Chapter 3
Sound Source Localization and Tracking

Kai Wu and Andy W.H. Khong

Abstract Sound source localization and tracking plays an important role in a teleconferencing system and social robot applications. Given the location of a sound, the social robot can be endowed with the capability of sound event awareness, which results in enhanced interaction with human beings. This chapter presents the problem of sound source localization and tracking, highlights their challenges, and reviews several existing techniques. In addition, a speech source tracking algorithm is proposed in order to achieve robust speaker tracking in the presence of sound interferers. Simulation is conducted and shows the effectiveness of the proposed method in a typical room environment.

3.1 Introduction

Sound source localization and tracking (SSLT) refers to the problem of estimating the location from which a sound signal originates with respect to the microphone array geometry. It plays an important role in a teleconferencing system and in social robot applications. In a teleconference scenario, a camera that is capable of automatic steering can be deployed to focus on the speaker given the estimated speaker position [22, 29]. In addition, source localization is often required and regarded as a preprocessing step before the enhancement of an acoustic signal from a particular location [20]. In the domain of social robotics, the localization technique is applied so that the robot can concentrate on a subject of interest or be made aware of where other sound events are coming from.

Multiple microphones are, in general, required in order to achieve SSLT. Different microphone array configurations have been used in the recent literature, e.g., binaural microphones [5], linear array [39], circular array [9] and distributed microphone

K. Wu (✉) · A.W.H. Khong
BeingThere Centre, Nanyang Technological University, Singapore, Singapore
e-mail: WU0001AI@e.ntu.edu.sg

A.W.H. Khong
e-mail: AndyKhong@ntu.edu.sg

N. Magnenat-Thalmann et al. (eds.), *Context Aware Human-Robot and Human-Agent Interaction*, Human–Computer Interaction Series,
DOI 10.1007/978-3-319-19947-4_3

arrays [13, 25]. The source position is estimated by exploiting the range differences from the source to the microphones. Although various algorithms have been developed in recent decades for SSLT applications, room reverberation, background noise, and sound interference are some of the key challenges that need to be addressed in a realistic environment. In the context of room acoustics, the microphones capture not only the direct-path propagation component of the source signal but also the multipath propagation component due to the reflections at the room boundaries. The multipath component, together with the background noise, distorts the time delay information contained in the microphone received signals and degrades the localization performance. In addition, one is often interested in localizing and tracking a desired source (e.g., human speech source) in the presence of certain sound interferers (e.g., fan noise, air-conditioner noise) which often exist in a room environment. These interferers may distract the system which, as a result, localizes the interferers rather than the desired source.

The organization of this chapter is as follows: in Sect. 3.2, mathematical formulation of the SSLT problem is introduced. Conventional localization and tracking methods are then reviewed. In Sect. 3.3, a proposed method that deals with the problem of speech source tracking in the presence of sound interference is discussed. The proposed method exploits the speech harmonicity feature so as ensure that only speech signals are used for tracking. The integration of SSLT for social robot application is discussed in Sect. 3.4. Finally, the future possible research directions and conclusions are presented in Sects. 3.5 and 3.6, respectively.

3.2 Overview of Sound Source Localization and Tracking Algorithms

SSLT algorithms can be classified into two categories: localization approach and tracking approach. The localization approach assumes independence between successive audio frames and estimates the source location independently across each data frame. The tracking approach exploits consistency between successive frames by assuming that the source is stationary or moving at a slow rate. In this section, the mathematical formulation for these two approaches is discussed.

3.2.1 Mathematical Formulation of Sound Source Localization

The SSLT problem is illustrated in Fig. 3.1. The speech signal $s(n)$ radiates away from the source position and propagates to the microphones. The received signals contains not only direct-path but also multipath components caused by reflection from

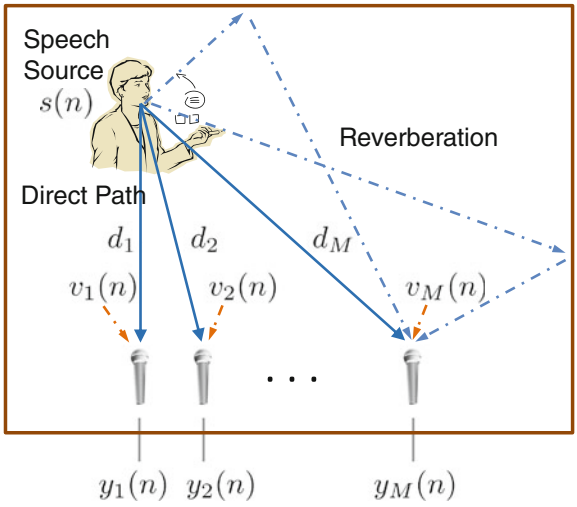

Fig. 3.1 Signal propagation model

the room boundaries. Within a short time frame, the channel from the source to the ith microphone can be considered as a linear time-invariant system and is represented by a channel impulse response $h_i(n)$. The ith microphone received signal can thus be formulated as [3]

$$y_i(n) = s(n) * h_i(n) + v_i(n), \quad i = 1, 2, \ldots, M, \tag{3.1}$$

where $*$ is the convolution operator, $v_i(n)$ is the additive noise, and M is the number of microphones. In order to infer the signal delay information, the impulse response $h_i(n)$ can be further decomposed into a direct-path component and a multipath component. The microphone received signal can thus be rewritten as

$$y_i(n) = a_i s(n - \tau_i) + s(n) * h_i'(n) + v_i(n), \quad i = 1, 2, \ldots, M, \tag{3.2}$$

where $0 \leq a_i \leq 1$ is the attenuation factor due to propagation, τ_i is the direct-path time delay from the source to the ith microphone, and $h_i'(n)$ denotes the remaining impulse response which is defined as the difference between the original response and the direct-path component. In (3.2), the time delay τ_i is dependent on the source position with respect to the microphone array. However, direct estimation of τ_i is not achievable since SSLT is a passive localization problem. Most of the algorithms exploit the relative time delay information among microphones and one such algorithm is introduced in the following section.

3.2.2 Sound Source Localization Using Beamforming-Based Approach

Given the microphone received signal $y_i(n)$, localization is usually performed using each data frame defined as

$$\mathbf{y}_i(k) = [y_i(kN)\ y_i(kN+1)\ \ldots\ y_i(kN+N-1)], \tag{3.3}$$

where N is the frame length and k is the frame index. Beamforming is one of the widely used approaches for sound source localization. In principle, the beamformer computes the spatial power spectrum for the whole region of interest and searches for the highest power corresponding to the source position estimate (see Fig. 3.2 for example). The family of beamforming techniques includes steered response power (SRP) [8, 10], minimum variance distortionless response [34], linearly constrained minimum variance [11, 34], etc.

The SRP beamformer gained popularity due to its simplicity. Considering M microphones, the SRP function defines the power

$$\mathcal{P}_k(\mathbf{r}') = \sum_{\omega_l \in \Omega} \left| \sum_{i=1}^{M} W_i(k,\omega_l) Y_i(k,\omega_l) e^{j\omega_l \|\mathbf{r}' - \mathbf{r}_i^{\mathrm{m}}\|_2 / c} \right|^2 \tag{3.4}$$

corresponding to the current steered location $\mathbf{r}'$ at time frame k, where $\mathbf{r}' = [x'\ y']^T$ is the steered location in the region of interest, $W_i(k,\omega_l)$ is a weighting function, $Y_i(k,\omega_l)$ is the short-time Fourier transform of the ith microphone received signal defined as $Y_i(k,\omega_l) = \mathcal{F}(\mathbf{y}_i(k))$, ω_l is the angular frequency of the lth bin index, c is the speed of sound, $\mathbf{r}_i^{\mathrm{m}}$ is the position of the ith microphone, $\|\mathbf{r}' - \mathbf{r}_i^{\mathrm{m}}\|_2$ is the distance from the steered location to the ith microphone position, and Ω is the interested frequency range over which the computation is carried out. In (3.4), the

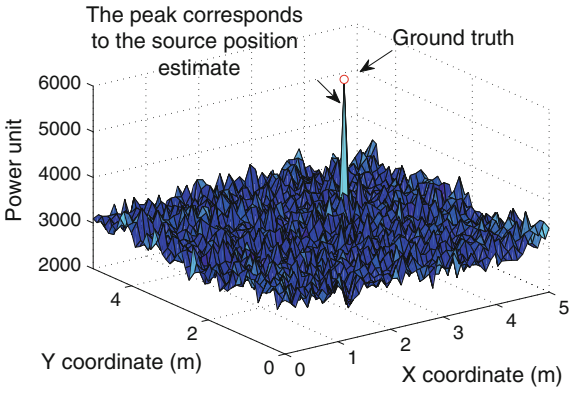

Fig. 3.2 The power spectrum when SNR = 20 dB, $T_{60} = 150$ ms. The ground truth of the source position is denoted by the *circle dot* which is plotted on *top* of the spectrum for clarity of presentation

SRP is performed by computing the time delay from the steered location $\mathbf{r}'$ to each microphone in the first step. The corresponding power is then calculated by time aligning the signals in the frequency domain according to the signal delays and summing over all the microphones. The weighting function $W_i(k, \omega_l)$ is important in power calculation. While different weighting functions can be used [24], the phase transform (PHAT) given as

$$W_i^{\text{PHAT}}(k, \omega_l) = \frac{1}{|Y_i(k, \omega_l)|} \tag{3.5}$$

remains one of the most commonly used weighting schemes. The corresponding beamformer is therefore named as SRP-PHAT. By substituting (3.5) into (3.4), it can be seen that the PHAT weighting is independent of the source energy and the computed SRP response is only dependent on the phase delay.

Furthermore, by steering the beamformer across the whole region of interest, one can obtain the power spectrum as shown in Fig. 3.2. Estimating the source position is therefore achieved by searching for the location that corresponds to the maximum power, i.e.,

$$\widehat{\mathbf{r}}_k = \arg\max_{\mathbf{r}' \in \mathcal{D}} \mathcal{P}_k(\mathbf{r}'), \tag{3.6}$$

where $\mathcal{D} = \{x, y | x_{\min} \leq x \leq x_{\max}, y_{\min} \leq y \leq y_{\max}\}$ is the considered searching domain.

It has been shown in [7] that the beamforming method achieves higher spatial resolution than other localization methods such as those based on time-difference-of-arrival method [3]. However, one drawback is the high computation complexity required for scanning the region of interest. Some researchers choose different resolution grids to reduce the computation burden [12]. In addition, a recently proposed work integrates the energy in each discrete grid to achieve better performance [4].

3.2.3 Sound Source Tracking Using Particle Filter-Based Approach

The localization algorithm discussed in Sect. 3.2.2 estimates the source position using each microphone data frame $\mathbf{y}_i(k)$ independently. The performance reduces when the background noise and reverberation increase since under these conditions, some of the data frames suffer from signal distortion and are therefore unable to provide reliable location estimates. However, if we assume that the source is stationary or moving at a low rate with respect to the convergence of the tracking algorithm, one possible approach to improve the performance is to exploit the temporal consistency of location measurements across successive frames.

We now consider successive data frames $\{\mathbf{y}_i(k) | \, k = 1, 2, \ldots, K\}$ where k is the frame index, and K is the total number of audio frames. The aim is to estimate the

source positions over all the time frames, leading to a source tracking problem. We first define the state variable as $\boldsymbol{\alpha}_k = [x_k \; y_k \; \dot{x}_k \; \dot{y}_k]^T$ at frame index k, where x_k and y_k correspond to the source position while $\dot{x}_k$ and $\dot{y}_k$ are the source velocities in x and y direction, respectively. Similarly, the measurement variable $\mathbf{z}_k = [\widehat{x}_k \; \widehat{y}_k]^T$ is defined. This measurement vector can be obtained from the SRP location estimate by evaluating (3.4)–(3.6) for the kth time frame data. Therefore, the state-space model can be written as

$$\boldsymbol{\alpha}_k = \mathcal{G}(\boldsymbol{\alpha}_{k-1}, \mathbf{u}_k), \tag{3.7a}$$

$$\mathbf{z}_k = \mathcal{H}(\boldsymbol{\alpha}_k, \mathbf{w}_k), \tag{3.7b}$$

where $\mathcal{G}(\cdot)$ is the process function defining the time evolution of the state, $\mathbf{u}_k$ is the process noise, $\mathcal{H}(\cdot)$ is the measurement equation defining the mapping from $\boldsymbol{\alpha}_k$ to $\mathbf{z}_k$, and $\mathbf{w}_k$ is the measurement noise.

To formulate $\mathcal{G}(\cdot)$ in (3.7a), the *Langevin* process model has been widely used as it provides a realistic model to simulate human source motion [13, 25, 31, 35, 37]. This model can be described using

$$\boldsymbol{\alpha}_k = \begin{bmatrix} 1 & 0 & aT & 0 \\ 0 & 1 & 0 & aT \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & a \end{bmatrix} \boldsymbol{\alpha}_{k-1} + \begin{bmatrix} bT & 0 \\ 0 & bT \\ b & 0 \\ 0 & b \end{bmatrix} \mathbf{u}_k, \tag{3.8}$$

where $\mathbf{u}_k \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the noise vector following Gaussian distribution, T is the time interval between consecutive frames, and $\boldsymbol{\mu} = [0 \; 0]^T$ and $\boldsymbol{\Sigma} = \mathbf{I}_{2\times 2}$ correspond to the mean vector and covariance matrix, respectively. In addition, the model parameters are defined as $a = \exp(-\beta T)$, $b = \bar{v}\sqrt{1-a^2}$, where $\bar{v} = 0.8\,\mathrm{m/s}$ is the steady-state velocity and $\beta = 10\,\mathrm{Hz}$ is the rate constant [25]. To formulate $\mathcal{H}(\cdot)$ in (3.7b), we note that $\mathbf{z}_k$ is defined as the two-dimensional location estimate obtained from SRP and hence, we can express

$$\mathbf{z}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \boldsymbol{\alpha}_k + \mathbf{w}_k, \tag{3.9}$$

where $\mathbf{w}_k$ represents the measurement error.

The process of sound source tracking is performed in a probabilistic manner. Statistically, the posterior probability density function (pdf) $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_{1:k})$ is used to denote the probability of state $\boldsymbol{\alpha}_k$ conditioned on the measurements up to time k and the measurement likelihood $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)})$ represents the probability of attaining measurement $\mathbf{z}_k$ conditioned on the state. Considering continuous data frames, the sound source tracking problem can therefore be formulated as follows: for each frame index k, given $\Pr(\boldsymbol{\alpha}_{k-1}|\mathbf{z}_{1:k-1})$ at the previous time frame, the objective is to estimate $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_{1:k})$ using the source motion model $\mathcal{G}(\cdot)$ and the new measurement $\mathbf{z}_k$.

While Kalman filtering has been proposed for source tracking [15, 18], the particle filter (PF) framework [1, 17] is deemed to be a better approach for the SSLT problem due to the absence of linearity and Gaussian distribution requirement in the state-space formulation. The PF was first introduced in SSLT in [35] and has gained great popularity [13, 14, 25, 27, 31, 37].

In the PF framework, the posterior density $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_{1:k})$ is approximated by a set of particles of the state space with associated weights $\{(\boldsymbol{\alpha}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$, i.e.,

$$\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_{1:k}) = \sum_{p=1}^{N_p} w_k^{(p)} \delta(\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^{(p)}), \tag{3.10}$$

where $p = 1, \ldots, N_p$ denotes the particle index, $\boldsymbol{\alpha}_k^{(p)}$ is the pth particle of state space, $w_k^{(p)}$ is its associated weight, and $\delta(\cdot)$ is the Dirac delta function. The bootstrap PF-based sound source tracking is performed as follows: suppose at time $k-1$, the set $\{(\boldsymbol{\alpha}_{k-1}^{(p)}, w_{k-1}^{(p)})\}_{p=1}^{N_p}$ is an approximation of the posterior density $\Pr(\boldsymbol{\alpha}_{k-1}|\mathbf{z}_{1:k-1})$, the set $\{(\boldsymbol{\alpha}_k^{(p)}, w_k^{(p)})\}_{p=1}^{N_p}$ at time index k corresponding to $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_{1:k})$ is then obtained by a propagation step

$$\boldsymbol{\alpha}_k^{(p)} = \mathcal{G}(\boldsymbol{\alpha}_{k-1}^{(p)}, \mathbf{u}_k), \tag{3.11}$$

followed by an update step,

$$w_k^{(p)} \propto w_{k-1}^{(p)} \Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)}). \tag{3.12}$$

Computation of $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)})$ is required in (3.12) and a *pseudo* likelihood approach has been proposed [25, 37] to reduce the computational load involved in the process of determining the SRP maximum corresponding to the source location measurement. In this formulation, the SRP map itself is used as an approximation of $\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)})$. To some extent the SRP can define the probability of the source being located in the steered positions within the room as it corresponds to the energy originating from those positions. The pseudo likelihood approach defines the likelihood as

$$\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \begin{cases} \mathcal{P}_k^{\gamma}(\boldsymbol{\ell}_k), & \text{for voiced frame} \\ \mathcal{U}_{\mathcal{D}}(\boldsymbol{\ell}_k), & \text{for unvoiced frame} \end{cases}, \tag{3.13}$$

where $\gamma = 2$ is a control parameter to regulate the SRP function for source tracking [25], $\mathcal{U}_{\mathcal{D}}(\cdot)$ is the uniform pdf over the considered enclosure domain $\mathcal{D}$, and $\boldsymbol{\ell}_k$ denotes the first two elements of $\boldsymbol{\alpha}_k$.

In practice, due to the proportionality in (3.12), the normalization process is always computed using

$$w_k^{(p)} \Leftarrow \frac{w_k^{(p)}}{\sum_{i=1}^{N_p} w_k^{(i)}}, \tag{3.14}$$

Table 3.1 Summary of the bootstrap PF

At time $k-1$, a set of particles $\{\boldsymbol{\alpha}_{k-1}^{(p)}, w_{k-1}^{(p)}\}_{p=1}^{N_p}$ is a discrete representation of posterior $\Pr(\boldsymbol{\alpha}_{k-1}|\mathbf{z}_{k-1})$.

For the kth frame:

1. *Particles propagation*: propagate each particle through the source dynamic model (3.7a),

$$\boldsymbol{\alpha}_k^{(p)} = \mathcal{G}(\boldsymbol{\alpha}_{k-1}^{(p)}, \mathbf{u}_k).$$

2. *Update*: the weight corresponding to each particle is updated according to the likelihood,

$$w_k^{(p)} = w_{k-1}^{(p)} \Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)}),$$

   followed by a normalization step $w_k^{(p)} \Leftarrow w_k^{(p)} (\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$.
3. *Resampling*: resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p} (w_k^{(p)})^2)^{-1}$.
4. *Result*: the particle set $\{\boldsymbol{\alpha}_k^{(p)}, w_k^{(p)}\}_{p=1}^{N_p}$ is obtained for approximation of $\Pr(\boldsymbol{\alpha}_k|\mathbf{z}_k)$. The state estimate at the kth frame is $\widehat{\boldsymbol{\alpha}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \boldsymbol{\alpha}_k^{(p)}$.

where $\Leftarrow$ denotes the assignment of a new value to the variable. In addition, the PF usually consists of a resampling stage which prevents the degeneration phenomenon where, after a few iterations, a majority of the particles would possess small weights incurring a waste of computation [1]. Finally the state estimate, at time frame index k, is given as

$$\widehat{\boldsymbol{\alpha}}_k = \sum_{p=1}^{N_p} w_k^{(p)} \boldsymbol{\alpha}_k^{(p)}, \tag{3.15}$$

and the first two elements of $\widehat{\boldsymbol{\alpha}}_k$ represent the position estimate from the tracking framework. A summary of the bootstrap PF-based sound source tracking algorithm can be found in Table 3.1.

3.3 Proposed Robust Speech Source Tracking

In Sect. 3.2, several approaches have been discussed for localizing and tracking a stationary or moving source. Significant progress has been made in recent decades for robust SSLT in different adverse environments. However, localizing or tracking a speech source in the presence of sound interferences is still an open problem. This is particularly important in robotic applications since the robots are expected to continue interacting with a human user in a noisy environment. It is also important to note that sound interferences may be nonstationary and unpredictable in nature. Take an office room for instance, the fan noise, air-conditioner noise, or a telephone ring may be located at different positions. Existing methods, in general, are unable

to distinguish between the desired speech source and interferers. The performance may be degraded when these interferers are present.

In this section, a speech source tracking method that is robust to interferers is introduced [38]. The proposed method incorporates a well-known speech feature in the frequency domain known as harmonicity. We first compare the speech spectrogram with some typical sound interference in Sect. 3.3.1 and illustrate the speech harmonic feature. Details of the proposed method will be introduced in Sect. 3.3.2. In Sect. 3.3.3, simulations are conducted to evaluate the performance of the proposed method in the presence of interference, noise, and reverberation.

3.3.1 The Harmonic Structure in the Speech Spectrogram

Figure 3.3 shows the spectrogram of a typical speech signal obtained from the TIMIT database [16] and that corresponding to different sound interferers obtained from the NOISEX-92 database [33]. The speech spectrogram, as shown in Fig. 3.3a, indicates that several harmonics (dark curves) corresponding to multiple integers of a pitch frequency are present. The pitch frequency represents the frequency of the vocal cord vibration, which normally ranges from 100 to 300 Hz , depending on whether

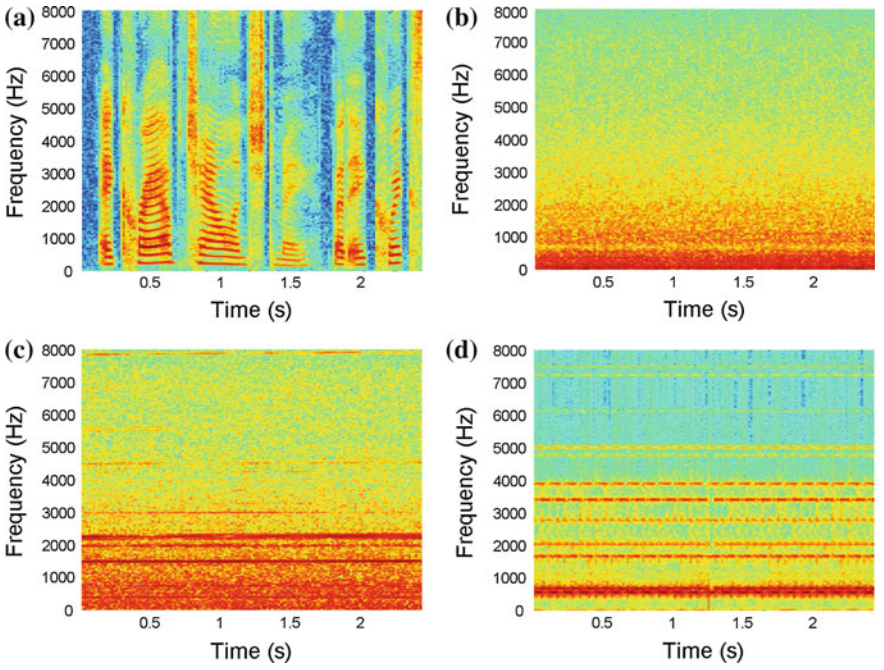

Fig. 3.3 Spectrograms of different signals. **a** Speech signal spectrogram. **b** Fan noise spectrogram. **c** Power drill noise spectrogram. **d** Telephone ring noise spectrogram

it is a male or a female voice [6]. This spectrogram indicates that speech energy is dominant on these harmonics. Figure 3.3b shows the spectrogram of a recorded fan noise where the energy is concentrated below 2 kHz. The spectrogram of a recorded power drill noise, shown in Fig. 3.3c, indicates a similar energy distribution in the low frequency range although high energy spectral lines appear at approximately 1.5, 2, and 2.2 kHz. These dominant frequencies may be caused by mechanical rotation or vibration. It is useful to note that no regular harmonic structure is exhibited in these two types of sound. In terms of the telephone ring sound, shown in Fig. 3.3d, a regular harmonic structure is caused by the presence of a single tone. However, the harmonics differ from that of the speech signal due to a difference in pitch frequency.

In the following, we therefore assume that the sound interference does not share the same harmonic bands as speech due to different pitch frequency, or that the interference does not possess any harmonic structure. The key objective of the proposed method is to estimate these harmonic bands corresponding to the speech components and to emphasize on the harmonic bands as they provide high signal-to-interference ratio (SIR). Other frequency regions are not used for tracking as these frequencies are contaminated by the sound interferers.

3.3.2 Speech Source Tracking in the Presence of Sound Interference

In the conventional sound source tracking framework, as introduced in Sect. 3.2.3, particles are propagated according to the source dynamic model before being weighted by the measurement likelihood. It computes the particle weights by employing a pseudo-likelihood that has been derived from SRP-PHAT measurements [13, 25, 37]. While this technique may achieve good tracking performance, the performance may significantly 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 toward that location away from the speech source (see Fig. 3.7a). The performance of these algorithms reduces significantly in low SIR, resulting in the SSLT losing track of the speech source.

To mitigate the degradation in performance, we exploit speech harmonicity such that the measurement likelihood is predominantly weighted by the speech signal as opposed to the interferers. The overall framework of the proposed method is as follows: (1) a prior source position is estimated using the assumed source dynamic model, (2) a beamformer is then applied to enhance the source signal from the prior estimated position in order to extract speech feature, (3) the reliable harmonic bands are estimated using the enhanced signal in the following step, (4) the new measurement likelihood is then derived by emphasizing these high SIR harmonic bands while discarding the other frequency regions.

3.3.2.1 Prior Prediction

In general, a clear source signal is often required in order to extract the corresponding speech features. However, due to the presence of interference and background noise, obtaining such a clear source signal is challenging. To improve the feature extraction performance, we propose a speech signal enhancement stage consisting of prior source position prediction and a beamformer. Considering the Langevin source dynamic model introduced in Sect. 3.2.3, for time frame index k, the prior source state can be estimated using (3.7a) and (3.8) as

$$\widehat{\boldsymbol{\alpha}}_k^- = \mathcal{G}(\widehat{\boldsymbol{\alpha}}_{k-1}^+, \mathbf{u}_k), \tag{3.16}$$

given the state estimate at the previous frame. Here, $\widehat{\boldsymbol{\alpha}}_{k-1}^+$ is the posterior state estimate at time frame index $k-1$. The prior source location estimate

$$\widehat{\mathbf{r}}_k^- = [\widehat{x}_k^- \;\; \widehat{y}_k^-]^T, \tag{3.17}$$

corresponds to the first two elements in $\widehat{\boldsymbol{\alpha}}_k^-$. Note that this prior estimate is based only on the assumed source motion. Its objective is to allow the beamformer to enhance the signal from this preliminary estimated source position. The feature-directed measurements, as will be described in the subsequent sections, will further refine the state estimate.

3.3.2.2 Feature Extraction

After obtaining a prior estimate of source position at each iteration, a beamformer can be employed to enhance the signal from that particular position. Note that the beamformer was used as a localization technique in Sect. 3.2.2. However, beamforming was initially used for enhancing the signal from a known source position and suppressing the interference and noise [34]. Various beamformers can be applied after a prior source location has been estimated. We consider, for example, the delay-and-sum beamformer [23] due to its simplicity although other forms of beamformers such as presented in [21, 32] may be used to enhance the speech signal. The delay-and-sum beamformer output for the prior estimated source location $\widehat{\mathbf{r}}_k^-$ is given as

$$S(\omega_l, \widehat{\mathbf{r}}_k^-) = \sum_{i=1}^{M} \Phi\left(D_i(\widehat{\mathbf{r}}_k^-)\right) Y_i(k, \omega_l) e^{j\omega_l D_i(\widehat{\mathbf{r}}_k^-)/c}, \tag{3.18}$$

where i is the microphone index, M is the number of microphones, and $Y_i(k, \omega_l)$ is the frequency-domain received signal from the ith microphone at kth frame. The variable ω_l is the angular frequency of lth frequency bin, c is the speed of sound, $D_i(\widehat{\mathbf{r}}_k^-) = \|\widehat{\mathbf{r}}_k^- - \mathbf{r}_i^{\mathrm{m}}\|_2$ is the distance from the prior estimated source position to the ith microphone, and $\Phi(\cdot)$ is a monotonic function that weighs the ith microphone

signal according to the source-sensor distance. In our simulations, we found that $\Phi\left(D_i(\widehat{\mathbf{r}_k})\right) = 1/D_i(\widehat{\mathbf{r}_k})$ performs well as it emphasizes the signal from the microphone that is closer to the source.

Figure 3.4 shows the signal enhancement result for a 6 s speech signal when a power drill interference is present at SIR = 5 dB and white Gaussian noise with signal-to-noise (SNR) ratio of 15 dB. These results were generated using the method of images [26] with $T_{60} = 200$ ms and eight microphones are placed 0.5 m away from the room perimeter (see Fig. 3.7). Figure 3.4a shows the spectrogram of the original speech signal where a clear harmonic structure can be found. Figure 3.4b shows the power drill interference spectrogram where no harmonic structure is present. In general, the source signal received by a single reference microphone is often distorted, especially when the interferer is close to the microphone, as shown in Fig. 3.4c. Extraction of speech harmonics from this received signal is therefore challenging. The beamformer enhanced signal, as shown in Fig. 3.4d, is indeed clearer than the microphone received signal. The speech harmonics are dominant across the whole spectrogram although certain interference energy leakage is visible. The beamformer enhanced signal will be used for feature extraction in the next step.

To extract the speech harmonics from a noisy spectrum, we use the multi-band excitation (MBE) fit method [2, 19]. As indicated in Fig. 3.5, the MBE model defines a voiced frame in the frequency domain as the product of spectrum envelop $H(\omega)$ and excitation spectrum $E(\omega, \omega_{\mathrm{p}})$ given by [19]

$$S_{\mathrm{spch}}(\omega) = H(\omega)E(\omega, \omega_{\mathrm{p}}), \tag{3.19}$$

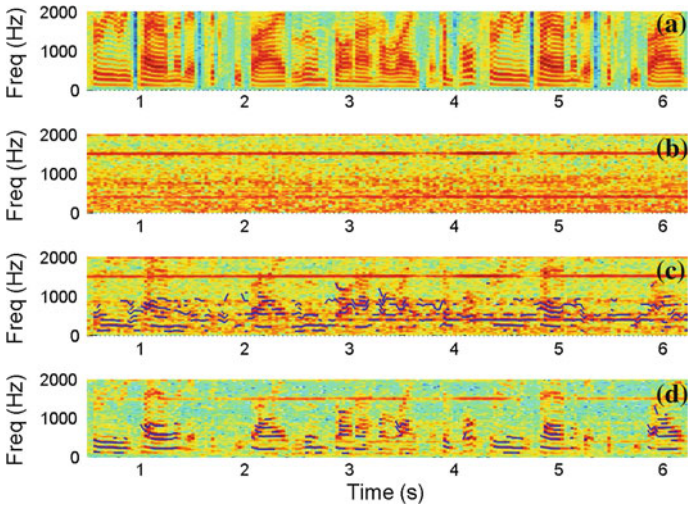

Fig. 3.4 Spectrogram and selected harmonic bands indicated in *blue lines*. **a** Clean speech. **b** Power-drill interference. **c** Reference microphone received signal and its selected harmonic bands (in *blue*). **d** Beamformer enhanced signal and its selected harmonic bands (in *blue*)

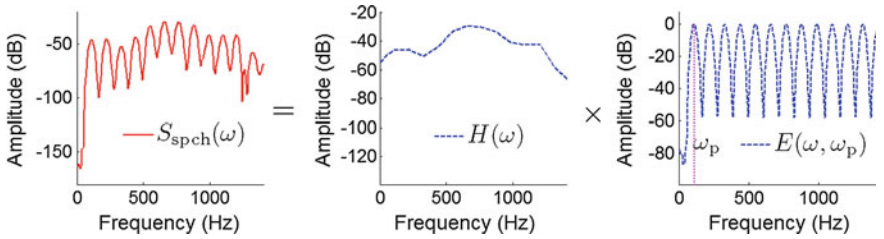

Fig. 3.5 MBE model for a speech signal. The voice frame can be modeled as a product of spectrum envelop $H(\omega)$ and excitation spectrum $E(\omega, \omega_{\text{p}})$ in the frequency domain

where ω_{p} is the pitch frequency, such that

$$E(\omega, \omega_{\text{p}}) = \sum_{q=1}^{Q} \Psi(\omega - q\omega_{\text{p}}), \tag{3.20}$$

where q is the harmonic index, Q is the number of harmonics, ω_{p} is the pitch frequency, and $\Psi(\omega)$ is the Fourier transform of the Hamming window.

We now consider extracting the harmonic information from the beamformer enhanced signal $S(\omega, \widehat{\mathbf{r}_k})$ via MBE model fitting. The harmonic information ω_{p} and $H(\omega)$ can be estimated via minimization of the fitting error between $S(\omega, \widehat{\mathbf{r}_k})$ and the MBE modeled signal

$$\begin{aligned} \varepsilon(\omega_{\text{p}}) &= \int_0^{2\pi} |S(\omega, \widehat{\mathbf{r}_k}) - S_{\text{spch}}(\omega)|^2 d\omega \\ &= \int_0^{2\pi} |S(\omega, \widehat{\mathbf{r}_k}) - H(\omega)E(\omega, \omega_{\text{p}})|^2 d\omega, \end{aligned} \tag{3.21}$$

where $S(\omega, \widehat{\mathbf{r}_k})$ has been defined in (3.18).

In practice, the above process is computed in discrete frequency domain where $\omega_l = 2\pi l/L$ denotes the angular frequency of lth frequency bins, L is the number of frequency bins, and ω_{p} is now computed from the discrete angular frequencies. In order to solve the nonlinear minimization problem in (3.21), the whole spectrum is further decomposed into several harmonic bands. The qth harmonic band ranges in the interval $[a_q,\ b_q]$, where the lower and upper limits are defined as $a_q = \lceil (q - 0.5)\omega_{\text{p}} \rfloor$ and $b_q = \lceil (q + 0.5)\omega_{\text{p}} \rfloor$, respectively, and $\lceil \cdot \rfloor$ denotes the selection of the nearest frequency bin. The variable $H(\omega)$ is also decoupled into complex amplitude H_q for each harmonic band q, so that the fitting error for each harmonic band is

$$\varepsilon_q(\omega_{\text{p}}) = \sum_{\omega_l = a_q}^{b_q} |S(\omega_l, \widehat{\mathbf{r}_k}) - H_q E(\omega_l, \omega_{\text{p}})|, \tag{3.22}$$

and the total error in (3.21) becomes

$$\varepsilon(\omega_{\mathrm{p}}) = \sum_{q=1}^{Q} \varepsilon_q(\omega_{\mathrm{p}}). \tag{3.23}$$

We note that there is a subtle difference between (3.23) and (3.21); in (3.23) we only sum over the Q harmonic bands of interest, while in (3.21) the whole spectrum is integrated.

The harmonic information is thus represented by two parameters, the pitch frequency ω_{p} and complex amplitude H_q for all harmonic bands. The variable H_q can be obtained by considering the derivative of (3.22) to be zero giving

$$H_q = \frac{\sum_{\omega_l=a_q}^{b_q} S(\omega_l, \widehat{\mathbf{r}}_k^{-}) E^*(\omega_l, \omega_{\mathrm{p}})}{\sum_{\omega_l=a_q}^{b_q} |E(\omega_l, \omega_{\mathrm{p}})|^2}, \tag{3.24}$$

where $*$ denotes conjugate operation. The pitch frequency ω_{p} can be estimated by the following steps: each fitting error $\varepsilon_q(\omega_{\mathrm{p}})$ is evaluated using the optimal value of H_q obtained in (3.24). The error function in (3.23) is then computed with respect to all pitch frequencies ω_{p} of interest. Finally, the global minimum of $\varepsilon(\omega_{\mathrm{p}})$ is determined and the corresponding ω_{p} is selected as the estimated $\widehat{\omega}_{\mathrm{p}}$ due to speech.

3.3.2.3 Feature-Directed Particle Weight Update

To obtain the feature-directed particle weight update, it is required to determine the most reliable harmonic bands and select them for computation of the likelihood. Two criteria are proposed to determine the reliability of the harmonic bands: (1) the normalized fitting error and (2) the normalized harmonic energy.

First, the normalized fitting error [2] is defined, for each harmonic, as the effectiveness of a given frequency band to be fitted with the speech harmonic model. It is computed as

$$\bar{\varepsilon}_q = \frac{\varepsilon_q(\widehat{\omega}_{\mathrm{p}})}{\sum_{\omega_l=a_q}^{b_q} |S(\omega_l, \widehat{\mathbf{r}}_k^{-})|^2}, \tag{3.25}$$

where the fitting error $\varepsilon_q(\widehat{\omega}_{\mathrm{p}})$ is computed by substituting the estimated pitch frequency $\widehat{\omega}_{\mathrm{p}}$ into (3.22). The fitting error is normalized by the energy of each corresponding harmonic band.

In the second step, the normalized harmonic energy, defined by the ratio of energy distributed on that harmonic over the total energy, i.e.,

$$P_q = \frac{\sum_{\omega_l=a_q}^{b_q} H_q E(\omega_l, \widehat{\omega}_\mathrm{p})}{\sum_{q=1}^{Q} \sum_{\omega_l=a_q}^{b_q} H_q E(\omega_l, \widehat{\omega}_\mathrm{p})}. \tag{3.26}$$

is computed. As the energy of the speech signal is expected to be concentrated in a harmonic structure, those harmonic bands with low $\bar{\varepsilon}_q$ and high P_q 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

$$G_q(\omega_l) = \begin{cases} \Psi(\omega_l - q\widehat{\omega}_\mathrm{p}), & \text{if } \bar{\varepsilon}_q \leq \zeta \ \& \ P_q \geq \eta, \ \omega_l \in [a_q, b_q] \\ 0, & \text{otherwise} \end{cases}, \tag{3.27a}$$

$$G(\omega_l) = \sum_{q=1}^{Q} G_q(\omega_l). \tag{3.27b}$$

Equation (3.27a) indicates that only harmonic bands that satisfy the thresholds are selected; the other frequency bands are discarded. Equation (3.27b) indicates that the selection process is carried out over all frequency bands of interest. The sum of the selected harmonic bands are denoted as $G(\omega_l)$.

Figure 3.6 shows extraction results of the speech harmonics using a frame of 32 ms. Figure 3.6a shows the MBE fitting result, computed using (3.22)–(3.24), 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 3.6b shows the result for the case where a power-drill signal is added to the speech signal at an SIR = 5 dB. The beamformer output $S(\omega_l, \widehat{\ell_k})$, 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 Fig. 3.6a, b, we note that the MBE fit shown in Fig. 3.6b is able to estimate the speech harmonics with reasonable accuracy, albeit with some distortion. The estimated reliable speech harmonic bands are shown with $G(\omega_l)$ and are denoted by the bold lines (which has been normalized to 0 dB for clarity).

The extraction discussed above considers a single data frame. By iterating the procedure over all the frames, $G(\omega_l)$ in (3.27b) can be extended to $G(k, \omega_l)$ which denotes the selected harmonic bands at the kth frame. The selected harmonics over all the frames are shown in Fig. 3.4d where a 6 s speech in the presence of power-drill interference is considered. We note that using the beamformer and MBE fit, speech harmonic bands can be estimated as indicated by the dark lines in the spectrogram.

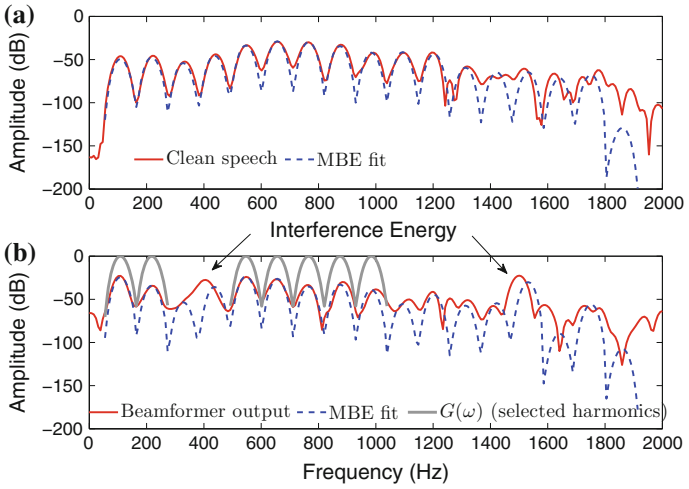

Fig. 3.6 MBE fitting result. **a** Clean speech and MBE fit. **b** Beamformer output, MBE fit, and $G(\omega)$ in the presence of a power drill signal

With $G(k, \omega_l)$, the new SRP function $\mathcal{P}_k(\boldsymbol{\ell})$ with weight $W_i(k, \omega_l)$ is given as

$$\mathcal{P}_k(\boldsymbol{\ell}) = \sum_{\omega_l \in \Omega} \left| \sum_{i=1}^{M} W_i(k, \omega_l) Y_i(k, \omega_l) e^{j\omega D_i(\boldsymbol{\ell})/c} \right|^2, \tag{3.28a}$$

$$W_i(k, \omega_l) = \frac{G(k, \omega_l)}{|Y_i(k, \omega_l)|}, \tag{3.28b}$$

where Ω is the frequency over which the SRP function is evaluated. Similar to the pseudo likelihood method [25, 37], the SRP function is used to define the measurement likelihood in the PF framework,

$$\Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k) = \begin{cases} \mathcal{P}_k^{\gamma}(\boldsymbol{\ell}), & \text{for voiced frame} \\ \mathcal{U}_D(\boldsymbol{\ell}), & \text{for unvoiced frame} \end{cases}, \tag{3.29}$$

where $\gamma = 2$ is a control parameter to regulate the SRP function for source tracking [25], 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 then used to update the particle weights of the particles. The proposed SSLT framework is summarized in Table 3.2.

3.3.3 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 200–300 ms. Eight microphones were distributed 0.5 m away from the perimeter of the room (see Fig. 3.7). An 8 s male speech signal sampled at 16 kHz from the TIMIT database [16] was used as the source signal. A power drill (PD) signal and a recorded telephone ring (TR) signal obtained from the NOISEX-92 database [33] were used as interferers. White Gaussian noise of 15 dB SNR was added to the microphone signals. The speed of source was approximately set at 0.6 m/s. The positions of speech source were estimated using a frame size of 512 samples with $N_p = 100$ particles. We also used an effective sample size threshold $N_{\text{thr}} = 37.5$, harmonic-band thresholds $\zeta = 0.6$ and $\eta = 0.03$. A total of 12 harmonic bands ($Q = 12$) were considered. The proposed method is compared with the conventional tracking method using SRP-PHAT as pseudo likelihood [25]. 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 (3.22)–(3.24). In this chapter, we quantify the performance using the average tracking error across all audio frames, i.e.,

Table 3.2 Summary of the proposed algorithm

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

For the kth frame:

1. *Prior prediction*: Propagate the previous state estimate through (3.16) to obtain prior estimate of the current state $\widehat{\boldsymbol{\alpha}}_k^{-}$.
2. *Feature extraction*: Apply beamformer according to (3.17), (3.18) to enhance the signal from the prior estimated position $\widehat{\mathbf{r}}_k^{-}$, and extract speech features using (3.22)–(3.24).
3. *Particles propagation*: Propagate each particle through the source dynamic model (3.7a),

$$\boldsymbol{\alpha}_k^{(p)} = \mathcal{G}(\boldsymbol{\alpha}_{k-1}^{(p)}, \mathbf{u}_k).$$

4. *Posterior weights update*: Obtain the feature directed particle likelihood using (3.25)–(3.29) and each particle is then assigned a weight according to its likelihood

$$w_k^{(p)} = w_{k-1}^{(p)} \Pr(\mathbf{z}_k|\boldsymbol{\alpha}_k^{(p)}),$$

followed by normalization $w_k^{(p)} \Leftarrow w_k^{(p)} (\sum_{i=1}^{N_p} w_k^{(i)})^{-1}$. The posterior state estimate is $\widehat{\mathbf{x}}_k^{+} = \sum_{p=1}^{N_p} w_k^{(p)} \boldsymbol{\alpha}_k^{(p)}$

5. *Resampling*: Resample the particles if the effective sample size is below a threshold, $N_{\text{eff}} < N_{\text{thr}}$, where $N_{\text{eff}} = (\sum_{p=1}^{N_p} (w_k^{(p)})^2)^{-1}$.

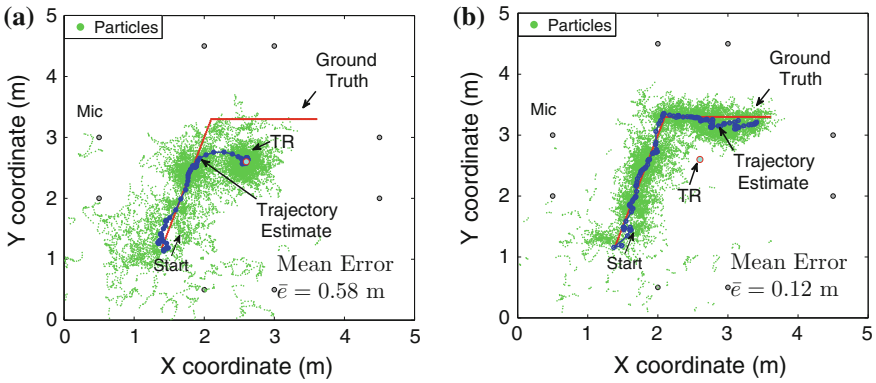

Fig. 3.7 Comparison of tracking results when TR is present at SIR $= -3$ dB, $T_{60} = 250$ ms. **a** Conventional SRP-PHAT tracking method. **b** Proposed tracking method

$$\bar{e} = \frac{1}{K} \sum_{k=1}^{K} ||\widehat{\mathbf{r}}_k^+ - \mathbf{r}_k||_2, \tag{3.30}$$

where $\widehat{\mathbf{r}}_k^+$ is the posterior estimated position at kth frame, $\mathbf{r}_k$ is the true source position, $||\cdot||_2$ is the L-2 norm, and K is the number of frames.

Figure 3.7 compares the tracking result for $T_{60} = 250$ ms in the presence of telephone ring at -3 dB SIR. Figure 3.7a shows that the tracking performance of the conventional SRP-PHAT approach is adversely affected by the interferer. Due to the high measurement likelihood of SRP-PHAT for the interferer region, the particles are "trapped" once they are propagated there, in this case the region near the telephone ring. The SRP-PHAT method has an average error of 0.58 m indicating that it does not converge to the speech source trajectory. On the other hand, Fig. 3.7b shows the tracking performance of the proposed method. This result shows that the proposed method is less significantly affected by the presence of the telephone ring achieving an average error of 0.12 m.

Figure 3.8 shows the tracking result when both power drill and telephone ring are present at 3 and 0 dB SIRs, respectively, with $T_{60} = 250$ ms. Again, Fig. 3.8a shows the conventional SRP-PHAT approach losing track of the speech source. The particles are "trapped" at the region near the power drill, leading to the average error of 0.61 m. On the other hand, the proposed method, shown in Fig. 3.8b, retains its robustness with an average error of 0.13 m.

Table 3.3 shows the average tracking error for various test conditions. The source trajectory and interference positions remain the same as the previous setup. 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 0.56 m when $T_{60} = 0.2$ s. The proposed method achieves an error of 0.11 m, which translates to an 80 % reduction of error over

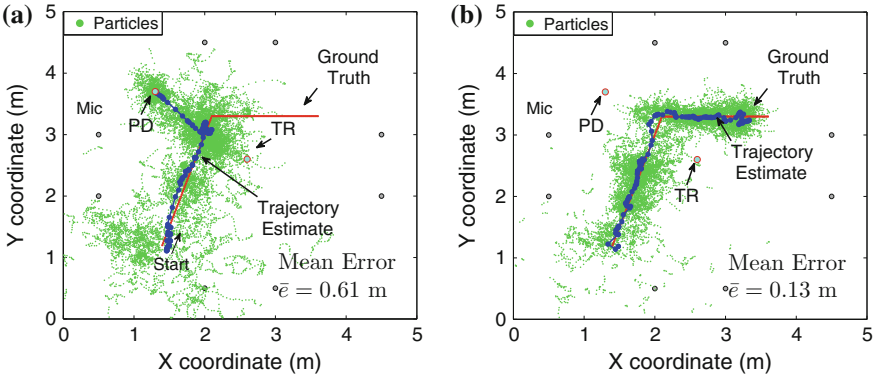

Fig. 3.8 Comparison of tracking results when both PD and TR are present at SIR = 3 dB, 0 dB, respectively, $T_{60} = 250$ ms. **a** Conventional SRP-PHAT tracking method. **b** Proposed tracking method

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. However, it is also observed that the performance of the proposed algorithm degrades modestly when reverberation time is increased. The proposed method may fail under adverse environments as indicated when $T_{60} = 0.3$ s, PD and PR are present at SIR of 3 and -6 dB.

Different source trajectory and interference configurations were also examined in Figs. 3.9 and 3.10. As before, these results show that the conventional SRP-PHAT approach is likely to be affected by interferers, while the proposed approach retains its robustness; the particles are propagated closely along the source trajectory.

Figure 3.11 shows the performance of both algorithms under different reverberation conditions. Figure 3.11a shows the results when power drill is present at an SIR = 0 dB. The SRP-PHAT tracking method, indicated by the dashed line, results in consistently high tracking errors of more than 1 m. The SRP-MBE tracking method, shown by the solid line, results in errors of less than 0.3 m when T_{60} is below 0.35 s.

Table 3.3 Comparison of mean tracking error $\bar{e}$ between the SRP-PHAT tracking method and the proposed tracking 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)	0.56 m	0.59 m	0.11 m	0.15 m
TR (SIR = 0 dB)	0.51 m	0.59 m	0.09 m	0.13 m
TR (SIR = -3 dB)	0.53 m	0.64 m	0.10 m	0.15 m
PD+TR (SIR = 3, 0 dB)	0.57 m	0.68 m	0.12 m	0.16 m
PD+TR (SIR = 3, -3 dB)	0.65 m	0.69 m	0.15 m	0.18 m
PD+TR (SIR = 3, -6 dB)	1.08 m	1.01 m	0.20 m	0.75 m

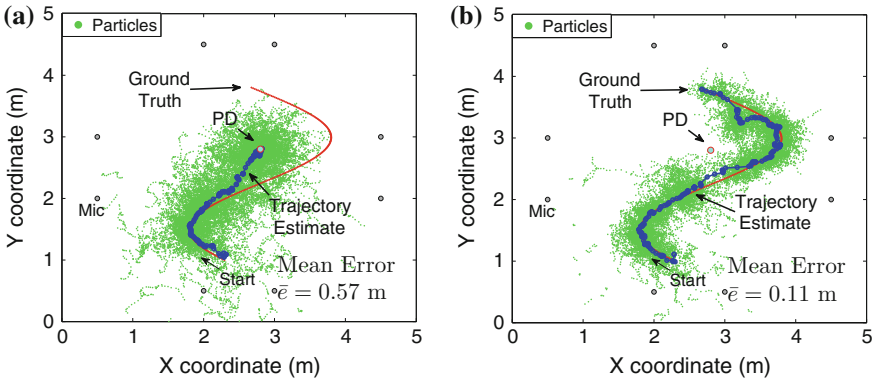

Fig. 3.9 Comparison of tracking results when PD is present at SIR = 3 dB, $T_{60} = 200$ ms. **a** Conventional SRP-PHAT tracking method. **b** Proposed tracking method

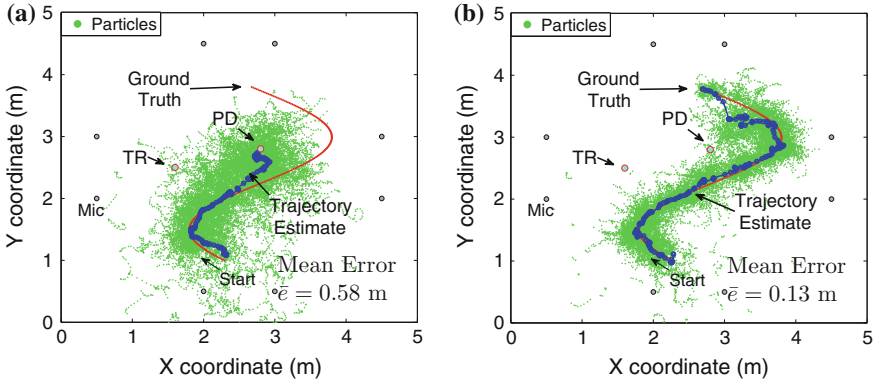

Fig. 3.10 Comparison of tracking results when both PD and TR are present at SIR = 3 dB, 0 dB, respectively, $T_{60} = 200$ ms. **a** Conventional SRP-PHAT tracking method. **b** Proposed tracking method

However, the performance deteriorates rather significantly when T_{60} is beyond 0.4 s. A similar conclusion can be drawn from Fig. 3.11b where the telephone ring is present at SIR $= -5$ dB. The SRP-PHAT tracking method consistently results in high tracking errors of more than 0.5 m, while the SRP-MBE deteriorates when T_{60} is higher than 0.3 s.

3.4 Integration with Social Robot

Sound source localization and tracking have been investigated in the previous sections. In this section, we describe a system where the SSLT module has been inte-

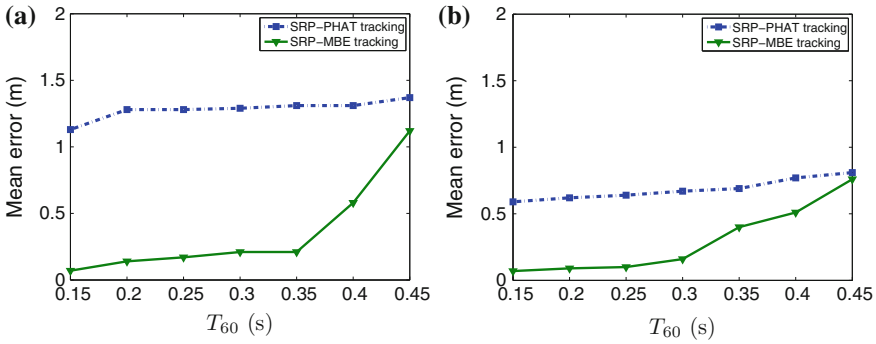

Fig. 3.11 Comparison of mean tracking error versus different reverberation time T_{60}. **a** Power drill is present at SIR $= 0$ dB. **b** Telephone ring is present at SIR $= -5$ dB

grated to the social robot and to the virtual human. Figure 3.12 shows the demo setup of a social robot system in the BeingThere Center, Nanyang Technological University. Microphones are employed linearly with known positions. The SSLT module estimates the position of a speaker within the room and delivers the position information through I2P connections to the server. The other modules (e.g., the head controller module) would therefore have access to the sound position information. Either the virtual human or the social robot is able to turn its head to a person who is speaking in the room. By focusing on the speaker, the interaction between robot and users is improved. The sound position information can also be combined with the face detection module, which allows the robot to be aware of all the users while focusing on the active speaking person.

Fig. 3.12 Integration setup with the social robot system

3.5 Future Avenues

This research focuses on SSLT problems in the meeting room environment and will continue to be the research focus in the near future. The following are some of the possible suggestions for future research:

1. **Improving the performance of SRP-MBE in the reverberant environment**. The performance of the proposed SRP-MBE tracking algorithm degrades when reverberation time increases. This is due to the fact that the harmonic bands are disturbed by a high amount of reverberation. The issue of how to recover or extract the time delay information from the degraded harmonic bands certainly requires future investigation.
2. **Tracking time-varying number of sources**. In recent years, tracking time-varying number of sources has gained much interest in the research community [14, 28, 30]. In a typical environment, there might be multiple speakers speaking at the same time, which results in speech signals overlapping. In addition, some speakers may become quiet after talking for a while. This practical situation requires an advanced probabilistic model such as random finite set [28, 36] to be incorporated in the particle filter framework to achieve multiple speaker tracking. In addition, it requires a mechanism to detect and initialize a newborn target and remove certain inactive targets from the state at a certain time instant [14].

3.6 Conclusions

In this chapter, we first reviewed the SSLT problem in a meeting room environment for teleconference purposes. The challenges include room reverberation, background noise, and sound interference. After reviewing some of the existing methods, a proposed SSLT framework was discussed for tracking a speech source in the presence of sound interference. This method is capable of estimating the speech harmonic bands for localizing and tracking. By only emphasizing the harmonic bands, better speech-sensitive measurement likelihood can be achieved resulting in better weight update for the particles. Simulation results show that the proposed method can achieve lower tracking error than the conventional SRP-PHAT method in the presence of multiple interferers.

References

1. Arulampalam MS, Maskell S, Gordon N, Clapp T (2002) A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process 50(2):174–188
2. Brandstein MS (1999) Time-delay estimation of reverberated speech exploiting harmonic structure. J Acoust Soc Am 105:2914–2919

3. Chen J, Benesty J, Huang YA (2006) Time delay estimation in room acoustic environments: an overview. EURASIP J Adv Signal Process 2006:1–1
4. Cobos M, Marti A, Lopez JJ (2011) A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling. IEEE Signal Process Lett 18(1):71–74
5. Deleforge A, Horaud, R (2012) The cocktail party robot: sound source separation and localisation with an active binaural head. In: Proceedings 7th ACM/IEEE international conference human-Robot interaction (HRI), pp 431–438
6. Deller JR, Proakis JG, Hansen JHL (2000) Discrete-time processing of speech signals. Wiley-IEEE Press, New York
7. DiBiase JH (2000) A high accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays. PhD thesis, Brown University
8. DiBiase JH, Silverman HF, Brandstein MS (2001) Robust localization in reverberant rooms. Microphone arrays: signal processing techniques and applications, pp 157–180
9. Dmochowski J, Benesty J, Affes S (2007) Direction of arrival estimation using the parameterized spatial correlation matrix. IEEE Trans Audio, Speech, Lang Process 15(4):1327–1339
10. Dmochowski J, Benesty J, Affes S (2007) A generalized steered response power method for computationally viable source localization. IEEE Trans Audio, Speech, Lang Process 15(8):2510–2526
11. Dmochowski J, Benesty J, Affes S (2008) Linearly constrained minimum variance source localization and spectral estimation. IEEE Trans Audio, Speech, Lang Process 16(8):1490–1502
12. Do H, Silverman HF, Yu Y (2007) A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array. In: Proceedings IEEE international conference on acoustics, speech and signal processing (ICASSP'07), vol 1, pp I–121–I–124
13. Fallon MF, Godsill S (2010) Acoustic source localization and tracking using track before detect. IEEE Trans Audio, Speech, Lang Process 18(6):1228–1242
14. Fallon MF, Godsill S (2012) Acoustic source localization and tracking of a time-varying number of speakers. IEEE Trans Audio, Speech, Lang Process 20(4):1409–1415
15. Gannot S, Dvorkind TG (2006) Microphone array speaker localizers using spatial-temporal information. EURASIP J Appl Signal Process (special issue on microphone arrays) 2006:1–17
16. Garofolo J, Lamel L, Fisher W, Fiscus J, Pallett D, Dahlgren N, Zue V (1993) TIMIT acoustic-phonetic continuous speech corpus. Philadelphia
17. Gordon NJ, Salmond DJ, Smith AFM (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In: Proceedings of IEE -F, radar and signal processing, vol 140, pp 107–113. IET
18. Grewal MS, Andrews AP (2011) Kalman filtering: theory and practice using MATLAB. Wiley, New York
19. Griffin DW, Lim JS (1988) Multiband excitation vocoder. IEEE Trans Acoust Speech Signal Process 36(8):1223–1235
20. Habets EAP, Benesty J (2012) A perspective on frequency-domain beamformers in room acoustics. IEEE Trans Audio, Speech, Lang Process 20(3):947–960
21. Habets EAP, Benesty J, Naylor PA (2012) A speech distortion and interference rejection constraint beamformer. IEEE Trans Audio, Speech, Lang Process 20(3):854–867
22. Huang Y, Benesty J, Elko GW, Mersereau RM (2001) Real-time passive source localization: a practical linear-correction least-squares approach. IEEE Trans Speech, Audio Process 9(8):943–956
23. Johnson DH, Dudgeon DE (1992) Array signal processing: concepts and techniques. Simon & Schuster
24. Knapp CH, Carter GC (1976) The generalized correlation method for estimation of time delay. IEEE Trans Acoust Speech Signal Process 24(4):320–327
25. Lehmann EA, Johansson AM (2007) Particle filter with integrated voice activity detection for acoustic source tracking. EURASIP J Adv Signal Process 2007

26. Lehmann EA, Johansson AM (2008) Prediction of energy decay in room impulse responses simulated with an image-source model. J Acoust Soc Am 124(1):269–277
27. Levy A, Gannot S, Habets EAP (2011) Multiple-hypothesis extended particle filter for acoustic source localization in reverberant environments. IEEE Trans Audio, Speech, Lang Process 19(6):1540–1555
28. Ma W-K, Vo B-N, Singh SS, Baddeley A (2006) Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach. IEEE Trans Signal Process 54(9):3291–3304
29. Marti A, Cobos M, Lopez JJ (2011) Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments. In: Proceedings of IEEE international conference on acoustics, speech, signal processing (ICASSP'11), pp 2592–2595
30. Morelande MR, Kreucher CM, Kastella K (2007) A bayesian approach to multiple target detection and tracking. IEEE Trans Signal Process 55(5):1589–1604
31. Talantzis F (2010) An acoustic source localization and tracking framework using particle filtering and information theory. IEEE Trans Audio, Speech, Lang Process 18(7):1806–1817
32. Timofeev S, Bahai ARS, Varaiya P (2008) Adaptive acoustic beamformer with source tracking capabilities. IEEE Trans Signal Process 56(7):2812–2820
33. Varga A, Steeneken HJM (1993) 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 12(3):247–251
34. Van Veen BD, Buckley KM (1988) Beamforming: a versatile approach to spatial filtering. IEEE ASSP Magazine 5(2):4–24
35. Vermaak J, Blake A (2001) Nonlinear filtering for speaker tracking in noisy and reverberant environments. In: Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'01), pp 3021–3024
36. Vo B-T, Vo B-N, Antonio C (2008) Bayesian filtering with random finite set observations. IEEE Trans Signal Process 56(4):1313–1326
37. Ward DB, Lehmann EA, Williamson RC (2003) Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Trans Speech Audio Process 11(6):826–836
38. Wu K, Goh ST, Khong AWH (2013) Speaker localization and tracking in the presence of sound interference by exploiting speech harmonicity. In: Proceedings of IEEE international conference on acoustics, speech, signal processing (ICASSP'13), pp 365–369
39. Zeng W-J, Li X-L (2010) High-resolution multiple wideband and nonstationary source localization with unknown number of sources. IEEE Trans Signal Process 58(6):3125–3136