INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Far-field Speaker Localization and Adaptive GLMB Tracking

Shoufeng Lin[1,†], *Zhaojie Luo*[2,†,*]

[1]School of Electrical Engineering, Computing & Mathematical Sciences, Curtin University, Australia
[2] Graduate School of Engineer Science, Osaka University, Japan

Shoufeng.lin@graduate.curtin.edu.au; luo@irl.sys.es.osaka-u.ac.jp

Abstract

In the speech signal processing area, far-field speaker localization using only the audio modality has been a fundamental but challenging problem, especially in presence of reverberation and a varying number of moving speakers. Many existing methods use speech onsets as reliable directional cues against reverberation and interference. However, signal processing can be computationally costly especially in time domain. In this paper, we present a computationally efficient implementation of the recently proposed Onset-Multichannel Cross Correlation Coefficient (MCCC) method. Instead of scanning the entire spatial grid, reverse mapping and linear interpolation are used. The proposed algorithm with better efficiency is referred to as the Onset-MCC in this paper. Performance of the Onset-MCC is studied over various reverberant and noisy scenarios. To further suppress outliers and address miss-detections, as well as for the adaptive tracking of a varying number of moving speakers, we present an adaptive implementation of the generalized labeled multi-Bernoulli (GLMB) filter. As shown in studied cases, the proposed system demonstrates reliable and accurate location estimates in far-field ($T_{60} = 1s$), and is applicable to tracking an unknown and time-varying number of moving speakers.

Index Terms: moving speaker, localization, tracking, reverberation, onset, multichannel cross-correlation, microphone array processing, CASA, GLMB filter

1. Introduction

Speaker localization using audio modality has been an important research topic. Obtaining accurate location estimates enables further signal processing, e.g. speaker tracking and diarization [1], speech separation [2, 3] via beamforming, as well as speech enhancement [4]. It also has a wide range of practical applications, e.g. automatic camera steering in smart environments, human-robot interaction, and virtual reality synthesis [5]. However, far-field localization in presence of strong reverberation and moving speakers, has remained a challenging topic.

In a recent localization paper [6], the author investigated the reverberation-robust localization using redundant information of multiple microphone pairs, and proposed the Onset-MCCC and multichannel cross correlation (MCC) - phase transform (PHAT) methods. The performance of the proposed methods has been evaluated for static and moving speakers in various reverberant and noisy scenarios, using sound recordings from a uniform circular array (UCA). Compared with some state-of-the-art location estimators, e.g. the MUSIC [7], SRP-PHAT [8], EB-ESPRIT [9], TF-CHB [10] and Neuro-Fuzzy [11], the proposed methods demonstrate encouraging capabilities for locating moving speakers with good spatial resolution. While the Onset-MCCC can produce reliable localization results for moving speakers in reverberant environments, it is computationally demanding to scan through entire spatial grid and calculate the multichannel cross-correlation coefficients. Moreover, like many other methods, localization results may contain miss-detections and outliers, especially for far-filed scenarios.

In this paper, inspired by the generalized steered response power method [12], the relative sample delays between microphone pairs are reverse-mapped to spatial grid, and cross-correlation coefficients of encoded subband onset signals are linearly interpolated across spatial grid, before they are accumulated over all subbands. This computationally efficient implementation is referred to as the Onset-MCC to make a distinction from the previously proposed method [6]. To address the imperfect location measurements (e.g. clutter and miss-detections), as well as to track the time-varying and nonlinear kinetic states of multiple speakers, particle filters of various Bayesian recursions can be applied [13, 14, 15, 16, 17, 18], including the Bayes random finite set (RFS) filters [17, 18, 19, 20]. The generalized labeled multi-Bernoulli (GLMB) filter is one of the latest Bayes RFS trackers, and provides an elegant closed-form solution to multi-object tracking. Recently a near-field GLMB multi-feature state filtering framework has been proposed in [21]. However, the applications of the GLMB filter in the far-field multi-speaker tracking problem using the audio modality deserve more attentions in the speech processing literature. In this paper, we implement the GLMB filter with the measurement driven birth (MDB) model [22] for adaptive tracking. The MDB-GLMB filter provides not only the filtered locations of speakers, but also the respectively associated identities. It can track the kinetic states of speakers and is applicable for the far-field scenarios with a time-varying number of moving speakers and in presence of strong reverberation.

The paper is organized as follows. Section 2 introduces the proposed Onset-MCC speaker localization. Section 3 proposes the GLMB multi-speaker adaptive tracking. Numerical evaluations are demonstrated in Section 4, and conclusions and discussions are provided in Section 5.

2. Localization Algorithm

2.1. Onset Detection and Cross-correlation

Distinct onsets are often used as direct-path cues for reverberation-robust localization [23, 24, 25]. Here we use the onset detection and encoding method as presented in [6]. The key steps are summarized as follows.

Denote the discrete signal acquired by microphone i as $x_i[n]$, $n \in \mathbb{Z}$. Following the CASA [26] approach, an auditory filterbank is used to decompose the wideband speech signal into subbands.

$$x_i^{(b)}[n] = x_i[n] * g^{(b)}[n], \tag{1}$$

where $*$ denotes convolution, $g^{(b)}[n]$ the gammatone filter in

[†] equal contribution
[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-1160

subband b, and $b \in [1, N_b]$ with N_b the number of subbands.

Distinct onsets can be found by comparing the signal amplitudes (local peaks) with the upper bound of the level of reflections [6]. The encoded onsets are found as $\hat{x}_i^{(b)}[n]$, $\forall\, n \in \hat{N}_b$, where the set of indices collects the onset time indices, by comparing the local peaks with the reverberation level upper bound, i.e. $\hat{N}_b = \{n \mid \frac{\hat{x}_i^{(b)}[n]}{\bar{x}_i^{(b)}[n]} \geq \pi\}$. Here $\hat{x}_i^{(b)}[n]$ denotes the local peaks of subband signals, and $\bar{x}_i^{(b)}[n]$ is the upper bound of the level of reflections, approximated via recursive averaging (see [6, 27, 28] for derivations)

$$\bar{x}_i^{(b)}[n] =\lambda \cdot \bar{x}_i^{(b)}[n-1] + (1-\lambda) \cdot \lfloor x_i^{(b)}(n/f_s) \rceil, \qquad (2)$$

where $\lambda \approx 10^{-\frac{3}{T_{60}\cdot f_s}}$ is a forgetting factor, depending on the reverberation time T_{60} and sampling rate f_s. The operator $\lfloor \cdot \rceil$ denotes half-wave rectification, i.e. $\lfloor x \rceil = \frac{1}{2}(x+|x|)$, $\forall x \in \mathbb{R}$.

After obtaining the reverberation-robust speech onsets, their cross-correlation coefficients from microphone pairs, e.g. (i, j) can be used to find the time-difference-of-arrivals (TDOAs) that correspond to speaker locations, i.e.

$$e_{ij}^{(b)}[\Delta^{(ij)}] = \mathrm{xcorr}(\hat{x}_i^{(b)}[n], \hat{x}_j^{(b)}[n - \Delta^{(ij)}]), \qquad (3)$$

where the function $\mathrm{xcorr}(\cdot,\cdot)$ obtains the standard cross-correlation coefficient, and $\Delta^{(ij)}$ is the relative sample delay between microphones i and j, which corresponds to the speaker location $\boldsymbol{\wp}$, i.e.

$$\Delta^{(ij)} = \mathrm{round}\big([\|\boldsymbol{\wp} - \boldsymbol{m}_j\| - \|\boldsymbol{\wp} - \boldsymbol{m}_i\|] f_s/v\big), \qquad (4)$$

where $\boldsymbol{\wp} = r_s(\cos\theta, \sin\theta)$, $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ denote microphone locations, $v = 343$m/s the velocity of sound, θ the azimuthal direction of arrival (DOA), and r_s denotes the distance from microphone array to the speaker. Here we assume that the speakers and microphones locate at the same azimuthal plane.

The straightforward way to find speaker locations is to scan the entire location space, and find the peaks of accumulated cross-correlation coefficients corresponding to each location. This however, is computationally expensive.

2.2. Reverse Mapping and Linear Interpolation

An efficient alternative is to reverse mapping from relative discrete sample delays to the location space [12]. Obviously, the maximum relative sample delay for a compact array geometry is considerably less than the number of possible spatial grid points. For example, for a UCA with a diameter of 0.1m, the maximum relative delay is about 14, for $f_s = 48000$Hz.

From (3) and (4), the cross-correlation coefficients is a function of the location $\boldsymbol{\wp}$. These coefficients are then linearly interpolated across the spatial grid. For example, if the azimuthal DOAs $\theta_m \in [0^\circ, 360^\circ)$ are desired and estimated at 1° grid steps, the locations can be simply denoted as $\boldsymbol{\wp}_m = \boldsymbol{\wp}(\theta_m)$, indexed by the DOA. Consequently, between any two consecutive spatial grid points $\boldsymbol{\wp}_{m_a}$ and $\boldsymbol{\wp}_{m_b}$, $(m_a \neq m_b)$, the cross-correlation coefficients can be obtained, i.e.,

$$e_{ij}^{(b)}[\boldsymbol{\wp}_m] = e_{ij}^{(b)}[\boldsymbol{\wp}_{m_a}]\frac{m - m_b}{m_a - m_b} - e_{ij}^{(b)}[\boldsymbol{\wp}_{m_b}]\frac{m - m_a}{m_a - m_b}. \quad (5)$$

The same process repeats for all the microphone pairs, excluding those far-apart to avoid spatial alias. Then for each spatial grid $\boldsymbol{\wp}_m$, or rather the DOA θ_m, the cross-correlation coefficients are accumulated:

$$e^{(b)}[\theta_m] = \prod_{(i,j)\in\mathbf{P}^{(b)}} e_{ij}^{(b)}[\theta_m], \qquad (6)$$

where $\mathbf{P}^{(b)}$ denotes the set of microphone pairs selected to avoid spatial alias in each subband [6, 11].

Following the analysis and synthesis approach, the subband results are used to form an overall localization function for the Onset-MCC method, i.e.

$$\epsilon^{\mathrm{onset-mcc}}[\theta_m] = \frac{1}{N_b}\sum_{b=1}^{N_b} e^{(b)}[\theta_m]. \qquad (7)$$

Compared with the Onset-MCCC developed in [6], the proposed Onset-MCC has improved computational efficiency, as the reverse mapping and linear interpolation significantly reduce the number of cross-correlation operations.

3. GLMB Multi-speaker Tracking

The DOA estimates from the Onset-MCC can be found from the peaks of (7), and denoted as $\hat{\Theta}_k \triangleq \{\hat{\theta}_{i_k} | i_k = 1, \cdots N_k\}$ for each time frame k ($N_k = 0$ when $\hat{\Theta}_k = \emptyset$). This set of *unordered* DOA estimates are used as the *measurements* for the multi-object tracking filter. To address the DOA estimation errors, clutter or miss-detections over time, the MDB GLMB filter can be used to adaptively [22] track the trajectories of speakers. It assumes at each time instant that each speaker either remains active with probability p_S and detected with probability p_D or miss-detected with probability $1-p_D$, or becomes silent with probability $1-p_S$. Spurious measurements are characterized with a clutter rate κ. Moreover, it assigns a unique label to each speaker, associated with its states (including locations).

The GLMB distribution is written as [19, 20]

$$\pi(\mathbf{S}_k) = \Delta(\mathbf{S}_k)\sum_{\xi\in\Xi} w^{(\xi)}(\mathcal{L}(\mathbf{S}_k)) \left[p^{(\xi)}\right]^{\mathbf{S}_k}, \qquad (8)$$

where $\mathbf{S}_k = \{\mathbf{s} | \mathbf{s} = (s_{j_k}, \ell_{\xi_k})\}$ is a set of labeled speaker states, s_{j_k} is a speaker state, ℓ_{ξ_k} is a label associated with a candidate speaker and is unique, $\Delta(\cdot)$ guarantees unique labels, ξ carries association map between targets and measurements, Ξ is a discrete space, $\mathcal{L}(\cdot)$ extracts labels, $w(\cdot)$ is the probability of hypothesis, and $\left[p^{(\xi)}\right]^{\mathbf{S}}$ is the probability density function of states. The GLMB distribution propagates via the Bayes prediction-update recursion, provided with the measurements, and produces the desired speaker states. The Langevin model [29] is used for the (first-order Markov) kinetic state transition of the speakers in the prediction step. The DOA measurements $\hat{\Theta}_k$ from microphone arrays are pre-converted to candidate Cartesian locations of speakers $\hat{P}_k = \{\hat{\wp}_{i_k}\}$ via straightforward triangulation, and used in the Bayes update step assuming normal distribution of measurements, i.e. $\hat{\wp} \sim \mathcal{N}(m(s), \sigma_{\wp}^2)$, where $\sigma_{\wp}^2$ denotes the variance of measurement, and $m(s)$ is a projection function from state space to measurement space.

4. Numerical Results

This section presents the localization results of the Onset-MCC method using sound signals from simulation [30, 31] and real-world recordings. State-of-the-art methods, e.g. the SRP-PHAT, MUSIC, Neuro-Fuzzy, MCC-PHAT, TF-CHB and EB-ESPRIT methods are used as benchmarks. Reverberation time varies from $T_{60} = 0.2s$ to 1s. Additive uncorrelated white noise is used for adjusting the signal-to-noise ratio (SNR).

For all tests of the Onset-MCC method here, $\lambda = 0.9998$ in (2). The gammatone filter [26, 32, 33] as the subband filter in (1) is used for its linear phase and frequency selectivity.

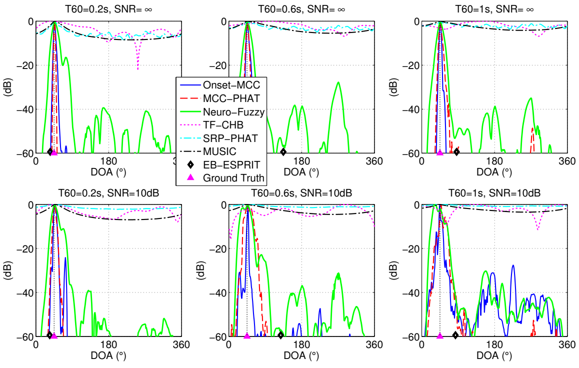

Figure 1: *One static source located at* 0.2m *from the wall, at DOA of* 45° *(marked as a triangle △).*

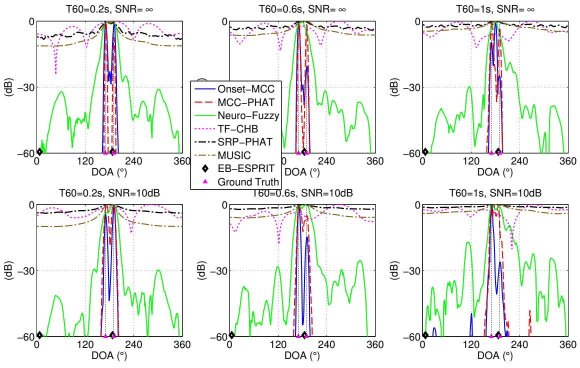

Figure 2: *Two closely located speakers. Ground truth DOAs are marked as triangles at* 170° *and* 190°.

The center frequencies of the filterbank range from 250Hz to 3600Hz, $r_s = 1$m, and the number of subbands $N_b = 16$. Three scenarios are studied. We find the speaker locations on the same azimuthal plane of the microphone array. The UCA uses 8 microphones with a radius of 0.05m. For the two simulated scenarios, the UCA is placed at the center of a rectangular room of 6m×8m×3m (width × length × height), and the speech segment used for each speaker is 4 seconds long. The speech signals are chosen from the TIMIT database [34].

4.1. Scenario 1: Single Static Speaker (simulated)

Most of the methods can work well for the localization of a single static speaker, when the speaker does not locate too close to an acoustically reflective surface. Here a more challenging scenario is used where a speaker locates at DOA 45°, and the distance to the closest wall is fixed to 0.2m.

Fig. 1 provides the normalized (and scaled by $10\lg(\cdot)$) DOA estimation histograms of localization functions from the proposed Onset-MCC method as well as that of the Neuro-Fuzzy method, the steered-response power of the TF-CHB, MCC-PHAT, SRP-PHAT, MUSIC methods, and the discrete estimates of the EB-ESPRIT method, respectively, over SNR and T_{60}. For the cases of static speakers, the EB-ESPRIT uses the overall average (4 seconds) of segmental covariance matrices to achieve best accuracy. It has discrete DOA estimates which are plotted in the diamond symbol on the horizontal axes. In this case, the early reflection from the closest wall is about 1ms behind the direct-path. With additive noise at SNRs from ∞ to 10dB, it is interesting to see that all the methods work fine at $T_{60} = 0.2$s, while as the reverberation increases, the EB-ESPRIT could not find the correct DOAs (even though given the *a priori* knowledge of one active source), and the dominant peak SRP of the TF-CHB also deviates significantly from 45°. While the strongest peaks of the Neuro-Fuzzy at SNR=10dB and $T_{60} \geq 0.6$s have deviations, those of the Onset-MCC and MCC-PHAT all correspond to the true speaker DOA.

4.2. Scenario 2: Two Close Static Speakers (simulated)

Fig. 2 plots the DOA localization results for the case when two static speakers locate at DOAs of 170° and 190°, respectively. Overall, from Fig. 2, the TF-CHB forms a wide peak at around 180° in most cases except $T_{60} = 1$s, indicating that the two speakers are fused. The EB-ESPRIT again assumes a known number of speakers (which avoids the errors due to estimation of the number of speakers at adverse conditions), and produces a DOA estimate at around 180° and a second estimate at close to 0°. This indicates that the EB-ESPRIT also has ambiguity to differentiate the two speakers. Except at $T_{60} = 0.2$s and high SNR, the SRP-PHAT, Neuro-Fuzzy and MUSIC also fuse the two sources into one. However, the proposed Onset-MCC method and the MCC-PHAT can reliably form two distinct peaks corresponding to the ground truth in most cases.

4.3. Scenario 3: Concurrent and Moving Speakers

The previous two sections presented static scenarios for baseline performance evaluation of the algorithms [6]. In this case, real-world recordings of moving speakers in a reverberant room ($T_{60} \approx 0.65s$) are used. The dimensions of the room are 3.4m×7.8m×2.7m. Three moving loudspeakers are used to play concatenated speech recordings from the TIMIT database [34]. Microphone arrays are placed close to the centre of the room: (1.2, 3.9, 1.5)m and (2.2, 3.9, 1.5)m, respectively. Omnidirectional electret microphones are used.

In such challenging scenarios, the root mean square error may not provide useful information, as there may be clutter or miss-detections. The OSPA metric [35] is thus used as the quantitative measure for evaluating localization methods. Here $p = 2$ and $c = 20°$ are chosen. Proper steered response power thresholds for respective methods are used for reasonably optimal performance.

The final DOA estimates from one of the microphone arrays and the corresponding OSPA results are given in Fig. 3. Each column shows the results from a particular method, i.e. from the left column to the right column, the methods used are respectively the Onset-MCC, MCC-PHAT, Neuro-Fuzzy, SRP-PHAT and MUSIC. The TF-CHB and EB-ESPRIT do not produce reliable results in this case of moving speakers, and hence are not plotted. In the top panel of each column, DOA estimates over time are plotted in black dots, while the ground truth locations of speakers are plotted in red triangles. It can be seen from panels (a) to (d) that although there are several spurious peaks and gaps, they produce close and clean DOA estimates in general. The SRP-PHAT seems to have the best overall OSPA results, but has some outliers. MUSIC has a significant amount of spurious estimates and the worst overall OSPA results, even though the ground truth number of speakers per each frame is provided for the algorithm. This indicates that subspace based methods may not work well for moving speakers. In general, the Onset-MCC, MCC-PHAT, Neuro-Fuzzy and

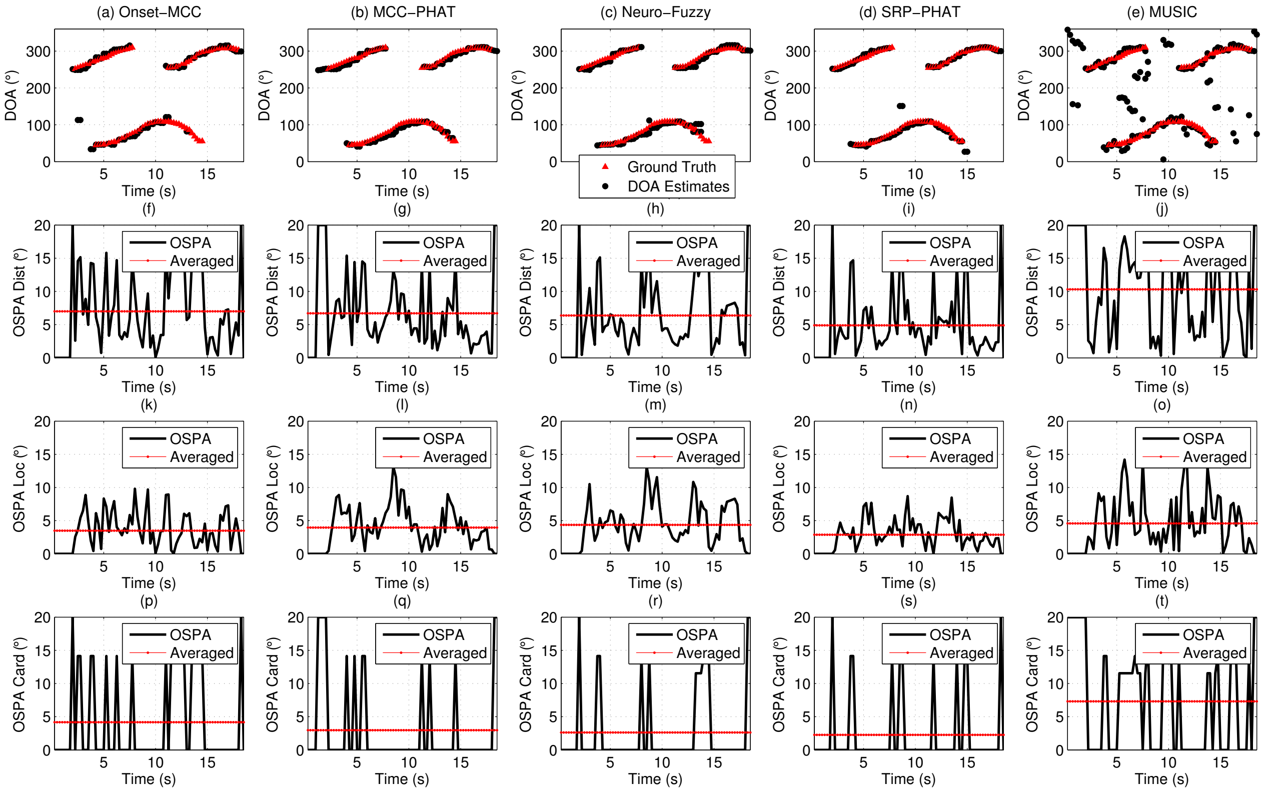

Figure 3: *DOA estimates (top row) and OSPA results from different methods. In the top row, ground truth locations are marked with red triangles, and estimates are plotted as black dots. Three speakers are moving and speaking in the real reverberant room ($T_{60} \approx 0.65s$).*

SRP-PHAT methods demonstrate similar overall performance in this case of moving speakers with strong reverberation. Note that the SRP-PHAT shows worse spatial resolution in Scenario 2, compared to the Onset-MCC method.

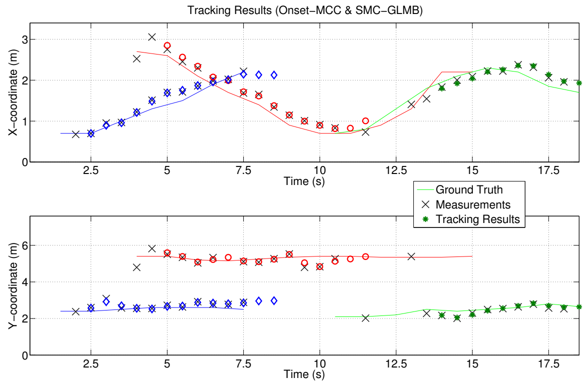

Figure 4: *Estimated Cartesian tracks of speakers using Onset-MCC measurements and the MDB GLMB filter. Tracks with different labels are plotted with different colors and symbols.*

Fig. 4 demonstrates the estimated trajectories of speakers from the GLMB using DOA localization results from the Onset-MCC method. The sequential Monte Carlo (SMC) implementation is used. $p_S = 0.98, p_D = 0.9, \kappa = 0.08$ and $\sigma_\wp = 0.15m$. Simple triangulation is used to convert the DOA estimates from the two microphone arrays to Cartesian locations. The time step used is 0.5s, which is sufficient for common speaker tracking applications. The number of speakers is time-varying and unknown *a priori* to the algorithm. The figure shows that the GLMB filter produces labeled (plotted in different colors) tracks for respective speakers. The brief miss-detection at time about 7s does not break the trajectory of the speaker in red. There is a period of miss-detection during 11s and 13s however, due to the gaps in DOA estimates from the two microphone arrays in this challenging scenario. In general, the tracking results are close to ground truth, despite the challenges.

5. Conclusions

This paper proposes a new far-field localization implementation referred to as the Onset-MCC. It provides reliable location estimates with better spatial resolution than baseline methods. To suppress localization errors as well as clutter and miss-detections, an adaptive multi-speaker tracking filter is also implemented based on the GLMB framework for tracking time-varying and moving speakers' kinetic states. Performance of the presented method is studied using not only simulated signals of reverberation time from 0.2 to $1s$, but also real recordings in an office room of $T_{60} \approx 0.65s$. Evaluation results show that the proposed method can reliably locate not only static speakers but also multiple concurrent and moving speakers, in presence of reverberation. Comparison with other baseline localization techniques in various reverberant conditions demonstrates the benefits of the proposed method.

6. References

[1] X. Qian, A. Brutti, O. Lanz, M. Omologo, and A. Cavallaro, "Multi-speaker tracking from an audio–visual sensing device," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2576–2588, 2019.

[2] S. Lin, S. Nordholm, H. H. Dam, and P. C. Yong, "An adaptive low-complexity coherence-based beamformer," in *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2013, pp. 263–266.

[3] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal multi-channel target speech separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541, 2020.

[4] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[5] S. Lin and X. Qian, "Audio-visual multi-speaker tracking based on the GLMB framework," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*. ISCA, 2020, pp. 3082–3086. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-1969

[6] S. Lin, "Reverberation-robust localization of speakers using distinct speech onsets and multichannel cross correlations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2098–2111, 2018.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[8] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001, pp. 157–180.

[9] H. Teutsch and W. Kellermann, "Acoustic source detection and localization based on wavefield decomposition using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2724–2736, 2006.

[10] A. M. Torres, M. Cobos, B. Pueo, and J. J. Lopez, "Robust acoustic source localization based on modal beamforming and time–frequency processing using circular microphone arrays," *The Journal of the Acoustical Society of America*, vol. 132, no. 3, pp. 1511–1520, 2012.

[11] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proc. Int. Workshop on Acoustic Echo and Noise Control, Tel Aviv, Israel*. Citeseer, 2010.

[12] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2510–2526, 2007.

[13] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[14] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[15] N. J. Gordon, D. J. Salmond, and A. F. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 140, no. 2. IET, 1993, pp. 107–113.

[16] S. S. Blackman, "Multiple hypothesis tracking for multiple target tracking," *IEEE Aerospace and Electronic Systems Magazine*, vol. 19, no. 1, pp. 5–18, 2004.

[17] I. R. Goodman, R. P. Mahler, and H. T. Nguyen, *Mathematics of data fusion*. Springer Science & Business Media, 2013, vol. 37.

[18] R. P. Mahler, *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[19] B.-T. Vo and B.-N. Vo, "Labeled random finite sets and multi-object conjugate priors," *IEEE Transactions on Signal Processing*, vol. 61, no. 13, pp. 3460–3475, 2013.

[20] B.-N. Vo, B.-T. Vo, and D. Phung, "Labeled random finite sets and the bayes multi-target tracking filter," *IEEE Transactions on Signal Processing*, vol. 62, no. 24, pp. 6554–6567, 2014.

[21] S. Lin, "Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-bernoulli framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2018. ICASSP 2018.*

[22] S. Lin, B. T. Vo, and S. E. Nordholm, "Measurement driven birth model for the generalized labeled multi-bernoulli filter," in *2016 International Conference on Control, Automation and Information Sciences (ICCAIS)*. IEEE, 2016, pp. 94–99.

[23] M. Kuhne, R. Togneri, and S. Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *IEEE signal processing letters*, vol. 16, no. 2, pp. 85–85, 2009.

[24] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*, vol. 120. Citeseer, 2006, pp. 133–137.

[25] N. T. N. Tho, S. Zhao, and D. L. Jones, "Robust doa estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2287–2291.

[26] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[27] S. Lin, "A new frequency coverage metric and a new subband encoding model, with an application in pitch estimation." in *INTERSPEECH*, 2018, pp. 2147–2151.

[28] ——, "Robust pitch estimation and tracking for speakers based on subband encoding and the generalized labeled multi-bernoulli filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 827–841, 2019.

[29] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3021–3024.

[30] E. A. Lehmann, A. M. Johansson, and S. Nordholm, "Reverberation-time prediction method for room impulse responses simulated with the image-source model," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2007, pp. 159–162.

[31] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[32] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[33] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic data consortium*, vol. 10, no. 5, p. 0, 1993.

[35] D. Schuhmacher, B.-T. Vo, and B.-N. Vo, "A consistent metric for performance evaluation of multi-object filters," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3447–3457, 2008.