LOCALIZATION OF ACOUSTIC SOURCES BASED ON THE TEAGER-KAISER ENERGY OPERATOR

Alexander Schasse and Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum
Universitätsstraße 150, 44780 Bochum, Germany
e-mail: {alexander.schasse, rainer.martin}@rub.de
web: www.rub.de/ika

ABSTRACT

This paper presents a new approach of microphone array sampling and processing for acoustic source localization. By sampling circular arrays in a round robin fashion, non-linear modulations are purposefully induced by means of the Doppler effect. The discrete time Teager-Kaiser Energy Operator is then used to analyze these modulations. It enables a batch-based localization of multiple sound sources at low complexity. The proposal system uses a small circular array but is suitable for other array geometries as well. In contrast to cross-correlation based localization techniques, we process only two signals while we maintain a circular symmetric system with no preferred look direction. Experiments are reported for up to 5 simultaneously active speech sources.

1. INTRODUCTION

The localization of multiple acoustic sources using a microphone array is an important field of current research. Its applications span from video conference systems over hearing aid processing to mobile communications. Common localization techniques in the case of single or multiple sound sources are adapted from beamforming, cross-correlation or subspace based methods and are reviewed for instance in [3, 6].

In this paper, we explore a new approach, which is based on a circular array sampling technique. The basic idea is to build a single signal by time-interleaved sampling of the microphone signals. In this way, we simulate a moving microphone and generate Doppler frequency shifts in the resulting signal. If the array geometry (e.g. linear or circular) is known, we can identify the source positions by analyzing the instantaneous frequency changes. In principle, arbitrary array geometries can be used, but in this paper we restrict the discussion to circular arrays to simplify the mathematical treatment. Our algorithm makes use of the Teager-Kaiser Energy Operator (TKEO), as introduced in [4]. Its most frequent application is a simultaneous amplitude and frequency demodulation [8], which has resulted in the DESA-algorithm for speech demodulation presented in [7]. This paper makes use of the TKEO itself to establish a batch-based algorithm for multiple source localization. In our context, the TKEO is more suitable since the DESA algorithm cannot handle fast changes in instantaneous frequency.

The remainder of this article is structured as follows. In Sec. 2, we introduce the proposed sampling technique and point out some fundamental characteristics. In Sec. 3 we present our approach for direction of arrival (DOA) estimation before simulation results are discussed in Sec. 4.

2. CIRCULAR SAMPLING

2.1 Geometric Analysis

For the following analysis, we relate this sampling technique to a single microphone, moving on a circular track, and sampled at distinct times, corresponding to the positions of the array sensors. Due to this movement and the Doppler effect the received signal \( x_D^{(n)}(t_n) \) shows periodic frequency shifts.

As a first step, we evaluate an expression for the instantaneous frequency of \( x_D^{(n)}(t_n) \) following the well known definition of the Doppler effect. For a single microphone, moving with velocity \( \text{vel} \) towards a sound source that emits a sinusoidal signal at frequency \( f_0 \), the frequency \( f_D \) of the received signal becomes

\[
f_D = f_0 \left( 1 + \frac{\text{vel}}{c} \right).
\]

Here, \( c \) is the speed of sound. At this point, we have to consider the circular sampling with regard to the array geometry as sketched in Fig. 1. Let \( (r_0, \phi_0) \) be the position of a single sound source in polar coordinates \((r, \phi)\) whose origin lies in the center of the microphone array. Then, assuming far field conditions, the wave propagation can be described by an unit vector in the form of

\[
e_{r_0} = \begin{pmatrix} \cos(\phi_0) \\ \sin(\phi_0) \end{pmatrix}.
\]

The velocity vector \( \mathbf{v}(t_n) \) that describes the speed of spatial sampling can be written as

\[
\mathbf{v}(t_n, \phi_0) = \begin{pmatrix} -\sin(c \cdot f_0 t_n + \phi_0) \\ \cos(c \cdot f_0 t_n + \phi_0) \end{pmatrix}
\]
As an alternative to (1), we can describe the circular sampling by a polyphase network as shown in Fig. 2, using the discrete Fourier transform (DFT). In this section, we look at the localization of sound sources using the proposed sampling technique. If the nonlinear effects were sufficiently small, one would have to calculate the instantaneous frequencies of all Q FM-signal shifted to a frequency $f_M$. However, a direct demodulation will not work due to the phase shift of the microphone signals. In addition to (6), the microphone signals should be limited to a bandwidth of $f_M / M$. These results become important when considering the frequency dependency of the proposed time domain algorithm. However, based on the closed form of $X_D^{(i)}(z)$ in (9), a frequency domain model and processing of the circular sampled signal can be established, too.

### 3. DOA ESTIMATION

In this section, we look at the localization of sound sources using the proposed sampling technique. If the nonlinearities were sufficiently small, one would have to calculate the instantaneous frequencies of all Q FM-signals mixed in $x_D^{(i)}(n)$. However, a direct demodulation will not work due to the different and, in particular, wideband carrier signals (source signals, e.g. speech, noise, etc). We present a time-domain framework based on the TKEO, which is defined in [4] for a discrete signal $x(t_n)$ as

$$\psi(x(t_n)) = x^2(t_n) - x(t_{n-1})x(t_{n+1}).$$

$\psi(x(t_n))$ is an estimate of the instantaneous energy of $x(t_n)$, and is approximately equal to the squared product of the signal amplitude and frequency. While the TKEO [4] allows for very efficient implementations its main drawbacks are, first, the center frequency of the processed signal $x(n)$ is limited to...
Depending on the number of microphones, the estimation error shows peaks for some frequencies (e.g. at 2 or 4 kHz). We made the observation that at these frequencies the localization becomes less distinctive and shows a noisy characteristic. We trace this effect back to the spatial sampling, since the samples used to build $x_D^{(n)}(t_n)$ do not cover the whole range of values of all $x_m(t_n)$.

Based on this analysis, we limit all microphone signals to the frequency band (b) by bandpass filtering before building the DTKEO signal, see Fig. 4. The widely deterministic shape of the localization results in Fig. 3 could also allow a wideband approach for source localization. In this work, however, we restrict ourselves to lower frequencies ($\Omega < \frac{\pi}{4}$).

Figure 4: Block diagram of the circular array signal processing for source localization.

3.2 Multiple Source Localization

In order to localize multiple sources, we apply a framewise processing to the $M$ microphone signals. For each time frame, we realize the source localization algorithm and collect the estimated values $\hat{\varphi}_q$. After a sufficient number of frames, the probability density function of $\tilde{\varphi}_q$ is estimated by using a histogram $\hat{p}(\tilde{\varphi}_q)$. If the sources are partially separated in time, then a sufficient number of valid estimations is obtained and the source positions can be detected as peaks in $\hat{p}(\tilde{\varphi}_q)$. To localize the sources automatically, a clustering (k-means or similar) or a probability mixture model could be applied. In this paper, we use the estimated distributions $\hat{p}(\tilde{\varphi}_q)$ as final results.

To improve the localization, we divide the frequency band of interest (labeled by (b) in Fig. 3) into $K$ subbands $\text{BP}_k$, $k = 1, \ldots, K$. This is based on two reasons. First, speech of different persons is sparse in the time-frequency-domain, and second, we obtain more estimations which makes the statistics more stable. For this step, we could also apply a short time DFT instead of conventional bandpass filtering. The localization results of all frequency bands and all time frames are aggregated into a single histogram.
3.3 Computational Complexity

The computational complexity of the presented algorithm is obviously small. In each time frame, the following simple operations have to be applied:

- Bandpass filtering of the sensor signals, which can be accomplished in frequency domain or via FIR-filtering.
- Build $x_D(t_n)$ and $x_D^2(t_n)$ by rearranging the sensor signal samples. This can be done efficiently, by calculating permutation matrices beforehand or via the polyphase representation in (8).
- The number of operations needed to calculate the TKEO is comparable to an FIR-filter with 3 coefficients.
- Due to the fact that only two signals are processed after the bandpass filtering and circular sampling, we reduce the number of processed samples by the factor $1/M$.
- The last step is then to estimate $\hat{\phi}_q$ by calculating the correlation in (13).

Since the computational complexity is very low, we can use small time frame shifts to get as many values for $\hat{\phi}_q$ as possible to improve the localization of multiple sources. Note that in contrast to cross-correlation based localization methods we process only two signals while we maintain a circularly symmetric system with no preferred look direction.

4. EXPERIMENTAL RESULTS

4.1 Simulation Setup

To evaluate the proposed approach, we define the positions of $Q = 5$ candidate point sources $q_j$ in polar coordinates at $r_{q_j} = 5m, \forall j = 0, \ldots, Q - 1$ and $\varphi_{q_0} = 0, \varphi_{q_1} = \pi / 3, \varphi_{q_2} = -3\pi / 8, \varphi_{q_3} = -\pi / 2$ and $\varphi_{q_4} = -3\pi / 8$. The circular array consists of $M = 4$ omnidirectional and equally spaced microphones at radius $r_0 = 2cm$ (solid line in Fig. 3) and $\varphi_m = \pi / 2$. The speech samples that we use as source signals $x_{q_0}(t_n), x_{q_1}(t_n)$ and $x_{q_4}(t_n)$, are shown in Fig. 7 (a) to (c), the remaining source signals $x_{q_3}(t_n)$ and $x_{q_4}(t_n)$ have a similar pattern of speech activity. The signals are sampled at $f_s = 16kHz$ and quantized with 16 bit in wav format. For processing, we use time frames of 256 samples (16ms) and a frame advance of 64 samples. The bandpass filters BP in Fig. 4 have a bandwidth of 500Hz with rising center frequencies (in steps of 500Hz) to extract the frequency range of 250 to 2250Hz.

4.2 Localization Results

We use the full length of the speech samples (8s) to collect the statistic. The results in form of $\hat{p}(\hat{\phi}_q)$ are shown in Fig. 6 an increasing number of active sources. The labels of the selected sources $q_j$ are written on the top of the plot. For $Q = 1$ (lowermost), we see a supergaussian distribution with a sharp maximum at the position of $\varphi_0$. For $Q > 1$, the variance increases, therefore the peaks become less distinctive (note the different axis labels of the single plots). We observe that the peaks match the true source positions fairly well (dashed red lines), with highest accuracy for $\hat{\phi}_q = \varphi_m = \pi / 2, i = 0, 1, 2, 3$. Despite their small azimuthal distance, we see that even the sources $q_1$ and $q_4$ can be separated if all $Q = 5$ speech sources are simultaneously active.

Fig. 7 (a) - (c) depicts the spectrograms of the source signals $x_{q_0}(t_n), x_{q_1}(t_n)$ and $x_{q_4}(t_n)$ and Fig. 7 (d) shows the framewise results of source localization when these three sources are simultaneously active. We use again time frames of 256 samples and a frame advance of 64 samples. The histogram is calculated in steps of 128 ms (29 frames) to perform a short time localization. Now we can point out some characteristics of the algorithm. At first, it is obvious that the localization gives random results if no source signal is active (3s to 5s). At the beginning and end of this pause, only one source is active and the localization works properly. If all three sources interfere with each other, the distinctiveness of the peaks depend on the power of the single sources in the analyzed frequency bands in each time frame.
sources in far-field conditions. We use small circular microphone arrays that are sampled in a round robin fashion to create nonlinear distortions in the sense of frequency modulation. The azimuth source position can be estimated by analyzing these nonlinearities with the discrete time Teager-Kaiser Energy Operator. In order to localize multiple sound sources simultaneously, a batch based processing is presented, which collects data for a distinct time to estimate the distribution of azimuth source positions. The localization algorithm has been validated with up to 5 speech sources. It can be extended to other array geometries.

REFERENCES


4.3 Comparison with Basic Source Localization Techniques

For the circular array of $M = 4$ microphones and radius $r_0 = 2$cm, other source localization techniques [3, 6] deliver equivalent or inferior results, compared to the presented approach. E.g. beamforming approaches like SRP or SRP-PHAT [1] are not suited for such small array dimensions. Estimating the time differences of arrival (TDOAs) using a cross-correlation [5] of the microphone signals would require a high subsample precision, while, for the proposed algorithm, we work at a reduced sampling rate. Comparable results to the presented localization technique are obtained using the MUSIC algorithm [9] as shown in Fig. 7 (e). Setting the temporal resolution to 128 ms, the MUSIC spectrum is less noisy, but the sources are located with much more variance over time, especially for $q_0$ and $q_1$. Moreover, subspace methods like MUSIC fail, if the number of active sources is larger than the number of microphones, i.e. $Q > M$, as at the top of Fig. 6.

5. CONCLUSION

We have presented an array sampling and processing technique which enables the localization of acoustic sound