THIS ARTICLE HAS BEEN ACCEPTED FOR PUBLICATION IN IEEE/ACM TRANS. AUDIO, SPEECH, LANG. PROCESS., DOI: 10.1109/TASLP.2016.2556282

Localization and Characterization of Multiple Harmonic Sources

Hannes Pessentheiner, *Student Member, IEEE,* Martin Hagmüller, *Member, IEEE,* and Gernot Kubin, *Member, IEEE*

***Abstract*—We introduce a new and intuitive algorithm to characterize and localize multiple harmonic sources intersecting in the spatial and frequency domains. It jointly estimates their fundamental frequencies, their respective amplitudes, and their directions of arrival based on an intelligent non-parametric signal representation. To obtain these parameters, we first apply variable-scale sampling on unbiased cross-correlation functions between pairs of microphone signals to generate a joint parameter space. Then, we employ a multidimensional maxima detector to represent the parameters in a sparse joint parameter space. In comparison to others, our algorithm solves the issue of pitch-period doubling when using cross-correlation functions, it estimates multiple harmonic sources with a signal power smaller than the signal power of the dominant harmonic source, and it associates the estimated parameters to their corresponding sources in a multidimensional sparse joint parameter space, which can be directly fed into a tracker. We tested our algorithm and three others on synthetic data and speech data recorded in a real reverberant environment and evaluated their performance by employing the joint recall measure, the root-mean-square error, and the cumulative distribution function of fundamental frequencies and directions of arrival. The evaluations show promising results: Our algorithm outperforms the others in terms of the joint recall measure, and it can achieve root-mean-square errors of one Hertz or one degree and smaller, which facilitates, e.g., distant-speech enhancement or source separation.**

***Index Terms*—Data association, direction of arrival (DOA), fundamental frequency, joint estimation, microphone array, pitch estimation, pitch-period doubling, sparse joint parameter space (SJPS).**

I. Introduction

In different fields of research (e.g., computational auditory [1], [2] or acoustic [3] scene analysis), signal parameters often need to be associated with their origin, e.g., a signal-emitting source. To describe an acoustic scene [4]–[6], we need to detect, localize, characterize, separate, and interpret these sources [7], [8]. To localize and characterize such sources, we jointly estimate multiple parameters to avoid data association requiring additional algorithms. Joint parameter spaces are a major issue in distant-speech enhancement during a meeting or in separating instruments of an orchestral recording. The larger the difference of the sources' parameters, the better the separator's performance. For example, if we separate two sources represented in a spectrogram, we still need to find independent parameters to succeed. Unfortunately, they are rare and hard to find. However, assuming the fundamental frequency (f_0) as the parameter of our choice, the overlap of a female and a male speaker's f_0s is small due to anatomical reasons [9]. What if two speakers with similar f_0 talk simultaneously? Then, we face crossings in the time-frequency domain that reduce the separator's performance. Moreover, they introduce uncertainty about the f_0s' association with the correct source. Another challenging problem is crossings followed by a discontinuous change of intonation. A tracker would need to decide which f_0 corresponds to which source—an ambiguous problem without a distinct solution. But when we extend this lower-dimensional problem to a higher-dimensional one by considering the direction of arrival (DOA), then we reduce the number of simultaneous crossings of both f_0 and DOA to a minimum or zero. This will increase the separator's ability to associate the f_0s to their origin.

Manuscript received September 24, 2015; revised February 8, 2016 and April 9, 2016; accepted April 17, 2016. The K-Project ASD is supported in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFW, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria, and The Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG). The DIRHA-Project was supported by the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement No. FP7-ICT-2011-7-288121. Furthermore, this project has received funding from the Marshall Plan Foundation. The Tesla K40 GPU-cards used for this research were donated by the NVIDIA Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Augusto Sarti. The authors are with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c/EG, 8010 Graz, Austria, e-mail: hannes.pessentheiner@tugraz.at; hagmueller@tugraz.at; g.kubin@ieee.org. Digital Object Identifier: 10.1109/TASLP.2016.2556282

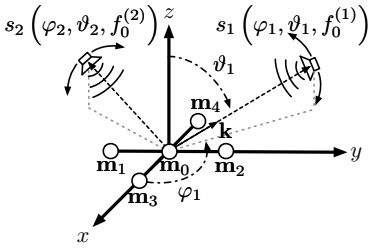


Fig. 1. Three-dimensional coordinate system with two moving harmonic sources, $s_1(\varphi_1, \vartheta_1, f_0^{(1)})$ and $s_2(\varphi_2, \vartheta_2, f_0^{(2)})$. Variables f_0, φ, and ϑ denote the fundamental frequency, the azimuth, and the elevation of a source, respectively; $\mathbf{k}$ is the spherical unit vector, and $\mathbf{m}_i$ denotes a microphone.

A. Problem Definition

In this article, we address the problem of jointly estimating the fundamental frequencies f_0, their respective amplitudes, and DOAs of moving and non-moving harmonic sources (see Fig. 1) by utilizing an intelligent non-parametric signal representation; hence, we bypass employing an explicit statistical estimator. We define the signal measured at the i-th microphone as

$$x_i[n_t] = \sum_{p=1}^{N_s} \mathcal{H}_i^{(p)}\{s_p[n_t]\} + \sum_{q=1}^{N_r} \mathcal{H}_i^{(q)}\{r_q[n_t]\} \qquad (1)$$

for $i = 1, ..., N_m$ microphones, where $s_p[n_t]$ denotes a harmonic signal and $r_q[n_t]$ represents an interfering noise source, which does not correlate with the harmonic sources. N_s is the number of the harmonic sources, and N_r is the number of the interfering noise signals. The system operator $\mathcal{H}$ denotes a source's spatialization in real reverberant conditions (in case of real-data experiments) and free-field conditions (in case of synthetic-data experiments). For instance, to model

a source's movement in free field, we consider the impulse response $h_i[n_t] = \delta\left[n_t - \tau_i[n_t]\right]$, where the time-delay of arrival (TDOA), τ_i, changes for varying sample indices n_t. For a harmonic signal $s[n_t]$ sampled at frequency f_s, which sweeps its instantaneous fundamental frequency from f_1 to f_2 within T_2 seconds and which consists of N_q harmonics with amplitudes α_k, we write

$$s[n_t] = \sum_{k=1}^{N_q} \alpha_k \cos\left(\frac{2\pi k f_1 T_2}{\ln(f_2/f_1)}\left[\left\{\frac{f_2}{f_1}\right\}^{\frac{n_t}{T_2 f_s}} - 1\right]\right) \quad (2)$$

for $0 \leq n_t \leq T_2 f_s$, where n_t is the sample index. We omitted index p for simplicity. These practically relevant signals are reasonable models for continuous changes of f_0 in voiced speech or glissandi played by a violinist during a concert. Throughout the article, the DOA is composed of the tuple (φ, ϑ), where φ is the azimuth and ϑ is the elevation.

B. Problem Solution

To jointly estimate DOAs, f_0s, and the respective amplitudes of harmonic sources, we first apply a bandpass filter bank to the signals captured by an array's microphones. Then we compute the cross-correlation function (CCF) of frames of the filtered signals and sample them by applying variable-scale sampling. (We do not sum any CCFs.) This allows to set up a joint parameter space (JPS), which we sparsify by employing a multidimensional maxima detector yielding a sparse joint parameter space (SJPS). The SJPS contains jointly estimated DOAs, f_0s, and the respective amplitudes of harmonic sources. To determine the algorithm's accuracy, we employ the root-mean-square error, the joint recall measure, and the cumulative distribution function of f_0s and DOAs.

C. Different Approaches

In the past two decades, several research teams developed approaches to jointly detect or estimate DOAs and f_0s. Based on their publications, we distinguish between two groups.

The first group represents pioneering approaches in source localization that consider both source parameters to estimate and represent them separately. For instance, [10] and [11] presented a robust method for speech-signal time-delay estimation in reverberant environments based on estimating f_0. A different groundbreaking approach for binaural signals is based on a multi-pitch tracking [12]. Other original approaches are based on time delay and frequency estimation of multiple sinusoids [13] for multiple speaker localization [14].

The second group consists of approaches that jointly estimate and represent parameters in a JPS. In this field, extensive research has been done, e.g., [15]–[18]. One of the most recent algorithms is based on a broadband minimum-variance distortionless beamformer [19]. The authors applied the algorithm to clean speech signals distorted by a non-moving interfering harmonic signal with five harmonics, white noise, and reverberation simulated by using the image method. Another remarkable contribution are the nonlinear least squares methods for joint DOA and f_0 estimation [20]. In [21], they jointly estimated the TDOA and f_0 of multiple sources by using the alternating direction of multipliers method (ADMM) optimization procedure. Another notable way to jointly estimate and represent f_0 and interaural time-difference is to apply extended recurrent timing neural networks [22].

In this article, we introduce a subclass of the second group which is composed of our algorithm that jointly estimates and represents both parameters in an SJPS. We sparsified the JPS by employing a multidimensional maxima detector, which enables us to estimate multiple harmonic sources. In [23], the authors suggested the idea of joint estimation and representation in an SJPS obtained by sampling a CCF. It is the cornerstone of the algorithm presented in this article.

What do all reported studies have in common? They did not explain how to solve the data association problem while estimating or detecting the parameters, and they did not sparsify the parameter space. Many approaches rely on estimating the global extremum of a cost function (which means that they are able to detect or estimate a single source only) or they rely on adaptive filters. Furthermore, most of them focused on testing their approaches on signals from musical instruments. All these issues led us to an innovative real-time capable solution for multiple sources based on [23] and tested on synthetic data and speech data recorded in a real reverberant environment.

D. The Predecessors' Roadmap

Before presenting our new algorithm, we summarize its direct predecessors to highlight the changes over time. Képesi et al. [23] introduced the idea of jointly estimating and representing both parameters in an SJPS in 2007 by means of extracting certain features from a biased CCF using two microphones only. Until 2013 several studies extended this idea. However, after reviewing their studies and conducting a vast number of experiments, we came to the conclusion that several modifications reported in [24]–[32] were unfavorable for our problem scenario. For instance, they estimated DOAs only and did not exploit their algorithms' (hidden) abilities to estimate f_0s. They considered broadband analysis in case of multiple sources which yielded accurate directional information but erroneous temporal information, as shown in our experiments. They analyzed summed CCFs, which introduced pitch-period doubling. They employed biased CCFs yielding estimates with varying amplitudes for signals whose sinusoidal components exhibit the same amplitude. The application of gammatone bandpass filters caused distorted estimates due to varying gains within a band and a missing group delay compensation. They did not consider a sparse representation of their estimates, which could have been directly fed into, e.g., a tracker. They also considered spectro-temporal fragments analysis [12], [33] and combined their existing algorithm with a spectro-temporal pre-processing module yielding a dramatic increase in computational costs.

E. Contributions

The proposed algorithm is based on [23] and inspired by [24]–[32], but it sparsifies a (quasi-continuous) JPS and estimates parameters of harmonic sources even with a signal power smaller than the signal power of the dominant harmonic source by employing filter banks and variable-scale sampling of CCFs. In comparison to all other approaches, we do not sum any CCFs. Using unbiased CCFs, considering bandpass filters that feature a manageable flat passband, processing each band separately, doing narrowband analysis, and representing the estimates in a sparse joint parameter space, we were able to solve all the problems mentioned above. By considering an intelligent non-parametric signal representation, we no

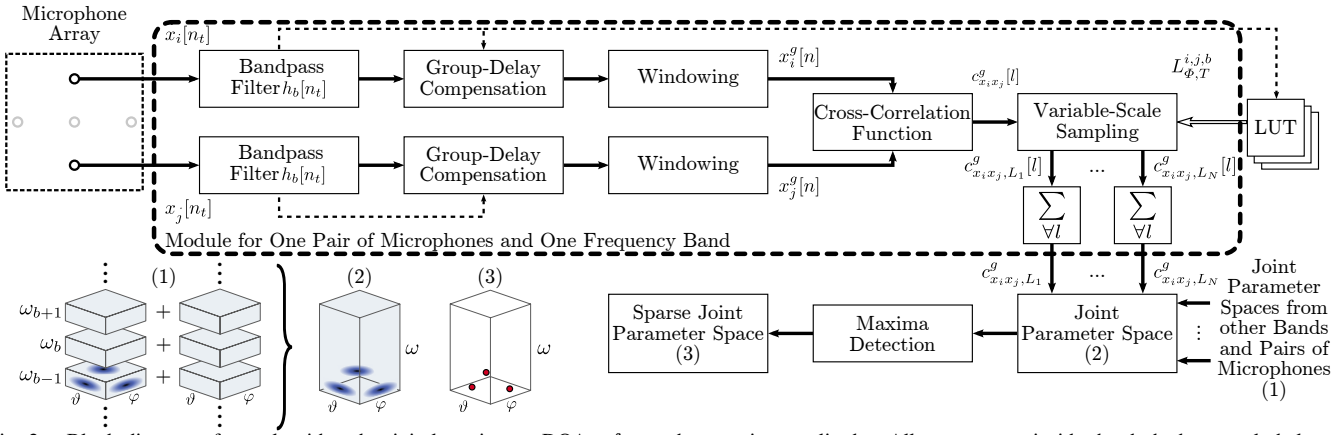


Fig. 2. Block diagram of our algorithm that jointly estimates DOAs, f_0s, and respective amplitudes. All components inside the dashed rectangle belong to a module for one pair of microphones and one frequency band. The number of modules depends on the number of available pairs of microphones and the number of frequency bands. The components labeled with 'Windowing' split the discrete-time signals $x_i[n_t]$ and $x_j[n_t]$ from microphones with index i and j into frames; n_t is the sample index of the whole captured signal and n is the sample index of a windowed signal. Variable $h_b[n_t]$ is the impulse response of the b-th bandpass filter, $c^g_{x_ix_j}[l]$ is the CCF of $x_i^g[n]$ and $x_j^g[n]$ with lag index l and frame index g, $c^g_{x_ix_j,L}[l]$ is the CCF sampled with a certain sampling period and sampling phase, $c^g_{x_ix_j,L}$ is the sampled CCF summed over all lags, $L^{(i,j,b)}_{\Phi,T}$ denotes the subset of sampling periods and sampling phases for the b-th band and microphones i and j, φ is the azimuth, ϑ denotes the elevation, and f represents the frequencies. The lookup table (LUT) contains all relevant indices for variable-scale sampling of a CCF.

longer require an explicit estimator. In comparison to [10]–[32], the new algorithm represents the signal parameters in an SJPS over time consisting of relevant information only, which can be directly fed into a tracker. Beyond that, we conducted a vast number of simulations with simultaneously active, synthetically generated sources featuring non-stationary harmonic signals causing intersections in the spatial and frequency domains. To conduct simulations with recorded speech signals from male and female speakers in a real reverberant environment, we compiled a unique speech corpus [34], [35]. In contrast with [10]–[32], we determined the ground truth for the recorded signals' instantaneous f_0s by analyzing their corresponding recorded glottograms, which enabled us to use a large variety of naturally produced fluent speech.

II. Proposed Algorithm

Fig. 2 shows the block diagram of our new algorithm. We will discuss its components in the following sub-/sections.

A. Parameters of Interest

The DOA [36] is the spatial angle of incidence [37]. It is the spatialization of the relative TDOA of a propagating monochromatic plane wave observed at two different locations. The fundamental frequency (f_0) [2] is the inverse of the fundamental period (T_0) of harmonic sources. It is the component of a harmonic structure exhibiting the lowest frequency. We do not refer to this as pitch, because pitch is a perceptual attribute and not a physical quantity [38], [39]. By using f_0 as a parameter, we can improve the performance of a speech separating system [40] and efficiently employ subband beamforming. Besides, it improves parameter estimation when two or more speakers share the same DOA [15], [31], [41].

B. Microphone Array

To sample the acoustic wave field at specific positions in space [36], we employ an array made up of omnidirectional microphones. In case of a linear or planar array, we recommend to mount it on the enclosure, e.g., the ceiling or the walls, of a room to reduce or avoid spatial ambiguity [37]. To jointly estimate the DOA and the f_0 using our algorithm, the array's maximum dimension d has to be large enough to decrease its omnidirectional behavior at lower frequencies but short enough so that the assumption of plane wave propagation remains valid [42] and no spatial aliasing occurs [37], [42]. According to [43], the minimum distance A between a source and the uniform linear array's center has to be

$$A(\gamma, \lambda, d) = d^2 \sin(\gamma)^2/(2\lambda) + d\,|\cos(\gamma)|/2 - \lambda/8, \quad (3)$$

where d is the array's maximum dimension, γ is the angle of incidence, and λ is the wavelength of interest. By assuming $0 \leq \gamma < 2\pi$, $\{\lambda | \lambda = v/f, 80 \leq f \leq 1000\}$, with f as the frequency of interest in Hz and v as the speed of sound in m/s, we can determine the minimum distance $\hat{A}$, which ensures plane wave propagation in any direction, according to

$$\hat{A} = \max_{\gamma,\lambda} A(\gamma, \lambda) \quad (4)$$

for a given d. In case of a uniform circular array (UCA), we assume d as the array's diameter.

C. Filter Bank

We use a filter bank to solve the multi-source problem by considering narrow-band analysis of multiple sources and to avoid pitch-period doubling [44], [45] when using the CCF [46], to reduce the influence of noise and narrow-band interfering sources, or to extract signal components within several frequency ranges. Even in case of acoustic beating caused by two superimposed signals with almost the same f_0 [47], narrow-band filters limit this effect to a small frequency range when using a CCF.

1) Bandpass Filter: We employ Kaiser window order-estimated bandpass filters [48]–[51] with predefined lower and upper cut-off frequencies. They exhibit impulse responses with decreasing lengths for higher bands and constant group delays, and we can attain reduced passband ripple and steep passband-stopband transitions with a manageable order. Common alternatives are, e.g., the Gammatone filter [52], the Butterworth filter [53], and the Cauer filter [53]. They feature a non-constant group delay that has to be compensated by phase reversed filtering. However, by using (anit-)symmetric FIR filters, we just need to properly delay the filtered signals.

2) Group Delay Compensation: We compensate the constant group delay to compare the estimated f_0 with its ground truth value and to provide time-synchronous f_0-estimates. In our case, we delay each bandpass signal by

$$\Delta N_h^{(b)} = (N_h^{(b)} - 1)/2 \tag{5}$$

samples, where $N_h^{(b)}$ is the number of samples of the b-th bandpass filter's impulse response $h_b[n]$. Hence, we introduce a delay for each band: $\overline{h_b}[n] = \delta[n - \Delta N_h^{(b)}]$.

3) Bandwidth: To avoid pitch-period doubling, we split the frequency range of interest into N_b bands with equal bandwidth smaller or equal than $\Delta f = f_l/2$ with f_l as the lowest fundamental frequency of interest (which is 75 Hz in our case). The number of bands is

$$N_b < (f_u - f_l)/\Delta f, \tag{6}$$

where f_u is the highest cut-off frequency.

D. Unbiased Cross-Correlation Function

The CCF [50],

$$c_{x_i x_j}[l] = \sum_{m=-\infty}^{+\infty} x_i[m] x_j^*[l+m], \tag{7}$$

is a function of time lag l, where $(\cdot)^*$ denotes a complex conjugation. To determine DOA and f_0, we calculate $c_{x_i x_j}[l]$ between $x_i[n]$ and $x_j[n]$, each with length N_x and $0 \leq n \leq N_x - 1$, for $-N_x + 1 \leq l \leq N_x - 1$. It represents the sampled wave field captured by two microphones with indices i and j over time. To speed up computations, we calculate the cross spectrum according to

$$c_{x_i x_j}[l] = \mathcal{F}^{-1}\{C_{x_i x_j}[k]\}, \tag{8}$$

where $\mathcal{F}^{-1}$ is the inverse discrete-time Fourier transform and $C_{x_i x_j}[k]$ is the cross spectrum of $X_i[k] = \mathcal{F}\{x_i[n]\}$ and $X_j[k] = \mathcal{F}\{x_j[n]\}$. The windowed CCF is

$$c_{x_i x_j}[l] = \begin{cases} w[l] \sum\limits_{m=0}^{N_x-1-l} x_i[m] x_j^*[m+l] & l \geq 0 \\ w[-l] \sum\limits_{m=0}^{N_x-1+l} x_j[m] x_i^*[m-l] & l < 0 \end{cases}, \tag{9}$$

where n is the time shift and N_w denotes the window's length. Considering the inverse window

$$w[l] = \begin{cases} \dfrac{1}{N_x - |l|} & -N_x + 1 \leq l \leq N_x - 1 \\ 0 & \text{else} \end{cases}, \tag{10}$$

to reduce the decrease in amplitude (for $|l| > 0$) yields the unbiased CCF. We compute the unbiased CCF frame-wise over time with a frame size of 0.032 s and an overlap of 0.010 s.

E. Sampling Phase and Sampling Period

The two major parameters to sample the CCF are the sampling phase and the sampling period.

The sampling phase $L_\Phi(\varphi, \vartheta)$ is an extrinsic parameter that is related to a source's location and its TDOA,

$$\tau_{i,j}(\varphi, \vartheta) = -(\mathbf{m}_i - \mathbf{m}_j)^T \mathbf{k}(\varphi, \vartheta)/v, \tag{11}$$

with $\mathbf{k}(\varphi, \vartheta) = (\sin(\vartheta)\cos(\varphi), \sin(\vartheta)\sin(\varphi), \cos(\vartheta))^T$ as the spherical unit vector, φ and ϑ as the azimuth and elevation of a source, $\mathbf{m}_i$ and $\mathbf{m}_j$ as the i-th and j-th microphone coordinates, and v is the speed of sound. To sample a CCF, we transform the TDOA into the sampling phase according to

$$L_\Phi^{(i,j)}(\varphi, \vartheta) = \lfloor \tau_{i,j}(\varphi, \vartheta)/T_s \rceil, \tag{12}$$

where $T_s = f_s^{-1}$ and $T_s \in \mathbb{R}$ and $\lfloor \cdot \rceil$ rounds its argument to the nearest integer to avoid fractional delays.

The sampling period $L_T(T_0)$ is an intrinsic parameter that is related to a source's f_0. We define it according to

$$L_T(T_0) = \lfloor T_0/T_s \rceil, \tag{13}$$

where $T_0 \in \mathbb{R}$. Considering a low sampling frequency and a big array size, errors caused by spatial aliasing, imperfectly optimized bandpass filters, and a decreasing frequency resolution for higher frequencies would predominate. Relative to those, rounding errors turn out to be negligible.

To localize and characterize one or more sources, we calculate sampling periods and sampling phases for all f_0s and directions of interest. We define the subset of sampling phases and sampling periods for the b-th band and a pair of microphones consisting of microphone i and j as

$$L_{\Phi,T}^{(i,j,b)} \subset \left(L_\Phi^{(i,j)}, L_T^{(b)} \right), \tag{14}$$

$$L_T^{(b)} = \{ L_T(T_0) \mid T_0 = f_0^{-1}, b f_l \leq f_0 \leq (2b+1) f_l/2 \}, \tag{15}$$

$$L_\Phi^{(i,j)} = \{ L_\Phi^{(i,j)}(\varphi, \vartheta) \mid 0 \leq \varphi < 360, 0 \leq \vartheta \leq 180 \}. \tag{16}$$

This yields $N_b \cdot N_g$ subsets, where N_g is the number of pairs of microphones. A single tuple of an arbitrary sampling period and sampling phase is defined as $L \triangleq \left(L_\Phi^{(i,j)}(\varphi, \vartheta), L_T^{(b)}(T_0) \right)$. A time-consuming step is the computation of all $N_b \cdot N_g$ subsets. Thus, we calculate them in advance and store them in a lookup table (LUT).

F. Variable-Scale Sampling of Cross-Correlation Function

The sampling of each CCF enables us to jointly estimate DOAs, f_0s, and the respective amplitudes of one or more harmonic sources. We sample the CCF with a limited number of sampling points at specific lags. Therefore, we define a discrete sampling function known as the Shah function [54],

$$\text{III}[l] = \sum_{m=-N_d}^{N_d} \delta \left[l - \left(m L_T^{(b)}(T_0) + L_\Phi^{(i,j)}(\varphi, \vartheta) \right) \right], \tag{17}$$

where $\delta[\cdot]$ is the Kronecker delta. The number of sampling points is $2N_d + 1$, where N_d is a small integer. We sample the CCF according to

$$\hat{c}_{x_i x_j}[l] = c_{x_i x_j}[l] \cdot \text{III}[l]. \tag{18}$$

Inserting (17) into (18) and summing over all lags l yields

$$\hat{c}_{x_i x_j} = \frac{1}{2N_d + 1} \sum_{\forall l} c_{x_i x_j}[l] \times \sum_{m=-N_d}^{N_d} \delta \left[l - \left(m L_T^{(b)}(T_0) + L_\Phi^{(i,j)}(\varphi, \vartheta) \right) \right], \tag{19}$$

for an arbitrary L. Now, we can construct a 3-tuple $(L_\Phi^{(i,j)}(\varphi, \vartheta), L_T^{(b)}(T_0), \hat{c}_{x_i x_j})$ that represents a point in a 4-dimensional parameter space (we can distinguish between sampling phases of different φ and ϑ). We compute the CCF for each pair of microphones' band and for lags l distributed symmetrically around $l = 0$. In order to justify the use of the

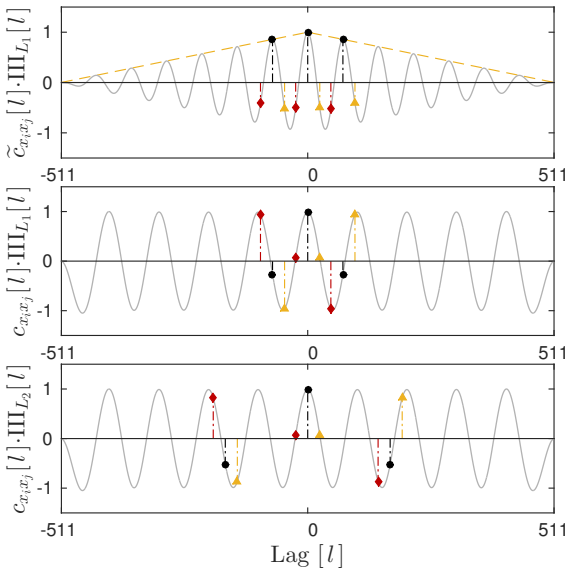


Fig. 3. Variable-scale sampling of the biased (top) and unbiased (middle, bottom) CCF by applying the Shah function with five ($N_d = 3$) and three ($N_d = 1$) sampling points, respectively, different sampling phases (red, black, yellow) and sampling periods based on the following frequencies: 7 Hz (top, middle) and 3 Hz (bottom). The yellow dashed lines (top) represent the decrease of amplitudes in case of a biased CCF.

unbiased CCF, we first analyze a variable-scale sampled biased CCF of a periodic signal. If we compute the biased CCF of a certain frequency band (see Fig. 3 top), do the variable-scale sampling, and sum over all lags, we can estimate the frequency components. However, if we would do this for a periodic signal with a lower frequency component, there would be—compared to the previous case—a remarkable difference in amplitudes. The resulting amplitude of the higher frequency component is larger than the amplitude of the lower one. However, the sampled amplitudes should be identical. By using the unbiased CCF, we overcome this problem. As shown in Fig. 3 (middle, bottom), the peaks of each unbiased CCF around $l = 0$ are almost identical due to the weighting described in (10).

G. Joint Parameter Space

The JPS is a joint representation of sampling periods, sampling phases, and respective amplitudes over time. Due to the joint estimation, these signal parameters are associated with each other. Fig. 4 shows a three dimensional JPS representing 3-tuples $(L_{\Phi}^{(i,j)}(\varphi, \vartheta = 90°), L_T^{(b)}(T_0), \hat{c}_{x_i x_j})$. We set up a JPS for each pair of microphones and add them together. For our purposes, the parameter space still contains irrelevant information. However, we are interested in tuples, i.e., points in the JPS, representing local maxima. Therefore, we sparsify this space (see Fig. 4) by employing an efficient multidimensional maxima detector to obtain a sparse representation of it, i.e., an SJPS, as shown in Fig. 5.

H. Multidimensional Maxima Detector

To detect local maxima in the JPS, we apply a real-time capable multidimensional maxima detector based on Lemire's streaming maximum-minimum filter [55], [56]. The detector sparsifies the JPS, which contains the associated parameters of interest for multiple sources, and is based on a sliding, hypercubic window. If the window size is too small, the detector might detect fluctuations caused by, e.g., noise, which would introduce undesirable local maxima. If the window size is too large, the detector might fail in detecting multiple sources, whose parameters are close together in the parameter space. A fundamental problem of extrema detection in bounded spaces is the detection of endpoint or boundary extrema [57], which can be true or false extrema. To solve the problem, we extend the sampling phases' domain according to $0 - N_v \leq \varphi < 360 + N_v$ and $0 - N_v \leq \vartheta \leq 180 + N_v$, and we extend each subset of sampling periods by N_v periods at both set boundaries. (In our experiments, we set $N_v = 1$.) Afterwards, we apply the extrema detector and eliminate those extrema detected in the extension. We sort the list of maxima according to their amplitude and select N_e maxima with the highest amplitudes. Variable N_e must be higher than the number of expected harmonic sources, N_s, times the number of harmonics, $\widehat{N}_q$, of a signal, where $\widehat{N}_q f_l \leq f_u$.

I. Joint Parameter Estimation

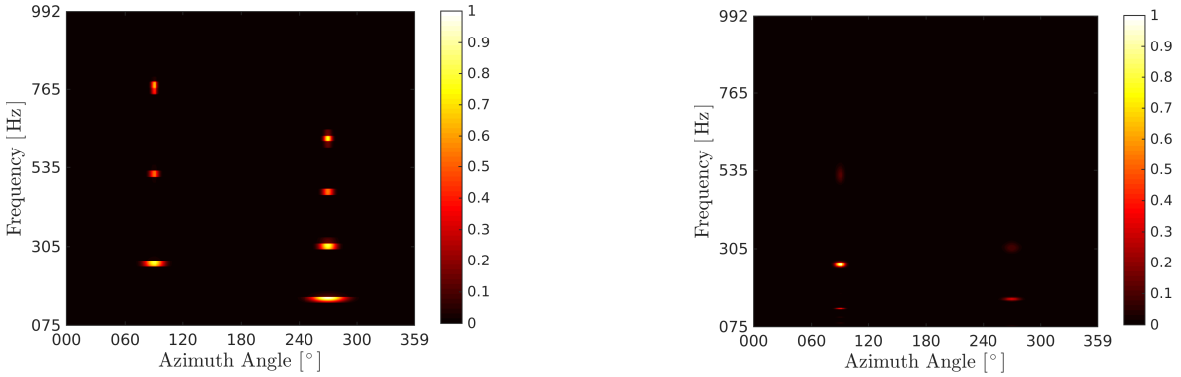
The SJPS is an intelligent non-parametric signal representation containing local maxima only. To jointly estimate the parameters of multiple harmonic sources, we need to know the general signal model (2) and analyze the SJPS. As shown in Fig. 5(a), the f_0's, the corresponding harmonics, and their respective amplitudes at a certain DOA belong to a single harmonic source. To determine the f_0 of this specific source without using an explicit estimator or detector, we pick its lowest estimated frequency within a narrow tolerance window around a certain DOA and ignore isolated clutter.

III. Experimental Design

Before we conducted our experiments, we specified the environmental parameters, i.e., we set the temperature to 20° Celsius yielding a speed of sound of $v \approx 343.2$ m/s. Then, we fixed the algorithmic parameters; we set the frame size to 0.032 s, the overlap to 0.010 s, $f_l = 75$ Hz, and $f_u = 1000$ Hz (which is high enough for our approach to estimate all f_0s). Additionally, we applied a Kaiser order-estimated filter bank with desired amplitudes of the stop and pass bands of $a_s = 0$, $a_p = 1$, and ripples of the stop and pass bands of $b_s = 0.01$ and $b_p = 0.05$. To limit the number of detected maxima, we considered a threshold of 10^{-5} and a maximum of $N_e = 16$ maxima, and we set the size of the maxima detector's search window to (6×6) indices. For our evaluations, we considered a tolerance of 10 Hertz and 10 degrees around the ground truth to define the root-mean-square errors and joint recalls, especially in case of double-source scenarios. Fig. 6 shows our algorithm's pseudo code.

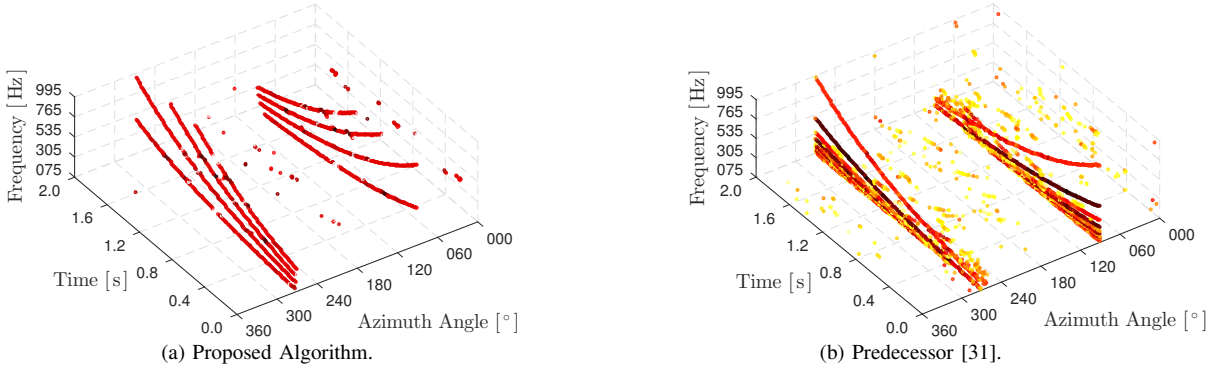
A. Synthetic-Data Experiments

We generated spatially non-moving and moving harmonic sources and noise sources in free field. We considered an angular grid in the interval $[000, 359]$ degrees azimuth with an angular step size of $\Delta\varphi = 1°$ and a fixed elevation of $\vartheta = 90°$. The distance between the center of the microphone array and the source was 3 m. The microphone array consisted of $N_m \in \{2, 4, 6, 8, 16\}$ microphones with maximum dimensions of $d \in \{0.20, 0.30, 0.40\}$ m. We selected a UCA for $N_m \geq 4$. Regarding our signal model in (2), we set $N_q = 4$, $f_2 = 500$ Hz, and $f_1 = 80$ Hz. In some scenarios, we added uniformly distributed white noise,



(a) Proposed Algorithm. (b) Predecessor [31].

Fig. 4. A normalized three-dimensional JPS (also known as PoPi plane [23]) representing the jointly estimated DOAs, f_0s and their corresponding second, third, and fourth harmonics with the respective CCF-values (half-wave rectified and normalized to achieve values between zero and one) computed (a) with our proposed algorithm and (b) its predecessor [31] without considering spectral fragments. By comparing both planes we can see in (b) that the predecessor exhibits pitch-period doubling (at approx. 120 Hz and 90°), fewer harmonics, and wrong amplitudes—all of them should be identical. The widening of the Gaussian-like kernels to lower frequency bands in (a) is due to the increase in a band's CCF's sampling period to lower frequency bands and the variable-scale sampling. The higher the band, the narrower the Gaussian-like kernel and vice versa.



(a) Proposed Algorithm. (b) Predecessor [31].

Fig. 5. Resulting trajectories after joint estimation of DOAs, f_0s and second, third, and fourth harmonics with the respective amplitudes with (a) our proposed algorithm and (b) its predecessor [31]. The predecessor's SJPS exhibits far more clutter, wrong amplitudes (yellow to red), missing higher harmonics, and spurious subharmonics. In both cases, we applied Lemire's extrema detector to sparsify the JPS.

Algorithm 1: Source Localizer and Characterizer

Data: Discrete-time multi-channel signals.
Result: Sparse joint parameter spaces.
initialization; // (5), (10), [51]
compute sets of indices for variable-scale sampling; // (11)-(16)
// store sets in lookup table
// consider extensions for maxima detector
apply bandpass filters;
split into frames;
while *getting frames* **do**
  foreach *pair of microphones* **do**
    split into frequency bands;
    foreach *frequency band* **do**
      transform into frequency domain;
      compute cross spectrum;
      transform into time domain; // (8)
      apply inverse windowing; // (9)
      apply sets of indices to unbiased cross correlation;
      sum samples related to each set's indices; // (19)
      // joint parameter space per pair and band
    end
    concatenate (≠sum) each frequency band's joint parameter space;
    // joint parameter space per pair
  end
  sum all pairs' joint parameter spaces;
  scale joint parameter space by number of pairs of microphones;
  // joint parameter space
  detect maxima; // [56]
  eliminate maxima in extension;
  // sparse joint parameter space
end

Fig. 6. Pseudo code of our algorithm. Two slashes indicate a comment.

$r[n] \sim \mathcal{U}(-0.4, +0.4)$ (rather than Gaussian noise to avoid clipping signals and to be able to precisely control the distribution's support). The test signals exhibited a duration of 2 s. In all double-source experiments we attenuated or amplified the target source signal yielding an amplitude for each harmonic according to $\alpha_k = \alpha = 0.4\sqrt{10^{\frac{\rho}{10}}}$, where ρ is the SIR (signal-to-interference ratio) or SNR (signal-to-noise ratio), which takes values in $\mathrm{SNR}, \mathrm{SIR} \in \{-10, 0, 10, 20, 30\}$ dB. To determine the best setting of algorithmic parameters, we conducted a vast number of Monte Carlo simulations in four different categories. For each simulation, we randomly chose the sampling frequency $f_s \in \{16, 32, 48, 64, 96\}$ kHz, the number of the Shah function's sampling points, $N_d \in \{1, 2\}$, and the parameters mentioned before. We initialized each source with a random DOA. In case of multiple sources, we set the minimum initial angular difference between each two sources to 20°. Mobile sources were moving along circular paths with an angular velocity of 1 m/s or 3 m/s clockwise or counter-clockwise causing intersections in directions and frequencies. If two microphones were selected only, we considered azimuth angles in the interval $[0, 180]$ degrees due to a linear array's spatial ambiguity [37]. For each scenario, we carried out 10^5 Monte Carlo runs to find the most robust setting of algorithmic parameters for all different categories. After doing so, we selected the most robust setting of parameters for further experiments to determine the algorithm's performance. These experiments were, again, Monte Carlo simulations because of varying initial DOAs, velocities, directions of moving sources, SNRs, and SIRs. We conducted experiments in four different categories with different algorithms for comparisons:

1) Single Harmonic Source: In the first scenario a non-moving source (see Fig. 7(a)) emitted an f_0-sweeping harmonic signal at varying locations. In the second scenario a moving source emitted an f_0-sweeping harmonic signal while moving along a circular path around the microphone array.
2) Single Harmonic Source Plus Noise Source: In one scenario a non-moving harmonic source emitted an f_0-sweeping harmonic signal at varying locations together with a non-moving noise source (see Fig. 7(b)) at different locations. In another scenario each source moved along a circular path around the array, which featured spatial intersections.
3) Two Harmonic Sources: In this category, we had again two different scenarios: two non-moving harmonic sources (see Fig. 7(c)) and two moving harmonic sources (see Fig. 8); in both scenarios the harmonic sources were emitting an f_0-sweeping harmonic signal at different locations. Our goal was to estimate the parameters of both sources.
4) Frequency-Hopping Harmonic Source Plus Noise Source: We simulated a trumpet emitting a sequence of tones [15], [17], [20] in a noisy environment by considering a non-moving randomly frequency-hopping harmonic source and a noise source at varying locations. The signal model was the same as described earlier, except that the fundamental frequency changed abruptly after time intervals of 500 ms.
5) Comparisons With Other Algorithms: We conducted experiments with four different algorithms: our algorithm denoted as VSS (variable scale sampling), POPI (position-pitch [23]), NLS (nonlinear least squares [20]), and aNLS (approximate nonlinear least squares [20]). We considered scenarios from the first three categories with non-moving sources. At this point, we need to clarify some issues regarding the NLS and the aNLS published in [20]. First, we implemented both algorithms following the description in [20] and realized that some relevant information was missing. The authors did not specify the line search algorithm to adapt the step size over iterations. Thus, we decided to implement a backtracking line search based on the Armijo-Goldstein condition [58]. Second, they did not mention which initial values they used for their step sizes and starting points. We set the initial step size, $\delta^{(\text{init})} = 1$, and we randomly selected the initial parameters (the DOA and the f_0) inside the domain. Third, almost all arguments of exponential functions in [20] feature a unit; however, all units in an exponential function's argument must cancel out. We realized that they did not multiply the affected arguments with the sampling period T_s. A workaround would be defining the time instances n_t in seconds instead of samples. Regarding the use of the NLS and aNLS algorithm, we set the model order to 4, the number of time instances per frame to 80, the number of iterations to 60, the line search method's contraction factor and slope modifier to 0.5 and 10^{-5}, respectively, and the sampling frequency, as suggested in [20], to 8 kHz.

B. Real-Data Experiments

To evaluate our algorithm's performance in real environments, we set up a large Austrian-German speech corpus named AMISCO (The Austrian German Multi-Sensor Corpus) [34], [35] containing multi-channel recordings (43 channels) labeled with a speaker's f_0, position, orientation, and other parameters. The corpus consists of 8.2 hours of read speech produced by 24 speakers, balanced male and female. The speakers read phonetically balanced sentences, commands, and digits at 16 different positions with 5 different orientations. Moreover, the corpus features glottograms of each speaker to determine f_0s, and it includes spatial trajectories of moving speakers determined by employing four Kinects. The recording environment consisted of a meeting room connected with a kitchen. Both rooms featured a reverberation time of $T_{60} \approx 0.5$ s. For our real-data experiments we used recordings of read items from speakers 08 (female) and 22 (male) (see Fig. 9). Besides, we only focused on two- and three-microphone recordings in the meeting room with maximum diameters $d = 0.30$ m and $d = 0.60$ m, respectively. We applied a short-term power estimation utilizing a first-order IIR smoothing of the signal's instantaneous power [59] to compute the speaker's SNR. To extract f_0s from all glottograms, we first computed a one-sided unbiased auto correlation of each glottogram's frame (with a frame length of 32 ms and a frame shift of 5 ms). Then, we employed a maximum detector to detect the lag of the auto correlation's global maximum between lags of 2 ms and 13 ms. The inverse of the global maximum's lag corresponds to the f_0, which we assume as our true f_0 [35].

IV. Experimental Results

A. Measures

We employed the joint recall (R), the root-mean-square error (RMSE), and the cumulative distribution function (CDF) to evaluate our experiments. By using R we determined the number of estimates within an interval, while we measured the distance between the estimated value and its ground truth by using the RMSE.
1) Joint Recall: It is the ratio of the number of correctly retrieved relevant parameters to the total number of relevant parameters, where a tuple (φ, f_0) is such a relevant parameter. Using the terminology of a confusion matrix, e.g., true positives (TP) and false negatives (FN), we defined the recall of jointly estimated DOAs and f_0s as

$$\mathrm{R}_\kappa(\varphi, f_0) = \frac{\mathrm{TP}_\kappa(\varphi, f_0)}{\mathrm{TP}_\kappa(\varphi, f_0) + \mathrm{FN}_\kappa(\varphi, f_0)} \tag{20}$$

with κ denoting the index of a Monte Carlo run and TP representing the true positives of all time frames per run. We assumed a tolerance window of 10 Hertz and 10 degrees. The average joint recall of N_c Monte Carlo simulations is

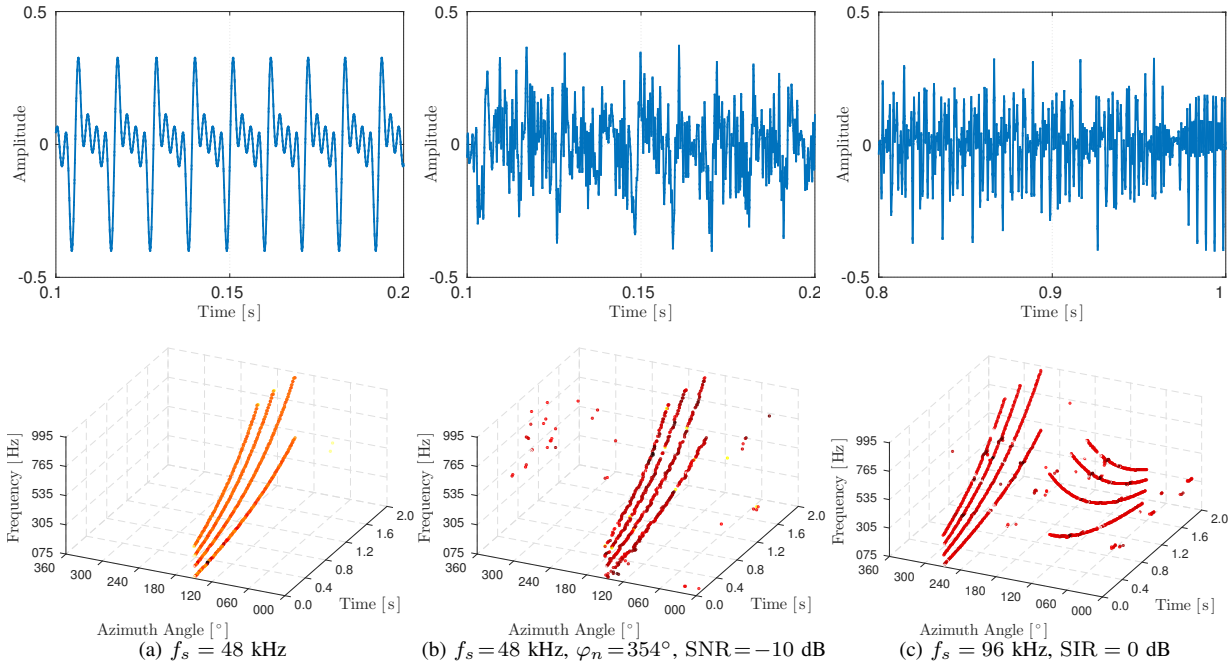
$$\overline{\mathrm{R}}(\varphi, f_0) = \frac{1}{N_c} \sum_{\kappa=1}^{N_c} \mathrm{R}_\kappa(\varphi, f_0). \tag{21}$$

2) Root-Mean-Square Error: It represents the difference between ground truth and observed values and is defined as

$$\mathrm{RMSE}_\kappa(\hat{\Theta}) = \sqrt{\frac{1}{N_{F,\kappa}} \sum_{g=1}^{N_{F,\kappa}} (\hat{\Theta}_g - \Theta_g)^2} \tag{22}$$

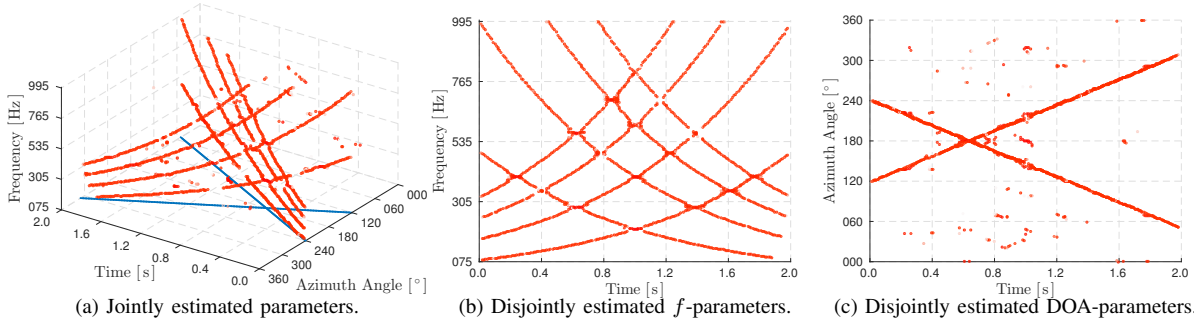
with $\hat{\Theta} \in \{\varphi, f_0\}$ and $\Theta \in \{\psi, f_0\}$ denoting the estimated values and the ground truth, respectively; g is the frame index and N_F is the total number of frames of a signal or recording in a single Monte Carlo run. It represents the RMSE of DOAs or f_0s, where ψ and f_0 are the ground truth parameters. The average RMSE of all Monte Carlo simulations is

$$\overline{\mathrm{RMSE}}(\hat{\Theta}) = \frac{1}{N_c} \sum_{\kappa=1}^{N_c} \mathrm{RMSE}_\kappa(\hat{\Theta}). \tag{23}$$



(a) $f_s = 48$ kHz (b) $f_s = 48$ kHz, $\varphi_n = 354°$, SNR $= -10$ dB (c) $f_s = 96$ kHz, SIR $= 0$ dB

Fig. 7. Jointly estimated f_0s, second, third, and fourth harmonics, as well as DOAs of a non-moving frequency-sweeping source. The experimental parameters are as follows: source direction $\varphi_t = 150°$, maximum dimension of microphone array $d = 0.40$ m, $N_m = 8$ microphones, five sampling points ($N_d = 2$), $N_e = 8$ is the maximum number of selected maxima, angular resolution $\Delta\varphi = 1°$, and $(3, 3)$ is the window size of the maxima detector. The sweeping harmonic signal starts with $f_0 = 75$ Hz and ends at $f_0 = 2000$ Hz. All harmonics exhibit the same amplitude. The distance between the virtual microphones and the source is 3 m. The first row illustrates the time signals, whereas the second row shows the respective SJPSs. In (a) we see a snapshot of a non-moving sweeping harmonic signal with four harmonics (top) and the corresponding SJPS (bottom). In (b) we considered additive white noise yielding SNR $= -10$ dB. Column (c) shows the time signal and SJPS of two non-moving sweeping harmonic signals with $(\varphi_t^{(1)}, \varphi_t^{(2)}) = (90°, 240°)$ and SIR $= 0$ dB.



(a) Jointly estimated parameters. (b) Disjointly estimated f-parameters. (c) Disjointly estimated DOA-parameters.

Fig. 8. Jointly (a) and disjointly (b,c) estimated f_0s and their second, third, and fourth harmonics, as well as DOAs of two moving frequency-sweeping sources. The experimental parameters are as follows: initial source directions $(\varphi_t^{(1)}, \varphi_t^{(2)}) = (120°, 240°)$, source velocity $v = 6$ m/s, distance between the center of the microphone array and the sources $r = 3$ m, sampling frequency $f_s = 96$ kHz, SIR $= 0$ dB, $d = 0.40$ m maximum dimension of microphone array, $N_m = 8$ microphones, five sampling points ($N_d = 2$), $N_e = 8$ is the maximum number of selected maxima, angular resolution $\Delta\varphi = 1°$, and $(3, 3)$ is the maxima detector's window size. The sweeping harmonic signals start with $f_0 = 75$ Hz and $f_0 = 2000$ Hz and end at $f_0 = 2000$ Hz and $f_0 = 75$ Hz, respectively. All harmonics exhibit the same amplitude. The plots in (b,c) illustrate the parameter spaces of disjoint estimates. It is impossible to associate the curves in (b) with their corresponding spatial trajectories in (c). However, in (a) these curves and trajectories are already associated to each other. The blue lines are the spatial trajectories of the angular components.

3) Cumulative Distribution Function: It shows a vast number of results in terms of a monotonic increasing curve. This allows to visualize the results of a big-data problem. We use X as a random variable whose individual outcomes are RMSE_κ, $\kappa = 1, ..., N_c$. Its CDF $F_X(\mathrm{RMSE})$ for a given RMSE is

$$F_X(\mathrm{RMSE}) = P(X \leq \mathrm{RMSE}). \tag{24}$$

Now, we choose Y as a random variable where its outcomes are the recall values R_κ, $\kappa = 1, ..., N_c$. The CDF is

$$F_Y(1 - \mathrm{R}) = P(Y \leq 1 - \mathrm{R}). \tag{25}$$

See the appendix for details about deriving $F_Y(1 - \mathrm{R})$.

B. Synthetic-Data Experiments

For each category of experiments described in the previous section, we first conducted Monte Carlo simulations with varying parameters to describe the algorithm's robustness for different settings; Figs. 10-12 show the corresponding results for all four categories, and Table I summarizes and highlights the important aspects of these figures. Then, we selected the best parameters, which are $f_s = 32$ kHz, $d = 0.40$ m, $N_m = 8$, $N_e = 16$, and $N_d = 2$, to do further experiments; Table II lists the corresponding results. Afterwards, we conducted experiments with the POPI, the NLS, and the aNLS algorithm. Table III lists the most important outcomes. It shows that our algorithm outperforms all the others in terms of $\mathrm{R}(\varphi, f_0)$.

Fig. 7(a) (bottom) and (b) (bottom) illustrates the SJPS over time of a non-moving harmonic source and a non-moving harmonic source plus an interfering noise source, respectively. Figs. 7(c) (bottom) and 8 shows the SJPS over time of two non-moving harmonic sources and two moving harmonic sources, respectively. As illustrated in Fig. 10,

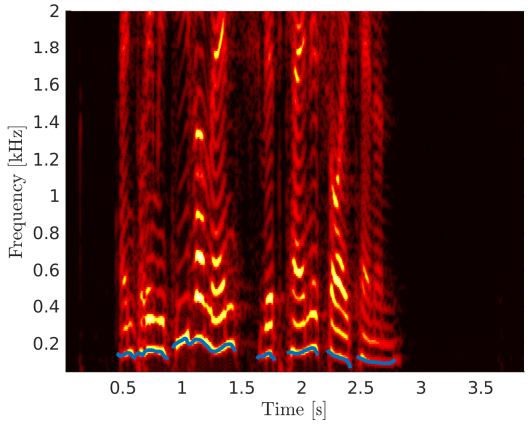


Fig. 9. Spectrogram of a male speaker's speech signal recorded with a headset. The (German-language) utterance is [aɱ priːmiːtiːvɐ mɛnʃ vɪrt kaɱə ʃɔʏ kɛnnən] (IPA). The f_0's ground truth values are marked with a blue line.

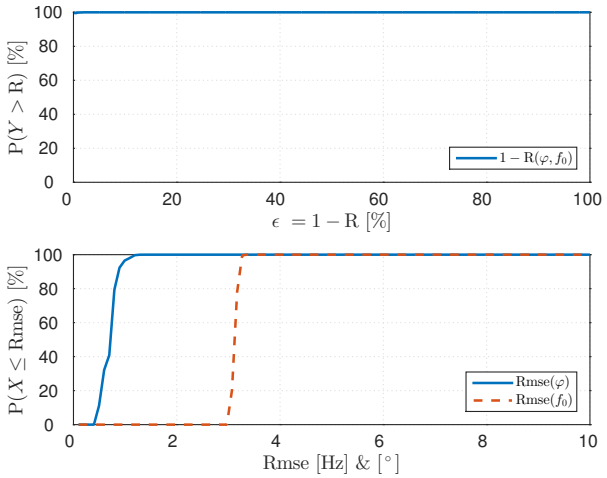


Fig. 10. Resulting CDFs of an experiment with synthesized moving harmonic signals. They describe the probability that the opposite of R in percent (top), i.e., ($\epsilon = 1 - \mathrm{R}$), and RMSE (bottom) of jointly estimated DOA and f_0 has a value equal to or less than $1 - \mathrm{R}$ and RMSE.

$P(Y > \mathrm{R}) = 100\%$ for $\epsilon \geq 0$. This means that each experiment resulted in $\mathrm{R}(\varphi, f_0) = 100\%$. In Fig. 10 we see that the $\mathrm{RMSE}(\varphi)$ and the $\mathrm{RMSE}(f_0)$ are around $0.8°$ and 3 Hz, respectively. We additionally considered scenarios with different SNRs and SIRs. Fig. 11(a) shows that the $\mathrm{R}(\varphi, f_0) \geq 90\%$ in 100% of all experiments and the $\mathrm{RMSE}(f_0)$ is similar to Fig. 10 but the range of $\mathrm{RMSE}(\varphi)$ is larger. As presented in Fig. 11(c), the $\mathrm{RMSE}(\varphi)$ decreases for increasing SNR. In Fig. 11(b) we see that $\mathrm{R}(\varphi, f_0) \geq 83\%$ in 100% of all experiments and the $\mathrm{RMSE}(f_0)$ is similar to Fig. 10 but the range of $\mathrm{RMSE}(\varphi)$ is larger but still smaller than in Fig. 11(a). Fig. 11(d) reflects these observations. The remarkable differences between Fig. 12 and Fig. 10-11 are the reduced $\mathrm{RMSE}(f_0)$s and $\mathrm{RMSE}(\varphi)$s.

C. Real-Data Experiments

We initially conducted Monte Carlo simulations to determine the best parameters, which are as follows: $f_s = 32$ kHz, $N_e = 16$, and $N_d = 2$. After selecting the best setting of parameters, we continued conducting Monte Carlo simulations by randomly selecting recordings. Due to a fixed setting of parameters and environmental properties, we show $P(X \leq \mathrm{RMSE})$ and $P(Y > \mathrm{R})$ only. In comparison to Figs. 10-12, Fig. 13 additionally feature $\mathrm{R}(\varphi)$ and $\mathrm{R}(f_0)$.

V. Discussion

A. Synthetic-Data Experiments

1) Single Harmonic Source: In this category our algorithm achieves a $\mathrm{R}(\varphi, f_0) = 100\%$ in each Monte Carlo simulation as shown in Fig. 10 and Table I-II. The algorithm perfectly solves the problem of jointly estimating the DOA and the f_0 of a single harmonic source while keeping the $\mathrm{RMSE}(\varphi)$ and $\mathrm{RMSE}(f_0)$ low.

2) Single Harmonic Source Plus Noise Source: Again, we achieve a $\mathrm{R}(\varphi, f_0) = 100\%$ for experiments with $\mathrm{SNR} \geq 0$ dB. As shown in Fig. 11 the recall starts decreasing for $\mathrm{SNR} < 0$ dB, which highlights the robustness against noise sources exhibiting the same and lower power as the harmonic source of interest. Table I supports this statement by showing that in 80% and 100% of all experiments our algorithm achieves a $\mathrm{R}(\varphi, f_0) = 100\%$ and $\mathrm{R}(\varphi, f_0) \geq 90\%$. Table I emphasizes the algorithm's robustness for scenarios with a $\mathrm{SNR} \geq 0$ dB; the $\mathrm{RMSE}(\varphi)$ and the $\mathrm{RMSE}(f_0)$ are still low.

3) Two Harmonic Sources: In this category our algorithm features, as shown in Fig. 11, lower RMSEs than in the previous one; however, $\overline{\mathrm{R}}(\varphi, f_0)$ is lower than in all other categories. This is due to the beating effect during crossings of frequencies shown in Fig. 7(c) (top) at 0.97 s and (bottom) at $240°$ and 0.5 s as well as in Fig. 8(b) at 0.4 s and 330 Hz, at 1.18 s and 730 Hz, and at 1.4 s and 550 Hz. This effect causes destructive interference of the superimposed signals. When estimating both harmonic sources, we achieve the highest $\overline{\mathrm{R}}(\varphi, f_0)$ when $\mathrm{SIR} = 0$ dB, because the signals of both sources exhibit the same power, i.e., they are equally present. In case of $\mathrm{SIR} = 30$ dB, one source dominates the other, which is problematic if both sources are spatially close to each other. The results in case of $\mathrm{SIR} = \pm 10$ dB are almost identical, because one source is dominating the other one.

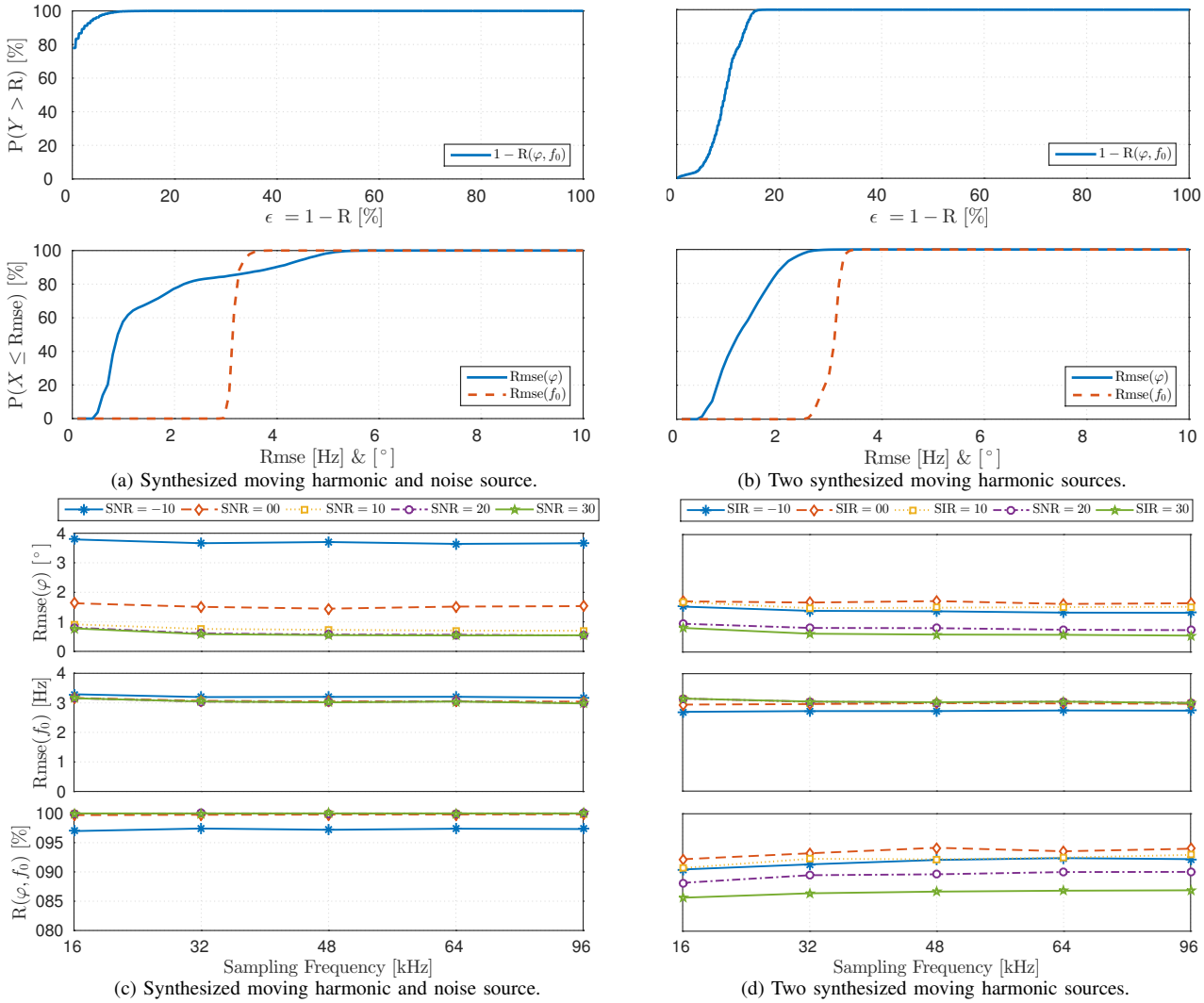
4) Non-Sweeping Harmonic Source Plus Noise Source: In comparison to the previous category and as shown in Fig. 12, the $\overline{\mathrm{R}}(\varphi, f_0)$ is lower at $f_s = 16$ kHz due to the signals' characteristics and the lower resolution at higher frequencies. In this category, the f_0s to-be-estimated are constant over a long period of time. The $\mathrm{RMSE}(f_0)$s increase if the ground truth of f_0 exhibits a value at higher frequencies and if the ground truth value is not an element of our frequency grid defined by (13). The RMSEs are smaller than in case of sweeping harmonic sources because the signals' f_0s are constant over certain time intervals and, though being uniformly distributed in frequency range, they occur more often in a range where the frequency resolution is approximately constant. Although we do not outperform the algorithm presented in [20], we show that our algorithm works with harmonic signals based on a musical instrument's signal model.

5) General Observations: The results of the previous section show that the $\mathrm{RMSE}(f_0)$ does not fall below 2.9 Hz in categories one, two, and three. This is due to several reasons: First, the finite number of sampling periods causes a quantization error. Second, increasing the sampling interval of the Shah function linearly and sample-by-sample in the lag domain corresponds to a nonlinear decrease of the frequency interval in the frequency domain ($f = 1/T$). Thus, at higher frequencies the quantization intervals in the frequency domain get larger. Third, we generated the source signals sample-by-sample using (2). However, we defined the value in the

TABLE I
RESULTS OF SYNTHETIC-DATA EXPERIMENTS WITH ALL PARAMETERS

Scenario	$P(Y>\mathrm{R})\approx 100\%$	$P(Y>\mathrm{R})\approx 90\%$	$P(Y>\mathrm{R})\approx 80\%$	$P(X\leq \mathrm{RMSE})\approx 100\%$	$P(X\leq \mathrm{RMSE})\approx 90\%$	$P(X\leq \mathrm{RMSE})\approx 80\%$
S	100	100	100	≤ 1.20 / ≤ 3.20	≤ 0.90 / ≤ 3.15	≤ 0.80 / ≤ 3.10
S+N	≥ 90	≥ 96	100	≤ 5.20 / ≤ 3.60	≤ 3.90 / ≤ 3.40	≤ 2.10 / ≤ 3.30
S+S	≥ 83	≥ 85	≥ 87	≤ 2.80 / ≤ 3.50	≤ 2.00 / ≤ 3.30	≤ 1.80 / ≤ 3.20
$\widehat{\mathrm{S}}$+N	≥ 70	≥ 90	≥ 96	≤ 5.50 / ≤ 3.75	≤ 3.60 / ≤ 1.80	≤ 2.00 / ≤ 1.60

The letters S, $\widehat{\mathrm{S}}$, and N denote sweeping harmonic, frequency-hopping harmonic, and noise signal, respectively. The values below $P(Y>\mathrm{R})$ are the recalls $\mathrm{R}(\varphi, f_0)$ in %. The values below $P(X\leq\mathrm{RMSE})$ are the root-mean-square errors $\mathrm{RMSE}(\varphi)$ in degrees (left) and $\mathrm{RMSE}(f_0)$ in Hertz (right). For instance, $\mathrm{R}(\varphi, f_0)\geq 96\%$ for $P(Y>\mathrm{R})\approx 90\%$ implies that the algorithm yielded a joint recall of 96% or higher in 90% of all experiments with a harmonic signal plus noise signal.



(a) Synthesized moving harmonic and noise source.
(b) Two synthesized moving harmonic sources.
(c) Synthesized moving harmonic and noise source.
(d) Two synthesized moving harmonic sources.

Fig. 11. Cumulative distribution functions (a,b) and root-mean-square errors and joint recalls (c-d) for different SNRs and SIRs.

center of each frame as our ground truth value, because our algorithm estimates f_0s frame-by-frame. Fourth, the rounding of sampling periods to the nearest integer causes increasing errors for higher frequencies. Furthermore, the results show that noise mainly affects the estimation of DOA.

6) Comparisons With Other Algorithms: As listed in Table III, our proposed algorithm outperforms or compares favorably with the other algorithms in all three categories in terms of $\overline{\mathrm{R}}(\varphi, f_0)$, especially in categories with two sources. The NLS as well as the aNLS algorithm are unable to estimate parameters of two or more sources. Their accuracy decreases for low SNR- and SIR-scenarios. The POPI algorithm performs better than the NLS and aNLS algorithm, however, as soon as one source dominates the other, its estimation performance decreases.

Focusing on $\overline{\mathrm{RMSE}}(\varphi)$, our algorithm outperforms all other algorithms in scenarios with two harmonic sources. In scenarios with a single harmonic source, the POPI algorithm achieves the smallest $\overline{\mathrm{RMSE}}(\varphi)$, which is due to the use of a single broadband CCF; it exhibits a narrow peak at the lag corresponding to the dominant source's DOA.

The NLS exhibits the smallest $\overline{\mathrm{RMSE}}(f_0)$, which corresponds to the findings reported in [20]. Our proposed algorithm achieves $\overline{\mathrm{RMSE}}(f_0)\approx 3$ Hz; this is mostly due to the

TABLE II
RESULTS OF SYNTHETIC-DATA EXPERIMENTS WITH BEST PARAMETERS

Scenario	$\overline{\mathrm{R}}(\varphi, f_0)$ [%]			$\overline{\mathrm{RMSE}}(\varphi)$ [°]			$\overline{\mathrm{RMSE}}(f_0)$ [Hz]		
non-moving S	100			0.26			3.03		
moving S	100			0.56			3.03		
	AVG	30 dB	-10 dB	AVG	30 dB	-10 dB	AVG	30 dB	-10 dB
non-moving S+N	100	100	98	1.22	0.29	3.48	3.07	3.03	3.20
moving S+N	100	100	98	1.45	0.56	3.66	3.08	3.03	3.21
non-moving $\widehat{\mathrm{S}}$+N	97	93	97	1.17	0.11	3.23	1.17	1.08	1.58
	AVG	30 dB	0 dB	AVG	30 dB	0 dB	AVG	30 dB	0 dB
non-moving S+S	91	87*	93*	1.09	0.35*	1.56*	2.96	3.04*	3.01*
moving S+S	91	87*	94*	1.25	0.60*	1.78*	2.95	3.03*	2.98*

The letters S, $\widehat{\mathrm{S}}$, and N denote sweeping harmonic, frequency-hopping harmonic, and noise signal, respectively. In case of mixed scenarios (S+N) or ($\hat{\mathrm{S}}$ + N), the first value in each column represents the averaged results of all scenarios with varying SNR, the second one for SNR = 30 dB, and the third one for SNR = −10 dB. In case of (S + S) scenarios, the second value in each column represents the results for SIR = 30 dB, the third for SIR = 0 dB; they are marked with a star. We set $d = 0.40$ m, $N_m = 8$, and $f_s = 32$ kHz.

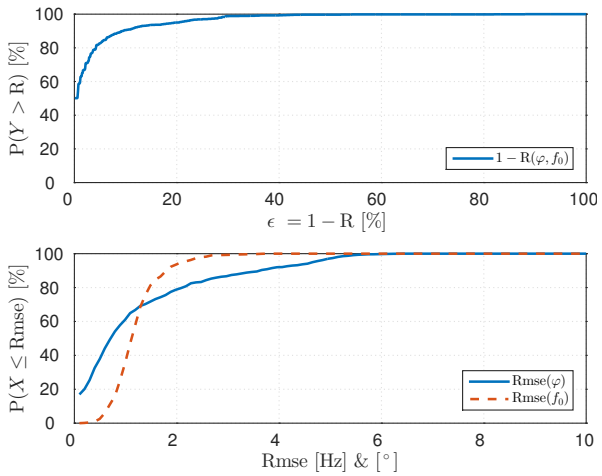
TABLE III
RESULTS OF SYNTHETIC-DATA EXPERIMENTS WITH ALL ALGORITHMS

Algorithm	$\overline{\mathrm{R}}(\varphi, f_0)$ [%]			$\overline{\mathrm{RMSE}}(\varphi)$ [°]			$\overline{\mathrm{RMSE}}(f_0)$ [Hz]		
VSS	100			0.26			3.03		
POPI	100			0.01			3.00		
NLS	51			3.80			0.39		
aNLS	41			3.53			1.23		
	AVG	30 dB	-10 dB	AVG	30 dB	-10 dB	AVG	30 dB	-10 dB
VSS	100	100	98	1.22	0.29	3.48	3.07	3.03	3.20
POPI	100	100	98	0.30	0.02	1.64	3.03	3.01	3.05
NLS	43	58	2	4.34	3.75	6.34	2.05	0.59	5.30
aNLS	32	44	2	4.13	3.48	5.54	2.24	1.27	5.74
	AVG	30 dB	0 dB	AVG	30 dB	0 dB	AVG	30 dB	0 dB
VSS	91	87*	93*	1.09	0.35*	1.56*	2.96	3.04*	3.01*
POPI	66	52*	92*	1.35	0.03*	2.26*	3.32	3.00*	3.17*
NLS	25	29*	10*	4.46	3.89*	5.21*	2.29	0.63*	3.56*
aNLS	19	22*	8*	4.21	3.52*	5.49*	2.33	1.23*	3.96*

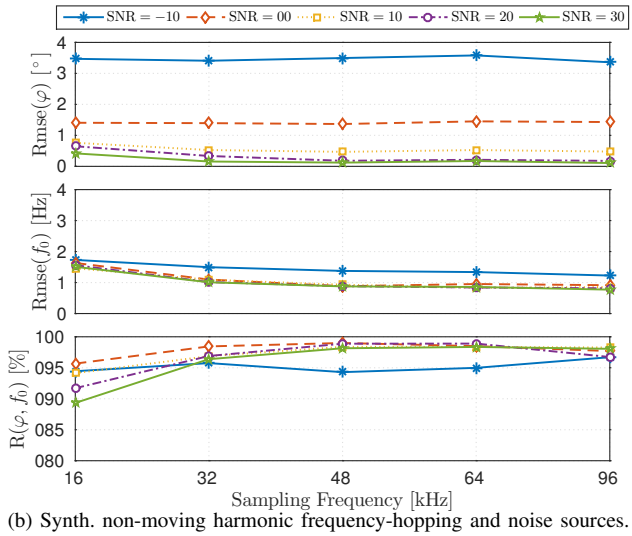
This table lists the results of synthetic-data experiments with different approaches: VSS (variable scale sampling), POPI (position-pitch [23]), NLS (nonlinear least squares [20]), and aNLS (approximate nonlinear least squares [20]). It consists of three sections covering the results of experiments with a single non-moving harmonic source, a single non-moving harmonic source plus noise source, and two non-moving harmonic sources, respectively. In the second section, the first value in each column represents the averaged results of all scenarios with varying SNR, the second one for SNR = 30 dB, and the third one for SNR = −10 dB. In the third section, the second value in each column represents the results for SIR = 30 dB, the third for SIR = 0 dB; they are marked with a star. We set $d = 0.40$ m and $N_m = 8$.

decreasing frequency resolution for increasing frequencies.

According to [20], the ideal NLS estimator is a maximum likelihood estimator that attains the Cramér-Rao bound in case of single-source scenarios with white Gaussian noise. In such scenarios, it should outperform all other algorithms, but this was not the case due to the following reasons: The aforementioned statement is true if we would evaluate the cost function for all f_0 and DOA candidates and search for the global maximum. However, the authors of [20] presented a version based on gradient ascent, which may converge to the global maximum (the true DOA and f_0) or to local maxima (with wrong f_0s) depending on the initial values and the employed line search algorithm. Furthermore, due to a finite number of iterations, the algorithm sometimes failed to reach the correct DOA. Moreover, we applied uniformly distributed white noise instead of white Gaussian noise; and, unlike [20], we employed signals with a frequency sweep.



(a) Synth. non-moving harmonic frequency-hopping and noise sources.



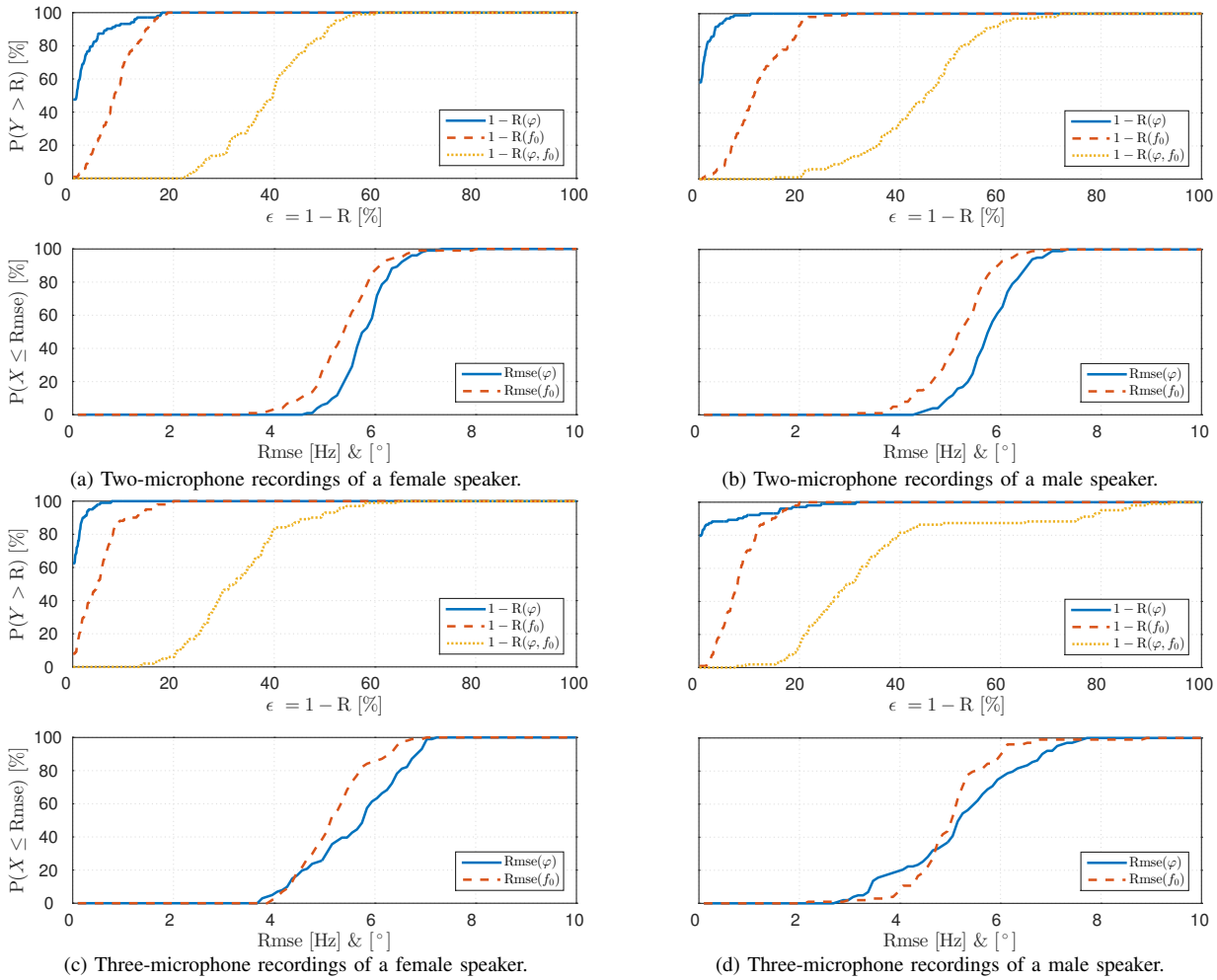
(b) Synth. non-moving harmonic frequency-hopping and noise sources.

Fig. 12. Cumulative distribution functions (a) and root-mean-square errors and joint recalls (b) for different SNRs and SIRs of an experiment with synthesized non-moving harmonic frequency-hopping sources and noise sources.

To sum it up, our proposed algorithm is able to jointly estimate the DOA and f_0 of two or more harmonic sources, whereas the others can cope with a single source only or they focus on the dominant source.

B. Real-Data Experiments

This set of experiments employs speech signals recorded in a real environment featuring, e.g., reverberation, (strong) multi-path components, and non-harmonic components like plosives, fricatives, and noise. Despite these challenging characteristics, the figures show that we can, however, successfully localize and characterize sources with just two or three microphones. Thus, the joint estimation and representation in an SJPS introduces new possibilities to further process the parameters in a higher-dimensional sense in a real environment. Besides, evaluating DOA and f_0 disjointly by using our algorithm yields even better results than estimating them jointly. However, estimating parameters disjointly requires an additional step, the data association, which in turn requires a certain amount of prior knowledge. In Fig. 13, we see that, sometimes, $P(Y > \mathrm{R}(\varphi)) = 100\%$ and $P(Y > \mathrm{R}(f_0)) = 100\%$, but $P(Y > \mathrm{R}(\varphi, f_0)) < 100\%$. This



(a) Two-microphone recordings of a female speaker.

(b) Two-microphone recordings of a male speaker.

(c) Three-microphone recordings of a female speaker.

(d) Three-microphone recordings of a male speaker.

Fig. 13. Cumulative distribution functions of experiments with two- and three-microphone recordings of a female and a male speaker.

is true because $\mathrm{R}(f_0) = \mathrm{R}(\varphi) = 100\%$ does not necessarily imply that $\mathrm{R}(\varphi, f_0) = 100\%$. For instance, if there is one reference and if there are two estimates, one matching the true f_0 only and the other one matching the true φ only, then $\mathrm{R}(\varphi, f_0) = 0\%$, although $\mathrm{R}(f_0) = \mathrm{R}(\varphi) = 100\%$. Now, we continue with estimating f_0s. A frame of the captured signal contained direct-path and multi-path components of a source. Due to the reverberation room's memory effect, the frequencies of the multi-path components slightly differed from the frequencies of the direct-path components within a time frame. Additionally, sometimes the multi-path components dominated in energy. These effects introduced small errors in f_0. Focusing on the joint estimation of DOAs and f_0s, both effects mentioned above decreased the R. Regarding RMSE we see that the $\overline{\mathrm{RMSE}}(f_0) < \overline{\mathrm{RMSE}}(\varphi)$, which is opposite to the experiments with synthetic data. This is again due to the multi-path components.

VI. Conclusion

Our new algorithm characterizes and localizes multiple acoustic sources by jointly estimating their fundamental frequencies and directions of arrival based on an intelligent non-parametric signal representation. It is real-time capable, deterministic, and it does not rely on an estimator, any machine learning algorithms, or data driven methods. Additionally, it is capable of solving the well-known issue of pitch-period doubling when using the cross-correlation function. We showed that we span a sparse joint parameter space (which can be directly fed into a tracker) by applying a filter bank, variable-scale sampling of cross-correlation functions, and a fast and accurate multidimensional maxima detector. We conducted almost 10^6 Monte Carlo simulations and evaluated its performance in free field and reverberant conditions (with a reverberation time of around 0.5 s) by using the joint recall measure, the root-mean-square error, and the cumulative distribution function of fundamental frequencies and directions of arrival. The algorithm is not as accurate as the adaptive approaches in single-source scenarios. However, it estimates directions of arrival, f_0s, and their respective amplitudes of multiple harmonic sources, it solves the data association problem, and it only relies on deterministic signal processing. Bringing it all together, we can claim that our algorithm is suitable and important in the field of spatio-temporal filtering and blind source separation to further increase the word accuracy ratio of a distant-speech recognition system.

Appendix
The Cumulative Distribution Function of Recalls

The cumulative distribution function for recalls is

$$F_Y(\mathrm{R}) = P(Y \leq \mathrm{R}). \tag{26}$$

However, as we are interested in values of R close to 100%, we may redefine $\mathrm{R} = 1 - \epsilon$ to obtain

$$F_Y(1 - \epsilon) = P(Y \leq 1 - \epsilon), \tag{27}$$

and finally, to make the graph reflect monotonically decreasing quality in a similar way as the CDF of the RMSE, we consider

$$1-F_Y(1-\epsilon)=1-P(Y\leq 1-\epsilon)=P(Y>1-\epsilon)=P(Y>\mathrm{R}) \quad (28)$$

and produce a graph shown in, e.g., Fig. 11(a) (top). In our article, it describes the probability that a certain percentage of all experiments yields an $1-\mathrm{R}$,

$$F_Y(1-\mathrm{R}) = P(Y \leq 1-\mathrm{R}), \quad (29)$$

of a certain value and smaller. For instance, in the latter case, $1-\mathrm{R}=0.75$ equals a joint recall of 0.25 or 25%.

We estimate the CDFs, $F_X(\mathrm{RMSE})$ and $F_Y(1-\mathrm{R})$, by (a) computing the total number of measurements, (b) sorting all measurements (e.g., RMSE_κ or $1-\mathrm{R}_\kappa$) in an ascending manner, (c) defining intervals from 0 to a non-negative number unequal zero, and (d) counting the measurements lying within those intervals. Employing a CDF to visualize our results yields several benefits, especially in case of Monte Carlo simulations. First, it reveals the whole range of outcomes, i.e., RMSEs and Rs or $1-\mathrm{R}$s, and their corresponding probabilities. The first plot in Fig. 11(a) shows that $\mathrm{R}=100\%$ in 80% of all experiments. At $1-\mathrm{R}=10\%$ there is already a probability of 100% that all Monte Carlo runs yield $\mathrm{R}=90\%$ and higher. Second, the slope of a CDF tells us in which interval most of the outcomes occur. For instance, in the second plot of Fig. 11(a) we see that most of the $\mathrm{RMSE}(f_0)$s are around 3.1 Hz. If we differentiate this CDF, we would get a probability density function with a mean of $\mu=3.1$ Hz and a small standard deviation and variance of around $\sigma=0.1$ Hz and $\sigma^2=0.01$ Hz^2, respectively.

ACKNOWLEDGMENT

H. Pessentheiner would like to express his gratitude to Prof. B. D. Rao (he is with the Department of Electrical and Computer Engineering, University of California, San Diego) who provided him an opportunity to join his team as a visiting scientist in 2013 and 2014. He would also like to give thanks to him for all his inspiring discussions and helpful advices. Moreover, authors would like to thank M. Gabbrielli for his assistance in analyzing, implementing, and testing the NLS and aNLS algorithm.

REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, reprint (09/29/1994) ed. Cambridge, MA, USA: MIT Press, May 1990.

[2] D. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, Sep. 2006.

[3] H. Teutsch, *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decomposition*. Berlin, Germany: Springer, Jan. 2007.

[4] K. Imoto, Y. Ohishi, H. Uematsu, and H. Ohmuro, "Acoustic scene analysis based on latent acoustic topic and event allocation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, Southampton, U.K., Sep. 2013, pp. 1–6.

[5] H. Kwon, H. Krishnamoorthi, V. Berisha, and A. Spanias, "A sensor network for real-time acoustic scene analysis," in *Proc. IEEE International Symposium on Circuits and Systems*, Taipei, Taiwan, May 2009, pp. 169–172.

[6] A. de Cheveigné and M. Slama, "Acoustic Scene Analysis Based on Power Decomposition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, Toulouse, France, May 2006, pp. 49–52.

[7] "SHINE: Speech-Acoustic Scene Analysis and Interpretation," Fondazione Bruno Kessler (FBK), Povo, Trento, Italy, Accessed on: Aug. 25, 2015. [Online]. Available: http://shine.fbk.eu

[8] "DIRHA: Distant-speech Interaction for Robust Home Applications," Fondazione Bruno Kessler (FBK), Povo, Trento, Italy, Accessed on: Aug. 25, 2015, Grant Agreement No. FP7-ICT-2011-7-288121. [Online]. Available: http://dirha.fbk.eu

[9] L. Thurman and G. Welch, *Bodymind & Voice: Foundations of Voice Education, Revised Edition*. Chicago, IL, USA: The Voicecare Network, 2000.

[10] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Munich, Germany, Apr. 1997, pp. 375–378.

[11] B. Yegnanarayana, S. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.

[12] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, Aug. 2007, pp. 2769–2772.

[13] G. Liao, H. C. So, and P. C. Ching, "Joint time delay and frequency estimation of multiple sinusoids," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, Salt Lake City, UT, USA, May 2001, pp. 3121–3124.

[14] L. Y. Ngan, Y. Wu, H. C. So, and P. C. Ching, " Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE International Symposium on Circuits and Systems*, vol. 3. Bangkok, Thailand: IEEE, May 2003, pp. 722–725.

[15] J. R. Jensen, M. G. Christensen, J. Benesty, and S. H. Jensen, "Joint Spatio-Temporal Filtering Methods for DOA and Fundamental Frequency Estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, p. 174, Jan 2015.

[16] J. R. Jensen, J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "On Frequency Domain Models for TDOA Estimation," in *Proc, 40th IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, Australia, Apr 2015, pp. 11–15.

[17] Y. Wu, A. Leshem, J. R. Jensen, and G. Liao, "Joint Pitch and DOA Estimation Using the ESPRIT Method," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 32–45, Jan 2015.

[18] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and Fundamental Frequency Estimation Methods based on 2-D Filtering," in *Proc. European Signal Processing Conference*, Aalborg, Denmark, Aug. 2010, pp. 2091–2095.

[19] S. Karimian-Azari, J. R. Jensen, and M. G. Christensen, "Fast joint DOA and pitch estimation using a broadband MVDR beamformer," in *Proc. 21st European Signal Processing Conference*, Marrakech, Morocco, Sep. 2013, pp. 1–5.

[20] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear Least Squares Methods for Joint DOA and Pitch Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 923–933, Jan. 2013.

[21] T. Kronvall, S. I. Adalbjornsson, and A. Jakobsson, "Joint DOA and multi-pitch estimation using block sparsity," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, pp. 3958–3962.

[22] S. N. Wrigley and G. J. Brown, "Recurrent timing neural networks for joint f0-localisation based speech separation ," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, USA, Apr. 2007, pp. 157–160.

[23] M. Képesi, F. Pernkopf, and M. Wohlmayr, "Joint Position-Pitch Tracking for 2-Channel Audio ," in *Proc. International Workshop on Content-Based Multimedia Indexing*, Bordeaux, France, Jun. 2007, pp. 303–306.

[24] M. Képesi, L. Ottowitz, and T. Habib, "Joint Position-Pitch Estimation for Multiple Speaker Scenarios," in *Proc. Joint Workshop on Hands-Free Speech Communication and Microphone Arrays*, Trento, Italy, May 2008, pp. 85–88.

[25] T. Habib, M. Képesi, and L. Ottowitz, "Experimental evaluation of the joint position-pitch estimation (POPI) algorithm in noisy environments," in *Proc. 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, Darmstadt, Germany, Jul. 2008, pp. 369–372.

[26] T. Habib, L. Ottowitz, and M. Képesi, "Experimental evaluation of multi-band position-pitch estimation (m-popi) algorithm for multi-speaker localization," in *Proc. 9th Annual Conference of the International Speech Communication Association*, Brisbane, Australia, Sep. 2008, pp. 1317–1320.

[27] T. Habib and H. Romsdorfer, "Comparison of SRP-PHAT and Multiband-PoPi Algorithms for Speaker Localization Using Particle Filters," in *Proc. 13th International Conference on Digital Audio Effects*, Graz, Austria, Sep. 2010, pp. 1–6.

[28] ——, "Combining multiband joint position-pitch algorithm and particle filters for speaker localization," in *Proc. IEEE Workshop on Sensor Array*

and Multichannel Signal Processing, Tel-Aviv, Israel, Oct. 2010, pp. 149–152.
[29] ——, "Concurrent speaker localization using multi-band position-pitch (m-popi) algorithm with spectro-temporal pre-processing," in *Proc. 11th Annual Conference of the International Speech Communication Association*, Makuhari, Chiba, Japan, Sep. 2010, pp. 2774–2777.
[30] ——, "Improving Multiband Position-Pitch Algorithm for Localization and Tracking of Multiple Concurrent Speakers by Using a Frequency Selective Criterion," in *Proc. 12th Annual Conference of the International Speech Communication Association*, Florence, Italy, Aug. 2011, pp. 2897–2900.
[31] ——, "Auditory inspired methods for localization of multiple concurrent speakers," *Computer Speech & Language*, vol. 27, no. 3, pp. 634–659, May 2013.
[32] M. Wohlmayr and M. Képesi, "Joint Position-Pitch Extraction from Multichannel Audio," in *Proc. 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, Aug. 2007, pp. 1629–1632.
[33] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, Apr. 2009, pp. 4593–4596.
[34] H. Pessentheiner, "AMISCO: The Austrian German Multi-Sensor Corpus," Graz University of Technology, Austria, 2015. [Online]. Available: https://www.spsc.tugraz.at/tools/amisco
[35] H. Pessentheiner, T. Pichler, and M. Hagmüller, "AMISCO: The Austrian German Multi-Sensor Corpus," in *Proc. 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia, to be published.
[36] J. Benesty, J. Chen, and Y. Huan, *Microphone Array Signal Processing*. Berlin, Germany: Springer, Mar. 2008.
[37] I. J. Tashev, *Sound Capture and Processing: Practical Approaches*. Chichester, U.K.: Wiley, Jul. 2009.
[38] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. New York, NY, USA: Springer, Feb. 2007.
[39] W. M. Hartmann, "Pitch, periodicity, and auditory organization," *The Journal of the Acoustical Society of America*, vol. 100, no. 6, pp. 3491–3502, Dec. 1996.
[40] M. Stark, M. Wohlmayr, and F. Pernkopf, "Source–Filter-Based Single-Channel Speech Separation Using Pitch Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 242–255, Feb. 2011.
[41] W. Zhang and B. D. Rao, "A Two Microphone-Based Approach for Source Localization of Multiple Speech Sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1913–1928, Nov. 2010.
[42] H. L. V. Trees, *Optimum Array Processing: Part IV of Detection, Estimation, and Modulation Theory.* New York, NY, USA: Wiley, May 2002.
[43] J. G. Ryan, "Criterion for the minimum source distance at which plane-wave beamforming can be applied," *The Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 595–598, Jul. 1998.
[44] L. Rabiner, " On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, Feb 1977.
[45] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, 1st ed. Berlin, Germany: Springer, Jun 1983.
[46] C. Roads, *The Computer Music Tutorial*. Cambridge, MA, USA: MIT Press, Feb. 1996.
[47] G. Weinreich, "Coupled piano strings," *Journal of the Acoustical Society of America*, vol. 62, no. 6, pp. 1474–1484, 1977.
[48] J. F. Kaiser, "Nonrecursive digital filter design using I_0-sinh window function," in *Proc. IEEE International Symposium on Circuits and Systems*, San Francisco, CA, USA, Apr. 1974, pp. 20–23.
[49] ——, "On the use of the I_0-sinh window for spectrum analysis," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-28, no. 1, pp. 105–107, 1980.
[50] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed. Englewood Cliffs, NJ, USA: Prentice-Hall, Aug. 2009.
[51] "MATLAB and Signal Processing Toolbox Release 2015a," Mathworks, Natick, MA, USA, Accessed on: Aug. 25, 2015.
[52] M. Slaney, "An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank," Apple Computer, Inc., Cupertino, CA, USA., Tech. Rep. 35, 1993.
[53] M. D. Lutovac, D. V. Tošić, and B. L. Evans, *Filter Design for Signal Processing Using MATLAB and MATHEMATICA*, 1st ed. Englewood Cliffs, NJ, USA: Prentice-Hall, Sep. 2000.
[54] R. Bracewell, *Fourier Analysis and Imaging*, 1st ed. New York, NY, USA: Springer, 2003.
[55] D. Lemire, "Streaming maximum-minimum filter using no more than three comparisons per element," *Nordic Journal of Computing*, vol. 13, no. 4, pp. 328–339, Dec. 2006.
[56] B. Luong, "The Min/Max Filter," Mathworks, Natick, MA, USA, Dec. Accessed on: Aug. 25, 2015. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/24705-min-max-filter
[57] D. G. Zill and W. S. Wright, *Calculus: Early Transcendentals*, 4th ed. Sudbury, MA, USA: Jones & Bartlett, Apr. 2011.
[58] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., ser. Springer Series in Operations Research and Financial Engineering. New York, NY, USA: Springer, 2006.
[59] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Hoboken, NJ, USA: Wiley, Jun. 2004.

Hannes Pessentheiner (S'15) was born in Austria, in 1985. He received the B.Sc. and M.Sc. degree (*cum laude*) in electrical engineering and audio engineering from the Graz University of Technology and the University of Music and Performing Arts, Graz, Austria, in 2011. Since 2012, he is a Ph.D. student attending the Doctoral School of Information and Communications Engineering, a Research Fellow, and a Teaching Associate at the Signal Processing and Speech Communication Laboratory (SPSC), Graz University of Technology. From 2013 to 2014, he was a Visiting Scholar at University of California, San Diego, CA, USA. His research interests include audio, array, and digital signal processing, adaptive systems, as well as close-talking and distant speech recognition.

Martin Hagmüller (S'04-M'10) was born in Austria, in 1974. He received the Dipl.-Ing. degree in sound engineering in 2001 and the Dr. techn. degree in electrical engineering in 2009 from the Graz University of Technology, Austria. He is Senior Scientist at the Signal Processing and Speech Communication Laboratory at Graz University of Technology, Austria. His research interests include speech signal processing, speech recognition, and disordered speech enhancement. He is currently the Deputy General Secretary of the Austrian Acoustics Association.

Gernot Kubin (S'84-M'91) received the Dipl.-Ing. degree in 1982 and the Dr.techn. degree (*sub auspiciis praesidentis*) in 1990 both in electrical engineering from Vienna University of Technology, Vienna, Austria. At Graz University of Technology, he has served as the Dean of Studies in Electrical and Audio Engineering 2004-2007, a Coordinator of the Doctoral School in Information and Communications Engineering, since 2007, and the Chair of the Senate 2007-2010 and since 2013. He is a Professor and the Founding Director of the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria, since 2000. Earlier international appointments include: CERN Geneva 1980, Vienna University of Technology 1983-2000, Erwin Schroedinger Fellow at Philips Research Labs Eindhoven 1985, AT&T Bell Labs Murray Hill 1992-1993 and 1995, KTH Stockholm 1998, Global IP Sound Stockholm 2000-2001 and San Francisco 2006, UC San Diego 2006, Danang UT 2009, and TU Munich 2015. He has held leading positions in the Vienna Telecommunications Research Centre FTW 1999-2016, the Christian Doppler Laboratory for Nonlinear Signal Processing 2002-2010, the Competence Network for Advanced Speech Technologies 2006-2010, the COMET Excellence Projects Advanced Audio Processing 2008-2013 and Acoustic Sensing and Design since 2013, the FWF National Research Network on Signal and Information Processing in Science and Engineering 2008-2011, and the Graz University of Technology Lead Project Dependable Internet of Things since 2016. He has co-authored more than 160 peer-reviewed publications and advised over 30 Ph.D. students. His research interests include nonlinear signal processing as well as speech and audio communication. He is a Member of the Board, Austrian Acoustics Association since 2000, an elected member of the IEEE Speech and Language Processing Technical Committee since 2011, and an elected member of the Speech Acoustics and Speech Processing committees of the German Information Technology Society since 2015. He received the 2015 Nikola Tesla medal for the highest number of patents awarded to a Graz University of Technology scientist in five years.