# SIMILARITY-BASED SOUND SOURCE LOCALIZATION WITH A COINCIDENT MICROPHONE ARRAY

*Karl Freiberger,* * †
Institute of Electronic Music and Acoustics,
University of Music and Performing Arts
Graz, Austria
karl.freiberger@gmx.at

*Alois Sontacchi,* †
Institute of Electronic Music and Acoustics,
University of Music and Performing Arts
Graz, Austria
sontacchi@iem.at

## ABSTRACT

This paper presents a robust, accurate sound source localization method using a compact, near-coincident microphone array. We derive features by combining the microphone signals and determine the direction of a single sound source by similarity matching. Therefore, the observed features are compared with a set of previously measured reference features, which are stored in a look-up table. By proper processing in the similarity domain, we are able to deal with signal pauses and low SNR without the need of a separate detection algorithm. For practical evaluation, we made recordings of speech signals (both loudspeaker-playback and human speaker) with a planar 4-channel prototype array in a medium-sized room. The proposed approach clearly outperforms existing coincident localization methods. We achieve high accuracy (2° mean absolute azimuth error at 0 dB SNR) for static sources, while being able to quickly follow rapid source angle changes.

## 1. INTRODUCTION

The task of acoustic source localization (ASL) is to estimate the location of one or several sound sources given acoustic information only. Typically, a microphone array is used as a sensor front-end. ASL can be used to determine the steering direction of a microphone array beamformer and/or to direct a video camera towards the estimated source direction [1]. Typical applications are hands-free communication, conferencing systems and human-like robots. Most of the established methods for ASL use spatially distributed microphones to capture the direction-dependent time difference of arrival (TDOA) [2, 3]. Since the magnitude of the TDOA is directly related to the microphone spacing, arrays well suited for TDOA-based ASL require more space than so called near-coincident microphone arrays (NCMAs).

NCMAs consist of two or more microphone capsules having their acoustic center as close to each other as possible. Instead of evaluating time differences, the key principle behind ASL with NCMAs is to use level differences between the microphone signals. These level differences can be caused either by dedicated directional capsules [4, 5, 6] or by omni-directional transducers which are differentially combined [7]. Established methods for coincident localization [4, 5, 6, 7] share in common that the source direction is determined by computing the active sound intensity vector.

In this paper, we propose a new coincident ASL-method based on supervised pattern recognition via a minimum distance classifier. The principle of our approach is depicted in Fig. 1. We derive specific features $\boldsymbol{Y}$ from the captured microphone signals and compare them to a set of pre-measured reference features, stored in a look-up table. This table consists of a number $Q$ of feature vectors $\boldsymbol{\mathcal{Y}}(\boldsymbol{\Theta}_q)$, each relating to a specific source position $\boldsymbol{\Theta}_q$. Basically, the source position is estimated as that position $\boldsymbol{\Theta}_q$ where $\boldsymbol{\mathcal{Y}}(\boldsymbol{\Theta}_q)$ is most similar to $\boldsymbol{Y}$.

To be able to track changing source locations, the processing is performed frame-wise. For each frame $l$, we obtain a similarity curve (SC) $C_l(\boldsymbol{\Theta}_q)$ via the Euclidean distance between $\boldsymbol{Y}_l$ and $\boldsymbol{\mathcal{Y}}(\boldsymbol{\Theta}_q)$. The SC is ought to peak at the position of the sound source. If the shape of the SC is however flat, without a clear, global maximum, it is likely that the current frame would produce a more or less random source location estimate. This is for instance typical for a speaking pause between two words. By using the shape of the SC for weighting the influence of the observed frame with respect to previous ones, we can however effectively suppress the influence of signal pauses. With that, we obtain a stable source position estimate without jumping away from the source in signal pauses which is important in many practical applications such as camera- or beam-steering.

Instead of smoothing the sequence of location estimates with a fixed time-constant, our approach is adaptive. This makes our position estimator able to quickly follow a sudden change of the source location and produce a very smooth result without outliers in case of a static source position. In contrast to a separate voice activity detector, our SC-shape based detection method is independent of the signal type, e.g. speech, narrow-band, noise, transient signals, and comes at virtually no additional computational cost.

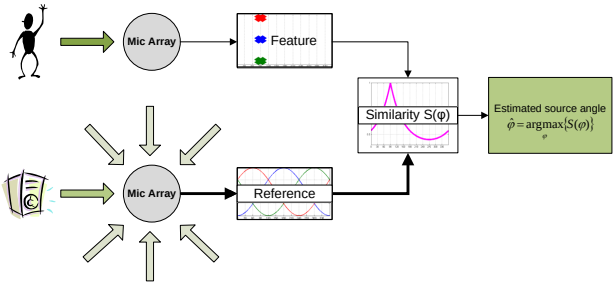
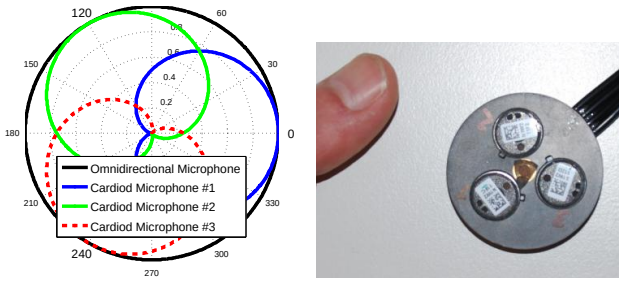



Figure 1: Supervised pattern classification principle.

* New affiliation: BCT - Electronic GesmbH, Saalachstrasse 88, 5020 Salzburg, Austria

† Thanks to AKG Acoustics, GmbH, Vienna, esp. Martin Opitz, Marco Riemann and Matthias Maly-Persy, for providing the prototype microphone array and supporting our work.




(a) Idealized polar plots (b) Photo of the prototype array

Figure 2: The microphone array used in this paper.

A good basis for localization of a single sound source is to track strong frequencies and compute an average SC over those frequencies. To extend our method to tracking of multiple sources, we suggest to perform clustering in the frequency-dependent SC instead of averaging. Multi-source tracking is however beyond the scope of this paper.

A basic property inherent to coincident localization is that two sources providing energy at the same frequency at the same time cannot be separated. Instead, if the sources have equal level, the mean direction is detected. However, multiple speakers are typically not always active at the exact same time and frequency and hence histogram or time-averaging approaches make multi-speaker localization possible [7].

In [7], it is suggested to use a measure of diffuseness as a reliability measure, which is conceptually very similar to our approach with rating the shape of the SC. As outlined above, we do however not use a threshold and do not use a fixed time constant for averaging, which makes our position estimator able to react either quick or smooth, depending on the situation. Another point relevant in practice, is the effect of microphone mismatch. Here, the coincident localization approach is likely to get less accurate, will however not completely fail, because the the overall characteristic of the level differences will not completely change due to manufacturing tolerances.

## 2. METHOD

### 2.1. Signal Model

Consider a single, acoustic point source $s$ located at a position $\boldsymbol{\Theta}_s = \left[\varphi_s, \quad \vartheta_s, \quad r_s\right]^T$. $\varphi$, $\vartheta$, and $r$ denote the spherical coordinates azimuth, elevation and radius, respectively. The coordinate system is centered at the center of a coincident microphone array. This array consists of $M$ microphone capsules indexed by $m = 0, \ldots, M-1$. Our goal is to estimate $\boldsymbol{\Theta}_s$ from the microphone array signals. In short time Fourier transform (STFT) domain, a linear, time invariant (LTI) model of the $m^{th}$ microphone signal is given as

$$X_m(l,f) = S(l,f) \cdot H_{m|\boldsymbol{\Theta}_s}(l,f) + V_m(l,f) \tag{1}$$

where $X_m(l,f)$, $S(l,f)$ and $V_m(l,f)$ represent the $m^{th}$ microphone, acoustic source and an additive disturbance (noise) signal, respectively. $l$ is the frame time index and $f$ the frequency index. $H_{m|\boldsymbol{\Theta}_s}(l,f)$ represents the frequency response of the $m^{th}$ microphone given a source position $\boldsymbol{\Theta}_s$.

Because the frequency response is dependent on the source position, we refer to $H_{m|\boldsymbol{\Theta}_s}(l,f)$ as the position dependent frequency response (PDFR). It models frequency dependence as well as the directivity and the proximity effect [8] of the microphone. The PDFR of an ideal first order microphone including the proximity effect, is given as [8]

$$H_{m|\boldsymbol{\Theta}_s}(f) = (1-\beta_m) + \beta_m \cos(\varphi - \varphi_m)\cos(\vartheta - \vartheta_m)\frac{1 + j\frac{2\pi f}{c}r}{j\frac{2\pi f}{c}r} \tag{2}$$

where $j = \sqrt{-1}$, $c$ is the speed of sound, $(\varphi_m, \vartheta_m)$ is the look-direction of the microphone and $\beta_m$ specifies the directivity. For high $2\pi f/c \cdot r$ (far-field), $\beta_m = 0.5$, $\beta_m = 0$, $\beta_m = 1$ yields the well-known cardioid, omni-directional and figure-8 polar pattern, respectively. To model a real microphone, the PDFR can be obtained by means of impulse response measurements, e.g. using the exponential sine sweep method [9].

### 2.2. Microphone array

For the following discussion and derivation of our ASL-algorithm we restrict to the planar, 4-channel NCMA configuration depicted in Fig. 2. This array consists of one omni-directional microphone capsule and three directional, first order microphones (cardioid polar pattern) respectively. The cardioids are oriented towards the azimuth angles $0^\circ$, $120^\circ$ and $240^\circ$, respectively, within the same plane $\vartheta = 0$. Instead of using a separate omni-directional capsule, the omni-characteristic can also be achieved by summing the cardioids [10]. In our experiments, the source localization performance was the same in both cases. Using a separate microphone can however produce a better low-end sound when coincident beamsteering is performed.

Due to the planar setup, robust estimation of the elevation angle $\vartheta_s$ is hardly possible. With a 3D-array such as the SFM it should however be possible to perform estimation of the elevation angle equally well as azimuth-estimation. The proximity effect provides a physical basis for estimation of the source distance $r_s$. However, first experiments and theoretical considerations indicate, that distance estimation is very sensitive to noise and limited to close (nearfield) sources [10]. Therefore, this paper restricts to tracking of the source azimuth angle. Hence, we use $\varphi$ instead of $\boldsymbol{\Theta}$ in (4) and all following equations. Furthermore, only tracking of a single sound source is considered. We do however suggest how to extend the presented method to allow for tracking of multiple sources at the same time.

### 2.3. Features

The basic idea behind our features is that there is a direction-dependent triplet of cardioid microphone gains (cf. Fig. 2a and Fig. 3, top). For our observed feature vector $\boldsymbol{Y}(l,f)$, we must try to obtain to these gains from the microphone signals $X_m(l,f)$.

$$\boldsymbol{Y}(l,f) = \left[Y_1(l,f), Y_2(l,f), Y_3(l,f)\right]^T \tag{3}$$

$$Y_m(l,f) = \frac{|X_m(l,f)|}{|X_0(l,f)|} = \frac{|S(l,f) \cdot H_{m|\varphi_s}(l,f) + V_m(l,f)|}{|S(l,f) \cdot H_{0|\varphi_s}(l,f) + V_0(l,f)|} \tag{4}$$

The directional microphones are indexed by $m = 1,2,3$, and $m = 0$ is the omni-directional channel. The normalization by

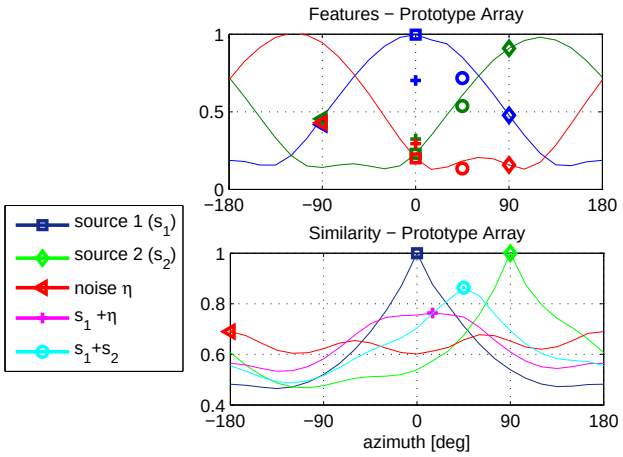



Figure 3: Effect of different source configurations (single sources, omni-directional noise, mix). Top: Reference features (solid lines) and feature vectors (markers) on basis of measured data of the prototype array (1 kHz, $r = 1\,\mathrm{m}$, $\vartheta = 0$). Bottom: Corresponding similarity curves (SCs). In case of a single sound source there is a sharp peak in the SC. If several directions contribute energy at the same frequency (extreme case: omni-directional noise, red curve) the SC gets flatter.

$|X_0(l,f)|$ is useful to become less dependent on the source signal: If $S(l,f)$ provides enough energy to suppress the influence of the noise terms $V_m(l,f)$, i.e. $S(l,f)\cdot H_{m|\varphi_s}(l,f) >> V_m(l,f)$, the source signal $S(l,f)$ in (4) cancels and only the ratio of the PDFRS remains. This ratio is known for a variety of source angles, because the PDFRs can be measured for a number $Q$ of angles $\varphi_q$, $q = 0,\ldots,Q-1$. The basic reference feature vector $\boldsymbol{\mathcal{Y}}(l,f,\varphi_q) = \left[\mathcal{Y}_1(l,f,\varphi_q),\mathcal{Y}_2(l,f,\varphi_q),\mathcal{Y}_3(l,f,\varphi_q)\right]^T$ is hence defined by

$$\mathcal{Y}_m(l,f,\varphi_q) = \frac{|H_m(l,f,\varphi_q)|}{|H_0(l,f,\varphi_q)|} \tag{5}$$

In case of multiple sources, reverberation or measurement noise the disturbance terms $V_m(l,f)$ in (4) cannot be neglected and the reference features in (5) are not appropriate. The difference between the cardioid channels decreases and hence the feature curves get compressed (cf. Fig. 3). With our planar array, the same thing happens for elevated sources (cf. (2)). To model all these effects, we extend our database with compressed versions of the clean features in (5).

$$\mathcal{Y}_m(l,f,\varphi_q,i) = \frac{|H_m(l,f,\varphi_q)| + G_i|\bar{H}_m(l,f)|}{|H_0(l,f,\varphi_q)| + G_i|\bar{H}_0(l,f)|} \tag{6}$$

where $G_i$, $i = 0,\ldots,I-1$ is a SNR-dependent weighting factor and $\bar{H}_m(l,f)$ is the mean of $H_m(l,f,\varphi)$ over $\varphi$. Compared to actually measuring features in noisy conditions, the advantage of the noisy reference feature model in (6) is that the measurement effort and memory requirements can be reduced significantly.
By focusing only on a number of $N_p$ strong, deterministic frequency components $f_p$, we can increase the performance in noisy environments, while reducing the computational complexity in the following processing steps. We obtain the peak-frequencies $f_p$ by peak-picking in $|X_0(l,f)|$.

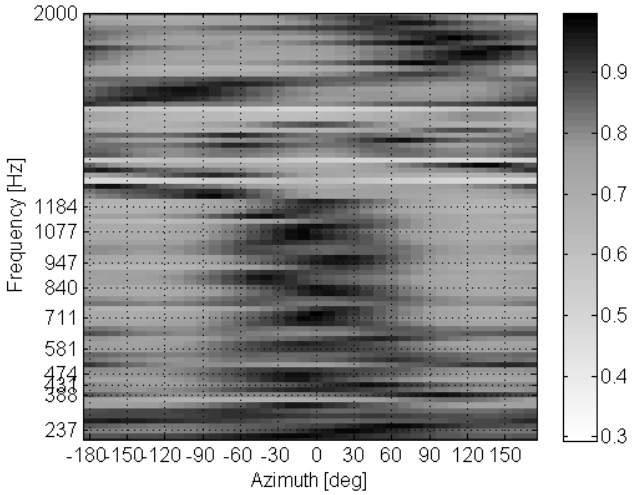



Figure 4: Similarity $C(f,\varphi_q)$. The source is a speech vowel signal located at $\varphi_s = 0^\circ$ mixed with omni-directional noise (6dB SNR). The peak-frequencies $f_p$ are indicated on the ordinate.

### 2.4. Similarity Matching

We use a similarity measure based on the Euclidean norm:

$$\mathrm{Sim}\left\{\boldsymbol{Y},\boldsymbol{\mathcal{Y}}\right\} = \frac{1}{1+\sqrt{\sum_{m=1}^{M-1}|Y_m-\mathcal{Y}_m|^2}} \tag{7}$$

Eq. (7) yields values bound between 0 (completely dissimilar) and 1 (vectors are the same). The similarity between the $l^{th}$ observed feature vector and the reference is computed for every reference position $\varphi_q$, peak frequency $f_p$ and SNR index $i$.

$$C(l,f_p,\varphi_q,i) = \mathrm{Sim}\left\{\boldsymbol{Y}(l,f_p),\boldsymbol{\mathcal{Y}}(l,f_p,\varphi_q,i)\right\} \tag{8}$$

Instead of simply searching for the global maximum, we propose the following procedure: First, we compute an index $i_{max}(l,f_p)$ that helps us to select the best matching SNR-version.

$$i_{max}(l,f_p) = \underset{i}{\mathrm{argmax}}\left\{\max_{\varphi_q}\left\{C(l,f_p,\varphi_q,i)\right\}\right\} \tag{9}$$

Then we average over frequency which yields a single SC:

$$C(l,\varphi_q) = \underset{f_p}{\mathrm{mean}}\left\{C(l,\varphi_q,f_p,i_{max}(l,f_p))\right\} \tag{10}$$

To illustrate why we focus only on strong frequency components $f_p$, an example of the frequency-dependent SC is shown in Fig. 4. At the peak-frequencies $f_p$, the SC peaks close to the true source angle. At other frequencies, where the source does not provide sufficient energy, noise prevails and there is an increased likelihood of having a flat SC without a clear peak or a peak at a wrong angle.

### 2.5. Reliability Filtering

The azimuth estimate could be computed directly from $C(l,\varphi_q)$ as follows:

$$\hat{\varphi}_s(l) = \mathrm{argmax}\{C(l,\varphi_q)\} \tag{11}$$

If the frame does however contain mainly background noise (e.g. in a speaking pause), the estimate is likely to be different from the

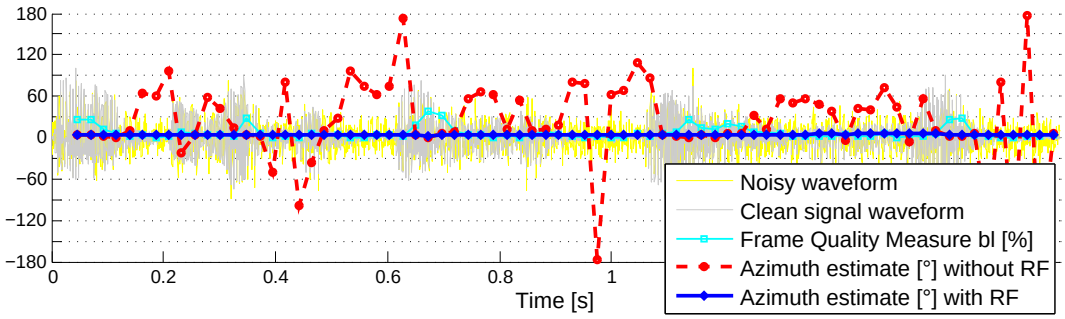



Figure 5: Effect of reliability filtering (RF) on a speech signal located at $0^\circ$ with added omni-directional pink noise (SNR=0dB). By rating the quality of each frame between 0 (unreliable) and 1 (reliable) a stable, correct localization result can be obtained. Because the SNR is low, the frame estimate without RF (red) looks erratic and random. For the same reason, the frame quality measure $b_l$ is close to zero most of the time. As can be seen from (12), the influence of such 'bad frames' is hence suppressed in the enhanced estimate (blue), i.e. we rely mainly on the previous estimate. Please note that at times where $b_l$ (cyan) is significantly higher than 0, the frame-estimator (red) delivers the true result. This is the basis for a correct enhanced estimate (blue).

actual source direction. Hence, we rate the "frame quality" and smooth the run of $C(l, \varphi_q)$ over time $l$. We use a 1-pole lowpass-filter with a time-varying coefficient $0 \leq b_l \leq 1$:

$$\tilde{C}(l, \varphi_q) = b_l \cdot C(l, \varphi_q) + (1 - b_l) \cdot \tilde{C}(l-1, \varphi_q) \qquad (12)$$

If $b_l = 0$ the SC is not updated, i.e. the previous SC is used. If $b_l = 1$ we rely only on the current frame and neglect the history. We compute $b_l$ by rating the shape of the SC with a sample variance like metric:

$$\tilde{b}_l = \frac{1}{Q-1} \sum_{q=0}^{Q-1} \left( C(l, \varphi_q) - \underset{\varphi_q}{\text{mean}} \{C(l, \varphi_q)\} \right)^2 \qquad (13)$$

A flat SC achieves a low value whereas a SC with a clear, single peak achieves a high value of $\tilde{b}_l$. To ensure that $b_l$ takes values close or equal to 1 under good conditions, we normalize and saturate $\tilde{b}_l$, i.e. $b_l = \max(\tilde{b}_l / \tilde{b}_{max})$, where $\tilde{b}_{max}$ is obtained from a recording under perfect conditions (single source, free-field, high SNR). With $\tilde{C}(l, \varphi_q)$ in (12), the enhanced position estimate is given as

$$\hat{\varphi}_s(l) = \underset{\varphi_q}{\text{argmax}} \{\tilde{C}(l, \varphi_q)\} \qquad (14)$$

Fig. 5 exemplifies the effect of reliability filtering (RF), i.e. the basic frame-level estimate in (11) is compared with the enhanced version in (14).

The azimuth estimate $\hat{\varphi}_s(l)$ in (14) is tied to the reference azimuth grid $\varphi_q$. To be able to produce results between the grid, interpolation between the maximum of the SC and its neighbors can be performed. We achieved good results with parabolic interpolation [10].

## 3. PRACTICAL EVALUATION

Recordings were carried out at the Institute of Electronic Music and Acoustics (IEM) in the "IEM-CUBE", an approximately 10 x 12 x 4 m large room usually used as a lab, for lectures and electro-acoustic music (reverberation time $RT_{60} \approx 0.7$ s ). The CUBE is equipped with an optical tracking system (OTS, a V624 data station and 15 M2 cameras by Vicon, cf. http://www.vicon.com) and a 24-channel hemispherical loudspeaker array (LSA).

As a sound source, we used 1) a loudspeaker and 2) a human speaker moving freely around the array. Both were tracked by the OTS for exact determination of the true source position (ground-truth) $\varphi_l$. The LSA was used for generation of omni-directional pink noise. To account for various SNRs, we added the appropriately weighted pink-noise recording to the clean target source recordings, i.e. the source recordings were made in quiet conditions (SNR between 25 and 45dB depending on the microphone, off/on-axis).

The reference database was obtained from impulse response (IR) measurements using a loudspeaker placed at different positions relative to the microphone array. To get smooth frequency responses and exclude noise and room reflections, these IRs were cut and windowed (to approx. 12ms) before transforming them to frequency domain. This makes the reference database more or less independent from the environment. We made experiments with a database recorded in a different room and achieved similar performance compared to a matched database.

The influence of diffuse reverberation is modeled via the noisy reference features in (6). It should however be noted that strong reflections from a dedicated direction may act as a competing source and can therefore impair the accuracy.

The placement of the array is not very critical because due to the normalization of the feature vector with the omni-directional channel, the characteristic pattern stays more or less the same. We compared placing the array on a desk with placement on the floor [10]. For the results shown in this paper, our array was placed on a small desk. We used a generic reference database (array placed on the floor, free-field) of our microphone with $10^\circ$ azimuth resolution. If the azimuth resolution is coarser the accuracy may be impaired due to imperfect interpolation.

We compared our similarity approach (SIM) with 1), a time- and 2), a frequency-domain intensity vector (IV) localization approach (TDIV, and FDIV, respectively). We used 512 samples long, hamming windowed frames with 50 % overlap at a samplerate of 11025 Hz. We considered $N_p = 10$ peak frequency components between 200 and 4000 Hz for the SIM.

For the TDIV, the energy of each channel is computed in time domain and transformed to IV components [4]. We used a smoothing pole $a = 0.9$ to average the IV over time to achieve better results.

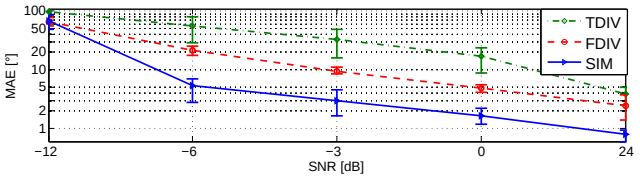



Figure 6: Static source position: Performance of different methods in terms of the mean (over $\varphi_s$) MAE. The errorbars indicate the first and the third quartile.

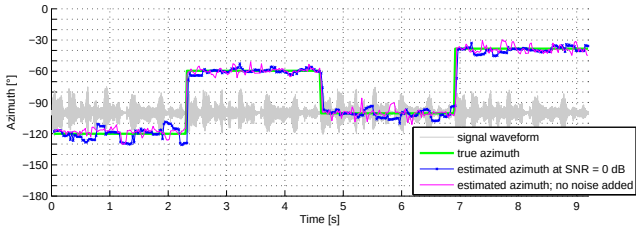



Figure 7: Source location jumps trough concatenation of different loudspeaker recordings (male speech). The estimate follows quickly.

The FDIV is based on a STFT and similar to the method described in [6]. Details on our implementation can be found in [10].

The estimation error is given as

$$\tilde{\varphi}_s(l) = \text{princarg}\left\{\hat{\varphi}_s(l) - \varphi_s(l)\right\} \quad , \tag{15}$$

where the principle argument function can be defined using the modulo operator mod: $\text{princarg}\,(\varphi) = \text{mod}\left\{\varphi + \pi, -2\pi\right\} + \pi$. As performance metrics, we used the mean absolute error MAE the root mean square error RMSE and the accuracy $\text{ACC}_\Delta$, where $\delta_\Delta(\tilde{\varphi}_s) = 1$, if $|\tilde{\varphi}_s| \leq \Delta$ and 0 otherwise.

$$\text{MAE} = \text{mean}\left\{|\tilde{\varphi}_s(l)|\right\} \tag{16}$$

$$\text{RMSE} = \sqrt{\text{mean}\left\{\tilde{\varphi}_s(l)^2\right\}} \tag{17}$$

$$\text{ACC}_\Delta = \frac{1}{L}\sum_{l=0}^{L-1}\delta_\Delta(\tilde{\varphi}_s(l)) \tag{18}$$

$\text{ACC}_5 = 90\%$ means for instance that $90\%$ of all frames achieve an estimation error $|\tilde{\varphi}_s(l)| \leq 5°$.

A short sentence (1.8s) of clean, male speech was played back from a loudspeaker positioned at 1m distance to the array. The elevation was $15°$ and the azimuth was varied between $-180°$ and $0°$ in steps of $10°$. A separate estimation result was computed for each angle. Fig. 6 shows the mean performance over all source angles in dependence of the SNR for the SIM,TDIV and FDIV. Our SIM-approach is very accurate and clearly superior to the TDIV and FDIV method. The exact values regarding the performance of our method are given in Table 1.
To assess the timing behavior of our algorithm, we concatenated the recordings from different azimuth angles, without pauses. Fig. 7 shows the true azimuth (optically tracked) and the estimate both for a recording with 0dB SNR and without added noise. Fig. 8 shows a similar plot, but for a male, human speaker walking around the array. More results can be found in [10].

| SNR | -12 | -6 | -3 | 0 | 3 | 6 | 12 | 24 |
|---|---|---|---|---|---|---|---|---|
| $ACC_5$ | 13 | 60 | 84.2 | 95.9 | 98.2 | 99.3 | 99.7 | 100 |
| $ACC_{10}$ | 18.9 | 85.6 | 97.3 | 100 | 100 | 100 | 100 | 100 |
| $ACC_{15}$ | 25.4 | 97.9 | 100 | 100 | 100 | 100 | 100 | 100 |
| $MAE$ | 68.5 | 5.3 | 3.0 | 1.6 | 1.4 | 1.2 | 1.0 | 0.8 |
| $RMSE$ | 84.9 | 6.0 | 3.4 | 2.1 | 1.8 | 1.4 | 1.3 | 1.0 |

Table 1: Performance of our SIM-approach with regard to static sources (Mean over $\varphi_s = (-180, -170, \ldots, 0)°$)).

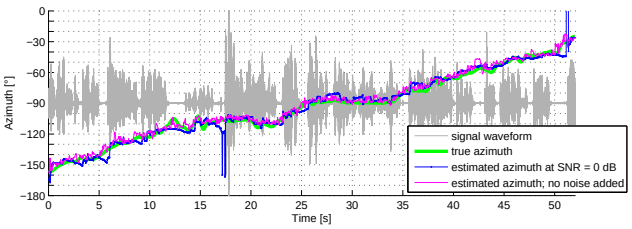



Figure 8: A male human speaks while walking around the array. The elevation was approx. $30° < \vartheta_s < 40°$, the radius $r_s \approx 1m$.

## 4. SUMMARY, CONCLUSION AND FUTURE WORK

We have presented a new method for tracking of a single sound source with a compact near-coincident microphone array (NCMA) that is cheap and handy. A reference feature database of the array has to be recorded once (free-field conditions, array placed on the floor). For source localization, a feature vector is computed frame-wise and compared to the database which yields a similarity curve (SC). We use a simple measure of the shape of the SC as a weight for the reliability of the current frame with respect to previous ones. With that, stable (no problems in signal pauses) and fast tracking can be achieved at the same time, without employing a separate detection algorithm. Conceptual advantages of our method are that we model the influence of noise and reverberation in our features and that we do not use fixed thresholds or time-constants. Practical experiments in a real room demonstrate the effectiveness of our approach, even in the presence of strong ($-6$ dB SNR) omni-directional noise.
The most obvious next working steps are a detailed study of the influence of microphone mismatch and evaluation of the performance in highly reverberant rooms. Our localization concept could be adapted to multi-source tracking and different array configurations, e.g. spherical arrays that also provide time-differences between the microphones. In contrast to our NCMA, such arrays allow for steering of higher order beam-patterns. Future work could also use the basic ideas behind our features and apply advanced pattern recognition approaches, e.g. a multi-class support vector machine.

## 5. REFERENCES

[1] C. Zhang, D. Florencio, D.E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 3, no. 3, pp. 538–548, 2008.

[2] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer, Berlin, 2008.

[3] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, chapter 8, pp. 157–180. Springer, Berlin, 2001.

[4] J. Merimaa and V. Pulkki, "Spatial impulse response rendering 1: Analysis and synthesis," *J. Audio Eng. Soc.*, vol. 53, no. 12, pp. 1115–1127, December 2005.

[5] B. Gunel, H. Hachabiboglu, and A.M. Kondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 4, pp. 154–157, May 2008.

[6] C. A. Dimoulas, K. A Avdelidis, G. M. Kalliris, and G. V. Papanikolaou, "Improved localization of sound sources using multi-band processing of ambisonic components," in *126nd AES Convention, Munich*, May 2009.

[7] O. Thiergart, R. Schultz-Amling, G. Del Galdo, D. Mahne, and F. Kuech, "Localization of sound sources in reverberant environments based on directional audio coding parameters," in *127th AES Convention, New York*, Oct. 2009.

[8] P. Cotterell, *On the Theory of the Second Order Soundfield Microphone*, Ph.D. thesis, University of Reading, 2002.

[9] M. Holters, T. Corbach, and U. Zölzer, "Impulse response measurement techniques and their applicability in the real world," in *Proc. of the* $12^{th}$ *Int. Conference on Digital Audio Effects (DAFx-09), Como, Italy*, September 1-4 2009.

[10] K. Freiberger, "Development and evaluation of source localization algorithms for coincident microphone arrays," M.S. thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria, 2010.