Thakallapalli, Sushmita; Kadiri, Sudarsana; Gangashetty, Suryakanth

Spectral Features derived from Single Frequency Filter for Multispeaker Localization

Published in:
National Conference on Communications (NCC) 2020

DOI:
10.1109/NCC48643.2020.9056007

Published: 01/01/2020

Document Version
Peer reviewed version

Spectral Features derived from Single Frequency Filter for Multispeaker Localization

Sushmita Thakallapalli
Speech Processing Laboratory
IIIT Hyderabad, India
sushmita.t@research.iiit.ac.in

Sudarsana Reddy Kadiri
Department of Signal Processing and Acoustics
Aalto University, Finland
sudarsana.kadiri@aalto.fi

Suryakanth V Gangashetty
Speech Processing Laboratory
IIIT Hyderabad, India
svg@iiit.ac.in

***Abstract*—In this paper, we present a multispeaker localization method using the time delay estimates obtained from the spectral features derived from the single frequency filter (SFF) representation. The mixture signals are transformed into SFF domain from which the temporal envelopes are extracted at each frequency. Subsequently, the spectral features such as mean and variance of temporal envelopes across frequencies are correlated for extracting the time delay estimates. Since these features emphasize the high SNR regions of the mixtures, correlation of the corresponding features across the channels leads to robust delay estimates in real acoustic environments. We study the efficacy of the developed approach by comparing its performance with the existing correlation based time delay estimation techniques. Both, a standard data set recorded in real-room acoustic environments and simulated data set are used for evaluations. It is observed that the localization performance of the proposed algorithm closely matches the performance of a state-of-the-art correlation approach and outperforms other approaches.**

***Index Terms*—Time delay estimation, Multispeaker localization, Single frequency filter, Cross correlation.**

I. Introduction

Speaker localization is a challenging task, especially in real-world environments. In most of the applications such as videoconferencing, hands-free voice communication, speech recognition and in hearing-aid devices, localization is a primary task to capture high-quality speech. One approach to source localization is to, first, estimate the time delays between microphone pairs and then use these estimates to find the direction of arrival of speech sources. In this two step approach, cross correlation based techniques are typically used for time delay estimation (TDE). The other methods are average magnitude difference function (AMDF) [5], adaptive eigenvalue decomposition [3] and information theoretic approaches [19]. Among the cross correlation methods, the cross correlation of filtered signals, called generalized cross correlation (GCC), is widely used [12]. GCC with phase transform (PHAT) weighting is proved to be the most robust among all the GCC weightings in low noise and reverberation environment [23]. However, in GCC-PHAT, the errors in direction of arrival (DoA) estimates increase as the SNR decreases. To address this issue, researchers proposed SNR based weights on GCC-PHAT to highlight the speech dominant time-frequency (TF) bins and to de-emphasize TF bins with noise or reverberant speech [11], [21]. The SNR based masks estimated by deep learning are applied on GCC-PHAT in [21]. In [11], both SNR variations and noise characteristics are considered for adaptive setting of the power normalization factor. GCC-PHAT is modified not only by SNR weighting but also by alternative techniques. GCC-PHAT is weighted by the reciprocal of AMDF to improve the accuracy of DoA estimates [4]. A smoothing filter is applied on the GCC-PHAT function to eliminate the frame wise fluctuations of the DoA estimates [8]. In GCC-PHAT, the sparsity in TF bins is ignored due to summing across frequencies. It is proposed to apply mel-scale filter bank on GCC-PHAT to use the sparsity of speech signals at sub-band level [7].

The central idea in most of the above mentioned approaches is to enhance the high SNR TF components while performing GCC-PHAT. Alternatively, we investigate the effect of GCC-PHAT on signals in which instants of significant excitation (high SNR regions) are emphasized. Earlier studies have shown that features extracted from single frequency filtering (SFF) time-frequency representation emphasize the high SNR regions in the speech signal [10]. However, these SFF based features were used for epoch extraction. In this paper, we explore the use of SFF based features for TDE.

A. Relation to previous work

A TDE approach that exploits the high SNR regions was proposed in an earlier study [20]. In [20], the signals in which GCIs are emphasized by linear prediction (LP) analysis are correlated for TDE. The obtained time-delay estimates are shown to be closer to the ground truth values than those obtained by GCC method. However, LP analysis deteriorates in low SNR conditions [22]. Recently, a more robust SFF based method of epoch extraction was proposed [10]. In addition, in [14], a method of TDE by SFF was proposed. It is a narrowband approach where the spectral envelopes at various frequencies are cross-correlated with the corresponding envelopes in the other channel. The drawback of this approach is that it has high computational complexity since cross-correlations are performed at all frequencies and at all instants. To address this issue, in our study we propose a SFF based broadband approach of localization with reduced computational complexity. SFF spectral features that highlight the high SNR regions are estimated. Subsequently, the SNR enhanced signals are correlated for TDE.

B. Organization of the paper

The paper is organized as follows: In Section II, we explain the motivation for the study and the methodology of using SFF for source localization. In Section III, the localization performance of the proposed method is compared with a few broadband approaches in various acoustic settings and a discussion on the results is presented. Section IV provides the summary and conclusions from our study.

II. Proposed Approach for Localization

The objective is to accurately estimate the time-delay from the mixtures collected at two microphones in a room acoustic environment. It is to be noted that the microphone signals are not the original source signals but are degraded source signals due to reverberation and noise. In most TDE techniques, the time-delay is estimated by cross-correlation of the signals at the microphones. The peak in the cross-correlation function gives accurate time-delay estimate if the signal at one microphone is a delayed version of the other. However, in real acoustic environments, the degraded microphone signals are not the time shifted versions of one another. Hence, cross-correlation of the microphone signals does not yield distinct and accurate peaks in the cross-correlation function. It is to be mentioned that the effects of noise and reverberation are reduced to some extent around the epoch locations in the speech signals, leading to correlated components around epochs [20]. In addition, the relative epoch locations in the production of speech are not modified by degradations [13]. Henceforth, cross-correlation of the signals in which the excitation source information (epoch locations) in the speech is emphasized will give clear peaks in the cross-correlation function. In this study, the robust excitation source features that we choose are the impulse-like events in the microphone speech signals because the location of epochs corresponds to the location of the impulse-like events in speech. The SFF representation (described in Sec. II-A) is used for detecting these events. Among various methods of epoch or impulse-like event extraction, SFF is chosen for the following reasons.

- In SFF time-frequency representation, the time resolution of impulses is high and also the spectral resolution of harmonics and resonances is high [10], [16].
  In Figure 1, better time-frequency resolution trade-off offered by SFF spectrum over STFT spectrum is demonstrated with the analysis of a synthetic signal consisting of a sequence of impulses and a few narrow band signals. Figure 1(a) and Figure 1(b) depict the analysis with STFT obtained with 512 DFT points, a small (frame size of 20 ms) and large (frame size of 64 ms) Hamming window respectively and hop size = 1 sample. Figure 1(c) shows SFF of the same signals with $r = 0.992$ and $K = 256$. It can be observed that, in Figure 1(a) the temporal resolution is good, while spectral leakage between the bands is seen. On the other hand, in Figure 1(b), closely spaced spectral components are visible but at the cost of decreased time resolution. In Figure 1(c) good spectral and temporal resolution (at impulses) is obtained. SFF, thus gives a good trade off between frequency and time resolution for $0.95 \leq r \leq 0.995$ [9].
- SFF is a filtering approach and hence no block processing artifacts are present.
- SFF method of epoch extraction is superior to traditional methods of epoch extraction [10].

A few of the SFF spectral features, such as mean and variance of the envelopes extracted at various frequencies emphasize the impulse-like events.

A. Single frequency filter (SFF) representation of speech signals

SFF output is a complex TF representation of a given speech signal. It is obtained by passing the speech signal through a single pole filter after frequency shifting. The pole is near $f_s/2$, where f_s is the sampling frequency. The amplitude envelope of SFF output has accurate values as it is calculated at the highest possible frequency $f_s/2$. The spectral resolution is high because the pole at $f_s/2$ is near the unit circle where the affect of other frequency components is minimum [6]. The steps in SFF method are:

1. The pre-emphasized speech signal $x[n]$ is frequency shifted by $\bar{f}_k$, where $\bar{f}_k = \frac{f_s}{2} - f_k$ and the resulting signal is given by:
$$\tilde{x}[k,n] = x[n]e^{-j\frac{2\pi \bar{f}_k}{f_s}n}, \tag{1}$$
for $n = 1, 2, \ldots, N$, and $k = 1, 2, \ldots, K$, where N is the total number of samples in the signal, and K is the total number of frequencies in SFF.
2. The signal $\tilde{x}[k,n]$ is passed through a single-pole filter with transfer function:
$$H(z) = \frac{1}{1 + rz^{-1}}, \tag{2}$$
where the value of r is less than 1 to ensure filter stability. Since SFF is a very low bandwidth filter, r is close to unity.
3. The output of the filter is given by:
$$y[k,n] = -ry[k,n-1] + \tilde{x}[k,n]. \tag{3}$$
where $y[k,n]$ is a complex number with real part $y_r[k,n]$ and imaginary part $y_i[k,n]$.
4. The envelope of the signal $y[k,n]$ is given by:
$$e[k,n] = \sqrt{y_r^2[k,n] + y_i^2[k,n]}. \tag{4}$$
where $e[k,n]$ is the SFF amplitude envelope of the filtered output at the k^{th} frequency.

The SFF output can be obtained for several frequencies at interval of $\triangle f$. That is,

$$f_k = k\Delta f, \qquad k = 1, 2, \ldots, K, \tag{5}$$

where $K = \frac{(f_s/2)}{\Delta f}$. From the amplitude envelope of SFF, $e[k,n]$, we can get the SFF spectrum of the signal at every instant of time.

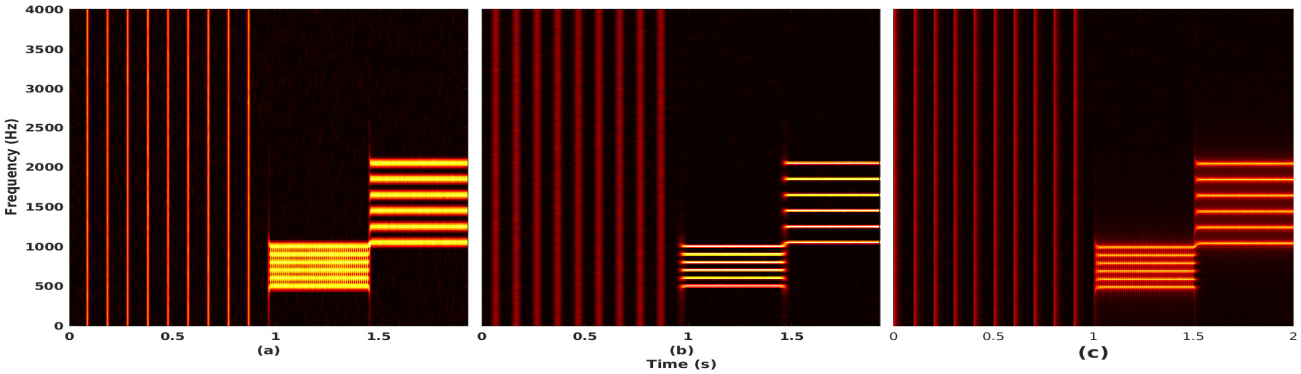

Fig. 1. (a) and (b) show STFT of synthetic signal (consisting of sequence of impulses and narrow band frequencies) for a frame size = 20 ms and 64 ms, respectively and hop size = 1 sample. (c) shows the SFF of the same signal for $r = 0.992$. In (a), (b) and (c) number of frequency components between 0 and $\frac{f_s}{2}$ is 256.

B. Spectral features of SFF for detection of impulse-like events

In this section we introduce two SFF spectral features, mean and variance, that emphasize impulse-like events of excitation in the given speech [10]. The usefulness of these features is that they highlight the individual speaker's significant characteristics in multispeaker waveform.

Spectral mean: Spectral mean, $\mu[n]$, is the mean of the envelopes and is given by:

$$\mu[n] = \frac{1}{K}\sum_{k=1}^{K} e[k,n]. \tag{6}$$

The high energy impulse-like events when averaged across frequencies lead to peaks in the spectral mean.

Spectral variance: The spectral variance, $\sigma^2[n]$, at the instants of glottal closure is low because an impulse-like excitation has a flat spectrum. Therefore, the valleys in the spectral variance depict the instants of impulse-like excitation. The spectral variance is calculated from normalized amplitude envelope, $\hat{e}[k,n]$ [10], and is given by:

$$\sigma^2[n] = \frac{1}{K}\sum_{k=1}^{K}(\hat{e}[k,n] - \hat{\mu}[n])^2, \tag{7}$$

where $\hat{e}[k,n] = \frac{e[k,n]}{\sum_{k=1}^{K} e[k,n]}$ and $\hat{\mu}[n] = \frac{1}{K}\sum_{k=1}^{K}\hat{e}[k,n] = \frac{1}{K}$, as the envelopes are normalized across frequency at each time instant.

Figure 2 gives an illustration of the SFF spectral mean property on a voiced segment of two concurrent speakers. In a simulated room environment, two concurrent speakers data is collected at a pair of microphones separated by 1 m. The plots in Figure 2 are obtained from one of the two microphones. Similar plots may be obtained at the other microphone. Figure Figure 2(a) is a voiced segment of a single speaker (Speaker 1). The SFF spectral mean of the speaker 1 is shown in Figure 2(b). The red marks indicate the instants of impulse-like events of speaker 1. Figure 2(c) corresponds to a voice segment of Speaker 2. In Figure 2(d), the SFF spectral mean of Speaker 2 is shown. The instants of significant excitation of Speaker 2 are depicted with blue marks. Figure 2(e) is the mixture of Speaker 1 and Speaker 2. Figure 2(f) is the SFF spectral mean of the mixture signal. The red and blue dots, marked at the peaks of the spectral mean, correspond to high SNR regions of speakers 1 and 2 respectively. It is interesting to note the differences in Figure 2(e) and Figure 2(f). While the peaks in Figure 2(f), corresponding to speakers' significant excitation, are prominent, no such peaks are present in Figure 2(e). Further, it can be clearly observed that the red and blue markers shown in Figure 2(f) are aligned across the plots in Figure 2(b) and Figure 2(d), respectively. This means that both the speakers' impulse-like events in Figure 2(f), closely match those of the individual speaker. Also, the derived spectral mean from the mixture signal, in Figure 2(f), clearly shows the time-delay of arrival of between the two speakers at a given microphone. From this, the relative locations of the speakers can be derived. Correlation of spectral means (Figure 2(f)) would yield distinct and prominent peaks in the angular spectrum than the correlation of original signals (Figure 2(e)). This is because, the spectral mean nicely captures the impulse-like events, where the effects of noise and reverberation are reduced. Similar conclusion can be drawn about SFF spectral variance feature, the only difference is that the impulse-like events are the valleys in spectral variance.

C. The proposed localization approach

Given the two mixture signals collected at two microphones, the first step in the proposed algorithm is to obtain the SFF representation of the two signals. For SFF estimation, r is set to 0.995. The parameter K is set to 150. We chose a lower value of K to reduce the computation time. Also, it is experimentally observed that an increase in K does not lead to a significant improvement in localization results. The next

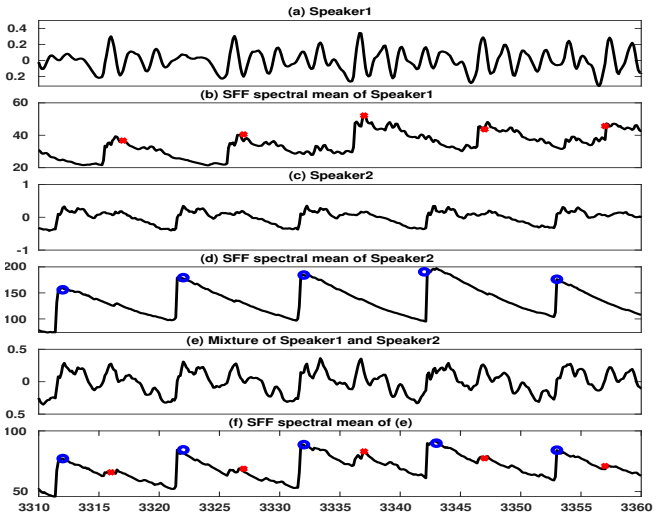

Fig. 2. (a) and (c) show a voiced segment of Speaker 1 and 2 respectively; (b) and (d) show the mean of SFF amplitude envelopes of Speaker 1 and 2 shown in (a) and (c); (e) shows the segment with both Speaker 1 and Speaker 2 active; and (f) shows the mean of SFF amplitude envelopes derived from (e). X-axis is Time in milliseconds.

step is to find the spectral mean and spectral variance of the two SFF representations as given by:

$$\mu_i[n] = \frac{1}{K}\sum_{k=1}^{K} e_i[k,n]; \tag{8}$$

$$\sigma_i^2[n] = \frac{1}{K}\sum_{k=1}^{K} (\hat{e_i}[k,n] - \hat{\mu_i}[n])^2, \tag{9}$$

where $i = 1, 2$; $e_i[k,n]$ is the amplitude envelope of the filtered output at the the k^{th} frequency and n^{th} time instant in i^{th} microphone channel; $\mu_i[n]$ is the SFF spectral mean of the signal at channel i; $\hat{e_i}[k,n]$ is the normalized envelope at channel i; $\hat{\mu_i}$ is the mean of normalized SFF envelopes at channel i and $\sigma_i^2[n]$ is the SFF spectral variance at channel i.

The spectral features obtained from 2 microphone channels in Equations (8) and (9) are correlated using GCC-PHAT, resulting in an angular spectrogram. The angular spectrograms are pooled over time, resulting in a averaged angular spectrum. The indices of the peaks in the averaged angular spectrum are the source time difference of arrival (TDoA) estimates. In Figure 3, angular spectrogram and the averaged angular spectrum obtained from the mean SFF envelopes on a mixture of 3 concurrent sources are shown. The target direction of arrival (DoA) estimate, which is the azimuth angle of arrival of the source with respect to the microphone axis is obtained by:

$$\hat{\theta} = \arccos\left(\hat{\tau}C/d\right); \tag{10}$$

where $\hat{\tau}$ is the estimated TDoA corresponding to the peak location in the mean angular spectrum; C is the speed of sound which is 340 m/s and d is the distance between the microphones in meters. The proposed speaker localization approaches using SFF spectral mean and SFF spectral variance are referred to as SFF-mean and SFF-var, respectively.

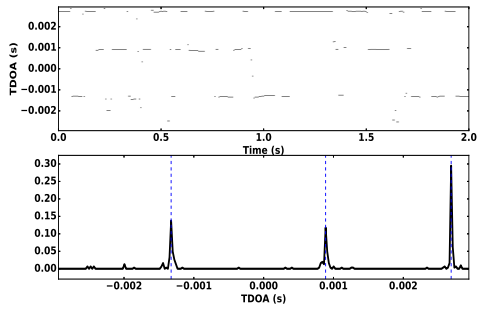

Fig. 3. Angular spectrogram and averaged angular spectrum (top and bottom figures respectively), which are obtained by correlation of mean of SFF envelopes across 2 channels. The mixture signals consist of 3 concurrent speakers. The dotted lines are the ground truth DoAs.

TABLE I
PARAMETER SETTINGS FOR THE BASELINE APPROACHES. – DENOTES "NOT APPLICABLE".

Algorithms	**GCC-PHAT**	**WC**	**HE**
Window size	50 ms	50 ms	50 ms
Hop size	1 sample	1 sample	1 sample
No of spectral components	513	-	-

III. PERFORMANCE EVALUATION AND DISCUSSION

Proposed SFF-based approaches (SFF-mean and SFF-var) are compared with three broadband approaches, namely: GCC-PHAT, waveform correlation (WC) and Hilbert envelope of LP residue (HE) [20]) on simulated and real data. WC is cross correlation obtained by time domain correlation of mixture signals. HE is cross correlation obtained by time domain correlation of Hilbert envelopes of LP residue of mixture signals. Localization is performed on 2.5 s data, whose f_s is 16 kHz. Maximum TDoA in 1 m microphones is 2.9 ms (47 samples). Number of TDoAs/DoAs used in all the approaches is 95, corresponding to a resolution of 2 degrees.

For a fair comparison, in all the approaches the temporal resolution is equal. The parameters chosen for the baseline approaches and SFF approach are shown in Table I. The reason for choosing lesser number of SFF spectral components is mentioned in Section II-C.

Root mean square error (RMSE) is reported for all test cases. In addition to RMSE, the number of missed detections (MD) is tabulated. If the estimated DoA is not within $\pm 5^o$ of the ground truth angle, then we consider it as a missed detection. For a better performing method, both RMSE (in degrees) and MD should be lower. We assume that that the number of sources to be localised are known priorly.

A. Simulated Data and Results

Two omnidirectional microphones placed 1 m apart in a simulated rectangular room with dimensions $5.6 \times 4.5 \times 2.6$ m are used as in [20]. The details of the simulator are available at [18]. Clean speech data (sampled at a rate of 16 kHz) of 4 male and 4 female speakers from the SiSEC *dev1* database are used as source speakers [2]. A source is placed at 2 m and at one of the 7 angles from the centre of the array. The 7 DoAs

considered in this study are [$30^o, 50^o, 70^o, 90^o, 110^o, 130^o$ and 150^o]. The room impulse response is calculated between 7 different locations of a given source and the fixed microphone pair for 4 reverberation times of [0.0 s, 0.1 s, 0.2 s and 0.3 s]. Thus for each source, 28 (7x4) mixture files are generated. In total we have 224 (8x7x4) mixture files for all combinations of speakers, speaker locations and reverberation times.

1) Effect of reverberation: For each reverberation time, 8 speakers at 7 different locations result in 56 mixtures. The subset of 56 mixtures from the 224 files are used to evaluate the various localization algorithms. Average RMSE (in degrees) and missed detections obtained over these 56 mixtures are reported in Table II. From the table, it can be observed that as the reverberation time increases, RMSE and the number of missed target sources increase for all the methods. Among the methods, the SFF-mean and GCC-PHAT methods are performing better in both RMSE and MD. SFF-var method is better than the HE and WC in terms of RMSE, and HE method is better than the WC and SFF-var in terms of MD. Overall, the approaches that are performing well with comparable results are GCC-PHAT, SFF-mean and HE. With a high number of missed detections, WC and SFF-var perform poorly.

2) Effect of noise: To study the effects of noise, we consider the 56 mixtures with 0.0 s reverberation from the 224 files. Three noise samples (pink, pub, and work) taken from ETSI noise database [1] are added at SNRs of 0 dB, 5 dB and 10 dB. For a given noise and a given SNR, simulations on 56 mixture files are evaluated and the resulting RMSE and MD are tabulated in Table III. From the table, it can be observed that as the SNR increases, RMSE and MD decrease. SFF-var is better than HE in terms of both RMSE and MD and it is better than WC in terms of RMSE only. In most cases, SFF-mean outperforms SFF-var in terms of both RMSE and MD. Overall, in all the noises and at all the SNRs, GCC-PHAT, SFF-mean and WC methods perform equally well with low RMSE and MD, while HE and SFF-var perform less.

To summarize, GCC-PHAT and SFF-mean give good localization results in both low SNR and reverberation conditions. While WC is robust only in noisy conditions, HE is robust in reverberation only.

B. Real Data and Results

The real datasets chosen are *dev1* and *dev2* development data in SiSEC [2], [15]. The SiSEC (*dev1* and *dev2*) live speech recording datasets consist of ten 10 s stereo mixtures of 3 female and 4 male speakers. The mixtures are recorded with reverberation times of 130 ms and 250 ms with microphone separation of 1 m. The ground-truth data is provided in the dataset. RMSE is calculated over all the files and the results are reported in Table IV for 3 speakers and 4 speakers, separately. From the table, it can be observed that RMSE and MDs of GCC-PHAT, HE and SFF-mean are the least and comparable. WC and SFF-var methods have not detected all the speakers. Further, between SFF-mean and SFF-var methods, SFF-mean method seems to be more robust and performs well.

TABLE II
AVERAGE RMSE (IN DEGREES) OBTAINED BY GCC-PHAT, WC, HE, SFF-MEAN AND SFF-VAR AT DIFFERENT REVERBERATION TIMES. TOP 2 BEST PERFORMING ALGORITHMS ARE IN **BOLD**. MISSED DETECTIONS (MD) ARE IN THE BRACKETS.

Rev. time	GCC-PHAT	WC	HE	SFF-mean	SFF-var
0.0 s	**0.43**	0.81	0.81	**0.43**	**0.43**
0.1 s	**0.5**	1.1(3)	1.0	**0.62**	0.75
0.2 s	**0.8**	1.26(8)	1.09	**0.69**	0.78(3)
0.3 s	0.8(1)	1.2(22)	**1.12**	**0.71(1)**	0.93(13)

TABLE III
AVERAGE RMSE (IN DEGREES) OBTAINED BY GCC-PHAT, WC, HE, SFF-MEAN AND SFF-VAR AT SNRS VARIED FROM 0 DB TO 10 DB BY ADDING DIFFERENT NOISES. TOP 2 BEST PERFORMING ALGORITHMS ARE IN **BOLD**. MISSED DETECTIONS (MD) ARE IN THE BRACKETS.

Noise	SNR	GCC-PHAT	WC	HE	SFF-mean	SFF-var
Pink	0 dB	**0.86**	**0.9**	1.57(4)	1.39 (4)	1.34(4)
	5 dB	**0.68**	**0.8**6	1.1	0.94	0.89
	10 dB	**0.54**	0.83	1.0	**0.47**	0.6
Pub	0 dB	**0.54**	0.9	0.6(2)	**0.64**	0.6(2)
	5 dB	**0.51**	0.84	0.94	**0.54**	0.62
	10 dB	**0.43**	0.84	0.9	**0.43**	**0.43**
Work	0 dB	**0.55**	0.92	0.99	**0.51**	0.62
	5 dB	0.54	0.88	1.0	**0.47**	**0.47**
	10 dB	**0.47**	0.81	0.83	**0.43**	**0.47**

TABLE IV
AVERAGE RMSE (IN DEGREES) OBTAINED BY GCC-PHAT, WC, HE, SFF-MEAN AND SFF-VAR ON SISEC *dev1* AND *dev2* DATASETS CONSISTING OF 3 AND 4 SPEAKERS. BEST PERFORMING ALGORITHMS ARE IN **BOLD**. MISSED DETECTIONS (MD) ARE IN THE BRACKETS.

Number of Speakers	GCC-PHAT	WC	HE	SFF-mean	SFF-var
3	**1.87**	2.01(3)	**1.8**	**1.87**	2.23
4	**1.29**	1.42(3)	**1.22**	**1.29**	1.29(2)

C. Discussion

The results of the baseline approaches are inline with the merits and demerits as mentioned in the literature. WC is a GCC without PHAT weighting. It is vulnerable to multiple reflections and does not perform well in reverberative conditions [17]. On the other hand, GCC with PHAT weighting gives good DoA estimates in low noise and reverberative conditions [23]. The performance of HE mainly depends on the performance of linear prediction (LP) analysis. The LP analysis deteriorates in low SNR conditions [22]. On the other hand, the proposed approaches exploits the high SNR property (impulse-like events) present in the speech production characteristics of the speakers for the localization. Hence, it works reasonably well in all the conditions, i,e., clean, noise and reverberation. Between the two proposed methods, SFF-mean seems to be better than SFF-var method.

IV. SUMMARY AND CONCLUSIONS

In this paper, we explored the benefits of SFF-based features for localization of multiple speakers in a room acoustic environment. The rationale for choosing the mean and variance

features of SFF were explained with illustrations. Various broadband cross-correlation approaches were compared to the proposed approaches by evaluating them on simulated as well as real data sets. It was observed that the performance of SFF-mean is better than that of SFF-var. Also, the performance of the SFF-mean closely matches the performance of GCC-PHAT (which is the state-of-the-art localization approach) and is better than the other localization methods. In future, other SFF based spectral features or combinations of SFF features may be explored to improve the performance. Several options for further exploration have opened up after this initial study and SFF features seem promising for localization.

V. Acknowledgements

The second author would like to thank the Academy of Finland (project no. 312490) for supporting his stay in Finland as a Postdoctoral Researcher.

References

[1] "ETSI, Eg 202 396-1: Speech processing, transmission and quality aspects (stq); speech quality performance in the presence of background noise," 2008.

[2] A. Liutkus, F. R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*. Cham: Springer International Publishing, 2017, pp. 323–332.

[3] J. Benesty, "Adaptive eignevalue decomposition algorithm for passive source localization," *Journal of the Acoustical Soceity of America (JASA)*, vol. 107, pp. 384–391, 2000.

[4] J. Chen, B. Jacob, and Y. Huang, "Performance of GCC-and AMDF-based Time-delay Estimation in Practical Reverberant Environments," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 25–36, Jan. 2005.

[5] R. Cusani, "Performance of fast time delay estimators," *IEEE Trans. Acoustics Speech and Signal Processing (TASSP)*, vol. 37, no. 5, pp. 757–759, 1989.

[6] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE Trans. Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 4, pp. 705–717, 2015.

[7] W. He, P. Motlicek, and J. Odobez, "Deep neural networks for multiple speaker detection and localization," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 74–79.

[8] L. Huang, S. Zhang, D. Zan, X. Zhang, and F. Li, "Speaker localization with smoothing generalized cross correlation based on naive bayes classifier," in *Proc. IEEE International Conference on Information Communication and Signal Processing (ICICSP)*, 2018, pp. 98–102.

[9] K. Gurugubelli and A. K. Vuppala, "Perceptually Enhanced Single Frequency Filtering for Dysarthric Speech Detection and Intelligibility Assessment," in *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, May 2019, pp. 6410–6414.

[10] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, 2017.

[11] H. Kang, M. Graczyk, and J. Skoglund, "On pre-filtering strategies for the GCC-PHAT algorithm," in *Proc. IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016, pp. 1–5.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoustics Speech and Signal Processing (TASSP)*, vol. 24, no. 4, pp. 320–327, 1976.

[13] P. S. Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant instants of excitation from speech signals," *IEEE Trans. Speech and Audio Processing (TSAP)*, vol. 7, no. 6, pp. 609–619, 1999.

[14] N. Murthy, B. Yegnanarayana, and S. R. Kadiri, "Time delay estimation from mixed multispeaker speech signals using single frequency filtering," *Circuits, Systems, and Signal Processing*, August 2019.

[15] N. Ono, Z. Koldovský, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013, pp. 1–6.

[16] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *Proc. International Conference on Speech Communication and Technology (INTERSPEECH)*, 2016, pp. 2155–2159.

[17] J. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas, and J. Escolano, "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments," *Applied Acoustics*, vol. 73, no. 8, pp. 698 – 712, 2012.

[18] S. Sivasankaran, "Room simulator in python," https://github.com/sunits/rir_simulator_python, [Last accessed:].

[19] F. Talantzis, A. G. Constantinides, and L. C. Polymenakos, "Estimation of direction of arrival using information theory," *IEEE Signal Processing Letters*, vol. 12, pp. 561–564, 2005.

[20] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech and Audio Processing (TSAP)*, vol. 13, no. 5, pp. 751–761, September 2005.

[21] Z. Wang, X. Zhang, and D. Wang, "Robust speaker localization guided by deep learning-based time-frequency masking," *IEEE Trans. Audio, Speech, and Language Processing (TASLP)*, vol. 27, no. 1, pp. 178–188, 2019.

[22] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, no. 1, pp. 25 – 42, 1999.

[23] C. Zhang, D. Florencio, and Z. Zhang, "Why does PHAT work well in lownoise, reverberative environments?" in *Proc. Int. Conf. Acoustics Speech and Signal Processing (ICASSP)*, 2008, pp. 2565–2568.