Low latency localization of multiple sound sources in reverberant environments

Marko Đurković,[a] Tim Habigt, Martin Rothbucher, and Klaus Diepold
Institute for Data Processing, Technische Universität München, Arcisstrasse 21, 80333 München, Germany
durkovic@tum.de, tim@tum.de, martin.rothbucher@tum.de, kldi@tum.de

Abstract: Sound source localization algorithms determine the physical position of a sound source in respect to a listener. For practical applications, a localization algorithm design has to take into account real world conditions like multiple active sources, reverberation, and noise. The application can impose additional constraints on the algorithm, e.g., a requirement for low latency. This work defines the most important constraints for practical applications, introduces an algorithm, which tries to fulfill all requirements as good as possible, and compares it to state-of-the-art sound source localization approaches.

PACS numbers: 43.60.Jn [JC]
Date Received: September 16, 2011 **Date Accepted:** October 6, 2011

1. Introduction

Hearing is an important ability of human beings, animals, and even technical systems. Even though humans rely heavily on their visual system, many events (e.g., the ringing of a doorbell) are purely acoustical in nature and cannot be detected by any other sense. In many cases, it is important to not only detect an acoustical event but also to know the location of the sound source. This may help to react to events that are not within the field of view of the observer. We are investigating algorithms, which localize multiple human speakers in a communication scenario. This localization step serves as a pre-processor for speech separation in a variety of applications like intelligent robotic systems or hearing aids. The scenario is subject to different constraints, which have to be handled by the localization algorithm.

One constraint is the physical size of the localization system, and in this work, we focus on lightweight mobile systems. In natural environments, rarely is only one source active, and a localization algorithm should therefore be able to localize multiple sources simultaneously. Additionally, the localization accuracy of an algorithm must not deteriorate significantly with reverberation because sound reflections and noise will be present at each microphone in realistic environments. Another constraint in many applications is that sound is processed in real time as it arrives, and therefore the computational complexity of an algorithm has to be low. Often a low latency is more important than this on-line processing capability because localization results are required immediately after a sound event occurs.

2. The compass algorithm

We developed COMPaSS (loCalization Of MultiPle Sound Sources) with all discussed constraints in mind, especially its usability in real environments and low latency. To enable a small physical size of the localization system, we chose a binaural approach where the hardware consists of only two microphones and two reflectors, which serve the same purpose as the pinnae of animals or humans. The reflectors have direction dependent transfer functions (TFs), which the algorithm uses to localize sources. The transfer functions have to be known *a priori* and are stored in a database of N_h

[a] Author to whom correspondence should be addressed.

direction dependent filters $h_{i,\eta}, i \in [1,2], \eta \in [1,...,N_h]$ where i denotes the microphone index and η denotes the TF index in the database. All source positions lie on a spatial sampling grid, and every index η is associated with one particular source direction.

2.1 Cross-correlation of the observations

COMPaSS assumes a realistic mixing model where each of the N active sound sources $s_n, n \in [1,...,N]$ is filtered with the filter pair $h_{i,\eta}, i \in [1,2], \eta \in [1,...,N_h]$ corresponding to the direction η of the sound source. We define the operator $p:\{1,...,N\} \rightarrow \{1,...,N_h\}, n \mapsto p(n)$, which maps a sound source index to the corresponding TF index defined by the sound source positions of an observed sound scene. The observations $x_1(t)$ and $x_2(t)$ for the left and right microphone are then given by

$$x_i(t) = \sum_{n=1}^{N} s_n(t) * h_{i,p(n),}\ i \in [1,2] \tag{1}$$

where $(*)$ denotes the convolution operator. Localization of sound sources is equivalent to the identification of the active filters in Eq. (1).

In the presence of only one source, the correct filter pair can be identified by inverse filtering or by a cross-convolution step.[1] This step convolves the observations with each filter pair from the database, yielding $y_{1,\eta}(t) = x_1(t) * h_{2,\eta}$ and $y_{2,\eta}(t) = x_2(t) * h_{1,\eta}$. The filter index η_0 indicating the correct position is retrieved by maximizing correlation between the cross-convolved signals

$$\eta_0 = \arg\max_{\eta} y_{1,\eta}(t) * y_{2,\eta}(t), \tag{2}$$

where $(*)$ denotes the cross-correlation operator. For more details, refer to MacDonald.[1]

COMPaSS extends the cross-convolution approach to multiple active sources. To this end COMPaSS assumes that the sound sources are sparse in some transform domain. Such sparsity assumptions have successfully been employed for techniques like underdetermined blind source separation.[2] In the presence of multiple sparse sources, the term W-disjoint orthogonality describes the fact that sparse signals can have disjoint sets of Fourier transform supports.[3]

If the source signals are W-disjoint in short time Fourier transform (STFT) domain, then at each time-frequency point there will be only one source active in the cross-convolved signals $y_{i,\eta}(t)$. Exploiting this fact, COMPaSS estimates for every possible filter the probability that it influenced the observations. The signals $y_{i,\eta}(t)$ are cut into N_k overlapping frames of length $L = 64$ ms with a shift of $L/2$ and each frame is stored as a vector $y_{i,\eta,k}$, where the subscript k indicates the frame number.

2.2 Similarity measurement

In the next step, the similarity between each pair of frames $\boldsymbol{y_{i,\eta,k}}, i \in [1,2]$ has to be calculated. For clarity, we will omit the η and k subscripts in this section.

COMPaSS transforms each frame into the STFT domain by subdividing it into N_s smaller overlapping frames. The subframes are transformed with an N_f-point discrete Fourier transform and stored as columns of the matrix $\boldsymbol{Y_i} \in \mathbb{C}^{N_f \times N_s}$. Each entry $\boldsymbol{Y_i}(f, l)$ of the matrix is the Fourier support of the fth frequency bin of the lth subframe.

Let $\boldsymbol{\gamma_{i,f}}$ denote the fth row of $\boldsymbol{Y_i}$. The similarity value $\boldsymbol{c}(f)$ is then calculated by

$$\boldsymbol{c}(f) = \left(\frac{\left| \boldsymbol{\gamma_{1,f}} \cdot \boldsymbol{\gamma_{2,f}^H} \right|}{\left\| \boldsymbol{\gamma_{1,f}} \right\|_2 \cdot \left\| \boldsymbol{\gamma_{2,f}} \right\|_2} \right)^2, \tag{3}$$

J. Acoust. Soc. Am. **130** (6), December 2011 Ðurković *et al.*: Localization of multiple sound sources EL393

which measures the linear dependence between two corresponding frequency bins over time. All entries of the vector $\boldsymbol{c}$ are in the interval [0,1] where a higher value indicates higher similarity at the corresponding frequency.

2.3 Filter scoring and extraction

COMPaSS stores the similarity values for each η and k of the K last frames as columns of the similarity matrix $\boldsymbol{C}_\eta \in \mathbb{R}^{N_f \times K}$. The entry $\boldsymbol{C}_\eta(f,k)$ is the similarity value achieved by filter pair $h_{i,\eta}$ for the kth frame at the fth frequency bin.

Using a winner-takes-it-all approach COMPaSS obtains the indices of the filters yielding the highest similarity values at one time-frequency point and stores them in the matrix $\boldsymbol{P} \in \mathbb{N}^{N_f \times K}$ with

$$\boldsymbol{P}(f,k) = \arg\max_{\eta} \boldsymbol{C}_\eta(f,k). \tag{4}$$

Let $\boldsymbol{B}_\eta \in \mathbb{N}^{N_f \times K}$ be a binary matrix indicating if filter η has the highest similarity in a specific bin and be obtained by

$$\boldsymbol{B}_\eta(f,k) = \begin{cases} 1 & \text{if} \boldsymbol{P}(f,k) = \eta \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

How much a time-frequency bin contributes to the final score of each filter is signal dependent. Chisaki *et al.*[4] suggest to give more importance to bins with higher signal energy because a higher SNR can be expected for those. COMPaSS uses a similar weighting of the frequency bins based on signal energy and the achieved similarity values. The filter- and signal-dependent weighting matrices $\boldsymbol{A}_\eta \in \mathbb{R}^{N_f \times K}$ are defined as

$$\boldsymbol{A}_\eta(f,k) = \left(\frac{|\boldsymbol{X}_\mathbf{1}(f,k)| + |\boldsymbol{X}_\mathbf{2}(f,k)|}{2}\right)^\alpha \cdot \left(\boldsymbol{C}_\eta(f,k)\right)^\beta, \tag{6}$$

where $\boldsymbol{X}_{\boldsymbol{i}}$ is the STFT representation of the observations $x_i(t)$. The parameters α and β set the influence of the signal energy and the similarity value on the final weight $\boldsymbol{A}_{\boldsymbol{\eta}}$. The choice of $\alpha = 1$ and $\beta = 1$ achieves good results in our experiments. If the signal energies of two active sources differ significantly, the influence of the energy term should be lowered with the parameter α. Finally, COMPaSS calculates the weighted histogram $\boldsymbol{p} \in \mathbb{R}^{N_h}$, the entries of which

$$\boldsymbol{p}(\eta) = tr\left(\boldsymbol{B}_\eta \cdot \boldsymbol{A}_\eta^T\right) \tag{7}$$

are proportional to the probability that a filter pair was active in the observations. The final stage of the localization algorithm extracts the most likely positions from the histogram iteratively. The histogram is modified in each iteration and is initialized to $\boldsymbol{p}_\mathbf{1} = \boldsymbol{p}$. The location of the nth source is extracted with

$$\tilde{\eta}_n = \arg\max_{\eta} \boldsymbol{p}_{\boldsymbol{n}}(\eta), \tag{8}$$

where $\tilde{\eta}_n$ denotes the index of the corresponding transfer function. Let the operator $\mathcal{D}(\eta_1,\ \eta_2)$ calculate the distance between the two source locations corresponding to η_1 and η_2. The histogram is then updated using the following rule

$$\boldsymbol{p}_{\boldsymbol{n}}(\eta) = \begin{cases} \boldsymbol{p}_{\boldsymbol{n}-1}(\eta) & \text{if } \mathcal{D}(\tilde{\eta}_{n-1}, \eta) > \delta_{\min} \\ 0 & \text{otherwise} \end{cases}, \tag{9}$$

where $\delta_{\min}$ enforces a minimal distance between two localized sources. By updating the histogram, the algorithm ensures that each location is extracted only once and sources with lower signal energy are not obscured.

3. Comparison

We compared COMPaSS to other state-of-the-art techniques which try to solve similar problems.

3.1 Compared algorithms

The first technique is sound source localization based on the self-splitting competitive learning (CSSCL) clustering technique.[5] This binaural technique was developed for robotic hearing systems, and according to Keyrouz *et al.*,[5] it has a lower complexity than comparable microphone array approaches. The technique was extended[6] to support more than two active sources.

The frequency domain binaural model (FDBM) localizes and separates sources exploiting interaural phase difference and interaural level difference in frequency domain. It was developed especially for speech sources and has been used as a front end for a speech recognition system.[7] Chisaki *et al.*[4] showed that FDBM is capable to localize two concurrent sound sources in azimuthal and elevation direction with high accuracy.

The third algorithm we compared is based on a combination of a steered response power beamformer using the phase transform and a particle filter (SRP-PHAT-PF)[8] and was designed explicitly for mobile robots. The particle filter approach solves the problem of assigning a localization result to the sources that are being observed and is also able to track moving sources. The system was tested in the presence of noise and reverberation and can run on-line on robot hardware.

3.2 Recordings and simulations

To compare the algorithms, we recorded a number of different sound scenes in an office room with dimensions 5.10 m × 3.49 m × 3.09 m (L × W × H) and a reverberation time RT_{60} of 0.64 s.

In this environment, we use a Knowles Electronic Manikin for Acoustic Research (KEMAR) and an array of eight microphones to record the test data. The microphones are arranged in the corners of a free standing cube with an edge length of 20 cm. Each sound scene is presented through loudspeakers to ensure invariant conditions. Due to spatial constraints, source locations were restricted to a semicircle. To cover more interesting sound scenarios, we used two different KEMAR orientations in the room. The setup of the experiments can be seen in Fig. 1.

For both KEMAR orientations, the TF database is measured at three elevations and 19 azimuths. The elevation planes are at $-10°$, $0°$, and $10°$. We recorded a total number of 945 sound scenes with a combined length of ≈1.3 h in the experiments.

In addition to the recordings, we also created exactly the same sound scenes with a simulator, which convolves the source signals with measured TFs according to Eq. (1). One interesting evaluation result will be the direct comparison between a recorded sound scene and its simulation.

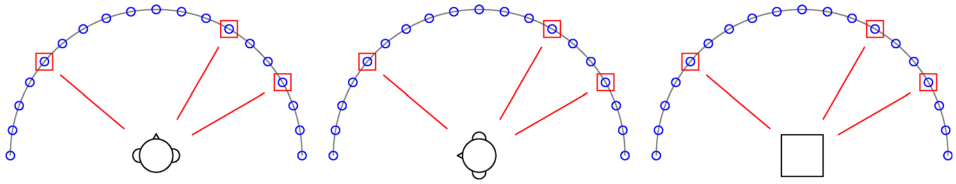

Fig. 1. (Color online) Top view of the experiment setup. The KEMAR's TFs are measured at three elevations and 19 azimuths on a circle with radius 1.3 m with a spatial grid resolution of 10°. Recorded scenarios consist of up to three active sources at different positions and all recordings are created with two KEMAR orientations and the microphone array.

3.3 Evaluation

Figure 2 shows the estimated angles of each compared algorithm over time for one three-source scenario. COMPaSS estimates the source positions almost 100% correctly. CSSCL misdetects a source at 0° instead of finding the real source at 40°. It is interesting to note that CSSCL does not return localization results for each frame but by design uses all 155 frames to calculate one result. The results of FDBM look much noisier than those of the other algorithms, but the estimated positions form visible clusters at the real source locations. SRP-PHAT-PF does not return any results until the 50th frame, which corresponds to an internal latency of ≈1.5 s caused by the particle filter. In contrast to the previous algorithms its results are continuous on the azimuth scale.

To evaluate the different localization algorithms, we use the localization success rate as a measure for the quality of the results. As we are using speech signals, our sources are not active in every frame, and we only evaluate the frames where the signal energy is above a threshold. The success rates are shown in Table 1. The columns denoted "exact" show the percentage of exact localizations. For the TF-based algorithms (COMPaSS, CSSCL, and FDBM), this is the number of correctly localized frames over the number of active frames. Due to the 10° spatial grid point distance, this measure has an implicit error tolerance of ±5°. The results of SRP-PHAT-PF are continuous, and therefore we consider them correct if they also lie within ±5° of the true position. The columns denoted "tolerance" show the localization success rate within a ±15° range. Additionally, the mean angular error (MAE) is given.

In the simulations, the TF-based algorithms show good numbers in the single source case and their performance drops as sources are added. In the three source case, COMPaSS loses only 3 percentage points, while CSSCL and FDBM lose over 25 percentage points. SRP-PHAT-PF has a lower exact accuracy than the TF-based algorithms, but its accuracy in the tolerance region is comparable. In the three source case, it surpasses FDBM and CSSCL. In the three source case, COMPaSS and the microphone array make only small mean angular errors (MAEs) of $\leq 3°$ and ≈7°, respectively. CSSCL and FDBM have an MAE of 20.28° and 28.93°, and therefore, estimated source positions will on average deviate significantly from the actual source locations.

The differences between the simulations and the real recordings are the presence of noise and possible deviations of the measured TFs. These deviations may arise due to changes in air pressure and room temperature and the finite length of the measured TFs. COMPaSS and SRP-PHAT-PF prove to be robust toward real conditions as their performance is not affected drastically. The accuracy of CSSCL and FDBM, however, disappoints in the real environment compared to the promising numbers in the simulations. In summary, COMPaSS and SRP-PHAT-PF are suitable candidates

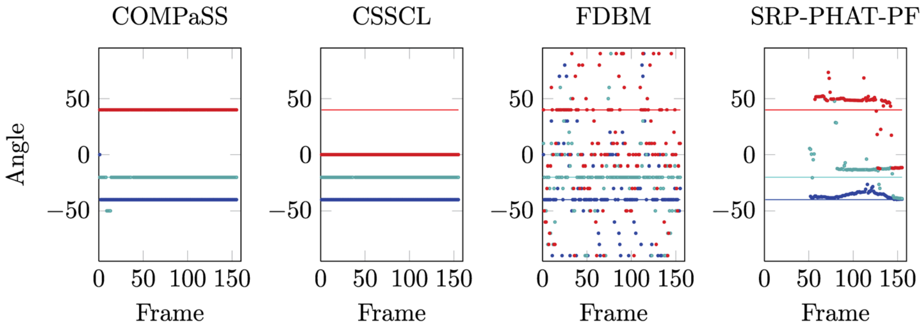

Fig. 2. (Color online) Each plot shows the detected azimuth angles of the three active sources at every time instance. The thin lines indicate the actual source position (−40°, −20°, and 40°) and the markers depict the respective algorithm's estimation.

Table 1. Localization results for identical sound scenes obtained by simulations and recordings in real environments. The percentage of correctly localized sources and the percentage of localizations lying in a 15° tolerance region are given. Additionally, the mean angular error (MAE) was measured.

	Simulations			Real recordings		
Algorithm	Exact	Tolerance	MAE (°)	Exact	Tolerance	MAE(°)
One source						
COMPaSS	1.00	1.00	0.00	1.00	1.00	0.00
CSSCL	1.00	1.00	0.00	0.72	0.72	18.25
FDBM	0.88	0.90	6.79	0.50	0.61	28.41
SRP-PHAT-PF	0.51	0.97	3.65	0.52	0.91	6.52
Two sources						
COMPaSS	0.97	0.97	2.17	0.93	0.94	5.64
CSSCL	0.90	0.90	8.13	0.42	0.46	44.57
FDBM	0.72	0.78	25.21	0.39	0.52	34.57
SRP-PHAT-PF	0.43	0.85	5.36	0.39	0.80	9.51
Three sources						
COMPaSS	0.97	0.97	2.84	0.84	0.86	11.85
CSSCL	0.71	0.73	20.28	0.35	0.41	43.82
FDBM	0.52	0.59	28.93	0.39	0.52	33.10
SRP-PHAT-PF	0.39	0.77	7.04	0.37	0.72	10.31

for practical use in real environments. Both show a high localization accuracy in the tolerance region paired with a small MAE. COMPaSS has an advantage of ≈10 percentage points in the tolerance region, and SRP-PHAT-PF has a slightly better MAE in the three-source case. If exact localization results are required, COMPaSS is the better choice as its accuracy is significantly higher in all cases.

4. Conclusions

In this paper, we presented COMPaSS, a low-complexity, low-latency multiple sound source localization technique for real environments. We compared it to three state of the art techniques that try to solve the same problem. COMPaSS is a TF-based approach and exploits the sparsity of sound sources. Our implementations of all four algorithms can process data on-line. COMPaSS and FDBM can calculate localization results for the very first frame, while SRP-PHAT-PF requires more sound information to produce reliable results. CSSCL is designed to process larger blocks of sound data and has the highest latency of all four algorithms. In a real world scenario with multiple sources, COMPaSS achieves the highest accuracy followed by SRP-PHAT-PF. CSSCL and FDBM perform significantly worse under these conditions.

Acknowledgments

This work is supported by the DFG excellence initiative research cluster Cognition for Technical Systems CoTeSys and the DFG collaborative research center SFB453 "High Fidelity Telepresence and Teleaction."

References and links

[1] J. A. Macdonald, "A localization algorithm based on head-related transfer functions," J. Acoust. Soc. Am. **123**, 4290–4296 (2008).

[2] R. Saab, O. Yilmaz, M. McKeown, and R. Abugharbieh, "Underdetermined anechoic blind source separation via lq basis-pursuit with q < 1," IEEE Trans. Signal Process. **55**, 4004–4017 (2007).

[3] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics Speech and Signal Processing* (2002), Vol. 1, pp. 529–532.

[4]Y. Chisaki, S. Kawano, K. Nagata, K. Matsuo, H. Nakashima, and T. Usagawa, "Azimuthal and elevation localization of two sound sources using interaural phase and level differences," Acoust. Sci. Technol. **29**, 139–148 (2008).

[5]F. Keyrouz, W. Maier, and K. Diepold, "Robotic localization and separation of concurrent sound sources using self-splitting competitive learning," in *IEEE Symposium on Computational Intelligence in Image and Signal Processing* (2007), pp. 340–345.

[6]F. Keyrouz, W. Maier, and K. Diepold, "Robotic binaural localization and separation of more than two concurrent sound sources," in *9th International Symposium on Signal Processing and Its Applications* (2007), pp. 1–4.

[7]Y. Chisaki, T. Nakanishi, H. Nakashima, and T. Usagawa, "Concurrent speech signal separation based on frequency domain binaural model," in *International Workshop on Acoustic Echo and Noise Control* (2003), pp. 255–258.

[8]J. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," Robot. Auton. Syst. **55**, 216–228 (2007).