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
email: {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, nonlinear 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. It's 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

In the following, we examine a circular microphone array with radius r_0 that contains M sensors and assume anechoic conditions. The microphone signals $x_{\mathrm{m}_i}(t_n)$, with $i = 0,\ldots,M-1$, are a mixture of Q source signals $x_{\mathrm{q}_j}(t_n)$, with $j = 0,\ldots,Q-1$. $t_n = {}^{n}/_{f_\mathrm{s}}$ is the discrete time, n the sample index ($n = 0,\ldots,N-1$) and f_s the sampling frequency. Now, circular sampling entails that we form a signal $x_\mathrm{D}(t_n)$ in the following manner: The first sample of $x_\mathrm{D}(t_n)$ is taken from the first microphone, the second one from the second microphone, and so on. To allow arbitrary starting positions of the round robin sampling, we introduce the superscript $^{(i_0)}$, with $i_0 \in \{0,\ldots,M-1\}$, which results in

$$\begin{aligned} x_\mathrm{D}^{(i_0)}(t_n) &= x_{\mathrm{m}_{i(n)}}(t_n) \ , \text{with} \\ i(n) &= (n+i_0)\mathrm{mod}M. \end{aligned} \tag{1}$$

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 $x_\mathrm{D}^{(i_0)}(t_n)$ shows periodic frequency shifts.

As a first step, we evaluate an expression for the instantaneous frequency of $x_\mathrm{D}^{(i_0)}(t_n)$ following the well known definition of the Doppler effect. For a single microphone, moving with velocity v_eff towards a sound source that emits a sinusoidal signal at frequency f_0, the frequency f_D of the received signal becomes

$$f_\mathrm{D} = f_0\left(1+\frac{v_\mathrm{eff}}{c}\right). \tag{2}$$

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 $(r_{\mathrm{q}_0},\varphi_{\mathrm{q}_0})$ be the position of a single sound source in polar coordinates (r,φ) 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

$$\mathbf{e}_{\mathrm{q}_0} = -\begin{pmatrix} \cos(\varphi_{\mathrm{q}_0}) \\ \sin(\varphi_{\mathrm{q}_0}) \end{pmatrix}. \tag{3}$$

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

$$\mathbf{v}(t_n,\varphi_0) = \begin{pmatrix} -\sin(c_\varphi t_n+\varphi_0) \\ \cos(c_\varphi t_n+\varphi_0) \end{pmatrix} \tag{4}$$

 ISSN 2076-1465

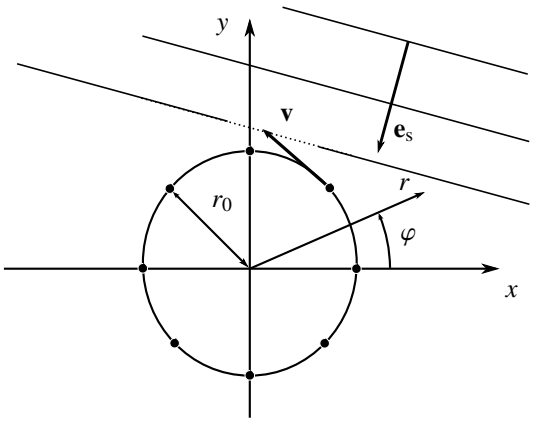

Figure 1: Measurement geometry - $\mathbf{v}$ describes the sampling direction, $\mathbf{e}_s$ is the normalized vector of the incoming wavefronts (far-field assumption).

where $c_\varphi = 2\pi f_s/M$ is the angular velocity, and $\varphi_0 = 2\pi\frac{i_0}{M}$ is the azimuth position from where the sampling starts (e.g. in Fig. 1 $\varphi_0 \in [0, \pi/4, \pi/2 \ldots 7\pi/4]$). Using the inner product $\langle\cdot,\cdot\rangle$, we calculate the effective velocity component by $v_{\text{eff}} = -\left\langle \mathbf{e}_{q_0}, \mathbf{v}(t_n, \varphi_0)\right\rangle$ and obtain an expression for the instantaneous frequency of $x_D^{(i_0)}(t_n)$

$$f_D^{(i_0)}(t_n) = f_0\left(1 - \frac{2\pi r_0}{Mc} f_s \sin\left(c_\varphi t_n + \varphi_0 - \varphi_{q_0}\right)\right) \tag{5}$$

using a far field approximation ($r_0 \ll r_{q_0}$). In (5), φ_{q_0} is the azimuth position of the sound source, i.e. the desired direction of arrival (DOA) of impinging waves.

Additionally, we can define a sampling theorem in the sense of the Nyquist condition

$$f_s > \frac{2f_0}{1 - 2\frac{2\pi r_0}{Mc} f_0} > 2f_0. \tag{6}$$

In relation to $2f_0$, the microphone signals need to be somewhat oversampled to acquire $x_D^{(i_0)}(t_n)$. For example, for $M = 4$, $r_0 = 2\text{cm}$, $c = 340\frac{\text{m}}{\text{s}}$ and $f_0 = 1\text{kHz}$, this results in $f_s > 2.5\text{kHz}$. As a first result, we see that the circular sampling is a transformation of the individual microphone signals that evokes nonlinear effects. The source signals become the carriers of a mixture of frequency modulated (FM) signals, whose instantaneous frequencies depend on the different DOAs φ_{q_j}.

2.2 Polyphase Representation

As an alternative to (1), we can describe the circular sampling by a polyphase network as shown in Fig. 2, using the discrete sampling function

$$w_M(n-\lambda) = \sum_{\nu=0}^{M-1} W_M^{\nu(n-\lambda)} = \begin{cases} 1, & n = mM + \lambda \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

with the Mth complex root of unity, $W_M = \sqrt[M]{1} = \mathrm{e}^{-j\frac{2\pi}{M}}$. The circularly sampled signal $x_D^{(i_0)}$ can now be expressed by single polyphase components of the microphone signals $x_{m_i}(n)$

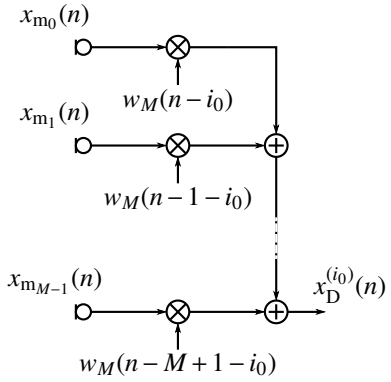

Figure 2: Circular sampling by means of a polyphase network.

following

$$x_D^{(i_0)}(n) = \sum_{i=0}^{M-1} x_{m_i}(n) w_M(n-i-i_0). \tag{8}$$

Based on the theory of polyphase networks [10, 2], the spectrum of $x_D^{(i_0)}(n)$ is given by

$$X_D^{(i_0)}(z) = \frac{1}{M}\sum_{i=0}^{M-1}\sum_{k=0}^{M-1} X_{m_i}(zW_M^k) W_M^{(i+i_0)k}. \tag{9}$$

Due to the phase shifted subsampling of the microphone signals in (8), the spectrum of the circular sampled signal is composed by the modulated spectra $X_{m_i}(zW_M^k)$ of $x_{m_i}(n)$ with a microphone dependent phase shift $W_M^{(i+i_0)k}$. Hence, we have the usual limitations concerning aliasing due to the subsampling of the microphone signals. In addition to (6), the microphone signals should be limited to a bandwidth of f_s/M. These results become important, when regarding the frequency dependency of the proposed time domain algorithm. However, based on the closed form of $X_D^{(i_0)}(z)$ 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 $x_D^{(i_0)}(t_n)$. 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

$$\psi(x(t_n)) = x^2(t_n) - x(t_{n-1})x(t_{n+1}). \tag{10}$$

$\psi(x(t_n))$ is an estimate of the instantaneous energy of $x(t_n)$, and is approximately equal to the squared product of the signals amplitude and frequency. While the TKEO [4] allows very efficient implementations its main drawbacks are, first, the center frequency of the processed signal $x(n)$ is limited to

$|\Omega| \leqslant \frac{\pi}{4}$ and, second, the more the amplitude and frequency differ from a constant, the larger is the estimation error.

However, applying the TKEO on $x_{\mathrm{D}}^{(i_0)}(t_n)$ delivers information about the changes in instantaneous frequency and, with it, a possibility for DOA estimation. This means that we need to find the changes in instantaneous frequency due to the sampling, and for this purpose, we need some reference. At this point, it is advantageous to compare the TKEO of two circularly sampled signals and calculate a differential TKEO signal (DTKEO)

$$\psi_{\mathrm{diff}}^{(i_1,i_2)}(t_n) = \psi(x_{\mathrm{D}}^{(i_1)}(t_n)) - \psi(x_{\mathrm{D}}^{(i_2)}(t_n)). \tag{11}$$

We assume identical envelopes of the circularly sampled signals, so the DTKEO signal is proportional to the difference of the squared instantaneous frequency changes of $x_{\mathrm{D}}^{(i_1)}(t_n)$ and $x_{\mathrm{D}}^{(i_2)}(t_n)$ as defined in (5). For a single sound source with its azimuth φ_{q_0} that emits a pure tone at frequency f_0, we get

$$\begin{aligned}\psi_{\mathrm{diff}}^{(i_1,i_2)}(t_n) \propto f_0^2 \Big[& 4\frac{r_0 c_\varphi}{c} \sin\Big(c_\varphi t_n - \varphi_{\mathrm{q}_0} + \frac{\pi(i_1+i_2)}{M}\Big) \\ & \sin\Big(\frac{\pi(i_2-i_1)}{M}\Big) - 2\Big(\frac{r_0 c_\varphi}{c}\Big)^2 \sin\Big(2c_\varphi t_n - 2\varphi_{\mathrm{q}_0} + \frac{2\pi(i_1+i_2)}{M}\Big) \\ & \sin\Big(\frac{2\pi(i_2-i_1)}{M}\Big)\Big]. \end{aligned} \tag{12}$$

By choosing $i_1 = 0$ and $i_2 = M/2$ with appropriate M, we eliminate the second term in (12) and a single amplified harmonic remains at frequency f_s/M ($c_\varphi = 2\pi f_s/M$). We can now estimate the desired DOA φ_{q_0} by recovering the phase of this harmonic where $\angle$ is the angle of a complex number

$$\widehat{\varphi}_{\mathrm{q}_0} = \frac{\pi}{2} - \angle\Big\{\sum_{n=0}^{N-1} \psi_{\mathrm{diff}}^{(0,\frac{M}{2})}(t_n) \mathrm{e}^{-J\frac{2\pi n}{M}}\Big\}. \tag{13}$$

3.1 Frequency Dependency

One important point is the influence of the signal frequency on the proposed localization technique. For this purpose, we take a look at a single sound source ($\varphi_{\mathrm{q}_0} = 0$), emitting a sinusoidal wave with varying frequency f_0. The DOA estimation error for two circular array geometries is shown in Fig. 3. We point out some principle characteristics in (a) to (d):

(a) For low frequencies, the localization does not work due to very small phase differences between the microphone signals.

(b) There is a frequency band, where the localization is exact. Its bandwidth depends principally on the sampling rate. This region matches the usable range of the TKEO as mentioned in [4].

(c) For frequencies above $\Omega = \frac{\pi}{4} \mathrel{\hat{=}} \frac{f_s}{8}$, the curves show specific deterministic shapes, depending on the number of microphones, array dimension and sampling frequency. Note that spatial aliasing occurs at 6 kHz for the four-microphone array, and at 3.7 kHz for the eight-microphone array as shown in Fig. 3.

(d) 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 $x_{\mathrm{D}}^{(i_0)}(t_n)$ do not cover the whole range of values of all $x_{\mathrm{m}_i}(t_n)$.

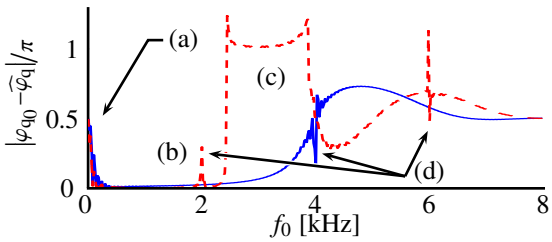

Figure 3: Illustration of the frequency dependent behavior of the source localization using the TKEO for different array dimensions at f_s = 16kHz (solid blue line: $M = 4$, r_0 = 2cm; dashed red line: $M = 8$, r_0 = 6cm).

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 ($\leqslant \frac{f_s}{8}$).

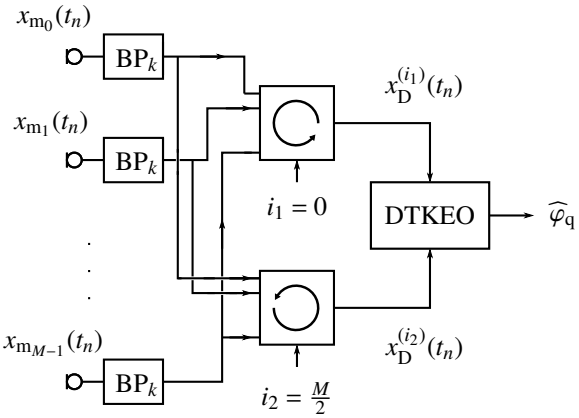

Figure 4: Block diagram of the circular array signal processing for source localization.

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 $\widehat{\varphi}_{\mathrm{q}}$. After a sufficient number of frames, the probability density function of $\widehat{\varphi}_{\mathrm{q}}$ is estimated by using a histogram $\widehat{p}(\widehat{\varphi}_{\mathrm{q}})$. 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 $\widehat{p}(\widehat{\varphi}_{\mathrm{q}})$. 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 $\widehat{p}(\widehat{\varphi}_{\mathrm{q}})$ as final results.

To improve the localization, we divide the frequency band of interest (labeled by (b) in Fig. 3) into K subbands 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 $x_{\mathrm{D}}^{(i_1)}(t_n)$ and $x_{\mathrm{D}}^{(i_2)}(t_n)$ 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 $2/M$.
- The last step is then to estimate $\widehat{\varphi}_{\mathrm{q}}$ 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 $\widehat{\varphi}_{\mathrm{q}}$ 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 $r_{\mathrm{q}_j} = 5\mathrm{m}, \forall\, j = 0,\ldots,Q-1$ and $\varphi_{\mathrm{q}_0} = 0$, $\varphi_{\mathrm{q}_1} = 2\pi/3$, $\varphi_{\mathrm{q}_2} = -3\pi/4$, $\varphi_{\mathrm{q}_3} = -\pi/2$ and $\varphi_{\mathrm{q}_4} = -3\pi/8$. The circular array consists of $M = 4$ omnidirectional and equally spaced microphones at radius $r_0 = 2\mathrm{cm}$ (solid line in Fig. 3) and $\varphi_{\mathrm{m}_i} = i\pi/2$. The speech samples that we use as source signals $x_{\mathrm{q}_0}(t_n)$, $x_{\mathrm{q}_1}(t_n)$ and $x_{\mathrm{q}_2}(t_n)$, are shown in Fig. 7 (a) to (c), the remaining source signals $x_{\mathrm{q}_3}(t_n)$ and $x_{\mathrm{q}_4}(t_n)$ have a similar pattern of speech activity.The signals are sampled at $f_{\mathrm{s}} = 16\mathrm{kHz}$ 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_k 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.

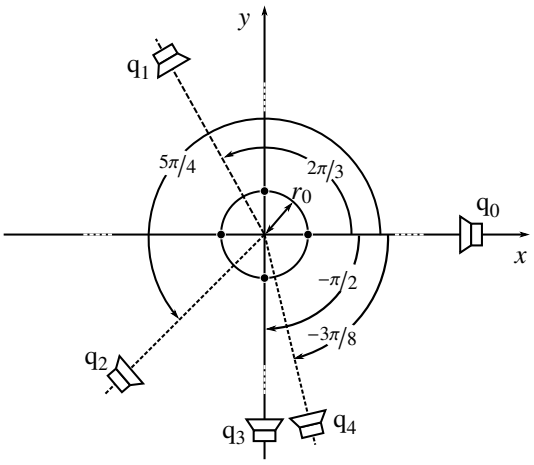

Figure 5: Simulation setup with up to $Q = 5$ sources q_0 to q_4. The circular array contains four equispaced sensors at a radius of $r_0 = 2\mathrm{cm}$.

4.2 Localization Results

We use the full length of the speech samples (8s) to collect the statistic. The results in form of $\widehat{p}(\widehat{\varphi}_{\mathrm{q}})$ 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 q_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 $\widehat{\varphi}_{\mathrm{q}} \approx \varphi_{\mathrm{m}_i} = i\pi/2, i = 0,1,2,3$. Despite their small azimuthal distance, we see that even the sources q_3 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 $x_{\mathrm{q}_0}(t_n)$, $x_{\mathrm{q}_1}(t_n)$ and $x_{\mathrm{q}_2}(t_n)$ 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.

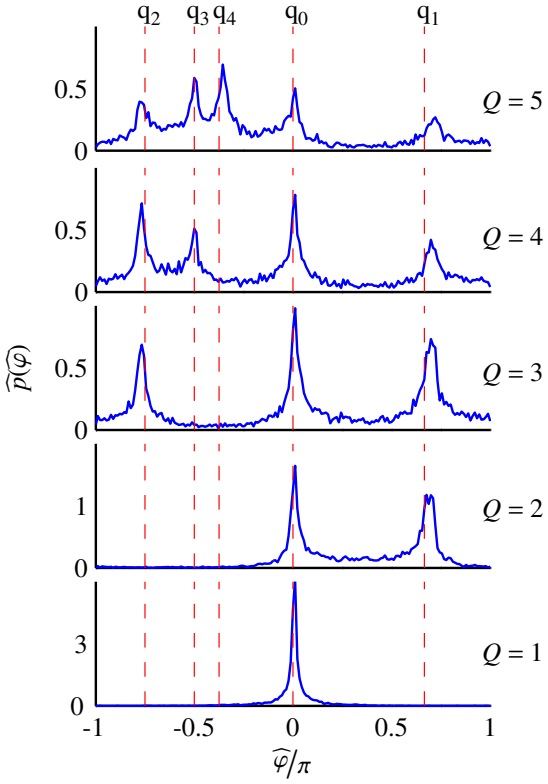

Figure 6: Estimated distributions of $\widehat{\varphi}_{\mathrm{q}}$ for rising number of active sources.

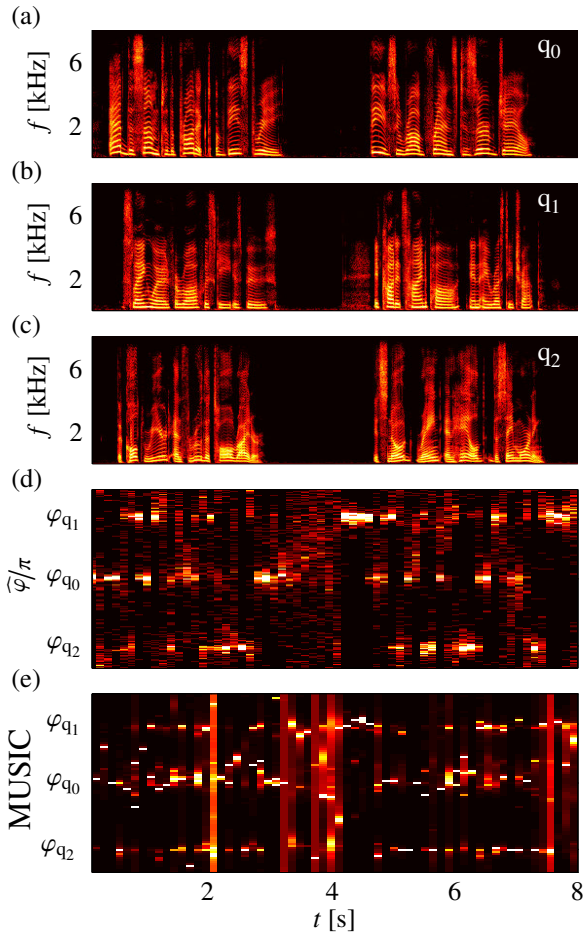

Figure 7: (d) shows the localization of three active sources ((a) at $\varphi_{q_0} = 0$, (b) at $\varphi_{q_1} = 2\pi/3$ and (c) at $\varphi_{q_2} = -3\pi/4$), using time frames of 128ms to collect statistics. (e) shows the results of the MUSIC algorithm for the same time resolution.

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 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

[1] J. DiBiase, H. Silverman, and M. Brandstein. Robust source localization in reverberant rooms. In M. Brandstein and D. Ward (eds.), *Microphone Arrays: Techniques and Applications*, pages 157–180. Springer-Verlag, 2001.

[2] H. G. Göckler and A. Groth. *Multiratensysteme*. Schlembach, 2004.

[3] Y. Huang, J. Benesty, and J. Chen. Time delay estimation and source localization. In J. Benesty, M. M. Soudhi, and Y. Huang (eds.), *Springer Handbook of Speech Processing*. Springer, 2007.

[4] J. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *Acoustics, Speech, and Signal Processing, International Conference on*, 1:381–384, Apr. 1990.

[5] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(4):320–327, January 2003.

[6] N. Madhu and R. Martin. Acoustic source localization with microphone arrays. In R. Martin, U. Heute, and C. Antweiler (eds.), *Advances in Digital Speech Transmission*. John Wiley, 2008.

[7] P. Maragos, J. Kaiser, and T. Quatieri. Energy separation in signal modulations with application to speech analysis. *Signal Processing, IEEE Transactions on*, 41(10):3024–3051, Oct 1993.

[8] P. Maragos, J. Kaiser, and T. Quatieri. On amplitude and frequency demodulation using energy operators. *Signal Processing, IEEE Transactions on*, 41(4):1532–1550, Apr 1993.

[9] R. Schmidt. Multiple emitter location and signal parameter estimation. *Antennas and Propagation, IEEE Transactions on*, 34(3):276–280, 1986.

[10] P. P. Vaidyanathan. *Multirate systems and filter banks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.