Robust Neuro-Fuzzy Speaker Localization Using a Circular Microphone Array

Axel Plinge, Marius H. Hennecke and Gernot A. Fink
Intelligent Systems Group, Robotics Research Institute, Technical University Dortmund, Germany

***Abstract*—A major application area of microphone array processing is the localization of sound sources, mainly of speaking persons. In contrast to most state-of-the-art approaches that are based on correlation measures, we propose a neurologically inspired system that generalizes findings about human spatial hearing to the multi-channel case. It mimics the processing in the human cochlea and the auditory mid-brain. To enhance the localization quality, a new spike generation approach is introduced, termed peak-over-average position (PoAP). A fuzzy combination is used to remove putative artifacts. In contrast to a human listener we employ multiple sensors to gain robustness in reverberant and noisy environments. Post-processing estimates the locations of concurrent speakers. The robustness of the proposed system is shown by comparison with the well-known steered response power approach. Finally, we show the applicability of our realtime neuro-fuzzy model to the concurrent speaker localization task using real reverberant recordings.**

***Index Terms*—microphone array, peak-over-average position, glimpsing model, precedence effect, speaker localization**

I. Introduction

The impressive ability of human listeners to locate and separate concurrent speakers in everyday situations has been a motor for research for over half a century. The fascinating mechanisms that allow us to create a rich auditory world from the movement of two eardrums have been partially demystified. Psychoacoustic experiments as well as biological and neurological research led to the popular "Auditory Scene Analysis" (ASA) theory [1]. It identifies atomic features, grouping "cues", and rules for their combination into "streams" over time in human auditory processing. With growing computing power, a variety of implementations were devised that use and advance our understanding of ASA [2].

Localization cues for humans are intensity difference and time difference between signals from both ears. Inspired by the human spatial hearing, the auditory processing in the cochlea is commonly modeled by a filter bank for frequency separation followed by a transformation mimicking the coding of neural pulses. Several implementations follow the classical cochlear model of Lyon [3] and use half-way rectified square-root compressed band signals to imitate the spike generation in the auditory nerve. The signals from both ears are correlated, modeling the binaural processing in the auditory mid-brain. This leads to a time delay; hence, to the direction of the acoustic source. However, half-way rectification leads to blurred correlation figures. Focusing on the phase-locking property of the cochlea, one-way zero-crossings of the band filtered signal itself yield better results [4].

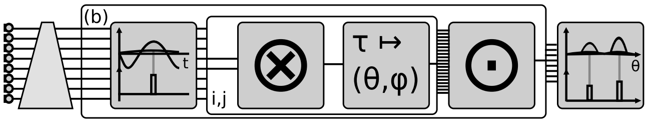

Figure 1. Processing structure: filter bank, spike generation, correlation, backprojection, combination and peak detection (from left to right).

Good results in speech separation and tracking have been achieved under anechoic or low reverberation conditions by ASA-based models [5]. Focusing on on-sets and imitating the "precedence effect", localization in reverberant conditions can be improved [6,7]. The recently proposed "glimpsing model" suggests that human speech perception is based upon sparse clear events with high signal-to-noise ratio (SNR) [8]. Still, reverberant and noisy environments remain a challenging task for ASA-models.

Fixing the number of sensors at two — aimed at strict imitation of the human prototype — is an unnecessary constraint for a technical system. The use of multiple sensors facilitates robust localization in noisy and reverberant environments, by exploiting the redundancy among all channels. Recently, hybrid approaches applying acoustic signal processing in combination with biologically inspired neural processing to subband or circular microphone arrays were proposed [9].

Here, such a hybrid approach is presented, that pursues two strategies to cope with reverberation and noise. First, it uses a small circular eight channel array, which provides non-aliased time delays and good coherence. Second, biologically inspired processing is implemented that singles out strong, clearly localized peaks to determine the speakers positions. We propose a refined cochlear model, employing a new and robust method of spike generation based on modulation maxima. Additionally, a fuzzy combination of modulated energy peaks forms a sharp representation of sound events in azimuth-time space. The systems performance and robustness in noisy and reverberant environments is demonstrated with real recordings with one and two concurrent speakers.

II. Neuro-Fuzzy Localization

The location hypotheses are generated in six steps as sketched in figure 1. First, each channel is split individually with a filter bank. Then, spikes are generated in each band. Microphone pairs are combined via cross-correlation to estimate time-delays, which are projected back into the spatial domain. Subsequently, these are combined using a fuzzy operation. Finally, modulated peaks in the azimuth domain are detected.

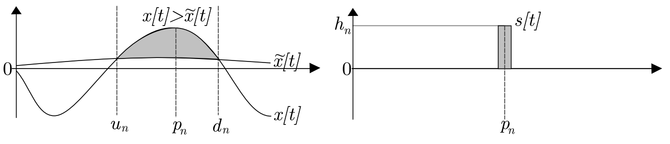

Figure 2. Peak-over-average spike generation. Input signal $x[t]$ and its average $\tilde{x}[t]$ with interval $[u_n, d_n]$ and maximum position p_n on the left, and the generated spike $s[t]$ on the right.

A. Filtering

According to the Patterson-Holdsworth model, the frequency selectivity of the cochlea is adequately modeled by a filter bank composed of n_b gammatone filters spaced according to the ERB scale [2, pp. 15–19]. 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 [10]

$$G^{(b)}[f] = (1 + \mathrm{j}(f - f_b)/w_b)^{-4}\,, \tag{1}$$

where j is the imaginary unit. For each band b, a time domain signal $x^{(b)}[t]$ is calculated using the parameters of Glasberg and Moore [2] with center frequency f_b and bandwidth w_b.

B. Spike Generation

In the mammalian cochlea, spikes are generated in a phase-locked manner. Intensities are coded relative to the overall amplitude [11]. The “precedence effect”, i.e. suppression of smaller secondary peaks following a strong first wavefront, is improving localization of reverberant signals [2,5]. Overall, only high SNR peaks or “glimpses” are used as reliable witnesses for speech. All these properties are incorporated in a single nonlinear time domain step.

Modulated peaks p_n in the signals $x[t]$ of each frequency band are detected as illustrated in figure 2. Here, n denotes the peak index and t the sample index, the band index (b) is omitted in this section for the sake of brevity. The signals moving average $\tilde{x}[t]$ is calculated with a $a = f_s \cdot 30\,\mathrm{ms}$ sample neighborhood to encompass two pitch periods, where f_s denotes the sampling frequency. Modulated intervals $[u_n, d_n]$ are identified between zero-crossings of the average subtracted signal $\hat{x}[t] := x[t] - \tilde{x}[t-d]$, shifted by $d = f_s \cdot 3\,\mathrm{ms}$ samples. The position

$$p_n = \underset{u_n \le t \le d_n}{\operatorname{argmax}}\, \hat{x}[t] \tag{2}$$

is termed the peak-over-average position PoAP_t. To encode the amplitude of the source signal, in accordance with the spike count in the auditory nerve, the amplitude

$$h_n = 2^{f_b/1000} \sum_{t=u_n}^{d_n} (\hat{x}_i[t])^{0.5} \tag{3}$$

is modulated by the signals square-root compressed peak-over-average values with a band dependent weighting to emphasize higher frequencies. Each PoAP_t event p_n triggers a rectangular impulse with a width of 50 µs and height h_n, as shown in figure 2. This timing encodes the phase of the source signal. A basic thresholding is applied to ignore peaks, that are less than $6\,\mathrm{dB}$ over the average.

C. Spike Correlation

The basic Jeffress-Colburn model [2, p. 162] argues, that time delay estimation between the ears in the auditory mid-brain can be modeled via a cross-correlation of the two signals. Here, two types of aliasing can occur: Spatial aliasing, where one wavelength spans less than the distance of the two microphones; and harmonic errors where multiple peaks fall into the correlation window. The gammatone filters attenuation is sufficiently high ($> 24\,\mathrm{dB}$) at $f_b \pm 2w_b$ to consider this the bands' boundaries. To avoid spatial aliasing, pairs (i, j) satisfying the condition

$$P^{(b)} = \left\{(i,j) \mid \|\boldsymbol{m}_i - \boldsymbol{m}_j\| < c \,/\, (f_b + 2w_b)\right\} \tag{4}$$

are used, where $\boldsymbol{m}_i$, $\boldsymbol{m}_j$ denote the microphone positions, $\|\cdot\|$ is the Euclidean distance and c the speed of sound. To reduce harmonic errors, we propose a band and pair dependent correlation frame size

$$K^{(b)}_{(i,j)} = (14\,\mathrm{ms} + \|\boldsymbol{m}_i - \boldsymbol{m}_j\|\,/c + 2\,/\,(f_b - 2w_b))\, f_s, \tag{5}$$

computed as the sum of the maximum pitch period of $14\,\mathrm{ms}$, the microphone pair distance and two wavelengths. The cross-correlations $e^{(b)}_{ij}[k,\tau]$, with τ denoting time delay, are calculated by matching all pairs of peaks inside the frames. Due to the spikes time-domain sparsity, this is faster than performing a correlation in the spectral domain.

D. Backprojection and Combination

Next, the time delays τ for each microphone pair are mapped back into the spatial domain. Spatial source positions

$$\boldsymbol{s} \in \{u(\theta,\phi) := (r\sin\theta\cos\phi, r\cos\theta\cos\phi, r\sin\phi)^T\} \tag{6}$$

are used to represent the domain as a fixed set of spherical coordinates. Under the far-field assumption for a small circular array, the radius can be fixed to an arbitrary distance, e.g., $r = 1.5\,\mathrm{m}$. 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 $\boldsymbol{s}$:

$$\tau_{ij}(\boldsymbol{s}) = (\|\boldsymbol{s} - \boldsymbol{m}_j\| - \|\boldsymbol{s} - \boldsymbol{m}_i\|) f_s \,/\, c. \tag{7}$$

Using linear interpolation, the energy $e^{(b)}_{ij}[k, \tau(\boldsymbol{s})]$ for each source position $\boldsymbol{s}$ is calculated by inverse mapping. Due to the spatial ambiguity, time delays for a single pair correspond to two azimuths, which is resolved by the combination of multiple pairs. In contrast to the common additive combination, we employ a Hamacher fuzzy t-norm based combination with parameter γ which has proven to be a robust method [12]:

$$h_\gamma(x,y) = \frac{xy}{\gamma + (1-\gamma)(x + y - xy)} =: x \odot y. \tag{8}$$

All desirable pairs are combined by iterated application of h_γ, $\bigodot_{i \in I} x_i := (((x_1 \odot x_2) \odot \ldots) \odot x_n)$, to calculate the peak energy distribution $e^{(b)}[k, \boldsymbol{s}]$ in time-location-frequency space:

$$e^{(b)}[k, \boldsymbol{s}] = \bigodot_{(i,j)\in P^{(b)}} e^{(b)}_{ij}[k, \tau(\boldsymbol{s})]. \tag{9}$$

E. Peak Localization

In the presence of strong reverberation, the number of clearly modulated events per second is small and secondary peaks are observable. However, assuming slow moving speakers, the true primary peaks prevail in a larger time context. The moving average over one second, i.e. $L = f_s \cdot 1\,\mathrm{s}$ is calculated over all data points with a shift of $f_s \cdot 250\,\mathrm{ms}$ samples:

$$\tilde{e}^{(b)}[k, \boldsymbol{s}] = \sum_{k'=k-L/2}^{k+L/2} e^{(b)}[k', \boldsymbol{s}]. \tag{10}$$

For natural speech, spectral magnitudes are dependent across frequency [13], which is exploited by multiple ASA grouping cues such as "common fate" and onset [2]. It may be assumed that no or only few $e^{(b)}[k, \boldsymbol{s}]$ values originating from different speech sources collide in frequency. If noise and reverberation are independent across frequency, common peak positions in the frequency bands provide independent "witnesses" for speech. Thus the sum over all frequency bands will likely produce peaks for speech energy peaks of a single source. Time domain aliasing in the correlation occurs frequency-dependent and therefore produces erroneous peaks at different source locations in different frequency bands; hence, noise and aliasing errors can be suppressed by counting the bands with energy peaks, $B[k, \boldsymbol{s}] := \{b \mid \tilde{e}^{(b)}[k, \boldsymbol{s}] > 0\}$, and discarding detections occurring in less than a third of the n_b frequency bands which corresponds to the typical spread of natural speech sounds [11]:

$$\tilde{e}[k, \boldsymbol{s}] = \begin{cases} \sum_{b \in B[k,\boldsymbol{s}]} \tilde{e}^{(b)}[k, \boldsymbol{s}] & \text{if } |B[k, \boldsymbol{s}]| \geq \lfloor n_b/3 \rfloor \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

In the typical tabletop placement, only shallow positive elevation angles are in the region of interest, and speakers can be separated by azimuth. If discrimination by elevation is not desired, $\tilde{e}[k, \boldsymbol{s}]$ can be summed over elevation

$$\tilde{e}[k, \theta] = \sum_{\phi} \tilde{e}[k, \boldsymbol{s} = u(\theta, \phi)]. \tag{12}$$

When considering larger time segments, the correlation results can be modeled as "true" peaks plus noise [14]. To incorporate the typical variations a 45° average, spanning the reverberation induced artifacts, is subtracted from a 5° average representing the signal. With this final angular peak-over-average evaluation, positions of modulated peaks are computed:

$$p^*[k, \theta] = \mathrm{PoAP}_\theta\, \tilde{e}[k, \theta]. \tag{13}$$

III. Results

The proposed system is implemented in C++ utilizing OpenMP and the FFTW, which allows for realtime performance on commodity hardware. Following, results on simulated reverberant signals and on real recordings are presented.

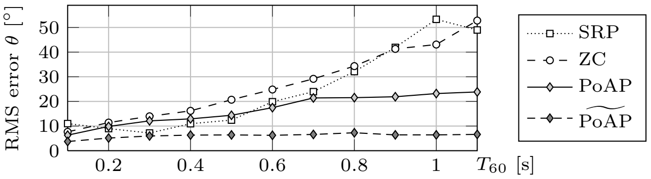

Figure 3. Comparison of argmax_θ localization performance by azimuth error average over positions in 1 to 1.5 m distance and 0° to 20° elevation with simulated 12dB SNR.

A. Simulation

Extensive tests were performed in a simulated reverberant environment to determine reasonable parameters. For the Hamacher t-norm $\gamma = 0.3$ lead to robust location estimates. Smaller values of γ make the processing more susceptible to noise, higher values produce sparser results. For the backprojection, azimuth and elevation are discretized as $\theta \in \{0°, 1°, \ldots, 359°\}$ and $\phi \in \{0°, 5°, \ldots, 35°\}$. The frame shift is fixed at $10\,\mathrm{ms}$. A sampling rate of $f_s = 48\,\mathrm{kHz}$ is used in conjunction with 16 bands with $f_b = 0.25, \ldots, 3.6\,\mathrm{kHz}$. The speed of sound is fixed to $c = 343\,\mathrm{m/s}$.

A rectangular $5 \times 6 \times 2.5\,\mathrm{m}$ room with speakers surrounding a circular eight channel microphone array with $r_a = 5\,\mathrm{cm}$ was simulated, employing the image source method [15] to allow for varying reverberation times T_{60}. Uncorrelated white noise was added to achieve a SNR of $12\,\mathrm{dB}$. The simulated speakers were positioned at three different azimuths; 0°, 10° and 20° elevation; and in 1 to 1.5 m distance around the array, separately uttering 5 s of speech.

In figure 3, the root-mean-square (RMS) error for $\mathrm{argmax}_\theta\, e[k, \theta]$ is plotted for the average over all positions. The localization quality of PoAP_t spike generation (diamonds) is compared to the zero-crossing method (circles); Half-way rectification produced too few usable estimates. Additionally, performance of the steered response power (SRP) approach with Hamacher t-norm combination [12] is shown (squares). To illustrate the effect of $L = f_s \cdot 1\,\mathrm{s}$ temporal averaging, the precision of $\mathrm{argmax}_\theta\, \tilde{e}[k, \boldsymbol{s}]$ is plotted as well (dark diamonds).

Compared with the zero-crossings and SRP, the PoAP_t detections are more robust against high reverberation with around 20° error for reverberation times over 0.7 s. The RMS azimuth error stays below 10° for all evaluated T_{60} times using temporal averaging.

B. AV16.3 corpus

The freely available AV16.3 corpus provides real world data and ground truth information for a variety of situations in a mildly reverberant meeting room [16]. Two circular eight channel microphone arrays with a radius of $r_a = 10\,\mathrm{cm}$ are used for recording at $f_s = 16\,\mathrm{kHz}$. The low sampling rate provides only coarse phase information. The larger diameter is responsible for lower spatial coherence and more spatial aliasing. The usable frequency range is reduced to under $2\,\mathrm{kHz}$, providing sparser peak events in time. $n_b = 12$ bands with $f_b = 0.2, \ldots, 1.6\,\mathrm{kHz}$ were used.

Sequence 1 is a recording of a single speaker taking up 16 positions in the conference room in proximity to the microphone arrays with a total length of 218 s. Figure 4 shows the detections for the first part of the sequence. A few additional detections due to foot fall sound etc. occurred. The RMS azimuth error over the whole sequence is 3.5° and 3.8° using array 1 and 2 respectively. Based on the average human head width in 1.5 m distance, detections within a margin of 6° of the true azimuth are considered accurate. Applying this border, the precision for array 1 and 2 is 92 % and 88 % respectively. The RMS energy sum of the filter bank output for the lapel microphone recording (figure 4 bottom) was used as ground truth for speech activity with a $-30\,\mathrm{dB}$ threshold. In respect to this, the $L = f_s \cdot 1\,\mathrm{s}$ frames detections achieve 99 % recall for both arrays.

C. FINCA Recordings

Several recordings were made in the conference room of our smart house, the FINCA [http://finca.irf.de]. The room is almost rectangular, about $3.7 \times 6.8 \times 2.6\,\mathrm{m}$ with a high reverberation time of approximately 0.6 s. A circular tabletop array with a radius of $r_a = 5\,\mathrm{cm}$ composed of eight omnidirectional microphones was used. Recordings were made with a sampling rate of $f_s = 48\,\mathrm{kHz}$. Using the smaller diameter and higher sampling frequency in comparison to the AV16.3 setup, speech in the full frequency range can be localized. Therefore, denser and more peaks are detected. $n_b = 16$ bands spanning the range of voiced speech components with $f_b = 0.25, \ldots, 3.6\,\mathrm{kHz}$ were used. Peaks in the range between 1.5 and 3.6 kHz are significantly sharper localized.

To test concurrent speaker separation, one person was sitting at the table, talking continuously, while the other was walking through the room, pausing shortly between positions. Resulting detections are depicted in figure 5. Both Speakers utterances form clearly visible tracks of $p^*[k, \theta]$ detections. The overall RMS azimuth error is 5.5°. Within a 6° margin, 80 % precision is achieved.

IV. Conclusion

The proposed application of biologically inspired signal processing methodologies for speaker localization employing a microphone array resulted in reliable short and long-term estimates for highly reverberant and low SNR conditions. Furthermore, the system is able to localize multiple concurrent speakers in real reverberant rooms. The robustness of our approach stems from the presented PoAP_t spike generation. The proposed frequency and microphone distance dependent window size reduces spatial aliasing and the influence of ambiguous correlation peaks. By employing the Hamacher t-norm for a fuzzy backprojection, the overall localization performance is improved. Temporal averaging sharpens the peak detections in a larger time context. The presented system performs favorably for the speaker localization task in comparison to state-of-the-art methods and runs in realtime on commodity hardware.

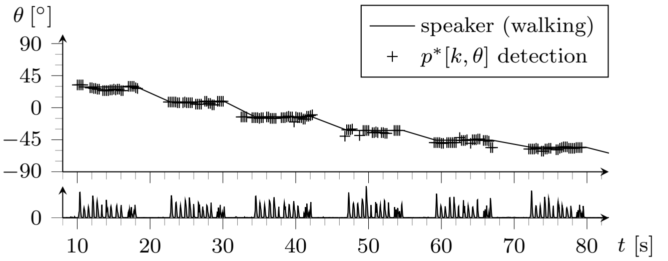

Figure 4. Localization for AV 16.3 Sequence 1: Detections (top), speech energy from lapel microphone (bottom).

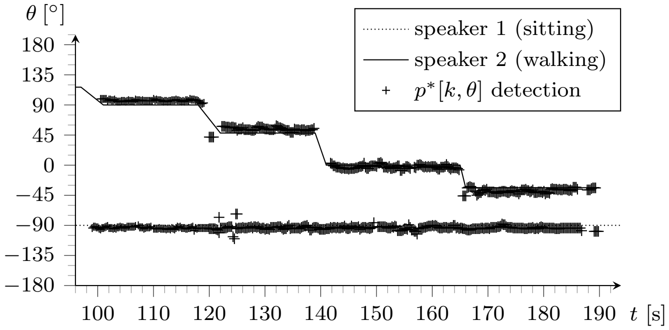

Figure 5. Localization of two concurrent speakers in the highly reverberant room of the FINCA.

References

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

[2] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, Algorithms, and Applications*. IEEE Press/Wiley Interscience, 2006.

[3] R. Lyon, "A computational model of binaural localization and separation," in *IEEE Int. Conf. Acoustics Speech & Signal Process.*, vol. 8, Boston, Massachusetts, USA, 1983, pp. 1148–1151.

[4] Y.-I. Kim, S. An, and R. Kil, "Zero-crossing based binaural mask estimation for missing data speech recognition," in *IEEE Int. Conf. Acoustics Speech & Signal Process.*, vol. 5, Toulouse, France, 2006.

[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] V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Korner, "A probabilistic model for binaural sound localization," *IEEE Trans. Systems, Man, & Cybernetics*, vol. 36, no. 5, pp. 982–994, 2006.

[7] 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*, 2009, pp. 208–217.

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

[9] R. M. Stern, E. B. Gouvea, and G. Thattai, ""polyaural" array processing for automatic speech recognition in degraded environments," in *Proc. INTERSPEECH*, 2007, pp. 926–929.

[10] M. Unoki and M. Akagi, "A method of signal extraction from noisy signal based on auditory scene analysis," *Speech Commun.*, vol. 27, no. 3, pp. 261–279, 1999.

[11] S. Handel, *Listening*. MIT Press, 1989.

[12] 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.

[13] J. Peterson and C. Kyriakakis, "Analysis of source localization in reverberant environments," in *IEEE SAM Workshop*, Waltham, Massachusetts, USA, 2006, pp. 672–676.

[14] G. Lathoud and J.-M. Odobez, "Short-Term Spatio-Temporal Clustering applied to Multiple Moving Speakers," *IEEE Trans. Audio Speech & Language Process.*, 2007.

[15] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[16] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "Av16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI '04 Workshop; LNCS*, vol. 3361, 2005, pp. 182–195.