Adaptive Beamforming Combined with Particle Filtering for Acoustic Source Localization

Ernst Warsitz, Reinhold Haeb-Umbach, Sven Peschke

University of Paderborn
Dept. of Communications Engineering
33098 Paderborn, Germany
{warsitz,haeb,peschke}@nt.uni-paderborn.de

Abstract

While the main objective of adaptive Filter-and-Sum beamforming is to obtain an enhanced speech signal for subsequent processing like speech recognition, we show how speaker localization information can be derived from the filter coefficients. To increase localization accuracy, speaker tracking is performed by non-linear Bayesian state estimation, which is realized by sequential Monte Carlo methods. Improved acquisition and tracking performance was achieved even in highly reverberant environments, in comparison with both a Kalman Filter and a recently proposed Particle Filter operating on the output of a non-adaptive Delay-and-Sum beamformer.

1. Introduction

Microphone array speech signal processing is an essential element for hands-free speech communication or recognition. While the primary objective is to obtain an enhanced speech signal for subsequent processing, an intermediate goal is often to localize the acoustic source, e.g. to steer a beamformer towards a target speaker. Other applications of acoustic source localization include automatic camera steering in video-conferencing applications, or enhancement of a human-robot interface by enabling the robot to localize the human speaker.

To estimate the location of an acoustic source in a room, typically a set of relative time delays among pairs of microphones has to be determined (TDE: time delay esimation). The generalized cross-correlation (GCC) method is the most popular technique to do so [1]. Several techniques have been proposed to obtain improved performance in reverberant environments, such as steered beamforming [5] and blind estimation of the impulse response of the direct sound field from the source to the microphones [2, 3].

While acoustic source localization can be used to steer a beamformer, we present the opposite approach here: We developed an adaptation algorithm for a Filter-and-Sum beamformer (FSB) to adjust the coefficients of the FIR filters to changing acoustic room impulse responses, e.g. due to speaker movement. The filter coefficients serve to align the direct sound field contributions in each microphone signal to obtain an enhanced speech signal for subsequent processing. Here we show how an estimate of the source location can be derived from the coefficients.

While the aforementioned approaches determine the current location of an acoustic source from data obtained at an array of sensors during the current time only, considerably more accurate estimates can be obtained by recursive estimation of the current source location using all previous data. Recently Ward et al. have proposed sequential Monte Carlo methods to approximate non-linear Bayesian estimation of the speaker position [4–6]. They devised Particle Filters for measurement data obtained from a GCC and a steered non-adaptive Delay-and-Sum beamformer (DSB) and demonstrated good tracking performance. Here, we develop a Particle Filter for the adaptive FSB and compare its performance with two alternatives: First, a Kalman Filter, the traditional solution to linear tracking problems, and, second, a Particle Filter operating on the output of a steered Delay-and-Sum (DSB) beamformer, similar to the one presented in [5]. It is shown that the FSB-based Particle Filter exhibits both improved acquisition and tracking performance.

2. Instantaneous Location Estimates

In a Filter-and-Sum beamformer, each microphone signal $x_m(n)$, $m = 1, \ldots, M$, is filtered in an FIR filter of impulse response $f_m(-n)$ to obtain the enhanced output signal

$$y(n) = \sum_{m=1}^{M} f_m(-n) * x_m(n). \qquad (1)$$

Here, M denotes the number of microphones. With (typically short) FIR filters fractional time delays can be realized in the discrete-time domain, which are required to steer the array in arbitrary directions-of-arrival (DOA). Further, FIR filters are capable of aligning the direct sound field contributions of the microphone signals and possible strong early reflections. In case of changes of the acoustic impulse responses from the source to the microphones, e.g. due to speaker movements, the FSB filter coefficients have to be adjusted. We adapt the filter coefficents by a stochastic gradient algorithm, which is derived from a constrained optimization problem, where the FSB output power is maximized while keeping the norm of the filter coefficient vector fixed, similar to adaptive principal component analysis [7].

While the purpose of acoustic front-end processing is to obtain an enhanced signal for subsequent processing stages (e.g. speech recognition) in the first place, the FIR filter coefficients obviously also contain information about the loca-

tion of the target speaker. Let $u(n) = e^{j\omega_k nT}$ be the single-frequency source signal at an hypothesized source position ℓ, where $nT, n = 1, 2, \ldots$, denote the sampling instants. In the absence of reverberation the m-th microphone signal is

$$x_m(n) = |X_m(\omega_k)| e^{j\omega_k (nT - \tau_m(\ell))}, \tag{2}$$

where the delay $\tau_m(\ell)$ is given by

$$\tau_m(\ell) = \frac{\parallel \ell - \ell_m \parallel}{c}. \tag{3}$$

Here, c denotes the speed of sound and ℓ_m is the location of the m-th microphone. Let $F_m^*(\omega_k)$ be the transfer function of the filter in the m-th microphone path. The output of the beamformer to the above single-frequency excitation can thus be written as

$$\begin{aligned} y(n) &= e^{j\omega_k nT} \sum_{m=1}^{M} F_m^*(\omega_k) |X_m(\omega_k)| e^{-j\omega_k \frac{\parallel \ell - \ell_m \parallel}{c}} \\ &= e^{j\omega_k nT} r(\ell, \omega_k) \end{aligned} \tag{4}$$

where

$$r(\ell, \omega_k) = \sum_{m=1}^{M} F_m^*(\omega_k) |X_m(\omega_k)| e^{-j\omega_k \frac{\parallel \ell - \ell_m \parallel}{c}} \tag{5}$$

is the FSB beamformer response, which we will call $r^{(FSB)}(\ell, \omega_k)$ in the following. The frequency-averaged beampattern

$$|r(\ell)|^2 = \sum_{\omega_k} W(\omega_k) |r(\ell, \omega_k)|^2, \tag{6}$$

where $W(k)$ is an arbitrary weighting function, can be viewed as a function of the source position ℓ. If ℓ_0 is the position of the source to which the FSB coefficients have been adapted, then one would expect $|r(\ell_0)|^2$ to be large. Thus, an instantaneous estimate of the source position can be obtained as

$$\hat{\ell} = \underset{\ell}{\operatorname{argmax}} \, |r(\ell)|^2. \tag{7}$$

In the case of a Delay-and-Sum beamformer as in [5] the filter coefficients are $F_m^*(\omega_k) = e^{-j\omega_k \|\ell_0 - \ell_m\|/c}$ and $r(l, \omega_k)$ becomes

$$r^{(DSB)}(\ell, \omega_k) = \sum_{m=1}^{M} e^{-j\omega_k \frac{\|\ell_0 - \ell_m\|}{c}} X_m(\omega_k) \tag{8}$$

where

$$X_m(\omega_k) = |X_m(\omega_k)| e^{-j\omega_k \frac{\|\ell - \ell_m\|}{c}} \tag{9}$$

is the microphone signal in the frequency domain.

3. Non-linear Bayesian Tracking

The noisy instantaneous location estimate (7) can be improved by tracking the speaker movement over time. To do so, the speaker localization problem has been formulated as a non-linear Bayesian state space estimation problem [6]. The unobservable state vector is

$$\boldsymbol{\alpha}_k = (\ell_x, \ell_y, v_x, v_y)^T, \tag{10}$$

where (ℓ_x, ℓ_y) and (v_x, v_y) are the location and velocity of the source in Cartesian coordinates. In Bayesian tracking, the a posteriori density $p(\boldsymbol{\alpha}_k | \boldsymbol{z}_{1:k})$ of the state vector $\boldsymbol{\alpha}_k$, given all measurements up to the current time $\boldsymbol{z}_{1:k} = (z_1, \ldots, z_k)$, has to be estimated. The Kalman Filter is an optimal solution to this problem if state and measurement equations are linear and if system and measurement noise are Gaussian. If these assumptions are not met, sequential Monte Carlo methods, also known as Particle Filters, are a promising suboptimal approach.

3.1. Model of Source Dynamics

We assume the following linear state space model of the source dynamics [8]:

$$\boldsymbol{\alpha}_{k+1} = \boldsymbol{\Phi} \boldsymbol{\alpha}_k + \boldsymbol{G} \boldsymbol{w}_k \tag{11}$$

where

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & \Delta T \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{G} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \Delta T & 0 \\ 0 & \Delta T \end{pmatrix}.$$

The system noise $\boldsymbol{w}_k = (w_{1k}, w_{2k})^T$ is normally distributed with $\boldsymbol{w}_k \sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_2)$, $\boldsymbol{I}_2$ denoting the identity matrix of order 2. ΔT is the time duration between two successive measurements, i.e. between two successive computations of the beampattern (6): $\Delta T = L/f_s$, where L is the frame shift and f_s the sampling frequency.

In this model the random acceleration from time frame k to $k+1$ is given by $a_k = (w_{1k}^2 + w_{2k}^2)^{1/2}$, which is a Rayleigh distributed random variable with expectation $E[a_k] = \sigma_w \sqrt{\pi/2}$. This allows for the computation of σ_w^2 by specifying an average acceleration $\bar{a}$ of the source, according to $\sigma_w^2 = 2\bar{a}^2/\pi$. We have chosen $\bar{a} = 2m/s^2$, which seemed to be a reasonable value for indoor speaker movement.

3.2. Measurement Model

If eq. (7) is taken as the measurement: $z_k := \hat{\ell}_k$, a linear measurement equation can be obtained

$$z_k = \boldsymbol{H} \boldsymbol{\alpha}_k + v_k, \tag{12}$$

with

$$\boldsymbol{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad v_k \sim \mathcal{N}(0, \sigma_v^2).$$

The variance σ_v^2 of the measurement noise can be estimated from observed values z_k in the usual way.

For the linear model given by (11) and (12) a Kalman Filter has been implemented. Note, however, that the beam pattern (6) contains more information than is preserved by the max-operation in (7). In a reverberant room $|r(\ell)|^2$ is typically a complicated multi-modal function. We observed that the position of the absolute maximum can jump to very different locations from one time frame to the next, while a local maximum, or at least a large value of $|r(\ell)|^2$ could always be observed at the true source location. Because of the multi-modality of $|r(\ell)|^2$ an extended Kalman Filter was not considered promising.

Instead, a Particle Filter has been developed to better incorporate the location information available in the beam pattern.

We adopted the approach taken in [5, 9] and defined a "pseudo-likelihood"

$$p(z_k|\boldsymbol{\alpha}_k) = L(|r(\ell|^2) \tag{13}$$

where $L(\xi) = \xi^i$, with values $i = 2, 3, 4$ is a shaping function which narrows the main beam and reduces the level of the sidelobes. In the experiments reported below we have chosen a value of $i = 3$.

4. Experimental Results

We analyzed the performance of the speaker localization both in simulations and with real-time experiments in our laboratory. Here, only the simulation results are presented.

Random trajectories of the acoustic source were generated with the state space model of eq. (11). We used a speech source signal of length 10 s, sampled at 16 kHz. The audio data at each sensor were obtained using the well-known image method [10]. For every 128 samples a new source position and the corresponding room impulse response was calculated to produce the microphone signals. Eight sensors were placed in a room of the size (6m) x (4m) x (3m); a pair in the center of each wall respectively. The distance between the two microphones of a pair was set to 0.2 m.

We simulated room reverberation times RT_{60} ranging from 0 to 0.8 s with an increment of 80 ms and added white Gaussian noise to each microphone signal at SNR levels of 5, 10, 15 and 20 dB. For every RT_{60}-SNR-combination we generated 10 speech files; altogether the test set of audio data consisted of 440 files.

In the Particle Filter experiments we simulated an overall of $N_p = 260$ particles, used the so-called "prior importance function" [11] and applied systematic importance-resampling once the effective number of samples was below the threshold of 235 samples. For a fair comparison, in the Kalman Filter experiments the number of grid points, over which the maximum is searched in (7), was also set to 260. Further, a weighting function $W(\omega_k)$, see eq. (6), was used which selects the frequency bins in the range from 1 to 4 kHz.

Fig.1 shows the RMS position error as a function of the room reverberation time for an SNR of 15 dB for Particle Filtering with different beamformer responses. "FSB" denotes a Particle Filter with $r^{(FSB)}(\ell, \omega_k)$ used in the pseudo-likelihood of eq. (13). For a given set of FSB coefficients $r^{(FSB)}(l, \omega_k)$ was evaluated for all particle positions $l = l^{(i)}$, $i = 1, \ldots, N_p$. To obtain the curves labeled "DSB", we used $r^{(DSB)}(\ell, \omega_k)$ of eq. (8), where, for a given microphone signal, the term was evaluated for all particle positions $l_0 = l^{(i)}$, $i = 1, \ldots, N_p$.

It can be seen that the baseline performance (curves denoted by "FSB/DSB") can be improved by:

- Computing the beam pattern over pairs of microphones, e.g. in the case of the FSB

$$|r(\ell, \omega_k)|^2 = \sum_{p=1}^{P} \left| \sum_{m=1}^{2} F^*_{p,m}(\omega_k) e^{-j\omega_k \frac{\|\ell - \ell_{p,m}\|}{c}} \right|^2 \tag{14}$$

  where P is the number of microphone pairs (here: P=4), and (p, m) is the index of the m-th microphone of the p-th pair (curves denoted by "FSB/DSB_Pair"),

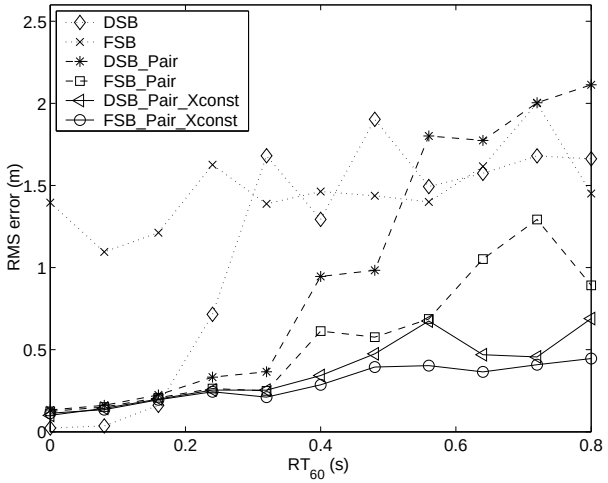

Figure 1: *RMS position error as a function of room reverberation time at an SNR of 15 dB for Particle Filters with different pseudo-likelihoods.*

- and, additionally, setting $|X_m(\omega_k)| = 1$ in eqs (5) and (8), respectively (curves denoted by "FSB/DSB_Pair_Xconst").

All further experiments reported are conducted with this last setup, which delivered the best results.

Figs. 2 and 3 show the RMS position error as a function of the room reverberation time for an SNR of 20 dB and 10 dB, respectively. Here we compare again FSB-based position estimation with DSB-based estimation, now for the three cases

- Instantaneous estimate, see eq. (7) (curves denoted by "FSB/DSB_inst").
- Postprocessing of the instantaneous estimate with a Kalman Filter (curves denoted by "FSB/DSB_KF").
- Particle Filtering with the best setup of the previous set of experiments (curves denoted by "FSB/DSB_PF").

As can be seen, filtering with an underlying state space model of the source motion greatly improves localization performance. However, the Particle Filter clearly outperforms the Kalman Filter. Further, our proposed FSB-based filter performs somewhat better than the DSB-based filter for the more noisy data. As can be seen, with Particle Filtering excellent localization performance can be achieved for a wide range of room reverberation times and noise levels.

Finally we studied the acquisition performance of the Particle Filters. While in the previous experiments the initial particle locations were at the true initial source location, we now draw the initial particle positions from a uniform distribution over the size of the room. For all combinations of SNR and RT_{60} values we observed the same trend: the FSB-based Particle Filter settled at the final accuracy much faster than the DSB-based filters. As an example, Fig. 4 shows the RMS position error as a function of time for SNR=15 dB and RT_{60}=0.4 s averaged over 10 audio files. The improved acquisition performance can be attributed to the fact, that the FSB beam pattern has more often

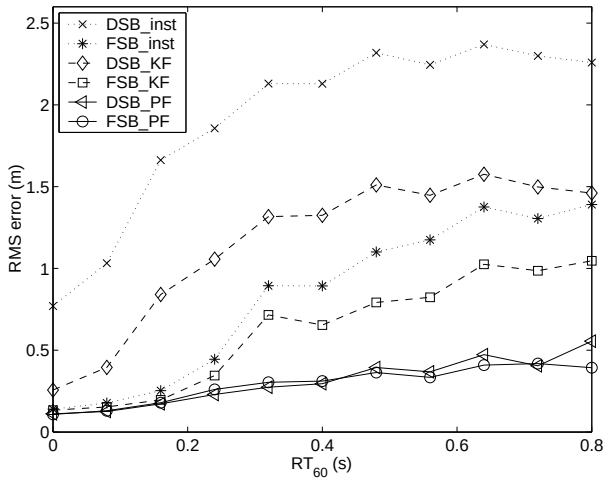

Figure 2: *RMS position error as a function of room reverberation time with additive noise at microphones (SNR=20 dB).*

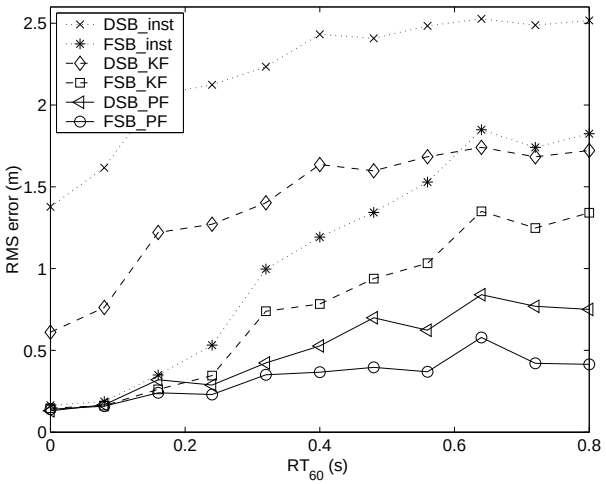

Figure 3: *RMS position error as a function of room reverberation time with additive noise at microphones (SNR=10 dB).*

high values in the vicinity of the true source location than the typically noisier DSB pattern.

5. Conclusions

Using an adaptive Filter-and-Sum beamformer for speech enhancement delivers acoustic source location information as a side effect with little computational overhead. This noisy location estimate can be greatly improved by non-linear state space Bayesian estimation employing Particle Filtering. The experimental results show good location accuracy for a wide range of room reverberation times and SNR values, both compared to a Kalman Filter and a recently proposed Particle Filter operating on the output of a non-adaptive Delay-and-Sum beamformer. The proposed FSB solution exhibits excellent acquisition performance.

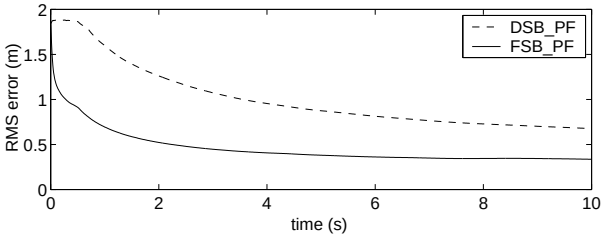

Figure 4: *RMS position error as a function of time, starting from a uniform distribution of the particles at time* $t = 0$ *(SNR=15 dB,* RT_{60}*=0.4 s).*

6. References

[1] C. H. Knapp and G. C. Carter "The generalized correlation method for estimation of time delay", *IEEE Trans. ASSP*, vol. ASSP-24, pp. 320-327, Aug. 1976

[2] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization", *J. Acoust. Soc. Amer.*, vol. 107, pp. 384-391, Jan. 2000

[3] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification", *IEEE Trans. Signal Processing*, vol. 51, pp 11-24, Jan. 2003

[4] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-01)*, Salt Lake City, May 2001

[5] D. B. Ward, and R. C. Williamson, "Particle filter beamforming for acoustic source location", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-02)*, Orlando, May 2002

[6] D. B. Ward, E. A. Lehmann, R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment", *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 826-836, 2003

[7] S. Haykin, *Neural Networks*, 2nd edition, Prentice Hall, 1999

[8] M. Hellebrandt, R. Mathar, "Location tracking of mobiles in cellular radio networks", *IEEE Trans. on Vehicular Technology*, vol. 48, Nr. 5, pp. 1558-1562, Sept. 1999

[9] E. A. Lehmann, D. B. Ward, R. C. Williamson, "Experimantal comparison of particle filtering algorithms for acoustic source localization in reverberant room", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-03)*, Hong Kong, April 2003

[10] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Amer.*, vol. 107, no. 4, pp. 943-950, 1979

[11] A. Doucet, N. de Freitas, and N. Gordon, eds. *Sequential Monte Carlo Methods in Practice*, Springer Verlag, 2001