Narayana Murthy, B. H.V.S.; Yegnanarayana, B.; Kadiri, Sudarsana Reddy

Time Delay Estimation from Mixed Multispeaker Speech Signals Using Single Frequency Filtering

Published in:
Circuits, Systems, and Signal Processing

DOI:
10.1007/s00034-019-01239-2

Published: 01/01/2019

Document Version
Publisher's PDF, also known as Version of record

Circuits, Systems, and Signal Processing
https://doi.org/10.1007/s00034-019-01239-2

Time Delay Estimation from Mixed Multispeaker Speech Signals Using Single Frequency Filtering

B. H. V. S. Narayana Murthy[1] · **B. Yegnanarayana**[2] · **Sudarsana Reddy Kadiri**[3]

Received: 29 August 2018 / Revised: 12 August 2019 / Accepted: 13 August 2019

Abstract

A method is proposed for time delay estimation (TDE) from mixed source (speaker) signals collected at two spatially separated microphones. The key idea in this proposal is that the crosscorrelation between corresponding segments of the mixed source signals is computed using the outputs of single frequency filtering (SFF) obtained at several frequencies, rather than using the collected waveforms directly. The advantage of the SFF output is that it will have high signal-to-noise ratio regions in both time and frequency domains. Also it gives multiple evidences, one from each of the SFF outputs. These multiple evidences are combined to obtain robustness in the TDE. The estimated time delays can be used to determine the number of speakers present in the mixed signals. The TDE is shown to be robust against different types and levels of degradations. The results are shown for actual mixed signals collected at two spatially separated microphones in a live laboratory environment, where the mixed signals contain speech from several spatially distributed speakers.

Keywords Speech analysis · Time delay estimation · Multispeaker speech · Number of speakers · Crosscorrelation · Single frequency filtering

✉ Sudarsana Reddy Kadiri
sudarsana.kadiri@aalto.fi

B. H. V. S. Narayana Murthy
bhvsnm@rcilab.in

B. Yegnanarayana
yegna@iiit.ac.in

1 Research Centre Imarat, Hyderabad 500069, India

2 Speech Processing Laboratory, International Institute of Information Technology, Hyderabad 500032, India

3 Department of Signal Processing and Acoustics, Aalto University, 00076 Espoo, Finland

Published online: 21 August 2019

1 Introduction

This paper proposes a method of estimating the time delay of a speaker's speech collected at two spatially separated microphones in a live laboratory environment. The method uses single frequency filtering (SFF) [1] analysis of speech for generating outputs at several individual frequencies. The time delay is estimated by combining the evidence obtained from each of the frequency components. The use of several frequency components provides robustness for the estimated time delay. If several speakers are speaking simultaneously, the time delay due to each of the speakers can be obtained from the actual mixed signals collected at the two microphones. The estimated time delays of the speakers in turn are used to determine the number of speakers from the multispeaker mixed signals as in [11].

Time delays of source signals arriving at two sensors are usually estimated using some form of crosscorrelation [10]. For a single source (speaker), the location of the peak in the crosscorrelation function of the speech signals at the two microphones corresponds to the delay of the source signal at the second microphone with respect to the signal at the first microphone. The peak and its location are affected if the signal and its delayed version do not match well. The mismatch can occur due to noise, multipath and reverberation. Several processing methods were explored on these degraded signals before computing the crosscorrelation function [6,7]. One of the methods proposed in [11] is to compute the Hilbert envelope of the linear prediction residual of the microphone signals and then compute the crosscorrelation function of the Hilbert envelopes of the two microphone signals. The Hilbert envelope highlights the impulse sequence characteristics in the signal, thus reducing the effects of waveform distortions.

The generalized crosscorrelation (GCC) method unifies several crosscorrelation methods into a general framework [6]. The time delay estimation by GCC is given by [6],

$$\hat{\tau}_{\mathrm{GCC}} = \operatorname{argmax} \ \psi_{\mathrm{GCC}}[m], \tag{1}$$

where

$$\psi_{\mathrm{GCC}}[m] = \sum_{k=0}^{K-1} \phi[k] S_{x_1x_2}[k] e^{j\frac{2\pi mk}{K}} \tag{2}$$

is the generalized crosscorrelation function of the two microphone signals $x_1[n]$ and $x_2[n]$. The $\phi[k]$ is the weighting function. The crossspectrum $S_{x_1x_2}[k]$ is given by

$$S_{x_1x_2}[k] = \mathbb{E}\{X_1[k]X_2^*[k]\}, \tag{3}$$

with $'*'$ denoting complex conjugate operator, $X_1[k]$ and $X_2[k]$ are the K-point discrete Fourier transforms (DFT) of $x_1[n]$ and $x_2[n]$, respectively. In practice, the expectation operator $\mathbb{E}\{.\}$ is replaced by the instantaneous value

$$\hat{S}_{x_1x_2}[k] = X_1[k]X_2^*[k]. \tag{4}$$

Weighting the crossspectrum has been extensively studied to overcome the effects of degradation due to noise and reverberation [3,5,13]. Among them, GCC-PHAT algorithm uses

$$\phi_{\mathrm{PHAT}}[k] = \frac{1}{|\hat{S}_{x_1x_2}[k]|}. \tag{5}$$

This weighting reduces signal dependencies. The peak in the GCC-PHAT function is used to estimate the time delay [8].

The GCC-PHAT may not give the best result for estimation of the time delays in the case of multiple source signals, as it considers all the frequencies equally, without considering the fact that speech signals have signal-to-noise ratio (SNR) as a function of both time and frequency [8]. To exploit this SNR dependency on the time–frequency bins, the GCC-PHAT functions are computed using mel-scale filter bank. The frequency band analysis allows estimation of the time delays depending on the strengths of the signal in each band. This enables separation of source with different delays due to high crosscorrelation values in different frequency bands [8].

Even for severely corrupted signals, there will be many time–frequency units dominated by speech. These time–frequency units with much clearer phase are sufficient to obtain a robust estimation of time delays [14,15]. Hence, novel algorithms based on time–frequency masking and deep learning were proposed to improve the crosscorrelation-based algorithms. The mask-weighted GCC-PHAT method was shown to improve the robustness of the time delay estimation in noisy and reverberant environments [14,15].

In a multispeaker multimicrophone scenario, for each speaker, there exists a fixed time delay of arrival of speech signals between a pair of microphones. The time delays corresponding to different speakers can be estimated using the crosscorrelation function of the mixed multispeaker signals. Locations of the dominant peaks in the crosscorrelation function of the mixed multispeaker signals give the time delays due to all the speakers. However, the crosscorrelation function of the mixed signals may not show unambiguous peaks at the time delays. This is due to damped sinusoidal components in speech corresponding to the resonances of the vocal tract and also due to the effects of reverberation and noise. These effects can be reduced by exploiting some speech-specific characteristics, mainly the impulse sequence of the excitation source in speech. In particular, speech exhibits relatively high SNR in the vicinity of the instants of significant excitation, i.e., the glottal closure instants, of the vocal tract in the voiced speech regions.

In this paper, a new method for time delay estimation (TDE) is proposed, which is based on the recently proposed SFF analysis of speech [1]. The advantage of SFF is that in some time segments, it gives high SNR component signals. Multiple evidences for time delay can be obtained, not only from successive frame segments of the signal, but also from the signal components at several frequencies. This provides robustness in the TDE. Section 2 gives a brief outline of the SFF analysis, highlighting the high SNR property of the resulting SFF outputs. Since the SFF output at each frequency contains information of the excitation impulse sequence, this information is exploited for TDE, as discussed in Sect. 3. The robustness of the proposed TDE method is

compared with the time delay estimated directly from the speech signal, from the Hilbert envelope of the linear prediction (LP) residual signal [11] and also using the GCC-PHAT method [4,10]. In Sect. 4, the proposed TDE method is illustrated for determining the number of speakers in the mixture signal data collected at two spatially separated microphones in a laboratory environment. In Sect. 5, the robustness of the proposed method is examined for different types and levels of degradations. Finally, in Sect. 6, the issues that need to be addressed further for the TDE and for determining the number of speakers from practical audio signals in room environments are discussed briefly.

2 Single Frequency Filtering of Speech Signals

In single frequency filtering (SFF), the envelope and the corresponding phase as a function of time are obtained at any desired frequency by passing the frequency-shifted speech signals through a near-ideal resonator located at $f_s/2$, where f_s is the sampling frequency. The steps involved in computing the SFF output at a given frequency f_k are as follows [1]:

1. The speech signal $s[n]$ is differenced to reduce any low- frequency trend in the recorded signal.

$$x[n] = s[n] - s[n-1]. \tag{6}$$

2. The differenced signal $x[n]$ is frequency shifted by multiplying it with $e^{j\bar{\omega}_k n}$, where $\bar{\omega}_k = \pi - \omega_k = \pi - \frac{2\pi f_k}{f_s}$. The frequency-shifted signal is given by

$$x_k[n] = x[n]e^{j\bar{\omega}_k n}. \tag{7}$$

   The Fourier transform of the frequency-shifted signal $x_k[n]$ is $X_k(\omega) = X(\omega - \bar{\omega}_k)$, where $X_k(\omega)$ and $X(\omega)$ are the Fourier transforms of the signals $x_k[n]$ and $x[n]$, respectively.
3. The signal $x_k[n]$ is passed through a single-pole filter,

$$H(z) = \frac{1}{1 + rz^{-1}}, \tag{8}$$

   where $r \approx 1$, if the root is on the negative real axis and is close to the unit circle. This corresponds to filtering the signal using a near-ideal resonator at $f_s/2$.
4. The filtered output is given by

$$y_k[n] = -ry_k[n-1] + x_k[n]. \tag{9}$$

5. Let $y_k[n] = y_{kr}[n] + jy_{ki}[n]$, where $y_{kr}[n]$ and $y_{ki}[n]$ are the real and imaginary parts of $y_k[n]$. The magnitude or envelope $v_k[n]$ and the phase $\theta_k[n]$ of the signal $y_k[n]$ are given by

$$v_k[n] = \sqrt{y_{kr}^2[n] + y_{ki}^2[n]}, \tag{10}$$

and

$$\theta_k[n] = \tan^{-1}\left(\frac{y_{ki}[n]}{y_{kr}[n]}\right). \tag{11}$$

To ensure stability of the filter, the value of r is chosen close to, but less than 1. In this study, $r = 0.995$ is used.

As discussed in [1], the important characteristic of $v_k[n]$ at different frequencies (f_k) is that it has some high SNR regions. This is due to correlation among the speech samples and lack of correlation among the noise samples. This high SNR property is exploited for TDE in this study. Note also that the time delay can be obtained from the SFF signal for a number of frequencies. Thus, there will be multiple evidences from several frequencies, which can be combined to improve the robustness of the estimation of the time delay.

3 Time Delay Estimation (TDE)

Time delay between two source signals arriving at two different microphone locations is estimated from the crosscorrelation function of the two signals. With a single source, there will be a peak in the correlation function due to the delay between the signals from the source at the two microphone locations. With multiple sources, there may be peaks in the crosscorrelation function at delays corresponding to all the sources, provided the signals from all the sources are present in the mixed signals being correlated. Crosscorrelation function of the signals directly may not yield a strong peak due to distortion of the waveforms received at the two microphones. Some of the factors causing distortion in a real room environment are: (a) propagation in the medium, (b) background noise, (c) multipath propagation and (d) reverberation.

The effects due to waveform distortion are reduced in the SFF outputs, as we consider the envelope of the signal at each frequency separately. Although filtering causes smearing of the signal in the time domain due to closeness of the root to the unit circle in the z-plane, the impulse sequence characteristics are preserved in the SFF output at each frequency [2,9]. Thus, the crosscorrelation of the SFF envelopes is not affected by the waveform distortion. Additional impulse sequences due to multipath propagation and reverberation do not match at the two microphone locations due to lack of coherence, whereas the impulse sequences due to the direct paths match well at the two microphone locations. Thus, the delay can be estimated from the crosscorrelation of the SFF envelopes of the two microphone signals at each frequency. The normalized crosscorrelation function is obtained for each frame from the corresponding segments of 50 ms from the two microphone signals. The choice of the segment duration depends on the maximum expected delay, which in turn depends on the spacing of the microphones. Typically, a spacing of 1 m between microphones can produce a maximum delay of about 3 ms (approximately), which corresponds to 48 samples at a sampling frequency of 16 kHz.

Let $x_1[n]$ and $x_2[n]$ be the two corresponding segments each of length N samples, from the two microphone signals, respectively. Then, the crosscorrelation function of the normalized signals is given by

$$c[m] = \sum_{n=0}^{N-1} \hat{x}_1[n]\hat{x}_2[n+m], \qquad m = -M \leq m \leq M, \tag{12}$$

where the normalized signals are given by

$$\hat{x}_1[n] = \frac{x_1[n]}{\left(\sum_{n=0}^{N-1} x_1^2[n]\right)^{\frac{1}{2}}}, \qquad n = 0, 1, 2, \ldots, N-1. \tag{13}$$

$$\hat{x}_2[n] = \frac{x_2[n]}{\left(\sum_{n=0}^{N-1} x_2^2[n]\right)^{\frac{1}{2}}}, \qquad n = 0, 1, 2, \ldots, N-1. \tag{14}$$

The crosscorrelation function $c[m]$ is obtained for a frame size of 50 ms and a frame shift of 5 ms. The number of crosscorrelation functions is equal to the number of frames in the signals. For example, for a 1 sec signal there will be 200 frames, since the frame shift is 5 ms. The location of the maximum of $c[m]$ with respect to the origin (zero lag) gives the time delay in samples between the signals arriving at the two microphones. If there is more than one source (speaker), the crosscorrelation function of the mixed signals displays peaks corresponding to the time delays between the microphones for all the sources (speakers). While the number of prominent peaks corresponds to the number of speakers, it may not happen in practice due to spurious peaks in the crosscorrelation function and also due to the fact that all speakers may not have speech with sufficiently high levels in the segments (frames) used for computing the crosscorrelation function. Hence, we use only the delay due to the most prominent peak in the crosscorrelation function.

Figure 1a shows the locations (time delays in ms) of the dominant peak in the crosscorrelation function, computed for every frame, of two mixed source signals collected at two microphones. The x axis is the frame index, and the y axis is the delay (in ms) of the highest peak in the crosscorrelation function for each frame. The plot shows the delays for each 50 ms frame of a 5 s signal, using a frame shift of 5 ms. The dots along a line correspond to one source. The number of such horizontal dotted lines corresponds to number of sources in the mixed signal. Note that there are a few spurious (not falling along a line) dots, which correspond to regions/frames of mixed speech signals where there is no speech or the speech signal-to-noise ratio is very low. The total number of dots for each delay in the entire signal is displayed as a histogram in Fig. 1b. Each strong peak in the histogram corresponds to a distinct source. Thus, the number of strong peaks in the histogram corresponds to number of speakers or sources in the mixed source signal. Ideally, the total number of values in all the peaks should be equal to the number of frames. But due to noise, and also due to the absence of any source data in some frames, there will be some spurious dots in Fig. 1a, resulting in very small number of frames with a delay in Fig. 1b. The challenge is to develop

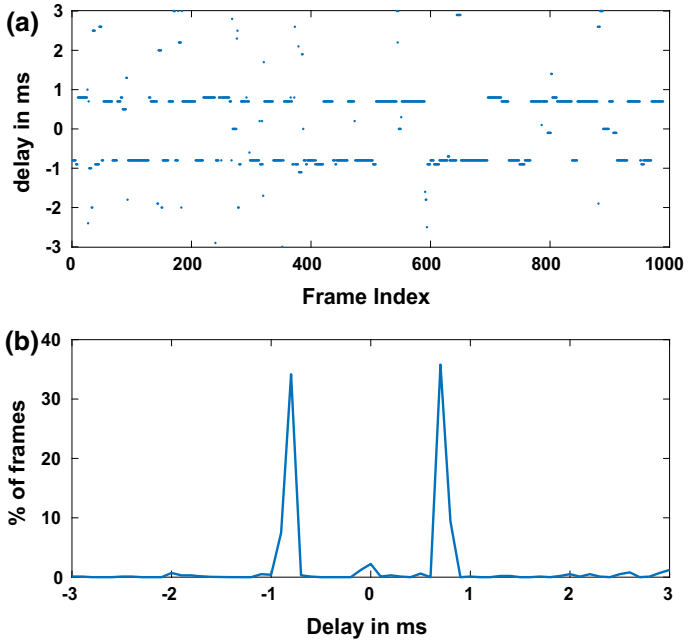

Fig. 1 Illustration of time delay estimated from the dominant peak in the crosscorrelation function and computation of histogram from it for two speaker data. **a** Dominant peak in the crosscorrelation function for each frame. **b** Histogram of **a** and it shows the total percentage of frames (α) for both the speakers

a time delay estimation method which gives the largest number of frames, i.e., sum of all values of the strong peaks in the histogram, for a given set of two mixed source signals at the two microphones. The percentage (%) of this sum in the total number of frames can be used as a measure of the performance of a TDE method. We denote this percentage as α in this paper. In this paper, we use this parameter (α) as a measure of performance to compare the proposed SFF-based method for time delay estimation with other methods reported in the literature, namely crosscorrelation of the signals directly [6], crosscorrelation of the Hilbert envelopes of the linear prediction residual signals [11] and the GCC-PHAT method [4,10].

Since different regions of speech may provide evidence for the delays corresponding to different speakers, the number of frames corresponding to each delay in the data collected at the two microphones helps in determining the number of speakers as well as their respective delays. Figure 2a shows the percentage of frames for each delay obtained from *actual* signals of two speakers collected at two spatially separated microphones. Figure 2b shows the percentage of frames for each delay obtained from the crosscorrelation on the Hilbert envelopes (HE) of linear prediction (LP) residual signals [11]. Figure 2c shows the percentage of frames for each delay obtained by the GCC-PHAT method [4,10]. A frame shift of 5 ms gives a total of 200 frames per second for the signals collected at the microphones. For illustration, 5 s of multispeaker data is considered, which gives a total of 1000 frames. The two prominent peaks in Fig. 2 correspond to the delays due to two speakers.

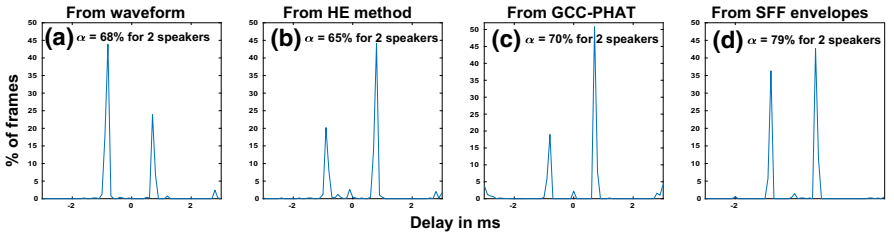

Fig. 2 Illustration of time delay estimation in terms of percentage of frames for each delay from two speaker data using **a** speech signal [6], **b** Hilbert envelopes (HE) of LP residual signal [11], **c** GCC-PHAT [4,10], and **d** SFF envelopes. The figure also shows the total percentage of frames (α) for both the speakers

We propose that the envelopes of the SFF outputs of these two signals at each frequency can be used for computing the crosscorrelation. The time-differenced envelopes are used to reduce the effect of low-frequency trend in the envelopes. The main advantage in using the SFF outputs at each frequency is that we can obtain as many independent crosscorrelation functions as the number of frequencies considered for a frame at a given instant. The evidence of the delay from the crosscorrelation functions at several frequencies can be combined to improve the robustness of TDE against degradation.

The crosscorrelation functions of the differenced SFF envelopes at many frequencies in the frequency range of interest are added to increase the evidence at the desired delays. At other delays, the values of the crosscorrelation functions will be low, thus contributing to small values in the averaged crosscorrelation function. If $c_k[m]$ is the crosscorrelation function of the envelopes of the signals at the two microphones corresponding to the frequency f_k, the average of the crosscorrelation functions across several frequencies is given by

$$\bar{c}[m] = \frac{1}{K} \sum_{k=1}^{K} c_k[m], \tag{15}$$

where K is the total number of frequencies considered in the SFF analysis.

We consider frequencies at intervals of 10 Hz, resulting in 801 frequencies in the frequency range of interest, namely 0 to $f_s/2$, where $f_s = 16$ kHz. The average crosscorrelation function $\bar{c}[m]$ is computed for each frame of 50 ms segment of the signals with a frame shift of 5 ms. The location of the maximum peak in $\bar{c}[m]$ for each frame is obtained. Figure 2d shows the percentage of frames for each delay obtained from $\bar{c}[m]$ for the mixed signal data considered in Fig. 2a. The plot in Fig. 2d clearly shows that the delays of the two strong peaks correspond to the two speakers as in Figs. 2a (speech signals), 2b [Hilbert envelopes (HE) of LP residual signal] and 2c (GCC-PHAT). It is also interesting to note that the percentage of frames (α) from the two prominent peaks in Fig. 2d (79%) is higher than the corresponding values 68, 65 and 70% obtained from Fig. 2a–c, respectively. This indicates the advantage of the proposed SFF-based method over other methods, especially the most popular GCC-PHAT method.

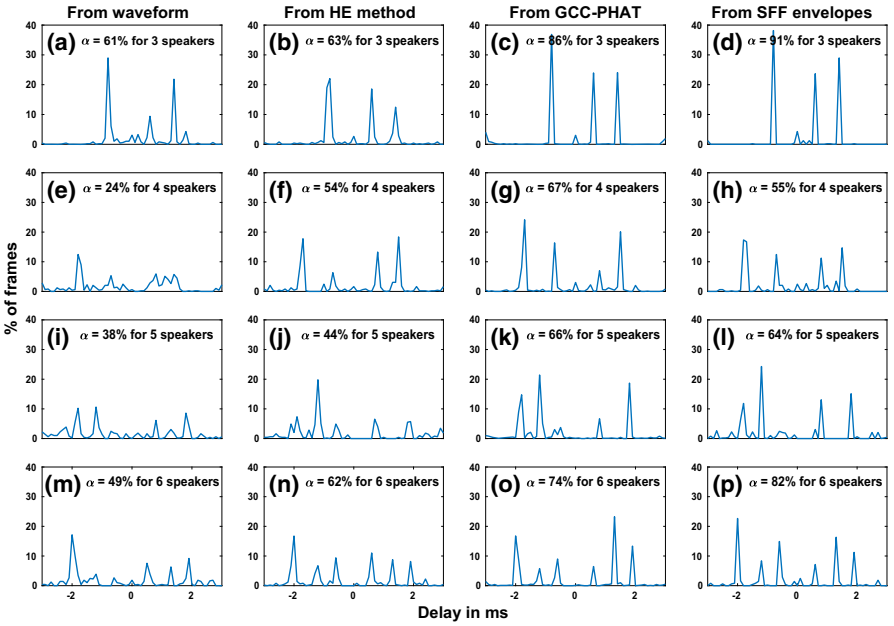

Fig. 3 Illustration of time delay estimation from multispeaker data using the speech signal [6], Hilbert envelope (HE) of LP residual signal [11], GCC-PHAT [4,10] and SFF envelopes. The figures show the percentage of frames for each delay in ms, for 3 speakers in **a**–**d**, 4 speakers in **e**–**h**, 5 speakers in **i**–**l** and 6 speakers in **m**–**p**. The total percentage of frames (α) at the delays corresponding to the speakers is also given in the figures as a measure of performance of the method

4 Studies on TDE from Multispeaker Data

Speech data for these studies were collected using two microphones separated by about 1 m in a laboratory environment, with an average (over the frequency range 0.5–3.5 kHz) reverberation time of about 0.5 s [11]. The data were collected using three, four, five and six speakers.

The recordings were made under the following conditions:

1. The speakers were seated at an average distance of about 1.5 m from the microphones, approximately along an arc of a circle. The heads of the speakers were approximately in the same plane as that of the microphones.
2. The speakers were seated in such a way that the delay is different for different speakers. In fact, any random placement of speakers with reference to the microphones satisfies this condition.
3. It is assumed that the level of the direct sound of speech at each of the microphones from each speaker is significantly higher relative to the noise and reverberation in the room.
4. All the speakers were stationary and spoke simultaneously by reading a text during recording, resulting in significant overlap.

During recording, the distances of the speakers from both the microphones were measured. The actual time delay of arrival τ of the speech signals collected at *Microphone-1* and *Microphone-2* from a speaker is given by

$$\tau = \frac{d_1 - d_2}{c}, \tag{16}$$

where c is speed of sound in air, and d_1 and d_2 are the distances of the speaker from *Microphone-1* and *Microphone-2*, respectively. A negative time delay (lead) indicates that the speaker is nearer to *Microphone-1* and relative to *Microphone-2*. The duration of the signals is about 15–20 s in each case.

When the number of sources (speakers) is two or more, then the average cross-correlation function $\bar{c}[m]$ should show the delay due to each speaker as a prominent peak. But due to relative strengths of speech at the two microphones for different speakers, sometimes only a few prominent ($<$ number of speakers) peaks may show up in the crosscorrelation function. The location of the maximum peak in the average crosscorrelation function is considered for each frame. The percentage of frames for each delay shows the peaks corresponding to all the speakers as in Fig. 2. Figure 3 shows the plots of percentage of frames as a function of delays for 3, 4, 5 and 6 speakers (corresponding to each row), for waveform-based method (column 1), HE-based method (column 2), GCC-PHAT method (column 3) and SFF-based method (column 4). In all the cases, the number of speakers can be identified from the number of strong peaks in these plots. In some cases as in Fig. 3i for 5 speakers case, the peaks due to some speakers (the third peak at around the delay of -0.7 ms) do not show up well. This is because the number of frames in the signal in which that speaker is present is very small. Since the delay is computed in integral multiples of samples, it is possible that the delay information is spread between two adjacent delays in samples. In all these studies, 5 s of the mixed signals were considered to obtain these plots. With sufficient duration of speech from each speaker, prominent peaks appear at the delays corresponding to all the speakers. It is to be noted that the peaks in the histograms are stronger in the SFF-based method in comparison with the waveform-based method and HE-based method. While the locations of the peaks in the histograms in Fig. 3 are same for all the methods, the evidence from the SFF-based method appears to be high in terms of percentage of total number of frames (α) corresponding to the speakers. Also, it is to be noted that the proposed method is comparable or better than the most popular GCC-PHAT method. Hence, in the next section, the performance of the proposed SFF-based method is compared with the GCC-PHAT method.

5 Robustness of the Proposed TDE Method for Different Types and Levels of Degradation

The robustness of the proposed SFF-based TDE method for different levels of babble noise and for different types of noises at 0 dB is discussed here.

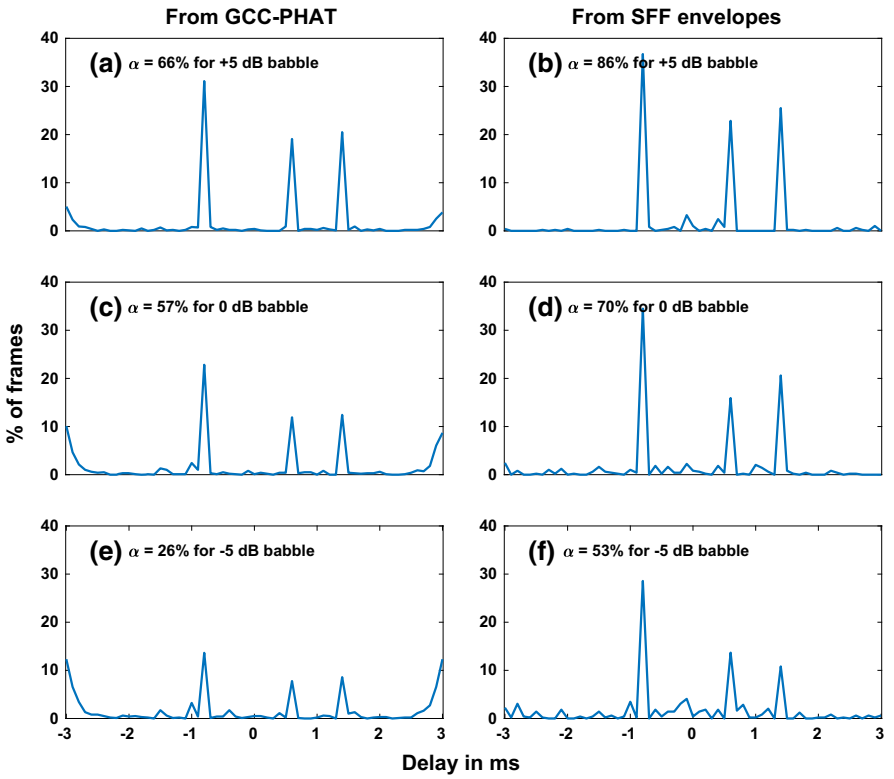

Fig. 4 Illustration of time delay estimation for three different levels of babble noise degradation for three speakers data using GCC-PHAT [4,10] and SFF envelopes. The figures show the percentage of frames for each delay in ms for + 5 dB in **a**, **b**, for 0 dB in **c**, **d** and for − 5 dB in **e**, **f**. The total percentage of frames (α) at the delays corresponding to the three speakers in each case is also given in the figures

5.1 Babble Noise at Different Levels

Babble noise at different SNR levels (5 dB, 0 dB and − 5 dB) is considered to study the effect of its speech-like characteristics in the degradation. As the waveform-based and HE-based methods performance is poorer, we compared the proposed method with GCC-PHAT method in this section.

Noise is added to the collected mixed signals at the two microphones for the 3 speakers case [12]. Figure 4 shows the plots of percentage of frames (α) for degradation due to additive babble noise at three different levels, namely 5 dB (shown in row 1), 0 dB (shown in row 2) and − 5 dB (shown in row 3). In all the cases, the peaks due to the 3 speakers are clearly visible. The plots in column 2 of Fig. 4 show the robustness of the proposed SFF-based method in comparison with the GCC-PHAT method (shown in column 1 of Fig. 4). In all the cases, the total percentage of frames (α) for the 3 speakers case is higher for the SFF-based method in comparison with the GCC-PHAT method.

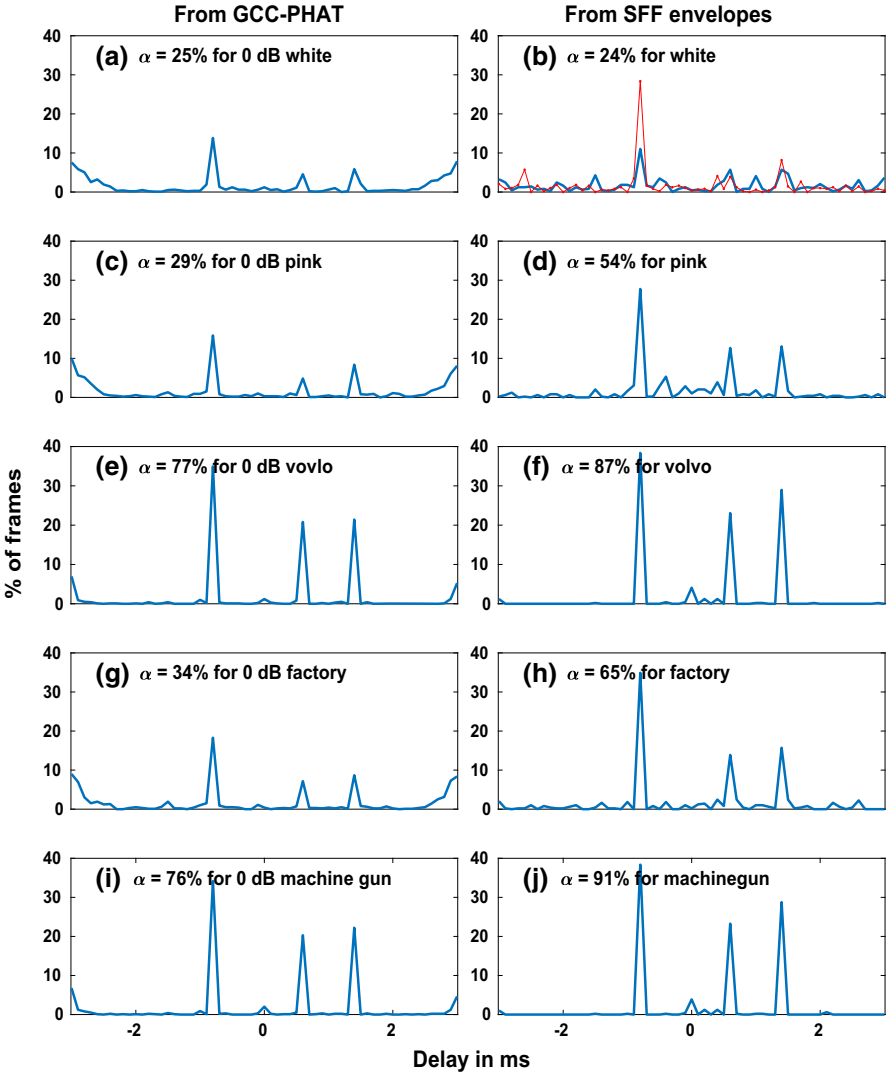

Fig. 5 Illustration of time delay estimation for different types of noises at 0 dB SNR level from three speakers data using GCC-PHAT [4,10] and SFF envelopes. The figures show the percentage of frames for each delay in ms for different types of degradations: White in **a**, **b**, pink in **c**, **d**, volvo in **e**, **f**, factory in **g**, **h** and machine gun in **i**, **j**. The total percentage of frames (α) at the delays corresponding to the three speakers in each case is also given in the figures

5.2 Different Types of Noises at 0 dB SNR

Similar to Figs. 4, 5 shows the plots of percentage of the frames as a function of delay for 5 different types of noises (white, pink, volvo, factory and machinegun) at 0 dB SNR level for the 3 speaker case. The plots in the column 1 are obtained from the GCC-PHAT method, and column 2 are obtained from the SFF envelopes. In all the

Table 1 Total percentage of frames (α) at the delays corresponding to the speakers for multispeaker case, using the waveform-based method [6], HE method [11], GCC-PHAT method [4,10] and SFF envelopes method

No. of speakers	Waveform (α)	HE (α)	GCC-PHAT (α)	SFF (α)
2 speakers	60	64	68	70
3 speakers	61	64	86	90
4 speakers	31	56	69	57
5 speakers	34	45	71	57
6 speakers	44	55	74	59

Table 2 Total percentage of frames (α) at the delays corresponding to the three speakers for the speech degraded by babble noise at SNR levels of 5 dB, 0 dB and − 5 dB, using the waveform-based method [6], HE method [11], GCC-PHAT method [4,10] and SFF envelopes method

Babble Noise	Waveform (α)	HE (α)	GCC-PHAT (α)	SFF (α)
+ 5 dB	57	40	71	81
0 dB	47	30	51	70
− 5 dB	32	25	25	41

cases, the proposed SFF-based method shows the three peaks corresponding to the 3 speakers, indicating the robustness of the method for different types of noises. It is worth noting that for white noise case, the total percentage of frames (α) is lower for the SFF method than for the waveform-based method. This is because the averaging of crosscorrelation function across all the frequencies reduced the robustness, as a majority of the frequencies correspond to very low SNR cases. In fact, if the average is taken over the frequencies in the range 0–1500 Hz, the α value goes up from 24 to 44%. The case of averaging for frequencies in the range 0–1500 Hz is shown as thin line (red in color) in Fig. 5b.

For the entire duration of the database, the total percentage of frames (α) at the delays corresponding to the speakers are tabulated for multispeaker data (two, three, four, five and six speakers) in Table 1, and for three speakers data degraded by babble noise at SNR levels of 5 dB, 0 dB and − 5 dB in Table 2. The α values for the three speakers data degraded by five different types of degradations at SNR level of 0 dB are given in Table 3. In all the cases of degradations (Tables 2 and 3), the higher percentage of frames (α values) clearly illustrate the superiority of the proposed SFF-based method over the waveform-based method, HE-based method and GCC-PHAT method. In the case of clean data (Table 1), the performance of the proposed SFF-based method is superior to the waveform-based method and HE-based method and comparable or better than the GCC-PHAT method.

Tables 4, 5 and 6 compare the time delays obtained from the waveform-based method (τ_1), HE method (τ_2), GCC-PHAT method (τ_3) and SFF envelopes method (τ_4) with the actual time delay τ obtained from the measured distances d_1 and d_2 [Eq. (16)] for multispeaker data (3, 4, 5 and 6), three speakers data degraded by the babble noise at three SNR levels (+ 5 dB, 0 dB and − 5 dB) and three speakers data degraded by different types of noises at 0 dB SNR level. From Table 4, it can be seen

Table 3 Total percentage of frames (α) at the delays corresponding to the three speakers for 5 different types of degradations at SNR level of 0 dB, using the waveform-based method [6], HE method [11], GCC-PHAT method [4,10] and SFF envelopes method

Noise at 0 dB	Waveform (α)	HE (α)	GCC-PHAT (α)	SFF (α)
White	46	11	19	18
Pink	48	17	27	48
Volvo	57	61	77	90
Factory	49	25	35	61
Machinegun	52	56	77	88

that the estimated time delays from different methods are in agreement with the actual time delays. In some cases, the time delays estimated from the SFF envelopes are in close agreement with the actual time delay. From Tables 5 and 6 , it can be seen that, compared to the time delays obtained from the waveform-based method, HE method and GCC-PHAT method, the time delays estimated from the SFF envelopes are in close agreement with the actual time delay, thus indicating the effectiveness and robustness of the proposed method for different SNR levels and different types of noises. Overall, Figs. 2, 3, 4, 5 and Tables 1, 2, 3, 4, 5 and 6 indicate the effectiveness of the proposed method in determining the number of speakers and their corresponding time delays from mixed multispeaker signals.

6 Summary and Conclusions

In this paper, a new approach for time delay estimation was proposed. The method is based on SFF analysis of speech signals, which is known to give signal components with high SNR in different regions in the time and frequency domains. The high SNR property, together with multiple evidences from the SFF outputs at different frequencies, yields reliable estimation of the time delay from the average crosscorrelation function. The method was also shown to be robust for additive babble noise degradation at different levels and also for different types of noises at 0 dB. The proposed time delay estimation method helps to identify the number of speakers from the mixed signals collected from spatially distributed microphones. In the present study, the speakers were stationary during recording, which ensures that the time delays are nearly constant.

The robustness of the method may be improved by suitably combining the evidence from several frequencies, instead of merely averaging the crosscorrelation functions. Further improvement in robustness may be achieved by combining evidence from several pairs of spatially distributed microphones.

The study can be extended to moving speaker scenario, by additionally tracking the variation in the time delays to determine the number of speakers. Some of the related issues that need to be addressed are the following. Large-scale evaluation over different room environments needs to be carried out to examine the utility of the proposed method for many real-world situations. The computational issues also need

Table 4 Comparison of estimated time delays obtained from the waveform-based method (τ_1) [6], HE method (τ_2) [11], GCC-PHAT method (τ_3) [4,10] and SFF envelopes method (τ_4) with the time delays τ computed from the measured distances d_1 and d_2 for multispeaker data

# Speakers	Speaker	d_1 (m)	d_2 (m)	Actual delay τ (ms)	Waveform τ_1 (ms)	HE τ_2 (ms)	GCC-PHAT τ_3 (ms)	SFF τ_4 (ms)
3	*Spkr-1*	0.46	0.76	− 0.87	− 0.81	− 0.75	− 0.81	− 0.81
	Spkr-2	0.98	0.74	0.63	0.56	0.62	0.62	0.62
	Spkr-3	0.97	0.49	1.38	1.44	1.38	1.38	1.38
4	*Spkr-1*	0.55	1.14	− 1.7	− 1.75	− 1.75	− 1.75	− 1.75
	Spkr-2	1.01	1.23	− 0.63	− 0.68	− 0.68	− 0.68	− 0.68
	Spkr-3	1.43	1.17	0.74	0.81	0.81	0.81	0.75
	Spkr-4	1.21	0.68	1.5	1.5	1.5	1.5	1.5
5	*Spkr-1*	0.6	1.24	− 1.83	− 1.87	− 1.81	− 1.81	− 1.81
	Spkr-2	0.88	1.29	− 1.2	− 1.12	− 1.18	− 1.18	− 1.18
	Spkr-3	1.30	1.49	− 0.54	− 0.62	− 0.62	− 0.68	− 0.62
	Spkr-4	1.42	1.14	0.80	0.81	0.81	0.75	0.75
	Spkr-5	1.16	0.54	1.77	1.81	1.81	1.81	1.81
6	*Spkr-1*	0.4	1.08	− 1.9	− 1.94	− 1.94	− 2.0	− 1.94
	Spkr-2	0.82	1.29	− 1.3	− 1.56	− 1.44	− 1.25	− 1.56
	Spkr-3	1.19	1.4	− 0.6	− 0.62	− 0.62	− 0.62	− 0.62
	Spkr-4	1.39	1.18	0.6	0.56	0.62	0.56	0.56
	Spkr-5	1.42	0.96	1.3	1.31	1.31	1.31	1.31
	Spkr-6	1.4	0.75	1.9	1.94	1.94	1.94	1.94

Table 5 Comparison of estimated time delays for three different levels of babble noise degradation for three speakers data using the waveform-based method (τ_1) [6], HE method (τ_2) [11], GCC-PHAT method (τ_3) [4,10] and SFF envelopes method (τ_4) with the actual time delays τ

Babble noise (SNR)	Speaker	Actual delay τ (ms)	Waveform τ_1 (ms)	HE τ_2 (ms)	GCC-PHAT τ_3 (ms)	SFF τ_4(ms)
+5 dB	*Spkr-1*	−0.87	−0.81	−0.75	−0.81	−0.81
	Spkr-2	0.63	0.56	0.62	0.62	0.62
	Spkr-3	1.38	1.44	1.38	1.38	1.38
0 dB	*Spkr-1*	−0.87	−0.81	−0.75	−0.81	0.81
	Spkr-2	0.63	0.50	0.62	0.62	0.62
	Spkr-3	1.38	1.44	1.27	1.38	1.38
−5 dB	*Spkr-1*	−0.87	−0.81	−0.87	−0.81	−0.81
	Spkr-2	0.63	0.26	0.62	0.56	0.62
	Spkr-3	1.38	1.44	1.19	1.38	1.38

Table 6 Comparison of estimated time delays for different types of noises at 0 dB SNR level from three speakers data using the waveform-based method (τ_1) [6], HE method (τ_2) [11], GCC-PHAT method (τ_3) [4, 10] and SFF envelopes method (τ_4) with the actual time delays τ

Noise (0 dB)	Speaker	Actual delay τ (ms)	Waveform τ_1 (ms)	HE τ_2 (ms)	GCC-PHAT τ_3 (ms)	SFF τ_4 (ms)
White	*Spkr-1*	− 0.87	− 0.81	− 0.75	− 0.81	− 0.87
	Spkr-2	0.63	0.56	0.46	0.62	0.62
	Spkr-3	1.38	1.44	1.19	1.38	1.38
Pink	*Spkr-1*	− 0.87	− 0.81	− 0.75	− 0.81	− 0.81
	Spkr-2	0.63	0.56	0.68	0.56	0.62
	Spkr-3	1.38	1.44	1.38	1.44	1.38
Vovlo	*Spkr-1*	− 0.87	− 0.81	− 0.75	− 0.81	− 0.81
	Spkr-2	0.63	0.56	0.62	0.62	0.62
	Spkr-3	1.38	1.44	1.38	1.38	1.38
Factory	*Spkr-1*	− 0.87	− 0.81	− 0.87	− 0.81	− 0.81
	Spkr-2	0.63	0.56	0.62	0.62	0.56
	Spkr-3	1.38	1.44	1.38	1.38	1.44
Machinegun	*Spkr-1*	− 0.87	− 0.81	− 0.75	− 0.81	− 0.81
	Spkr-2	0.63	0.56	0.62	0.62	0.56
	Spkr-3	1.38	1.44	1.38	1.38	1.44

to be addressed for the method to be applicable in real-time conditions. Finally, the accuracy in the current method is limited to the delay expressed in integer number of samples and hence by the sampling frequency. On the other hand, the time delay values in practice are real numbers. Hence, methods need to be developed to derive the real values of the time delays from the sampled values of the signals. One can also extend this study to sources distributed in three-dimensional space, instead of assuming that the sources and microphones are in the plane.

Acknowledgements Open access funding provided by Aalto University. The second author would like to thank the Indian National Science Academy (INSA) for their support. The third author would like to thank the Academy of Finland (Project 312490) for supporting his stay in Finland as a Postdoctoral Researcher.

Compliance with Ethical Standards

Conflict of interest The authors declare no competing financial interests.

References

1. G. Aneeja, B. Yegnanarayana, Single frequency filtering approach for discriminating speech and non-speech. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(4), 705–717 (2015)

2. G. Aneeja, S.R. Kadiri, B. Yegnanarayana, Detection of glottal closure instants in degraded speech using single frequency filtering analysis, in *Proceedings of INTERSPEECH* (2018), pp. 2300–2304
3. S. Bulek, N. Erdol, Effects of cross-spectrum estimation in convolutive blind source separation: a comparative study, in *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE)* (2011), pp. 122–127
4. G.C. Carter, Coherence and time delay estimation. Proc. IEEE **75**(2), 236–255 (1987)
5. B. Champagne, S. Bedard, A. Stephenne, Performance of time-delay estimation in the presence of room reverberation. IEEE Trans. Speech Audio Process. **4**(2), 148–152 (1996)
6. J. Chen, J. Benesty, Y. Huang, Time delay estimation in room acoustic environments: an overview. EURASIP J. Appl. Signal Process. **2006**, 026503 (2006)
7. J. Chen, J. Benesty, Y. Huang, Time-delay estimation via linear interpolation and cross correlation. IEEE Trans. Speech Audio Process. **12**(5), 509–519 (2004)
8. W. He, P. Motlicek, J. Odobez, Deep neural networks for multiple speaker detection and localization, in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)* (2018), pp. 74–79
9. S.R. Kadiri, B. Yegnanarayana, Epoch extraction from emotional speech using single frequency filtering approach. Speech Commun. **86**, 52–63 (2017)
10. C. Knapp, G. Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
11. B. Yegnanarayana, S.R.M. Prasanna, R. Duraiswami, D. Zotkin, Processing of reverberant speech for time-delay estimation. in *IEEE Transactions on Speech and Audio Processing*, vol. 13 (2005), pp. 1110–1118
12. A. Varga, H.J. 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. **12**(3), 247–251 (1993)
13. J. Vermaak, A. Blake, Nonlinear filtering for speaker tracking in noisy and reverberant environments, in *International Conference on Acoustics, Speech, and Signal Processing* (2001), pp. 3021–3024
14. Z.Q. Wang, X. Zhang, D. Wang, Robust speaker localization guided by deep learning-based time-frequency masking. IEEE Trans. Audio Speech Lang. Process. **27**(1), 178–188 (2019)
15. Z.Q. Wang, X. Zhang, D. Wang, Robust TDOA estimation based on time-frequency masking and deep neural networks, in *Proceedings of INTERSPEECH* (2018), pp. 322–326