adverse conditions, the GCC function is not always superior to the CC function in finding the true DOA of a signal correctly.

3. ROBUST DOA ESTIMATION BY PARTICLE FILTERING

3.1. Motivation

Generally, assuming that the SNR of the target signal is positive, conventional CC function gives rough estimates of signal directions, although they might not be perfectly correct due to the blunt main-lobe. It is helpful for the GCC function to give the accurate DOA, if the rough DOA is provided in advance. This paper proposes to fuse both advantages of the CC and GCC functions by particle filtering in the manner of a sequential Monte Carlo approach. The particle filter is flexible on modeling the system compared with other Bayesian filters such like a traditional Kalman filter and an extended Wiener filter, because it does not require any linearity or Gaussianity on the model.

3.2. Problem statement

Signal model:

Signals received by i-th and j-th microphones are devided into each short-term frame through the Hanning window, and the short-term Fourier transforms, $X_{i,k}(f)$ and $X_{j,k}(f)$, are prepared. Both CC and whitened CC (WCC) functions are calculated locally in each frame k and frequency f.

$$\text{Cross-correlation:} \quad R^{(CC)}_{x_i x_j,\,k}(\tau, f) \tag{5}$$

$$\text{Whitened cross-correlation:} \quad R^{(WCC)}_{x_i x_j,\,k}(\tau, f) \tag{6}$$

System model:

When a sound source is still or moves slowly and continuously, frame-based DOA $\tau_{k,f}$ is modeled by a Markovian, non-linear, non-Gaussian state-space model in each frequency f.

$$\tau_{k,f} = \tau_{k-1,f} + \nu_{k,f}, \tag{7}$$

where $\nu_{k,f}$ represents a system noise. The system noise is modeled by a random walk process without any knowledge on the behavior of the sound source.

Observation model:

Both CC and WCC functions are calculated straightforwardly from the signals $x_i(t)$ and $x_j(t)$ received by a pair of microphones. The observed CC and WCC functions are represented as follows.

$$R^{(CC)}_{x_i x_j,\,k}(T_k, F_k \mid \tau_k, f_k), \tag{8}$$

$$R^{(WCC)}_{x_i x_j,\,k}(T_k, F_k \mid \tau_k, f_k), \tag{9}$$

where T_k and F_k represent the observed time lag and frequency in the k-th frame. Observation noises include both measuring noises and undesired acoustic interferences.

State estimation:

The time lag in each frequency is sequentially estimated by cross-correlation-based functions with observed signals.

$$p(\ \tau_{1:k},\ f_{1:k} \mid x_i(1:k),\ x_j(1:k)\ ) \tag{10}$$

3.3. State estimation by particle filtering

Sequential state estimation is approximated as updating weighted particles on a linear state space model. Concerning the conventional DOA estimation or source tracking applications, states are usually modeled as frame-based DOAs. On the other hand, we consider that normalized cross-correlation-based functions are likelihoods for states, and the likelihoods are represented as the weights $\{w_k\}$ for distributed particle $\{m_k\}$.

$$\{(m^{(r)}_k,\ w^{(r)}_k)\} \approx \tfrac{1}{N} \max\{R^{(r)}_{x_i x_j,\,k}(\tau, f),\ 0\}, \tag{11}$$
$$N = \textstyle\sum_f \sum_\tau \max\{R^{(r)}_{x_i x_j,\,k}(\tau, f),\ 0\},$$

where function selector $r = \{CC, WCC\}$.

The proposed method performs two-step bidimensional particle filtering. Firstly, CC functions give the weights and filter out particles in the frequency direction. Next, other weights are prepared by the subband WCC functions locally in the lag direction. The two-step particle filtering is implemented as follows.

$$\begin{aligned} &p(\ R_{x_i x_j,\,k}(\tau, ; f) \mid R^{(observed)}_{x_i x_j,\,1:k}(\tau, f)\ ) \\ &= p(\ R^{(WCC)}_{x_i x_j,\,k}(\tau, f) \mid R^{(CC)}_{x_i x_j,\,k}(\tau, f)\ ) \\ &\cdot\, p(\ R^{(CC)}_{x_i x_j,\,k}(\tau, f) \mid R_{x_i x_j,\,1:k-1}(\tau, f)\ ) \end{aligned} \tag{12}$$

Practically, the particles and the weights are updated sequentially in each short-term frame according to the observed CC and WCC functions as follows.

$$\begin{aligned} &\{(m^{(WCC)}_k, \quad w^{(WCC)}_k)\} \\ &\overset{Updating}{\longleftarrow} \quad \{(m^{(CC)}_k,\ w^{(CC)}_k)\} \\ &\overset{Updating}{\longleftarrow} \quad \{(m^{(WCC)}_{k-1},\ w^{(WCC)}_{k-1})\} \end{aligned} \tag{13}$$

In the first stage in the k-th frame, assuming that SNR of the target signal is positive, sub-band CC functions gives a weight $\{w^{(CC)}_k\}$ based on global energy distribution including both the target and noise signals. Then, particles $\{m^{(WCC)}_{k-1}\}$, which are delivered from the previous frame, are filtered out with the weights. In the second stage, subband WCC functions form more distinct particle distribution $\{m^{(WCC)}_k\}$ locally in each frequency.

The two-step filtering and resampling are performed frame by frame as follows.

STEP 1: (particle initialization) Particles are prepared in a bidimensional spectro-spatial domain. In the first frame, uniform distribution is adopted to arrange particles initially. Afterwards, the particles obtained in the previous frame give rough prediction of both the spectrum and the signal direction of the target signal in the current frame.

STEP 2: (1st filtering by CC weight) The particles are filtered out with the weights obtained by the half-wave rectified normalized CC function locally in each frequency.

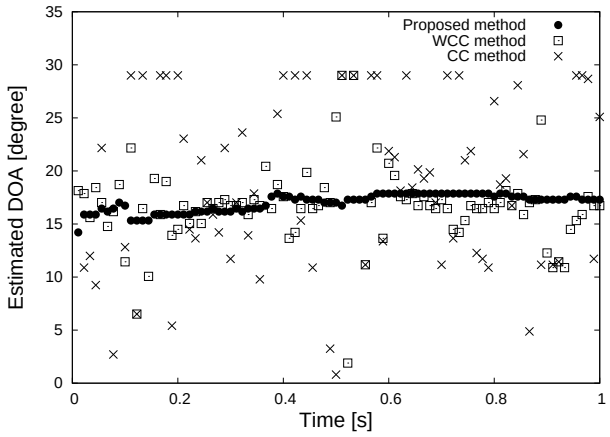

Fig. 1. DOAs estimated in each short-term frame by the proposed ('•'), WCC ('□'), and CC ('×') methods in 10 dB SNR condition. The true DOA was fixed and set at 16 degrees.

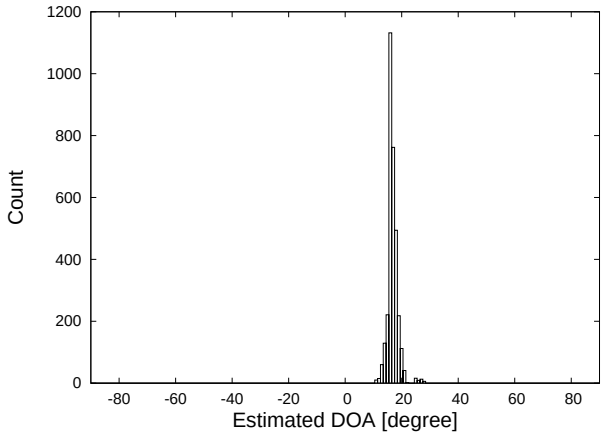

Fig. 2. Histogram of the estimated DOAs over 3,240 frames in 10 dB SNR condition by the proposed method. The true DOA was set at 16 degrees in the whole frame.

STEP 3: (resampling) The weighted particles are resampled in a spectro-spatial domain. The resampled particle distribution reflects the energy distribution. Resultant distribution is considered proposal particle distribution in the next step.

STEP 4: (2nd filtering by WCC weight) GCC function is half-wave rectified, and is used as the weight for filtering the particles resampled in **STEP 3**. The particles are delivered to the next frame besides the next step.

STEP 5: (resampling) The particles are summed up along frequency, and a global correlation function is obtained throughout the whole frequency.

STEP 6: (finding DOA) Finally, DOA is estimated from the time lag that corresponds the peak in the global correlation function.

4. PERFORMANCE EVALUATION

4.1. Experimental condition

Performance of the proposed method was evaluated by using speech data acquired in a real room with computer-generated noise signals. Connected digit speech utterances, which were partially selected from the TI-digit speech database [13], were played through loud speakers (BOSE 101) in a meeting room (7.0 m $\times$ 4.1 m $\times$ 2.6 m). In the room, two microphones (audio-technica AT805F) were placed with a spacing of 0.15 m, and two loud speakers were placed at the direction of 0 (the front) or 16 degrees toward the paired-microphone. The received signals were recorded at 48 kHz with 16 bit accuracy by a DAT (SONY TCD-100). The recorded signals were slightly distorted by background noises and room reverberation, since the room was not soundproofed and reverberation time was 0.23 s approximately. Afterwards, in a computer, independent white Gaussian noises were generated and were added to the received speech signals as non-directional distributed measuring noises. The signals were band-limited in between 300 Hz and 3,400 Hz, and DOA estimation was performed with 3,498 (53 [in lag direction] $\times$ 66 [in frequency direction]) particles every 21.3 ms not only by the proposed method but also by CC and WCC methods as references.

4.2. Single source condition

A loud speaker was 1.0 m away from the microphones at 16 degrees. Figure 1 shows the parts of estimated DOAs by the proposed (marked by '•'), WCC (marked by '□'), and CC (marked by '×') methods in 10 dB SNR conditions. Histograms of the whole DOAs estimated in 3,240 frames are also shown in Figs. 2-4. The experimental results indicate that the proposed method outperforms conventional CC and WCC methods over local and global accuracy. Figure 1 demonstrates that Markov modeling of frame-based DOA estimates contributes to make the DOA trajectory smooth with no fatal error. The accurate and smooth DOA trajectories provide considerable benefits for practical applications.

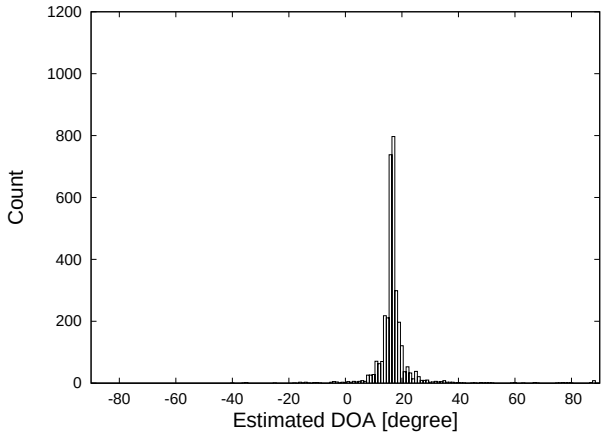

Fig. 3. Histogram of the estimated DOAs over 3,240 frames in 10 dB SNR condition by the WCC method. The true DOA was set at 16 degrees in the whole frame.

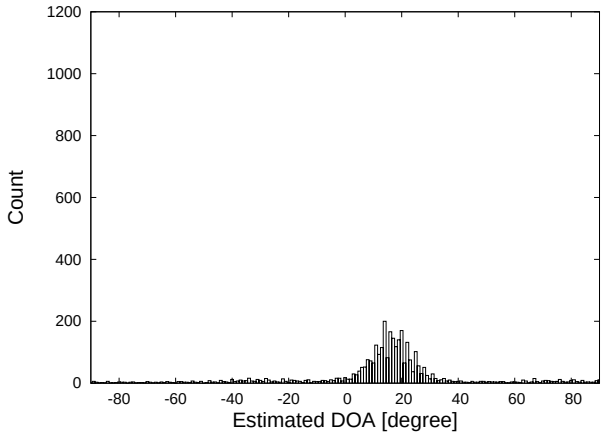

Fig. 4. Histogram of the estimated DOAs over 3,240 frames in 10 dB SNR condition by the CC method. The true DOA was set at 16 degrees in the whole frame.

4.3. Multiple source condition

We consider estimating DOA of the signal from a single active source in multiple source condition. Two loud speakers were 1.0 m away from the microphones at 0 and 16 degrees, and the either loud speaker played TI-digit utterances alternately. Figure 5 shows DOA estimates by the proposed (marked by '$\bullet$'), WCC (marked by '$+$'), and CC (marked by '$\times$') methods in 10 dB SNR conditions. It is hard for the proposed method to follow the sudden change of signal direction without delay.

The proposed method is improved to solve the problem in multiple source conditions. Histograms of DOA estimates are also formed as well as that in a single source condition. Then, the histogram is employed as proposal distribution in STEP 1. Considering an in-car situation, both a driver and a navigator are seated, but they move their heads freely while speaking. The histogram of previous DOA estimates helps to know rough directions of both a driver and a navigator.

Preparing the histogram of enough DOA estimates, preliminary performace of the improved method is shown in Fig. 6. The improved method achieves the rapid pursuit against sudden DOA changes.

5. CONCLUSION

This paper proposes a robust DOA estimation approach by two-step particle filtering under noisy environments, for example, in-car situations. Spectro-spatial cross-correlation-based function, that is, a set of subband cross-correlation functions, is modeled as a non-Gaussian state space model. The subband cross-correlation-based functions are considered as likelihoods, and determine the weights for particle filtering. The proposed method has significant advantage over conventional CC and GCC methods in terms of both accuracy and stability. Markov modeling of frame-based DOA trajectory enables the smooth and stable estimation under noisy environments In multiple source conditions, further improvement is achieved when histogram of DOA estimates is adopted as proposal distribution.

Acknowledgment
The authors would like to acknowledge Prof. Norikazu Ikoma (Kyushu Institute of Technology) for their generous suggestions. 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.

6. REFERENCES

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

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech, Signal Processing, vol. 24, pp. 320–327, 1976.

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

[4] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-

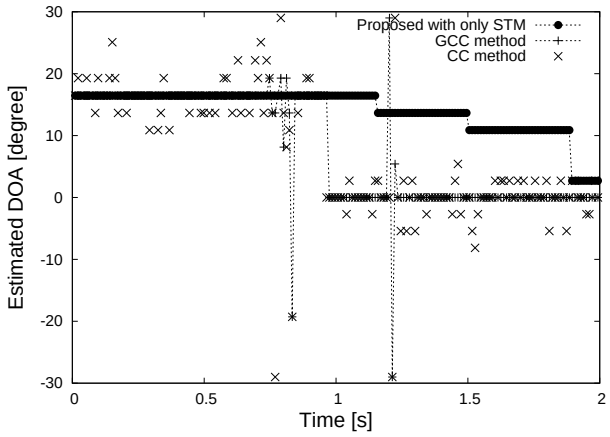

Fig. 5. DOAs estimated in each short-term frame by the proposed ('•'), WCC ('+'), and CC ('×') methods in 10 dB SNR condition. The true DOA was changed from 16 degrees to 0 degree at 1 [s].

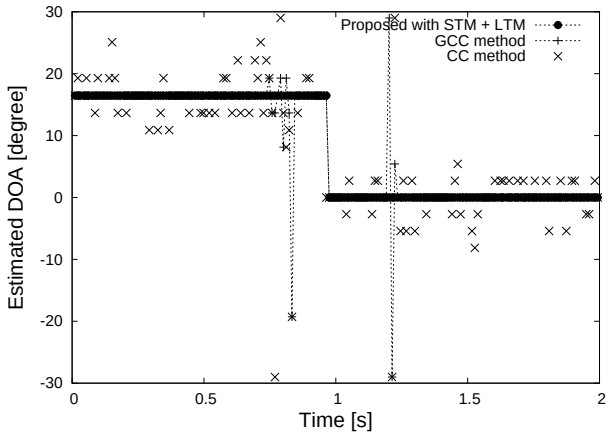

Fig. 6. DOAs estimated in each short-term frame by the proposed and improved ('•'), WCC ('+'), and CC ('×') methods in 10 dB SNR condition. The improved method employs proposal distribution given by histgram of the whole DOA estimates. The true DOA was changed from 16 degrees to 0 degree at 1 [s].

line nonlinear/non-Gaussian Bayesian tracking," IEEE Trans. Signal Processing, Vol. 50, Issue 2, pp. 174–188, 2002.

[5] N. Ikoma, N. Ichimura, T. Higuchi, and H. Maeda, "Maneuvering target tracking by using particle filter," Proc. IFSA World Congress and 20th NAFIPS Intl. Conf., Vol. 4, pp. 2223–2228, 2001.

[6] K. V. Tangirala and K. R. Namuduri, "Object Tracking in Video Using Particle Filtering," Proc. Intl. Conf. on Acoust., Speech, and Signal Processing (ICASSP '05), Vol. 2, pp. 657–660, 2005.

[7] S. Herman and P. Moulin, "A particle filtering approach to FM-band passive radar tracking and automatic target recognition," Proc. IEEE Aerospace Conf., Vol. 4, pp. 1789–1808, 2002.

[8] J. Vermaark and A. Blake, "Nonlinear Filtering for speaker tracking in noisy and reverberant environments," Proc. Intl. Conf. on Acoust., Speech, and Signal Processing (ICASSP '01), Vol. 5, pp. 3021–3024, 2001.

[9] J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Particle methods for Bayesian modeling and enhancement of speech signals," IEEE Trans. Speech and Audio Processing, Vol. 10, Issue 3, pp. 173–185, 2002.

[10] D. Ward, E. Lehmann, and R. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," IEEE Trans. Speech Audio Processing, Vol. 11, pp. 826–836, 2003.

[11] H. Asoh, F. Asano, T. Yoshimura, Y. Motomura, N. Ichimura, I. Hara, J. Ogata, and K. Yamamoto, "An application of a particle filter to bayesian multiple sound source tracking with audio and video information," Proc. Proc. 7th Intl. Conf. on Information Fusion, pp. 805–812, 2004.

[12] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," IEEE Trans. Antennas Propagation, Vol. AP-34, pp. 276–280, 1986.

[13] R. G. Leonard, "A database for speaker independent digit recognition," Proc. Intl. Conf. on Acoust., Speech, and Signal Processing (ICASSP '84), Vol. 9, pp. 328–331, 1984.