Speech Source Tracking Based on Particle Filter under non-Gaussian Noise and Reverberant Environments

Ruifang Wang[1, a], Xiaoyu Lan[1, b]
[1]*School of Electronic and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China*

Keywords: Speech Source Tracking, non-Gaussian Noise, Particle Filter, Generalized Correntropy Function.

Abstract: Tracking a moving speech source in non-Gaussian noise environments is a challenging problem. A speech source tracking method based on the particle filter (PF) and the generalized correntropy function (GCTF) in non-Gaussian noise and reverberant environments is proposed in the paper. Multiple TDOAs are estimated by the GCTF and the multiple-hypothesis likelihood is calculated as weights for the PF. Next, predict the particles from the Langevin model for the PF. Finally, the global position of moving speech source is estimated in term of representation of weighted particles. Simulation results demonstrate the vadility of the proposed method.

1 INTRODUCTION

Tracking a speech source accurately in reverberant environments is desirable for teleconferencing system (B. Kapralos, M. R. M. Jenkin and M. Evangelos, 2003), robots (K. Nakadai, et al, 2006), and human-machine interaction (T.P. Spexard, M. Hanheide, and G. Sagerer, 2007). Acquiring the position of the speech source plays an important role in speech signal processing region. The environmental noise and reverberation of the speech signal are two challenging problems for speech source tracking. In conventional speech source localization and tracking approaches (E. T. Roig, F. Jacobsen and E. F. Grande, 2010), (M. F. Fallon, and S. J. Godsill, 2012), they only depend on the current observations to estimate the positions of the speech source. To improve tracking performance, Bayesian filtering algorithms are used to track the moving speech source, which employs not only current observations but also previous observations. The particle filter (PF) is an approximation of the optimal sequential Bayesian estimation via Monte Carlo simulations for non-linear and non-Gaussian system. The PF incorporated multiple-hypothesis model was applied to the speaker tracking problem based upon TDOA observations (abbreviated to PF) (D. B. Ward, E. A. Lehmann and R. C. Williamson, 2003). A novel framework of PF based on information theory was discussed for speaker tracking (F. Talantzis, 2010). A non-concurrent multiple talkers tracking based on extended Kalman particle filtering (EKPF) was proposed (X. Zhong, and J. R. Hopgood, 2014). In (X. Zhong, A. Mohammadi, et al, 2013), a distributed particle filter (DPF) was proposed in speaker tracking in a distributed microphone network, in which each node runs a local PF for local posteriors fused to obtain a global posterior probability (abbreviated to DPF-EKF). In (Q. Zhang, Z. Chen, and F. Yin, 2016), a distributed marginalized auxiliary particle filter was proposed for speaker tracking.

For above-mentioned speech source tracking methods, the background noise is assumed to be Gaussian noise. However, the practical background noise may be non-Gaussian noise such as knock on the door, sudden phone ringing and a fit of couching, which is impulsive in essence and would lead to poor tracking performance for these speech source methods. To remedy impacts of non-Gaussian background noise on tracking performance, a PF based speaker tracking method under non-Gaussian noise environments is proposed. First, the symmetric alpha-stable (SαS) distributions (M. Shao and C. L. Nikias, 1993) are employed to model the non-Gaussian noise and TDOA observations of speech signals received between a microphone pair at each node are approximated via a generalized correntropy function (GCTF) (W. Liu, P.P. Pokharel, et al, 2007). Next, the Langevin model (D. B. Ward, E. A. Lehmann and R. C. Williamson, 2003) is used to

Wang, R. and Lan, X.
Speech Source Tracking based on Particle Filter under non-Gaussian Noise and Reverberant Environments.
DOI: 10.5220/0008874204610466
In *Proceedings of 5th International Conference on Vehicle, Mechanical and Electrical Engineering (ICVMEE 2019)*, pages 461-466
ISBN: 978-989-758-412-1

model the time-varying states of a moving speech source to predict the particles and a multiple-hypothesis model is introduced to calculate the likelihood function as weights corresponding to the particles of the PF. Finally, the global time-varying position estimations at each time step are obtained in terms of weighted particles.

2 FUNDAMENTAL ALGORITHM

2.1 Particle Filter for Tracking Problem

Considering time-dependent state vector $\mathbf{x}_k$ in a distributed sensor network, where k being a discrete time index. The state-space model of the system and observation models at node j are given as (D. B. Ward, et al, 2003).

$$\begin{cases} \mathbf{x}_k = f_k(\mathbf{x}_{k-1}) + u_k \\ \mathbf{z}_k = h_k(\mathbf{x}_k) + v_k \end{cases} \quad (1)$$

Where $\mathbf{z}_k$ is observation vector of $\mathbf{x}_k$, f_k and h_k are the system dynamics function and the observation function, respectively, u_k and v_k are the process noise and observation noise with known probability density function, respectively.

The Bayesian filter for tracking problem is to calculate the posterior probability density $p(\mathbf{x}_k \left| \mathbf{z}_{1:k} \right.)$. Particle filter estimates the Bayesian recursion via the Monte Carlo simulation and works in the principle of sequential importance resampling (SIR) algorithm. In the prediction step, N particles $\left\{ \mathbf{X}_k^n \right\}_1^N$ are drawn from a suitable chosen proposal function $q(\mathbf{x}_k \left| \mathbf{X}_{1:k-1}^n, \mathbf{z}_{1:k} \right.)$ at time k. In the update step, the weight w_k^n corresponding to the nth particle $\mathbf{X}_k^n$ is calculated based on the prior transition density as the proposal function, i.e., $q(\mathbf{x}_k \left| \mathbf{X}_{1:k-1}^n, \mathbf{z}_{1:k} \right.) = p(\mathbf{x}_k \left| \mathbf{X}_{k-1}^n \right.)$, written as

$$w_k^n = \frac{p(\mathbf{z}_k \left| \mathbf{X}_k^n \right.) p(\mathbf{x}_k^n \left| \mathbf{X}_{k-1}^n \right.)}{p(\mathbf{x}_k^n \left| \mathbf{X}_{k-1}^n \right.)} = p(\mathbf{z}_k \left| \mathbf{X}_k^n \right.) \quad (2)$$

Where $p(\mathbf{z}_k \left| \mathbf{X}_k^n \right.)$ is likelihood function.

The PF is to represent posterior probability $p(\mathbf{x}_k \left| \mathbf{z}_{1:k} \right.)$ by a set $\left\{ \mathbf{X}_k^n, w_k^n \right\}_{n=1}^N$, given as

$$p(\mathbf{x}_k \left| \mathbf{z}_{1:k} \right.) \approx \sum_{n=1}^{N} w_k^n \delta(\mathbf{x}_k - \mathbf{X}_k^n) \quad (3)$$

Where $\delta(\cdot)$ denotes the multi-dimensional Dirac delta function. Finally, the MMSE estimate of the state $\mathbf{x}_k$ is estimated as

$$\hat{\mathbf{x}}_k = \sum_{n=1}^{N} w_k^n \mathbf{X}_k^n \quad (4)$$

2.2 TDOA Estimation under non-Gaussian Environments

For non-Gaussian noise, symmetric alpha-stable (SαS) processes can model the impulsive noise better than other processes (M. Shao and C. L. Nikias, 1993), (W. Liu, P.P. Pokharel, et al, 2007) which does not have finite second order statistics and a closed-form probability density function unfortunately. Normally, alpha-stable processes can be described with characteristic functions, written as

$$\varphi(\boldsymbol{t}) = \exp\left\{ \boldsymbol{jbt} - \gamma \left| \boldsymbol{t} \right|^{\alpha} \left[1 + \boldsymbol{j}\beta \operatorname{sign}(\boldsymbol{t}) \omega(\boldsymbol{t}, \alpha) \right] \right\} \quad (5)$$

$$\omega(t, \alpha) = \begin{cases} \tan(\alpha\pi/2) & , \quad for\ \alpha \neq 1 \\ (2/\pi)\log\left| t \right| & , \quad for\ \alpha = 1 \end{cases} \quad (6)$$

Where $\alpha \in (0, 2]$ is the characteristic exponent.

When speech source signals received by a pair of microphones is polluted by non-Gaussian noise, accurate TDOA estimations is difficult to be obtained via typical TDOA estimation methods for example generalized cross-correlation (GCC) (C. Knapp, and G. C. Carter, 1976). To solve the problem, a generalized correntropy function (GCTF) (W. Liu, P.P. Pokharel, et al, 2007) based TDOA estimation method is presented for speech source tracking under the non-Gaussian noise environment. The GCTF $D_k^j(\tau_j)$ at node j is defined as

$$D_k^j(\tau_k^j) = E\left[\kappa(s_{j,1}(k) - s_{j,2}(k + \tau_k^j)) \right] \quad (7)$$

$$\kappa(\bullet) = \frac{1}{\sqrt{2\pi}\sigma'} \exp\left(-\frac{(\bullet)^2}{2\sigma'^2} \right) \quad (8)$$

Where $s_{j,1}(k)$ and $s_{j,2}(k)$ denote the two signals received at two microphones of node j, $E[\bullet]$ represents mathematical expectation operation, $\kappa(\bullet)$ is the Gaussian kernel and σ' $(\sigma'>0)$ is the kernel size.

The TDOA observations at node j can be estimated by a GCTF estimator

$$\hat{\tau}_k^j = \underset{\tau_k^j \in [-\tau^{j\max}, \tau^{j\max}]}{-\arg\max} (D_k^j(\tau_k^j)) \tag{9}$$

Where $\tau^{j\max}$ denotes the maximal probable value of the TDOA at node j.

Considering the noise and reverberation, generally, N_m TDOAs selected from first N_m local maxima of $D_k^j(\tau)$ constitute the TDOA observation vector $\mathbf{z}_k^j = \left[\hat{\tau}_{k,1}^j, \hat{\tau}_{k,2}^j, \cdots, \hat{\tau}_{k,N_m}^j\right]^T$ at node j (D. B. Ward, E. A. Lehmann and R. C. Williamson, 2003).

3 SPEECH SOURCE TRACKING UNDER NON-GAUSSIAN ENVIRONMENTS

3.1 Speech Source Dynamic Model

The Langevin model is simple and has worked well in practice to represent the time-varying locations of a speech source moving trajectory which is denoted as (D. B. Ward, E. A. Lehmann and R. C. Williamson, 2003), (Q. Zhang, Z. Chen, and F. Yin, 2016).

$$\mathbf{x_k} = \begin{bmatrix} 1 & 0 & a\Delta T & 0 \\ 0 & 1 & 0 & a\Delta T \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} b\Delta T & 0 & 0 & 0 \\ 0 & b\Delta T & 0 & 0 \\ 0 & 0 & b & 0 \\ 0 & 0 & 0 & b \end{bmatrix} \mu_k \tag{10}$$

Where $\mathbf{x}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$ denotes the speech source's state at time step k in the Cartesian coordinates, ΔT is the discrete time interval, $\mu_\mathbf{k}$ is the time-uncorrelated Gaussian white noise vector, and the parameters a and b are defined as

$$a = \exp(-\beta\Delta T) \qquad b = \bar{v}\sqrt{1-a^2} \tag{11}$$

Where β is the rate constant, and $\bar{v}$ is the root-mean-square velocity.

3.2 Multiple-Hypothesis Likelihood Model

Consider the local likelihood function $p(\mathbf{z}_k^j | \mathbf{x}_k)$ at node j based on N_m TDOA observations in $\mathbf{z}_k^j$. Due to noise and reverberation, among N_m TDOAs at most one associated with the true speech source, whereas the others correspond to the spurious speech source (X. Zhong, and J. R. Hopgood, 2014). Thus, the multiple-hypothesis likelihood model is employed as the local likelihood function or local weight for particles at node j, written as (X. Zhong, and J. R. Hopgood, 2014).

$$p(\mathbf{z}_k^j | \mathbf{x}_k) = \frac{q_0}{(2\tau^{j\max})^{N_m}} + \frac{1}{(2\tau^{j\max})^{(N_m-1)}} \sum_{i=1}^{N_m} q_i \mathcal{N}(\hat{\tau}_{k,i}^j; \tau_k^j(\mathbf{x}_k), \sigma^2) \tag{12}$$

Where q_0 is the prior probability that none of N_m TDOA observations corresponds to the true speech source, $q_i (i = 1, \cdots N_m)$ is the prior probability that only the ith TDOA corresponds to the true source, $\sum_{i=0}^{N_m} q_i = 1$ and $\mathcal{N}(\bullet)$ denotes the Gaussian distribution.

3.3 Speech Source Tracking Method Based on PF and GCTF

Under non-Gaussian noise environments, the PF is employed for speech source tracking. Assume that observation vectors $\mathbf{z}_k^j (j = 1, 2, \cdots, J)$ in the distributed microphone network with J nodes are conditionally independent given a particle $\mathbf{X}_k^n$. Then the global likelihood function $p(\mathbf{z}_k | \mathbf{X}_k^n)$ in Eq. (2) can be factorized into all local likelihood functions $p(\mathbf{z}_k^j | \mathbf{X}_k^n)$ in Eq. (12), written as

$$p(\mathbf{z}_k | \mathbf{X}_k^n) = \prod_{j=1}^{J} p(\mathbf{z}_k^j | \mathbf{X}_k^n) \tag{13}$$

Then a global MMSE estimate $\hat{\mathbf{x}}_k$ of the speech source state $\mathbf{x}_k$ can be obtained from Eq. (4).

The speech source tracking method based on PF and GCTF under non-Gaussian noise environments

is described as follows (abbreviated as PF-GCTF). Firstly, the TDOA observations of speech signals with non-Gaussian noise received by microphone pair are estimated from the GCTF according to Eq. (7). Taking into account the reverberation, multiple TDOA candidates are selected as observation vector at each node and based on them the multiple-hypothesis likelihood model is performed to calculate the local likelihood function. Next, predict the particles according to the dynamic model in Eq. (10), and global likelihood functions, i.e., weights, corresponding to the particles are computed from Eq. (13). Finally, a global position estimate of the speech source state can be obtained in form of weighted particles in Eq. (4).

4 SIMULATIONS AND DISCUSSIONS

4.1 Simulation Setup

To verify the performance of a speech source tracking method, the Root Mean Square Error (RMSE) is given as

$$\mathrm{RMSE}=\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{l}_{\mathbf{x}_k}-\hat{\boldsymbol{l}}_{\mathbf{x}_k}\right\|^2} \tag{14}$$

Where $\hat{\boldsymbol{l}}_{\mathbf{x}_k}$ and $\boldsymbol{l}_{\mathbf{x}_k}$ represent the position estimate and ground true position at time k, respectively, M denotes the number of Monte Carlo simulations.

In the SαS noise environment, the generalized signal noise ratio (GSNR) is used to describe the different non-Gaussian environments

$$\mathrm{GSNR}=10\log_{10}\frac{\sigma_s^2}{\gamma}(\mathrm{dB}) \tag{15}$$

Where σ_s^2 is the signal variance and γ is the dispersion parameter of the SαS noise.

In simulation experiments, a female speech source with the length about 4s and 16 kHz sampling frequency moves along a semicircle trajectory in a room which size is $5\mathrm{m}\times5\mathrm{m}\times3\mathrm{m}$, and microphone network has been constructed in advance with J=12 pairs of omni-direction microphones shown in Fig.1. The heights of microphones and speech source are set 1.5 m. The spacing distance of two microphones in each node is 0.6 m. The speech signal is split into 120 frames and each frame length is 32ms. The signal received by each microphone is captured by Image method (E. A. Lehmann, A. M. Johansson, et al., 2007), setting the size of room, reverberation time T60 and the microphone coordinates, then different impulsive noise is added to each microphone, generating different GSNRs and reverberations signals.

The simulation parameters are set as follows. For the Langevin model, $\beta=10\mathrm{s}^{-1}$ and $\bar{v}=1\mathrm{ms}^{-1}$; for the PF, the number of particles is N =500 and the initial states of speech source state are considered randomly; for the TDOAs, the number of the TDOAs is Nm=4; for the GCTF, the kernel size σ' is set as 0.5; for the multi-hypothesis model, q_0 =0.25 and the observations standard deviation is σ=5×10^{-5}.

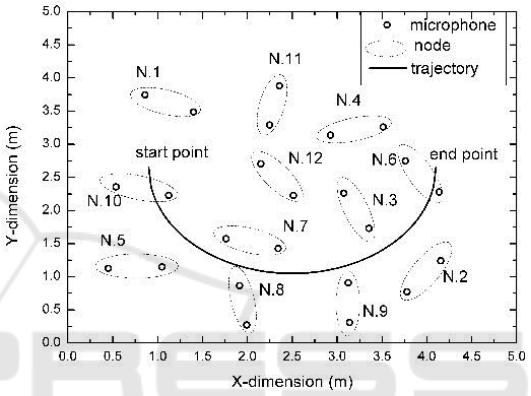

Figure 1. Speech source trajectory and layout of the 12 microphone pairs in X-Y plane.

4.2 Result Discussions

To evaluate the proposed method (PF-GCTF), some comparative experiments with the existing speech source tracking methods are conducted, i.e., the PF (D. B. Ward, E. A. Lehmann and R. C. Williamson, 2003) and the DPF-EKF (X. Zhong, A. Mohammadi, et al, 2013). These methods are evaluated in the RMSE results in Eq. (14), and the tracking results are averaged over 50 Monte Carlo simulations based on the same speech signal and the simulation setup.

4.2.1 Speech Source Tracking Results with Different Reverberation Time T60

Table 1 shows that the RMSE results of all methods with different reverberation time T60 from 100 ms to 300 ms, when GSNR=6 dB and α=0.8. It can be observed that the RMSE values of all methods become larger when reverberation gets heavier. Obviously, the DPF-EKF almost cannot track the moving speech source under different reverberations and the PF method has better tracking performance

only when T60 < 200 ms. It can be seen from Table 1 that the tracking performance of the PF-GCTF method is better than the PF and DPF-EKF with smaller RMSE values. It illustrates that the proposed method is robust to the environmental reverberations.

Table 1. Average RMSE results versus different reverberation times T60.

T_{60} (ms)	PF-GCTF (m)	PF (m)	DPF-EKF (m)
100	0.1062	0.1614	1.405
150	0.0921	0.1792	1.4544
200	0.1101	0.3987	1.7067
250	0.3385	0.63	1.8741
300	0.3974	0.9043	1.9738

4.2.2 Speech Source Tracking Results with Different GSNR

Fig.2 illustrates that the RMSE results of all methods with different GSNRs from -4 dB to 8 dB, when the reverberation time T60 = 100 ms and α=0.8. It can be seen from Fig.2 that with the rise of the GSNR the RMSE values of all methods become smaller. We can find that the tracking performance of the DPF-EKF is the worst with larger RMSE values and the PF method owns better tracking accuracy only when GSNR > 6 dB. However, the proposed method can successfully track the moving speech source under different GSNR conditions with smaller RMSE values. It implies that the PF-GCTF is a valid speech source method for non-Gaussian background noise of different GSNRs.

4.2.3 Speech Source Tracking Results with Different Characteristic Exponents α

The RMSE results of all methods with different characteristic exponents from 0.6 to 1.6 are illustrated in Table 2 when T60 = 100 ms and GSNR=0 dB. We can find that the PF and DPF-EKF have better tracking accuracies only when α > 1. Nevertheless, the RMSE values of the proposed method in Table 2 are smaller when 0.6 < α < 1.6 which implies the PF-GCTF is an effective speech source tracking method under non-Gaussian noise environments.

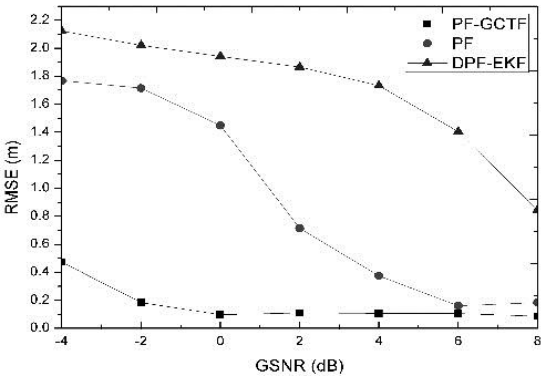

Figure 2. Average RMSE results versus different GSNRs.

Table 2. Average RMSE results versus different characteristic exponents α.

α	PF-GCTF (m)	PF (m)	DPF-EKF (m)
0.6	0.1565	2.2503	2.1875
0.8	0.1301	1.4484	1.9428
1	0.0967	0.1552	1.3202
1.2	0.0892	0.0853	0.2833
1.4	0.0846	0.0672	0.1517
1.6	0.09	0.0641	0.1302

5 CONCLUSIONS

In the paper, a tracking method based on PF and GTPF is proposed to estimate the positions of the moving speech source under non-Gaussian noise and reverberant environments. Since the generalized correntropy function is employed to estimate TDOAs, the proposed method based on PF can track a moving speech source successfully in non-Gaussian noise environments. Simulation results illustrate that the PF-GCTF outperforms other comparative methods and is robust against non-Gaussian background noise and room reverberations.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation for Young Scientists of China (Grant No.61801308).

REFERENCES

B. Kapralos, M. R. M. Jenkin and M. Evangelos, Audiovisual localization of multiple speakers in a video teleconferencing setting, Int. J. Imaging Syst. Technology, pp. 13 (1): 95-105 (2003).

C. Knapp, and G. C. Carter, The generalized correlation method for estimation of time delay, IEEE Trans. Acoust., Speech, Signal Process., pp. 24(4): 320-327 (1976).

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., pp. 11 (6):826-836 (2003).

E. T. Roig, F. Jacobsen and E. F. Grande, Beamforming with a circular microphone array for localization of environmental noise sources, J. Acoust. Soc. Am., pp. 128(6):3535-3542 (2010).

E. A. Lehmann, A. M. Johansson, et al., Reverberation-time prediction method for room impulse responses simulated with the image-source model, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 159-162 (2007).

F. Talantzis, An acoustic source localization and tracking framework using particle filtering and information theory, IEEE Trans. Audio Speech Lang. Process., pp. 18 (7):1806-1817 (2010).

K. Nakadai, et al., Robust tracking of multiple sound sources by spatial integration of room and robot microphone arrays, in: IEEE Int. Conf. Acoust., Speech, Signal Process., pp. IV- 929- IV-932 (2006).

M. F. Fallon, and S. J. Godsill, Acoustic source localization and tracking of a time-varying number of speakers, IEEE Trans. Audio Speech Lang. Process., pp. 20(4):1409-1415 (2012).

M. Shao and C. L. Nikias, Signal processing with fractional lower order moments: Stable processes and their applications, Proceedings of the IEEE, pp. 81 (7): 986-1010 (1993).

Q. Zhang, Z. Chen, and F. Yin, Distributed marginalized auxiliary particle filter for speaker tracking in distributed microphone networks, IEEE Trans. Audio Speech Lang. Process., pp. 24(11): 1921-1934 (2016).

T.P. Spexard, M. Hanheide, and G. Sagerer, Human-oriented interaction with an anthropomorphic robot, IEEE Trans. Robotics, pp.23 (5):852- 862 (2007).

W. Liu, P.P. Pokharel, et al., Correntropy: properties and applications in non-Gaussian signal processing, IEEE Trans. Signal Process., pp. 55 (11): 5286-5298 (2007).

X. Zhong, and J. R. Hopgood, Particle filtering for TDOA based acoustic source tracking: Non-concurrent multiple talkers, Signal Process., pp. 96(5):382-394 (2014).

X. Zhong, A. Mohammadi, et al., Acoustic source tracking in a reverberant environment using a pairwise synchronous microphone network, in: 16th Int. Conf. Information Fusion, pp. 953-960 (2013).