MULTIPLE SPEAKER TRACKING USING A MICROPHONE ARRAY BY COMBINING AUDITORY PROCESSING AND A GAUSSIAN MIXTURE CARDINALIZED PROBABILITY HYPOTHESIS DENSITY FILTER

Axel Plinge* *Daniel Hauschildt*$^{\sharp}$ *Marius H. Hennecke*† *Gernot A. Fink*†

* Leibniz Research Centre for Working Environment and Human Factors at TU Dortmund
$^{\sharp}$ Section Information Technology, Robotics Research Institute, TU Dortmund
† Intelligent Systems Group, Robotics Research Institute, TU Dortmund
{*axel.plinge, daniel.hauschildt, marius.hennecke, gernot.fink*}@*tu-dortmund.de*

ABSTRACT

Tracking speakers is an important application in smart environments. Acoustic tracking using microphone arrays is a challenging task due to two major reasons: On the one hand, multiple persons may speak simultaneously and thus the number of speakers varies over time; on the other hand, due to the nature of reverberated speech, the provided position hypotheses contain many gaps and clutter. In the proposed approach, the "glimpsing model" is realized by neurobiologically inspired calculation of robust but sparse position hypotheses in combination with a Gaussian mixture cardinalized probability hypothesis density filter. By iteratively applying the filter to the position hypotheses from multiple frequency bands, good results are achieved. Using a statistical speech model derived from recordings, a real-time capable implementation is used to track multiple speakers in a conference room with significant reverberation.

Index Terms— speaker tracking, glimpsing model, cochlear model, Peak-over-Average-Position, Gaussian mixture cardinalized PHD filter

1. INTRODUCTION

Human listeners show an impressive ability to locate and separate concurrent speakers by hearing in everyday situations. A popular theory, based on psychoacoustic experiments as well as biological and neurological research, is the "Auditory Scene Analysis" (ASA) [1]. It identifies atomic features and rules for their combination into objects or "streams" over time. To that end, both bottom-up feature-driven processes and top-down model driven processes are employed.

Bottom-up localization cues for humans are the intensity difference and the time difference between signals of both ears [2]. Encouraging results in speech separation and tracking have been achieved by ASA-based computer models in anechoic or low reverberation conditions [3]. By simulation of the "precedence effect" – the suppression of smaller secondary peaks following a strong first wavefront – the negative effect of reverberation can be reduced [4, 5]. Basic research oriented computational ASA applications use two sensors of an artificial human head, while technical localization solutions often employ circular or t-shaped microphone arrays with eight or more sensors. Recently, hybrid approaches applying biologically inspired neural processing to microphone arrays were proposed [6].

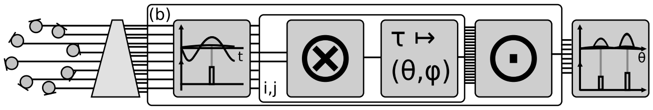


Fig. 1. Bottom-up processing: filter bank, spike generation, correlation, backprojection, combination and peak detection (f.l.t.r.).

In reverberant and noisy environments, only a few clearly localized time-frequency events can be found. The "glimpsing model" suggests that human speech perception in adverse conditions is based upon sparse clear events with high signal-to-noise ratio (SNR) [7]. To localize and track multiple speakers in reverberant environments, clearly localized "glimpses" can be integrated to continuous speaker tracks. This poses several challenges: The number of speakers varies over time and due to the nature of human speech there will be missed detections, clutter and no association between the measurements and the speakers.

Recently, several multi-target tracking algorithms like the multiple hypothesis tracker [8] or joint probabilistic data association [9] have been proposed to handle these problems. Unfortunately, both algorithms are computationally expensive since they need to handle the data association between measurements and tracks. A new approach to multi-target tracking based on random finite set (RFS) theory has been proposed by Mahler: The probability hypothesis density (PHD) filter [10] and a generalized version, the cardinalized PHD (CPHD) filter [11], in order to handle the multi target tracking more efficiently. In a nutshell, the efficiency is achieved due to the fact that the spatial multi-target probability distribution is only approximated by its first order moment, denoted as the PHD, thereby avoiding the combinatorial problem arising from the data association. In comparison to the standard PHD filter, the CPHD filter not only propagates the spatial probability distribution but also the cardinality probability distribution, therefore providing a much better cardinality estimate. Several implementations of the PHD and CPHD filters exist; most of them are either particle based [12] or Gaussian mixture (GM) [13] based approximations. In this paper, a GM CPHD filter variant [14] is chosen.

The input of the GM CPHD filter are position hypotheses generated in six steps as sketched in Figure 1. A circular eight microphone array is used for signal acquisition. The auditory processing in the cochlea is modeled by a filter bank for frequency separation followed by a nonlinear transformation mimicking the coding of neural pulses [2]. The robust Peak-over-Average-Position (PoAP) of spike generation method is employed with phase-locking to signal max-

This work was in part supported by the German Research Foundation (DFG) under contract number Fi 799/5-1.

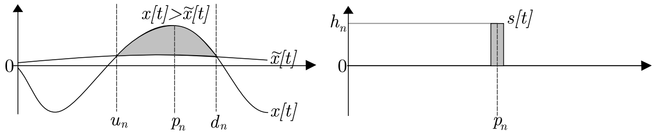


Fig. 2. Peak-over-average-Position (PoAP) spike generation.

ima, thus focusing on on-sets in order to imitate the "precedence effect" and "glimpsing" only strong events [15]. The result of neural pulse correlation is projected back to spatial coordinates. Subsequently, modulated peaks in the azimuth domain are detected, which are then used to generate the input for a GM CPHD filter. Using a statistical model derived from an initial recording, the CPHD filter provides the multi target tracking results.

2. NEURO-FUZZY LOCALIZATION

The frequency selectivity of the basilar membrane is modelled by a $B = 16$ band filter bank composed of gammatone filters equidistantly spaced on the equivalent rectangular bandwidth (ERB) scale between 200 Hz and 3.6 kHz. The filter bank is implemented via discrete Fourier transformation and overlap-add for time-domain reconstruction to provide online capability and avoid phase distortions at the same time. The filters are defined in the spectral domain using a gammatone approximation. To model the neural spike generation in the organ of Corti, rectangular pulses are generated phase-locked to signal maxima using the PoAP spike generation method [15] illustrated in Figure 2. The input signal $x[t]$ is compared to its 30 ms average $\tilde{x}[t]$. For each modulated interval $[u_n, d_n]$, where $x[t] > \tilde{x}[t]$, the maximum position

$$p_n = \underset{u_n \leq t \leq d_n}{\operatorname{argmax}} \, x[t] - \tilde{x}[t] \tag{1}$$

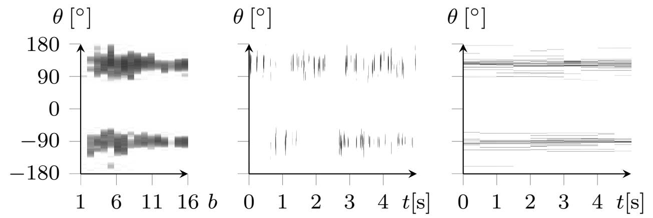
is determined and a spike of height h_n, computed from the sum of the Peak-over-Average amplitudes in the interval, is generated. The output $s[t]$ can be modeled as a vector sequence $(p_n, h_n)_n =: S_i^{(b)}$ for each microphone signal with index i in each frequency band with index b. By shifting the average relative to the signal, a basic simulation of the "precedence effect" is achieved [5]. Only high SNR peaks or "glimpses" are used as reliable witnesses for speech by accepting only peaks more than a threshold t_g above the average.

Time delay estimation between the ears in the auditory midbrain can be modeled via a cross-correlation of two signals in each frequency band b in accordance with the basic Jeffress-Colburn model [3, 2]. To reduce harmonic errors, a band and pair dependent correlation frame size is computed. The cross-correlations $e_{ij}^{(b)}[k, \tau]$, with τ denoting time delay, are calculated in 10 ms steps by matching all pairs of peaks inside the short-time frames. Due to the spikes time-domain sparsity, this is faster than performing a correlation in the spectral domain [15].

In order to represent the domain as a discrete set of coordinates, spherical spatial source positions $\boldsymbol{s} = u(\theta, \phi)$ with $\theta = -180°, -179°, \ldots, 179°$ and $\phi = 0°, 5°, \ldots, 45°$ are used. For speaker separation by azimuth with the circular array, coarse elevation precision is sufficient and under the far-field assumption the source distance can be neglected. The time delay of arrival is given by the difference of distances from the microphone positions $\boldsymbol{m}_i$, $\boldsymbol{m}_j$ to the source position $\boldsymbol{s}$

$$\tau_{ij}(\boldsymbol{s}) = (\|\boldsymbol{s} - \boldsymbol{m}_j\| - \|\boldsymbol{s} - \boldsymbol{m}_i\|) f_s \,/\, c. \tag{2}$$

The sampling frequency f_s is 48 kHz and the speed of sound $c =$ 343 m/s is assumed to be constant. Using linear interpolation, the energy $e_{ij}^{(b)}[k, \tau(\boldsymbol{s})]$ for each source position $\boldsymbol{s}$ is calculated by inverse mapping. Product-like combination of likelihoods using a Hamacher fuzzy t-norm h_γ yields robust estimates without "ghosts" [16]



(a) $\Sigma_t \Sigma_\phi e$ (b) $\Sigma_b \Sigma_\phi e$ (c) Σ_b $\mathrm{PoAP}_\theta \tilde{e}$

Fig. 3. Spatial likelihood for two concurrent speakers.

$$h_\gamma(x, y) = \frac{xy}{\gamma + (1 - \gamma)(x + y - xy)} =: x \odot y. \tag{3}$$

The microphones pairs' likelihoods are combined by iterated application of h_γ to calculate a joint pseudo likelihood

$$e^{(b)}[k, \boldsymbol{s}] = \bigodot_{(i,j)} e_{ij}^{(b)}[k, \tau(\boldsymbol{s})] \tag{4}$$

in time-location-frequency space over all microphone pairs (i, j).

Reverberant speech is found to produce Gaussian distributed peaks over longer time periods [17, 18]. Thus, a difference-of-Gaussian like post-processing can be applied to sharpen the localization results [15]. After calculating the average $\tilde{e}^{(b)}[l, \theta]$ over all elevations ϕ and a window of $L = 2$ s frames with index l shifted by 0.5 s, both a 5° and 45° average are calculated over azimuth, and the latter is subtracted form the first. The azimuth PoAPs

$$P^{(b)} = \left\{(\theta_n, l) | \theta_n \in PoAP_\theta \left(\tilde{e}^{(b)}[l, \theta]\right)\right\} \tag{5}$$

are extracted. Figure 3 shows the spatial likelihood for a recording of two concurrent speakers at fixed positions. On the left it is plotted against azimuth and frequency band. For higher frequencies the decreasing variance is clearly visible (a). The middle plot against time and azimuth illustrates the time domain sparsity (b). The PoAP_θ position hypotheses $P^{(b)}$ are shown on the right (c).

Due to the nature of reverberated speech, the provided position hypotheses still contain many gaps and clutter. Consequently, further post-processing is required. Additionally, multiple persons may speak simultaneously, thus the number of speakers varies over time. In such cases, GM CPHD filters are known to provide good results.

3. TRACKING WITH A GM CPHD FILTER

The tracking of multiple speakers localized by the aforementioned auditory processing inspired system is achieved by a GM CPHD filter variant following [14] which is in turn extended to handle multiple frequency bands. Beginning with an RFS of n individual states $x^{(i)} = (\theta^{(i)}, \dot{\theta}^{(i)})$

$$X = \left\{x^{(1)}, \ldots, x^{(n)}\right\} \tag{6}$$

at time k, the motion of a single speaker is modeled as a linear Gaussian model with mean $\zeta_{k-1} = \overline{x}_{k-1}$ and covariance Q

$$f_{k|k-1}(x_k | x_{k-1}) = \mathcal{N}(x_k; A\zeta_{k-1}, Q) \tag{7}$$

$$A = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}. \tag{8}$$

The time step Δt is set to 0.5 s. Given an RFS

$$Z_b = \left\{ z_b^{(1)}, \ldots, z_b^{(n)} \right\} \tag{9}$$

of single band measurements z_b and due to the fact that the azimuth PoAPs $P^{(b)}$ provide Gaussian distributed values for θ [18, 17], the single sensor measurement model is also a linear Gaussian model

$$\begin{aligned} g_{b,k}(z_b|x_k) &= \mathcal{N}\left(z_b; H\zeta_k, \sigma_b^2\right) & (10) \\ H &= \begin{pmatrix} 1 & 0 \end{pmatrix} & (11) \end{aligned}$$

where σ_b^2 is the per band measurement variance depicted in Figure 4. The $B = 16$ bands are modeled as individual sensors $Z = [Z_1, \ldots, Z_B]$ each having its own set of band specific detection and clutter probabilities. The main idea of fusing the per band measurements in order to obtain the speaker positions is to apply the CPHD filter update equation sequentially. Since the result will be largely dependent on the order of the sensors a prioritization by detection probability $p_{b,D}$ is performed for the multi-band sensor update.

In case of the GM CPHD filter, a birth intensity and cardinality distribution needs to be provided. Since a simple sensor likelihood model is used, a birth intensity model can easily be derived such that new states are generated at the measurement position with an angular speed of zero. As a result, the number of new born Gaussian components N_n equals the sum of all measurements of all bands per timestep k, in particular $N_n = \sum_{b=1}^{B} |Z_b|$. Consolidating all CPHD filter steps, the proposed algorithm works as follows:

1. Estimate GM components with the CPHD filter prediction.
   (a) Update all the individual GM components with the standard Kalman filter equation.
   (b) Create $N_n = \sum_{b=1}^{B} |Z_b|$ new GM components according to the birth model.
2. Prioritize sensors according to the detection probability.
3. Iteratively update GM components with measurements Z_i according to the CPHD filter update equation.
   (a) Apply the CPHD filter update equation.
   (b) Prune GM components according to [13].
4. Determine cardinality N from the *a posteriori* multi-object cardinality distribution by choosing the number of objects with the highest probability.
5. Extract states by choosing the N GM components with the highest weights.

Speaker labels are assigned using a basic time-to-live (TTL) approach inspired by [17]. For each state, the label of the nearest speaker from the last 5 s is selected, if existing, otherwise a new label is chosen.

4. EVALUATION

Recordings of multiple speakers were made in the conference room of our smart house, the FINCA [http://finca.irf.de]. The room is almost rectangular, about $3.7 \times 6.8 \times 2.6\,\mathrm{m}^3$, with strong reverberation. A circular array with a radius of $r_a = 5$ cm composed of eight omni-directional microphones was used.

In order to derive a model of the room, speech and speaker, a recording of a single speaker at fixed positions was made. From this data, a reverberation time of 624 ± 54 ms over all microphone signals was calculated using a blind estimation algorithm [19]. For the bottom-up processing, the "glimpsing threshold" t_g was set to 9 dB, the Hamacher parameter γ to 0.3 by inspection and the window size for the final temporal averaging was set to 2 s. Sensor statistics over the $P^{(b)}$ position hypotheses were calculated as parameters for the CPHD filters prioritization step. Each of the 16 bands differ greatly in terms of detection probability, variance and clutter. In Figure 4 azimuth deviation σ_b and detection probability $p_{b,D}$ are plotted per band. From the statistics, the birth p_Γ and clutter p_K cardinality distributions are assumed to be Poisson distributed with an average number of birth and clutter measurements of $\lambda_{p_\Gamma} = 0.2$ and $\lambda_{b,p_K} = 0.001$. The per band clutter intensity distributions $\kappa_b(z)$ are all chosen to be uniform with probability $\kappa(z) = 0.05$. For the state prediction process, the state independent survival probability and the process noise variance are $p_s = 0.99$ and $Q = 4°$.

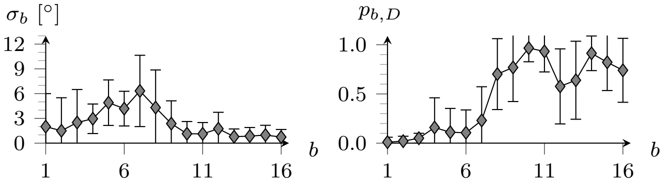


Fig. 4. $P^{(b)}$ statistics for a recording of a fixed single speaker.

For the tracking task, two speakers were speaking with considerable overlap. The first speaker was sitting at a fixed position of $\theta = 160°$ at a distance of 1.4 m to the array and started speaking immediately. The second speaker started after 30 s while walking around the array from $-135°$ to $135°$ and back again in a 1.25 m radius. The first speaker stopped talking 12 s after the rendezvous point. Both speakers wore a lapel-microphone to provide a ground truth for speech activity, presented in the $e_{1,2}$ plot of Figure 5.

The TTL conjunction forms two continuous tracks, as shown in Figure 5 (top). In the $|\tilde{Z}_b|$ plot, the average (solid) and maximum (dotted) number of position hypotheses is visible. This illustrates the sparsity of the position hypotheses, which hampers the tracking task. The overall Wasserstein distance [20] is depicted in the W_2 plot. The average value is $8.93°$. Representing the wrong speaker count, the cardinality error is shown in the ε_N plot. The overall root mean square error for the azimuth estimates θ is $5.64°$; $1.73°$ for the first and $7.46°$ for the second speaker. Allowing a deviation of one head width of 25 cm in speaker distance, a precision of 96.0 % (speakerwise 100 % and 92.7 %) was achieved, the recall was 99.1 % (98.1 % and 100 %). Continuous tracks are derived despite the noncontinuous occurrences of speech "glimpses" over the frequency bands. Comparing the plots, it can be seen that the larger errors in the Wasserstein distance correspond to the cardinality errors and speech gaps, for example at 50 s. The relative movement discrepancy to the linear ground truth assumption is a possible cause for the higher inaccuracy for the second speaker.

5. CONCLUSION

The GM CPHD filter was applied to the position hypotheses in the multiple frequency bands provided by the neurobiologically inspired bottom-up processing. By restricting the hypotheses to the most reliable ones and incorporating a sensor model into the GM CPHD filter, the "glimpsing model" could be realized for the tracking task. The real-time capable implementation performed favourably for temporal and spectral non-stationary speech signals. The applicability of the proposed multiple concurrent speaker tracking approach to real-world data recorded in a highly reverberant room was shown.

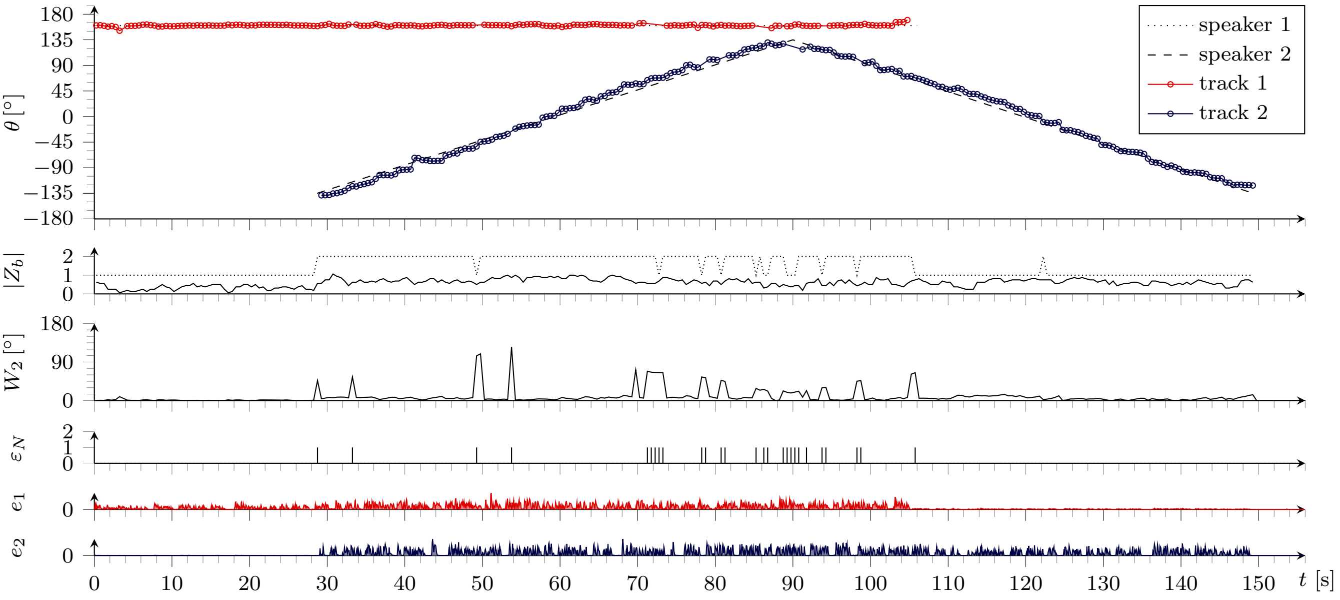


Fig. 5. Tracks for two speakers plotted on top of the linear ground truth, average number of position hypotheses over all bands $|\tilde{Z}_b|$, Wasserstein distance W_2 and cardinality error ε_N and lapel microphone speech energies $e_{1,2}$ (from top to bottom).

6. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[2] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*, MIT Press, 1996.

[3] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, IEEE Press/Wiley Interscience, 2006.

[4] J. Liu, D. Perez-Gonzalez, A. Rees, H. Erwin, and S. Wermter, "Multiple sound source localisation in reverberant environments inspired by the auditory midbrain," in *Proc. ICANN*, Limassol, Cyprus, 2009, pp. 208–217.

[5] K. J. Palomäki, G. J. Brown, and D. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, 2004.

[6] R. M. Stern, E. Gouvea, C. Kim, K. Kumar, and H.-M. Park, "Binaural and multiple-microphone signal processing motivated by auditory perception," in *Proc. HSCMA Workshop*, Trento, Italy, 2008, pp. 98–103.

[7] M. P. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.

[8] S. Thrun, "A probabilistic online mapping algorithm for teams of mobile robots," *Int. J. Robotics Research*, vol. 20, no. 5, pp. 335–363, 2001.

[9] J. Vermaak, S. J. Godsill, and P. Perez, "Monte Carlo filtering for multi-target tracking and data association," *IEEE Trans. Aero. Elec. Sys.*, vol. 41 (1), pp. 309–332, 2005.

[10] R. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aero. Elec. Sys.*, vol. 39, no. 4, pp. 1152–1178, 2003.

[11] R. Mahler, "PHD filters of higher order in target number," *IEEE Trans. Aero. Elec. Sys.*, vol. 43, no. 4, pp. 1523–1543, 2007.

[12] J. Kemper and D. Hauschildt, "Passive Infrared Localization with a Probability Hypothesis Density (PHD) Filter," in *Proc. WPNC*, Dresden, Germany, 2010.

[13] Ba-Ngu Vo and Wing-Kin Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4091–4104, 2006.

[14] Ba-Tuong Vo, Ba-Ngu Vo, and A. Cantoni, "Analytic implementations of the cardinalized probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3553–3567, 2007.

[15] A. Plinge, M. H. Hennecke, and G. A. Fink, "Robust neuro-fuzzy speaker localization using a circular microphone array," in *Proc. IWAENC*, Tel Aviv, Israel, 2010.

[16] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," *EURASIP J. Audio Speech & Music Process.*, vol. 2008, pp. 1–14, 2008.

[17] N. Madhu and R. Martin, "A scalable framework for multiple speaker localization and tracking," in *Proc. IWAENC*, Seattle, WA, USA, 2008.

[18] M. Cobos, J. J. Lopez, and S. Spors, "Analysis of room reverberation effects in source localization using small microphone arrays," in *Proc. ISCCSP*, Limassol, Cyprus, 2010.

[19] H. W. Löllmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in *Proc. IWAENC*, Tel Aviv, Israel, 2010.

[20] J.R. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst., Man, Cybern.*, vol. 34, no. 3, pp. 327 – 336, 2004.