SPEAKER LOCALIZATION AND TRACKING IN THE PRESENCE OF SOUND INTERFERENCE BY EXPLOITING SPEECH HARMONICITY

Kai Wu, Shu Ting Goh, Andy W. H. Khong

School of Electrical & Electronic Engineering,
Nanyang Technological University, Singapore,
Email: wu0001ai@e.ntu.edu.sg, {shuting, andykhong}@ntu.edu.sg.

ABSTRACT

The performance of conventional acoustic source localization and tracking system reduces significantly when reverberation, noise, and acoustic interference are present. In this paper, a robust speaker tracking algorithm for an enclosed environment in the presence of interference and noise is proposed. We exploit the harmonic structure which is a distinctive feature in speech to enhance the robustness against acoustic interference. In order to extract the speech harmonic information, a beamformer is employed to enhance the signal from a prior estimated source location. A new particle weight update is then computed based on the steered response power function given the estimated speech harmonic information. Simulation results show that the proposed method achieves robustness in localization and tracking of a speech source in the presence of interference, noise and reverberation.

Index Terms— Acoustic localization and tracking, particle filter, speech harmonics, microphone array

1. INTRODUCTION

Acoustic source localization and tracking (ASLT) using a microphone array has been an active research area for applications including teleconferencing, automatic camera steering and surveillance [1]. However, room reverberation, background noise, and sound interference are some of the challenges that need to be addressed for realistic applications. Therefore, there is a high demand for developing robust algorithms that operate well under an adverse environment.

Conventional localization methods can be classified into single-step or dual-step approaches. In single-step approaches, the source location is directly estimated from the received signal by scanning across all possible source locations [2–4]. In dual-step approaches, the time-difference-of-arrivals (TDOAs) are first estimated and subsequently used to locate the source given the microphone array geometry [5–7]. However, these approaches estimate the source location by assuming it is independent across each time frame.

More recently, state-space approaches which exploit the fact that the measurements of the source signal retain temporal consistency across successive frames have been proposed [8–14], leading to a tracking scenario. The particle filter (PF) [15] was introduced in acoustic source tracking to achieve robustness against reverberation and noise [8,9]. A voice activity detection module was subsequently integrated into the ASLT framework to deal with silent periods in non-stationary speech [10]. In addition, a track-before-detect framework that is capable of reducing the computation load for a large number of particles has been presented in [11]. The problem of tracking for multi-targets and time-varying number of targets has also been investigated in [11, 12].

Although significant progress has been made, the problem of speech source tracking in the presence of interference remains an open problem. Existing tracking methods employ the steered response power beamformer with phase-transform (SRP-PHAT) as a measurement likelihood for particle weight update [9–12]. However, since these approaches are non-discriminative in nature, the presence of interference will result in consistently high likelihood at the location(s) of the interferer(s) which, in turn, causes particles to converge at the wrong location away from the speech source.

In this paper, a speech harmonicity based ASLT algorithm is proposed to deal with the effect of interference. Although the authors of [16, 17] proposed localization methods by jointly estimating the pitch frequency and source position, their primary aim is not to reject interference signals. In addition, conventional tracking methods do not consider any source spectrum feature [9–12]. Our proposed method, on the other hand, incorporates distinctive speech harmonics and a PF framework to achieve robust speech source tracking in the presence of interference. First, a beamformer is employed to enhance the signal from a prior estimated source location to extract the speech harmonic information. A new SRP function is then constructed by considering the fact that speech energy is concentrated on the harmonic bands, while the interference energy may be distributed over different frequency regions. Finally, a new particle weight update scheme is derived based on the new SRP function to achieve speech-sensitive source tracking. Simulations are conducted to compare the tracking performance between the proposed and the conventional method in the presence of interference, noise and reverberation.

2. THE PROPOSED FRAMEWORK

2.1. Source dynamic model

The bootstrap PF is commonly used in the ASLT due to its low computational complexity [18]. The source state $\boldsymbol{\alpha}_k$ is defined as $\boldsymbol{\alpha}_k = [x_k, y_k, \dot{x}_k, \dot{y}_k]^T$ at time instant k, where x_k and y_k correspond to the source position while $\dot{x}_k$ and $\dot{y}_k$ are the source velocity in x and y direction, respectively. We also define the observation variable $\mathbf{z}_k = [\widehat{x}_k, \widehat{y}_k]^T$ which contains the source location estimate. In the ASLT framework, the source dynamic is described using a first-order Markov process given by

$$\boldsymbol{\alpha}_k = g(\boldsymbol{\alpha}_{k-1}, \mathbf{u}_k), \tag{1}$$

This research, which is carried out at BeingThere Centre, is supported by the Singapore National Foundation under its International Research @ Singapore Funding Initiative and administered by the IDM Programme Office.

where $g(\cdot)$ denotes the state-transition process and $\mathbf{u}_k$ denotes the process noise. Similar to [8–11, 14], we employ the Langevin process which had been proposed as a source-dynamic model for simulating a realistic human motion. The state-transition $g(\cdot)$ in (1) can be represented by

$$\boldsymbol{\alpha}_k = \begin{bmatrix} 1 & 0 & aT & 0 \\ 0 & 1 & 0 & aT \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix} \boldsymbol{\alpha}_{k-1} + \begin{bmatrix} bT & 0 \\ 0 & bT \\ b & 0 \\ 0 & b \end{bmatrix} \mathbf{u}_k, \quad (2)$$

where $\mathbf{u}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the noise variable, T is the time interval between consecutive frames, $\boldsymbol{\mu} = [0, 0]^T$ and $\boldsymbol{\Sigma} = \mathbf{I}_{2\times 2}$ correspond to the mean vector and covariance matrix, respectively. The parameters a and b are defined as

$$a = \exp(-\beta T), \quad (3a)$$

$$b = \bar{v}\sqrt{1 - a^2}, \quad (3b)$$

where $\bar{v}$ is the steady-state velocity and β is the rate constant. In this paper, we have used, similar to [10], $\bar{v} = 0.8$ m/s, $\beta = 10$ Hz.

2.2. Prior Prediction

In the ASLT framework, particles are propagated according to the source-dynamic model before being weighted by the particle likelihood. Existing approaches compute the particle weights by employing a pseudo-likelihood that has been derived from the SRP-PHAT measurements [9–12]. While these techniques achieve good localization and tracking performance, their performance may reduce in the presence of interference. This is due to the inability of SRP-PHAT to discriminate between the speech source and the acoustic interference in general. It implies that any acoustic interference will result in a dominant peak occurring at the interferer's position, and the particles are likely to propagate towards that location away from the speech source. The performance of these algorithms reduces significantly in low signal-to-interference ratio (SIR), resulting in the ASLT losing track of the speech source.

To mitigate the degradation in performance, we exploit speech features such that the likelihood measurement is predominantly weighted by the speech signal as opposed to the interferers. We propose to first estimate the prior source position using (1) such that this prior state estimate is given by

$$\widehat{\boldsymbol{\alpha}}_k^- = g(\widehat{\boldsymbol{\alpha}}_{k-1}^+, \mathbf{u}_k), \quad (4)$$

where $\widehat{\boldsymbol{\alpha}}_{k-1}^+$ is the posterior state estimate at time instant $k-1$. The prior source location estimate

$$\widehat{\boldsymbol{\ell}}_k^- = [\widehat{x}_k^-, \widehat{y}_k^-]^T, \quad (5)$$

corresponds to the first two elements in $\widehat{\boldsymbol{\alpha}}_k^-$. This prior estimate is based only on the knowledge of the source motion. The feature-directed measurements, as will be described in subsequent subsections, will further refine the state estimate.

2.3. Feature Extraction

Various techniques can be employed to enhance the signal after a prior source location has been estimated. We consider the delay-and-sum beamformer [19] due to its simplicity although other forms of beamformer such as presented in [20, 21] can be used to enhance the speech signal. The delay-and-sum beamformer output for the prior estimated source location $\widehat{\boldsymbol{\ell}}_k^-$, is given by

$$S(\omega, \widehat{\boldsymbol{\ell}}_k^-) = \sum_{m=1}^{M} \Phi\left(D_m(\widehat{\boldsymbol{\ell}}_k^-)\right) F_m(\omega) e^{\jmath\omega D_m(\widehat{\boldsymbol{\ell}}_k^-)/c}, \quad (6)$$

where m is the microphone index, M is the number of microphones, $F_m(\omega)$ is the frequency-domain received signal from the mth microphone, ω is the angular frequency, c is the speed of sound, $D_m(\widehat{\boldsymbol{\ell}}_k^-) = \|\widehat{\boldsymbol{\ell}}_k^- - \boldsymbol{\ell}_m\|_2$ is the distance from prior estimated source position to the mth microphone, and $\Phi(\cdot)$ is a monotonic function that weighs the mth microphone signal according to the source-sensor distance. In our simulations, we found that $\Phi\left(D_m(\widehat{\boldsymbol{\ell}}_k^-)\right) = 1/D_m(\widehat{\boldsymbol{\ell}}_k^-)$ performs well since it emphasizes the signal from the microphone that is closer to the source.

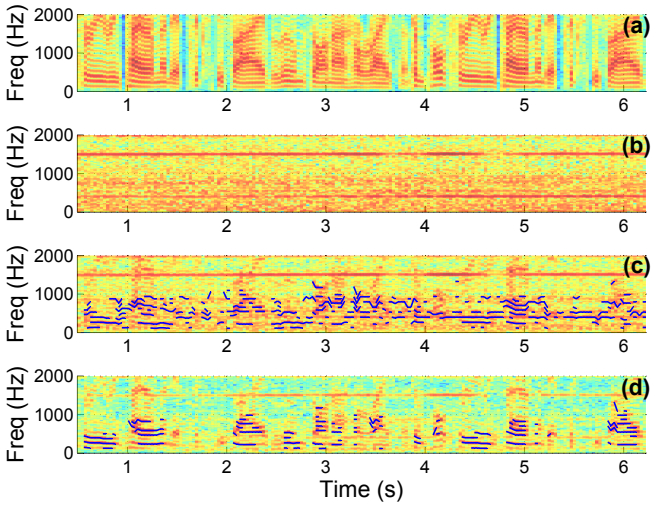

Fig. 1. Spectrogram and selected harmonic bands indicated in blue lines. (a) Clean speech. (b) Power-drill interference. (c) Reference microphone received signal and its selected harmonic bands (in blue). (d) Beamformer enhanced signal and its selected harmonic bands (in blue).

It is well-known that the speech spectrum possesses a harmonic structure [22] (see Fig. 1 (a)), where the harmonics correspond to multiple integers of a pitch frequency. We assume, in this work, that the interferers either do not share the same harmonic bands as the speech spectrum (due to differences in pitch frequency) or the spectrum of the interferer(s) do not exhibit any harmonicity. Figure 1 (b) shows the spectrum of the power drill extracted from the NOISEX-92 database [23]. We note from this spectrum that although there are dominant energies at 400 and 1500 Hz, no harmonic structure is present.

The purpose of the proposed method is to discriminate these spectral components and extract harmonic bands corresponding to the human speech for particle likelihood computation since these regions provide a high SIR. In general, the spectrum corresponding to the signal received from one channel is often distorted, especially for the case where the interferer is close to the microphone as shown in Fig. 1 (c). Extraction of speech harmonics is therefore challenging. In the proposed method, the signal is first enhanced by the proposed beamforming technique using (6), and the enhanced spectrum, as shown in Fig. 1 (d), will be used for speech feature extraction.

To extract the speech harmonics from a noisy spectrum, we employ the multi-band excitation (MBE) fit method [24, 25]. This model defines a voiced frame in the frequency domain as a product

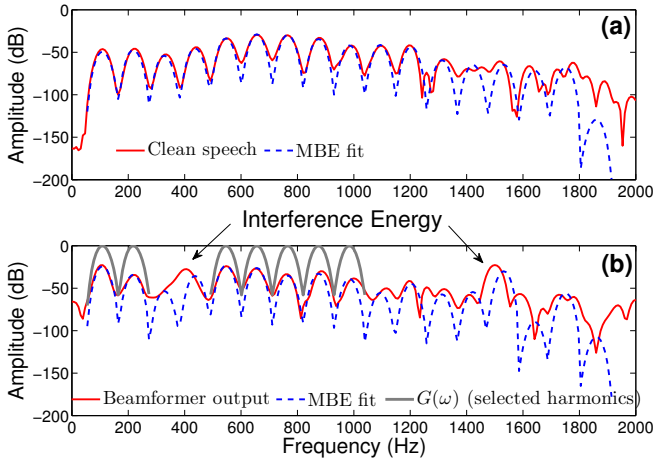

Fig. 2. MBE fitting result. (a) Clean speech and MBE fit. (b) Beamformer output, MBE fit and $G(\omega)$ in the presence of a power drill signal.

of excitation spectrum $E(\omega, \omega_0)$ and spectrum envelope $H(\omega)$ given by [24]

$$S_{\text{spch}}(\omega) = H(\omega)E(\omega, \omega_0), \tag{7a}$$

$$E(\omega, \omega_0) = \sum_{q=1}^{Q} \Psi(\omega - q\omega_0), \tag{7b}$$

where q is the harmonic index, Q is the number of harmonics, ω_0 is the pitch frequency, and $\Psi(\omega)$ is the Fourier transform of the Hamming window. For a distorted speech signal, the parameters can be estimated by minimizing the fitting error summed over all the harmonic bands, i.e.,

$$\varepsilon(\omega_0) = \sum_q \varepsilon_q(\omega_0), \tag{8}$$

where the fitting error for each harmonic band $\varepsilon_q(\omega_0)$ is given by

$$\varepsilon_q(\omega_0) = \frac{1}{2\pi} \int_{a_q}^{b_q} |S(\omega, \widehat{\boldsymbol{\ell}}_k^-) - H_q E(\omega, \omega_0)|^2 d\omega. \tag{9}$$

Here, $S(\omega, \widehat{\boldsymbol{\ell}}_k^-)$ has been defined in (6), $H(\omega)$ from (7a), is decoupled into several complex amplitude H_q for each harmonic band q, and the interval $[a_q, b_q]$ is the frequency band centered on the qth harmonic, where $a_q = (q - 0.5)\omega_0$ and $b_q = (q + 0.5)\omega_0$.

The complex amplitude for each harmonic band H_q can be obtained by considering the derivative of (9) to be zero giving

$$H_q = \frac{\int_{a_q}^{b_q} S(\omega, \widehat{\boldsymbol{\ell}}_k^-) E^*(\omega, \omega_0) d\omega}{\int_{a_q}^{b_q} |E(\omega, \omega_0)|^2 d\omega}. \tag{10}$$

Therefore, feature extraction is performed using the following steps: each fitting error $\varepsilon_q(\omega_0)$ is evaluated using the optimal value of H_q obtained in (10). The error function in (8) is then computed with respect to all pitch frequencies ω_0 of interest. Finally, the global minimum of $\varepsilon(\omega_0)$ is determined and the corresponding ω_0 is selected as the estimated $\widehat{\omega}_0$.

2.4. Feature-directed Particle Weight Update

To obtain the feature-directed particle weight update, it is important to determine the most reliable harmonic bands and select those harmonic bands for the computation of the likelihood. Two criterions are proposed to measure the reliability of the harmonic bands: (1) the normalized fitting error and (2) the normalized harmonic energy. The normalized fitting error [25],

$$\bar{\varepsilon}_q = \frac{\varepsilon_q(\widehat{\omega}_0)}{\frac{1}{2\pi}\int_{a_q}^{b_q} |S(\omega, \widehat{\boldsymbol{\ell}}_k^-)|^2 d\omega} \tag{11}$$

is defined, for each harmonic, as the effectiveness of a given frequency band to be fitted into the speech harmonic model. The normalized harmonic energy, on the other hand, is defined by the ratio of energy distributed on that harmonic over the total energy, i.e.,

$$P_q = \frac{\int_{a_q}^{b_q} H_q E(\omega, \widehat{\omega}_0) d\omega}{\sum_{q=1}^{Q} \int_{a_q}^{b_q} H_q E(\omega, \widehat{\omega}_0) d\omega}. \tag{12}$$

Since the energy of the speech signal is expected to be concentrated in a harmonic structure, those harmonic bands with low fitting error and high energy ratio are more likely to retain most of the speech components, while other regions are expected to contain the interference signal. We therefore set two harmonic-band thresholds ζ and η for selecting the reliable (speech) harmonic bands such that

$$G_q(\omega) = \begin{cases} \Psi(\omega - q\widehat{\omega}_0), \text{ if } \bar{\varepsilon}_q \le \zeta \;\&\; P_q \ge \eta,\ \omega \in [a_q, b_q] \\ 0, \text{ otherwise} \end{cases} \tag{13a}$$

$$G(\omega) = \sum_q G_q(\omega). \tag{13b}$$

Figure 2 shows extraction results of the speech harmonics using a frame of 32 ms. Figure 2 (a) shows the MBE fitting result, computed using (8)-(10), for the case of clean speech where no interferer is present. We note that the MBE approximation, shown by the dotted line, is capable of estimating the harmonics of clean speech. Figure 2 (b) shows result for the case where a power-drill signal is

Table 1. Summary of the proposed algorithm.

At time $k-1$, given that a set of particles $\{\boldsymbol{\alpha}_{k-1}^{(n)}, w_{k-1}^{(n)}\}_{n=1}^{Ns}$ is a discrete representation of posterior $p(\boldsymbol{\alpha}_{k-1}|\mathbf{z}_{k-1})$, the posterior state estimate is $\widehat{\boldsymbol{\alpha}}_{k-1}^{+} = \sum_{n=1}^{Ns} w_{k-1}^{(n)} \boldsymbol{\alpha}_{k-1}^{(n)}$.

For the kth frame:

1. *Prior prediction*: Propagate the state estimate through (4) to obtain prior estimate of the current state $\widehat{\boldsymbol{\alpha}}_k^-$.
2. *Feature extraction*: Apply (5) (6) to enhance the signal from $\widehat{\boldsymbol{\ell}}_k^-$, and extract speech features using (8)-(10).
3. *Particles propagation*: Propagate each particle through the source dynamic model (1), $\boldsymbol{\alpha}_k^{(n)} = g(\boldsymbol{\alpha}_{k-1}^{(n)}, \mathbf{u}_k)$.
4. *Posterior weights update*: Obtain the feature directed particle likelihood using (11)-(15) and each particle is then assigned a weight according to its likelihood $\widetilde{w}_k^{(n)} = w_{k-1}^{(n)} p(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(n)})$, followed by normalization $w_k^{(n)} = \widetilde{w}_k^{(n)} (\sum_{i=1}^{N} \widetilde{w}_k^{(i)})^{-1}$. The posterior state estimate is $\widehat{\boldsymbol{\alpha}}_k^{+} = \sum_{n=1}^{Ns} w_k^{(n)} \boldsymbol{\alpha}_k^{(n)}$.
5. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{n=1}^{N} (w_k^{(n)})^2)^{-1}$.

Table 2. Comparison of mean tracking error ($\bar{e}$) between the SRP-PHAT method and the proposed method.

	SRP-PHAT tracking method		Proposed tracking method	
	$T_{60} = 0.2$ s	$T_{60} = 0.3$ s	$T_{60} = 0.2$ s	$T_{60} = 0.3$ s
PD (SIR = 3 dB)	0.55 (m)	0.57 (m)	0.09 (m)	0.15 (m)
TR (SIR = -3 dB)	0.52 (m)	0.56 (m)	0.07 (m)	0.10 (m)
PD+TR (SIR = 3, 0 dB)	0.50 (m)	0.69 (m)	0.08 (m)	0.13 (m)
PD+TR (SIR = 3, -3 dB)	0.58 (m)	0.65 (m)	0.08 (m)	0.14 (m)

added into the speech signal at an SIR=5 dB. The beamformer output $S(\omega, \widehat{\boldsymbol{\ell}_k})$, shown by the solid line, therefore consists of spectral components corresponding to the power drill at 400 and 1500 Hz and the speech signal. Comparing Figs. 2 (a) and (b), we note that the MBE fit shown in Fig. 2 (b) is able to estimate the speech harmonics with reasonable accuracy albeit with some distortion. Estimation of reliable speech harmonic bands is shown with $G(\omega)$ denoted by the bold lines (which has been normalized to 0 dB for clarity.) Speech harmonics that are selected using $G(\omega)$ are shown in Fig. 1 (d) where a 6 s speech with the presence of power-drill interference is considered. We note that employing the beamformer and MBE fit, speech harmonic bands can be estimated as indicated by the dark lines of the spectrogram.

With $G(\omega)$ in (13b), the new SRP function $P(\boldsymbol{\ell})$ with weight $W_m(\omega)$ is given by

$$P(\boldsymbol{\ell}) = \int_{\Omega} \left| \sum_{m=1}^{M} W_m(\omega) F_m(\omega) e^{\jmath \omega D_m(\boldsymbol{\ell})/c} \right|^2 d\omega, \qquad (14a)$$

$$W_m(\omega) = \frac{G(\omega)}{|F_m(\omega)|}, \qquad (14b)$$

where Ω is the frequency over which the SRP function is evaluated. Similar to the pseudo likelihood method [9, 10], the SRP function is used to define the measurement likelihood in the PF framework,

$$p(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \begin{cases} P^r(\boldsymbol{\ell}), & \text{for voiced frame} \\ \mathcal{U}_D(\boldsymbol{\ell}), & \text{for unvoiced frame} \end{cases}, \qquad (15)$$

where r is a control parameter to regulate the SRP function for source tracking [10], and $\mathcal{U}_D(\cdot)$ is the uniform PDF over the considered enclosure domain $D = \{x_k, y_k | x_{\min} \leq x_k \leq x_{\max}, y_{\min} \leq y_k \leq y_{\max}\}$. The likelihood function is used as weights to update the particles. The proposed ASLT framework is summarized in Table 1.

3. SIMULATION RESULTS

Simulations were conducted using synthetic impulse responses generated by the method of images [26]. The dimension of the room was 5 m × 5 m × 2.5 m, and the reverberation time T_{60} were 200 and 300 ms. Eight microphones were distributed along the perimeter of the room. (see Fig. 3). An 8 s male speech sampled at 16 kHz from the TIMIT database [27] was used as a source signal. A power drill (PD) signal obtained from the NOISEX-92 database [23] and a recorded telephone ring (TR) signal were used as interferers. White Gaussian noise (WGN) of 15 dB SNR was added to the microphone signals. The positions of speech source were computed using a frame size of 512 samples with $N = 100$ particles. We also used an effective sample size threshold $N_{\text{thr}} = 37.5$, a nonlinear exponent $r = 2$, harmonic-band thresholds $\zeta = 0.6$ and $\eta = 0.03$. Total of 12 harmonic bands ($Q = 12$) was considered. The proposed method is compared with the conventional method using SRP-PHAT [10]. Both methods were evaluated using $0 \leq \Omega \leq 2$ kHz from which, for the proposed algorithm, speech pitch frequency was estimated from 100 to 300 Hz using (8)-(10). In this paper, we quantify the performance using $e_k = ||\widehat{\boldsymbol{\ell}}_k^+ - \boldsymbol{\ell}_k||_2$, where the $\widehat{\boldsymbol{\ell}}_k^+$ is the posterior estimated position at kth frame, and $\boldsymbol{\ell}_k$ is the true source position. The average tracking error $\bar{e} = \frac{1}{K}\sum_{k=1}^{K} e_k$ quantifies the performance across all audio frames.

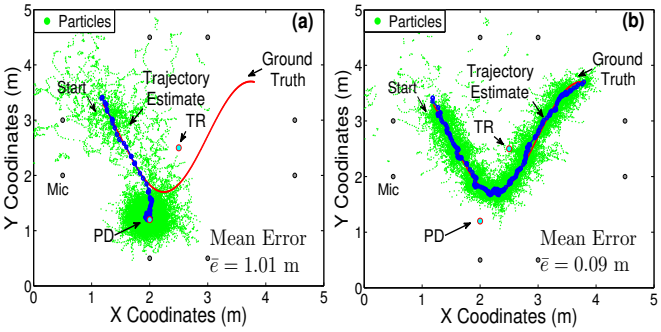

Fig. 3. Comparison of tracking results when both PD and TR are present at SIR = 0 dB, $T_{60} = 200$ ms. (a) Conventional SRP-PHAT tracking method. (b) Proposed tracking method.

Figure 3 compares the tracking result for $T_{60} = 200$ ms with both telephone and power drill at 0 dB SIR. Figure 3 (a) shows that the tracking performance of the conventional SRP-PHAT approach is adversely affected by the interferers. Due to the high measurement likelihood of SRP-PHAT for the interferer regions, the particles will be 'trapped' once they are propagated there, in this case the region near the power drill. The SRP-PHAT method has an average error of 1.01 m indicating that it does not converge to the speech source trajectory. On the other hand, Fig. 3 (b) shows the tracking performance of the proposed method. This result shows that the proposed method is less significantly affected by the presence of interferers achieving an average error of less than 0.1 m.

Table 2 shows the average tracking error for various test conditions. These results show that the proposed algorithm can achieve better accuracy than the SRP-PHAT method. For instance, in the presence of power drill at 3 dB SIR, the SRP-PHAT method exhibits a large tracking error of 0.55 m when $T_{60} = 0.2$ s. The proposed method achieves an error of less than 0.1 m which translates to an 80% reduction of error over the SRP-PHAT method. Furthermore, the proposed method maintains its robustness in localization and tracking in the presence of two interferers while the SRP-PHAT approach suffers from large tracking error under low SIR condition.

4. CONCLUSION

A speech harmonic extraction based ASLT framework is proposed. This method is capable of estimating the speech harmonic bands for localizing and tracking. By only emphasizing the harmonic bands, a better speech-sensitive measurement likelihood can be achieved resulting in a better weight update for the particles. Simulation results show that the proposed method can achieve a lower tracking error than the conventional SRP-PHAT method in the presence of multiple interferers.

5. REFERENCES

[1] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Magazine*, vol. 28, pp. 20–32, Jan. 2011.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas and Propagation*, vol. 34, no. 3, pp. 276 – 280, Mar. 1986.

[3] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327 –1339, May 2007.

[4] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2510 –2526, Nov. 2007.

[5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 24, no. 4, pp. 320 – 327, Aug 1976.

[6] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP J. on Adv. Signal Process.*, vol. 2006, 2006.

[7] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 943 –956, Nov. 2001.

[8] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'01)*, 2001, pp. 3021–3024.

[9] 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.

[10] E. A. Lehmann and A. M. Johansson, "Particle filter with integrated voice activity detection for acoustic source tracking," *EURASIP J. on Adv. Signal Process.*, vol. 2007, 2007.

[11] M. F. Fallon and S. Godsill, "Acoustic source localization and tracking using track before detect," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1228–1242, 2010.

[12] 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.*, vol. 20, no. 4, pp. 1409–1415, 2012.

[13] A. Levy, S. Gannot, and E. A. P. Habets, "Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1540 –1555, Aug. 2011.

[14] F. Talantzis, "An acoustic source localization and tracking framework using particle filtering and information theory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1806 –1817, Sep. 2010.

[15] N. J. Gordon, D. J. Salmond, and A. F. M. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation," in *Proc. IEE -F, Radar and Signal Process.* IET, 1993, vol. 140, pp. 107–113.

[16] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint doa and multi-pitch estimation based on subspace techniques," *EURASIP J. on Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–11, 2012.

[17] M. Kepesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Commun. and Microphone Arrays, 2008*, May, pp. 85–88.

[18] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174 –188, Feb. 2002.

[19] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*, Simon & Schuster, 1992.

[20] S. Timofeev, A. R. S. Bahai, and P. Varaiya, "Adaptive acoustic beamformer with source tracking capabilities," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 2812–2820, 2008.

[21] E. A. P. Habets, J. Benesty, and P. A. Naylor, "A speech distortion and interference rejection constraint beamformer," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 854–867, 2012.

[22] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Wiley-IEEE Press, 2000.

[23] A. Varga and 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[24] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 36, no. 8, pp. 1223 –1235, Aug. 1988.

[25] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2914–2919, 1999.

[26] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, no. 1, pp. 269–277, July 2008.

[27] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, 1993.