EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Joint estimation of pitch and direction of arrival: improving robustness and accuracy for multi-speaker scenarios

Stephan Gerlach[1,4*], Jörg Bitzer[1,2], Stefan Goetze[1,4] and Simon Doclo[3,4]

Abstract

In many speech communication applications, robust localization and tracking of multiple speakers in noisy and reverberant environments are of major importance. Several algorithms to tackle this problem have been proposed in the last decades. In this paper, we propose several extensions to a recently presented joint direction of arrival (DOA) and pitch estimation method, increasing its robustness in multi-speaker scenarios, noise, and reverberation. First, a spectral comb filter is added to the original algorithm to better cope with concurrent speakers. Second, the well-known generalized cross-correlation with phase transform (GCC-PHAT) is used as an additional weighting function to improve the DOA estimation accuracy in terms of correct hits. Third, using multiple microphone pairs, the multi-channel cross-correlation approach is incorporated to improve the robustness against noise and reverberation. In order to improve tracking for moving and even intersecting speakers, a particle filter is used. Experiments with real-world recordings in realistic acoustic conditions show that the proposed extensions increase the DOA hit rate by about 33% compared to the original algorithm for two step-wise moving sources at a signal-to-noise ratio (SNR) of 15 dB and a reverberation time RT_{60} of 560 ms.

Keywords: Joint DOA and pitch estimation; Spectral comb; GCC-PHAT; Multi-channel cross-correlation; Particle filter

1 Introduction

Automatic detection, localization, and tracking of speaker are of high interest in several applications such as hands-free speech communication and video conferencing, as well as for computational auditory scene analysis and human-machine interfaces. For example, in current high-quality video-conferencing systems, the users are typically not located close to the microphones, and furthermore, several users may be talking simultaneously.

To distinguish between multiple concurrent speakers, it is desirable to be able to differentiate between their directions of arrival (DOAs) and their voice characteristics. This information can then be used to, e.g., enhance automatic speech recognition, indicate active speakers, steer the camera of a video-conferencing system, or to suppress undesired acoustic disturbances.

*Correspondence: stephan.gerlach@idmt.fraunhofer.de
[1]Project Group Hearing, Speech and Audio Technology (HSA), Fraunhofer Institute for Digital Media Technology (IDMT), Oldenburg 26129, Germany
[4]Cluster of Excellence Hearing for All, Oldenburg 26129, Germany
Full list of author information is available at the end of the article

A common method for DOA estimation is to first estimate the time difference of arrival (TDOA) between different microphone pairs. An overview of these methods, as well as related references, can be found in [1,2]. A well-known TDOA estimation method is the generalized cross-correlation with phase transform (GCC-PHAT), first introduced in [3] and intensively investigated for speech signals in, e.g., [4-7]. The dual delay line algorithm in [8] is another method to estimate the azimuths of sound sources by analyzing the coincidences along two-channel delay-line pairs. Other methods for DOA estimation are based on blind channel identification, such as the adaptive eigenvalue decomposition algorithm (AEDA) [9,10], or subspace methods such as multiple signal classification (MUSIC) [11]. Another category of DOA estimation algorithms are energy-based methods which only use the measured signal energy at each microphone [12], or combined methods using both TDOA and energy information [13,14].

The spectro-temporal characteristics of speech signals, e.g., the fundamental frequency (pitch), can also be

analyzed to distinguish between concurrent speakers [15]. Traditional pitch estimation methods are based on, e.g., zero crossing rate analysis, detection of harmonics in the autocorrelation function, and cepstrum analysis [16]. Recently, a pitch estimation filter with amplitude compression (PEFAC) in the spectral domain has been proposed in [17], and methods for joint pitch and model order estimation have been proposed in [18-20]. Multi-pitch estimation has also become a topic of research, and several approaches are summarized in [21].

DOA and pitch estimation are typically treated separately, and only a couple of attempts have been made for joint DOA and pitch estimation. A possible solution is a two-step approach. In the first step, the position of a source is estimated from (multiple) microphone pairs. In the second step, the microphone signals are combined using a beamformer to obtain a single-channel output signal which is used to estimate the pitch using a single-channel pitch estimation method. In [22], a joint position and pitch (PoPi) estimation method has been proposed which is based on either cross-correlations or cross-power spectral densities (CPSDs). Several extensions have been proposed using cepstral weighting [23], gammatone-like weighting [24], time-domain GCC-PHAT replacement [25], particle filtering [26], and speaker-dependent subgrouping [27]. In [28], a different method based on a recurrent timing neural network is used for joint DOA and pitch estimation. Other methods make use of the 2D-Capon method [29,30], a subspace approach termed as multi-channel multi-pitch harmonic MUSIC (MC-HMUSIC) [31], a minimum variance distortionless response (MVDR) beamformer [32] which additionally estimates the model order to determine the number of harmonics of the source signal, and a non-linear least squares (NLS)-based method [33], all using a harmonic signal model to jointly estimate the DOA and pitch. When jointly estimating DOA and pitch, the parameter estimation typically mutually benefits from each other. Although these joint estimation methods perform quite well for clean speech signals (i.e., without noise and no reverberation), their performance typically degrades considerably in adverse acoustic environments.

The focus of this paper is to improve joint DOA and pitch estimation for multiple speakers in terms of accuracy and robustness in realistic acoustic situations. We have taken the CPSD-based method proposed in [22] combined with cepstral weighting [23], gammatone-like weighting [24], and a subsequent particle filtering [26] as the core algorithm, and we propose several extensions to improve both accuracy and robustness in this paper. As a first extension, a frequency-domain comb filter is introduced to improve the performance for simultaneously active speakers. As a second extension, a GCC-PHAT weighting function is introduced, resulting in an improved DOA estimation accuracy. As a third extension, instead of simply averaging the multiple microphone pair results, the multi-channel cross-correlation (MCCC) method, presented in [34], is adapted to the joint DOA and pitch estimator, leading to a robustness improvement especially for noisy conditions.

This paper is structured as follows: In Section 2, we introduce the core algorithm for joint DOA and pitch estimation and describe each of the proposed extensions. In Section 3, the core algorithm and its extensions are evaluated for different amounts of reverberation and signal-to-noise ratios (SNRs). Finally, the paper concludes with the most relevant findings from the proposed extensions in Section 4.

2 Algorithm

Figure 1 gives an overview of the complete proposed algorithm, depicting the different processing steps which can be divided into three parts (pre-, main, and post-processing). The proposed extensions are highlighted by gray-shaded areas. Since we are interested in joint DOA and pitch estimation, the main feature of the algorithm is the computation of a two-dimensional (2D) pattern for DOA and pitch. As core algorithm, the CPSD-based method proposed in [22] combined with cepstral weighting [23] and gammatone-like weighting [24] is used. To enable speaker tracking, a subsequent particle filter [26] is also part of the core algorithm. In Section 2.1, we introduce the considered scenario and notation. The core algorithm is described in detail in Section 2.2, while each of our extensions is explained separately in Section 2.3.

2.1 Scenario and notation

We consider a acoustic scenario where Q speech sources are recorded using M microphones in a noisy and reverberant environment. The ith microphone signal $y_i[k]$, with k the discrete time index, is first transformed to the frequency domain using the short-time Fourier transform (STFT), i.e.,

$$\mathsf{y}_i[n,\lambda] = \mathrm{STFT}\{y_i[k]\}, \qquad i = 1 \dots M, \tag{1}$$

with frequency index $n = 1 \dots N$ and frame index λ. The STFT spectra can be modeled as

$$\mathsf{y}_i[n,\lambda] = \underbrace{\mathbf{h}_i^{\mathrm{T}}[n,\lambda]\,\mathbf{s}[n,\lambda]}_{\mathsf{x}_i[n,\lambda]} + \mathsf{v}_i[n,\lambda] \tag{2}$$

$$\mathsf{x}_i[n,\lambda], \tag{3}$$

where $\mathbf{h}_i[n,\lambda] = \left[\mathsf{h}_{i1}[n,\lambda], \dots, \mathsf{h}_{iQ}[n,\lambda]\right]^{\mathrm{T}}$ denotes the acoustic transfer functions between the speech sources $\mathbf{s}[n,\lambda] = \left[\mathsf{s}_1[n,\lambda], \dots, \mathsf{s}_Q[n,\lambda]\right]^{\mathrm{T}}$ and the ith microphone, and $\mathsf{x}_i[n,\lambda]$ and $\mathsf{v}_i[n,\lambda]$ represent the speech and noise components in the ith microphone signal, respectively. The superscript T denotes the transpose operation.

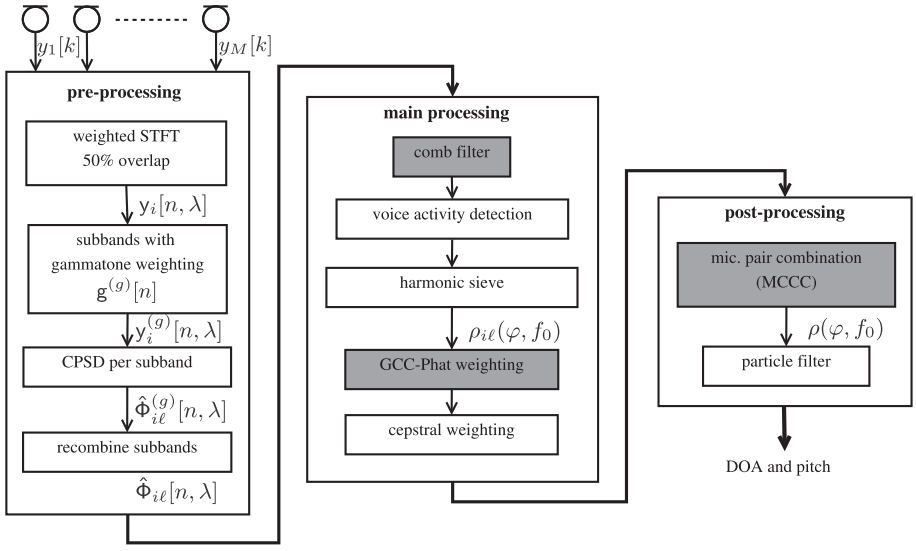

Figure 1 Overview of the proposed method. Gray-shaded areas are the extensions proposed in this paper. The signal-flow path starts with the microphone signals and continues through the pre-processing followed by the main processing and is completed with the calculation of a two-dimensional pattern $\rho(\varphi, f_0)$ in the post-processing. A particle filter is used on this 2D pattern to determine the DOA and pitch estimate.

Each acoustic transfer function $\mathsf{h}_{iq}[n,\lambda]$ can be expressed as

$$\mathsf{h}_{iq}[n,\lambda] = A_{iq}[n,\lambda]\, e^{-j\psi_{iq}[n,\lambda]}, \qquad q = 1 \ldots Q, \tag{4}$$

where $A_{iq}[n,\lambda]$ and $\psi_{iq}[n,\lambda]$ represent the amplitude and the phase of the acoustic transfer function, respectively.

As proposed in [24], a subgrouping of the spectra $\mathsf{y}_i[n,\lambda]$ is applied, in order to improve multi-speaker detection, i.e.,

$$\mathsf{y}_i^{(g)}[n,\lambda] = \mathsf{y}_i[n,\lambda] \cdot \mathsf{g}^{(g)}[n], \qquad g = 1 \ldots G, \tag{5}$$

where $\mathsf{y}_i^{(g)}[n,\lambda]$ denotes the weighted spectrum, and the superscript g indicates the frequency group number, which results in G (partially overlapping) spectra. We used a gammatone-like weighting function $\mathsf{g}^{(g)}[n]$ as depicted in Figure 2.

In addition, the CPSD

$$\Phi_{i\ell}^{(g)}[n,\lambda] = \mathrm{E}\left\{\mathsf{y}_i^{(g)}[n,\lambda]\ \mathsf{y}_\ell^{(g)*}[n,\lambda]\right\} \tag{6}$$

between the ith and the ℓth microphone is computed for each subgroup g, where $\mathrm{E}\{\cdot\}$ denotes the expectation operator, and complex conjugate terms are marked by the operator $(\cdot)^*$. In practice, the CPSD is estimated using a recursive smoothing procedure corresponding to a first-order low-pass filter [35], i.e.,

$$\hat{\Phi}_{i\ell}^{(g)}[n,\lambda] = \alpha\hat{\Phi}_{i\ell}^{(g)}[n,\lambda-1] + (1-\alpha)\mathsf{y}_i^{(g)}[n,\lambda]\ \mathsf{y}_\ell^{(g)*}[n,\lambda], \tag{7}$$

where the symbol $\hat{(\cdot)}$ indicates an estimated value, and $0 \leq \alpha < 1$ is a smoothing factor. Please note that in our case, the CPSD calculation in Equation 7 is performed in G subspectra. Afterwards, the CPSDs are normalized by the maximum of each subspectrum and recombined, i.e.,

$$\hat{\Phi}_{i\ell}[n,\lambda] = \frac{1}{G}\sum_{g=1}^{G}\frac{\hat{\Phi}_{i\ell}^{(g)}[n,\lambda]}{\max_n\left\{\left|\hat{\Phi}_{i\ell}^{(g)}[n,\lambda]\right|\right\}}, \tag{8}$$

where $\max_n\{\cdot\}$ denotes the maximum operator over index n. The normalization of each subspectrum in Equation 8 attempts to emphasize all speech source components in the recombined representation, as described in [24]. This is because in multi-speaker scenarios, harmonic speech sources have a different influence on the subspectra, and the narrowband CPSD $\hat{\Phi}_{i\ell}^{(g)}[n,\lambda]$ may be dominated by different signal components.

A CPSD-based voice activity detection (VAD) [36] is used to determine speech segments. Only the time frames in which voice activity has been detected are considered in the following processing. Please note that in the remainder of this paper, we will omit the frame index λ for simplification where it is not needed.

2.2 Joint DOA and pitch estimation

Assuming free field condition, plane waves, and a single source signal $\mathsf{s}_1[n]$ impinging with DOA φ on a uniform linear array (ULA), as shown in Figure 3a,b, the relationship between the ith and lth microphone signal is equal to

$$\mathsf{x}_i[n] = \mathsf{x}_\ell[n]\, e^{-j\psi_{i\ell}[n]} \tag{9}$$

$$\psi_{i\ell}[n] = 2\pi f_n \frac{d_{i\ell}\cdot\cos(\varphi)}{c}, \tag{10}$$

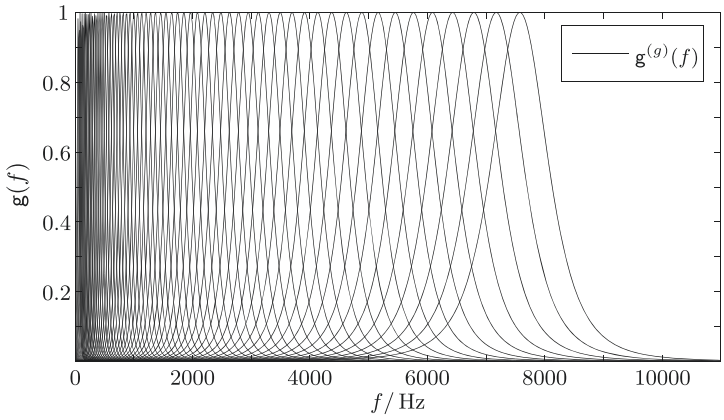

Figure 2 Sixty-four gammatone-like weighting functions.

where $\psi_{i\ell}$ describes the phase, depending on the center frequency f_n at frequency index n, the distance $d_{i\ell}$ between microphones i and ℓ, and the speed of sound c. Figure 3b depicts the definition of the DOA relative to the microphone array used throughout this paper. Without additional noise, the CPSD can then be understood as

$$\hat{\Phi}_{i\ell}[n,\lambda] = \alpha\hat{\Phi}_{i\ell}[n,\lambda-1] + (1-\alpha)\,|\mathrm{y}_i[n]|^2\, e^{j\psi_{i\ell}[n]}. \tag{11}$$

For the joint DOA and pitch estimation, a 2D DOA/pitch pattern will be computed using the CPSD [22]. Only voiced signals will be considered as relevant sources, where it is assumed that these speech signals consist of a fundamental frequency f_0 (pitch) and multiple harmonics. We use a harmonic sieve in order to estimate the pitch of the speech signal. This is shown in Figure 4, where the underlying concept of a harmonic sieve is presented, assuming different pitch values up to the fourth harmonic. The indices of the analyzed frequency bins of the harmonic sieve are defined as

$$n_p = \underbrace{\left\lfloor p \cdot \frac{f_0}{f_\mathrm{s}} \cdot N + 0.5 \right\rfloor}_{\text{round}}, \qquad p = 1 \ldots P, \tag{12}$$

where p denotes the harmonic, N is the frame size, and f_s is the sampling frequency. Only signal components corresponding to the harmonic sieve will be considered for the estimation. The harmonic sieve is computed for all values in the considered pitch range. For the exemplary harmonic sieve in Figure 4, the third example ($f_0 = 200$ Hz) would result in the best estimate, since the pitch of the signal and the harmonic sieve match. In [22], two different types of harmonic sieves are proposed, either based on cross-correlation or based on the CPSD. In this paper, we will only consider the CPSD-based version.

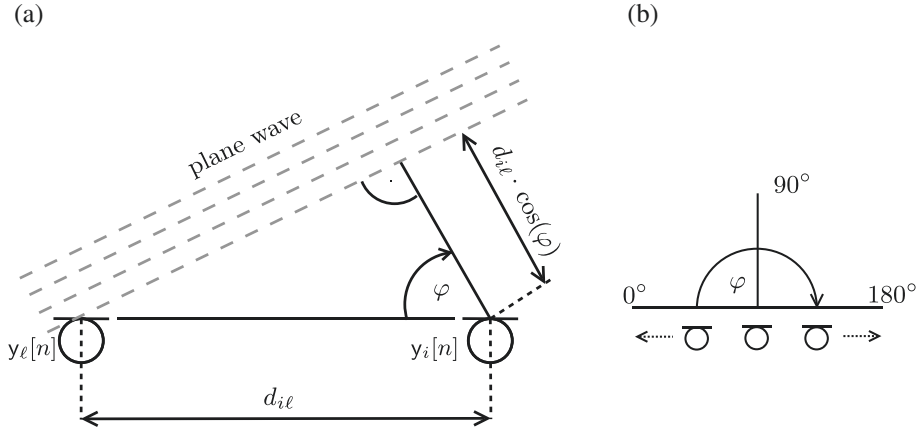

Figure 3 Direction of arrival. (a) Geometrical interpretation of the relation between direction of arrival φ and distance $d_{i\ell}$ between microphones i and ℓ, assuming a single speech source and a plane sound wave in free field condition. **(b)** Definition of the direction of arrival φ relative to the microphone array.

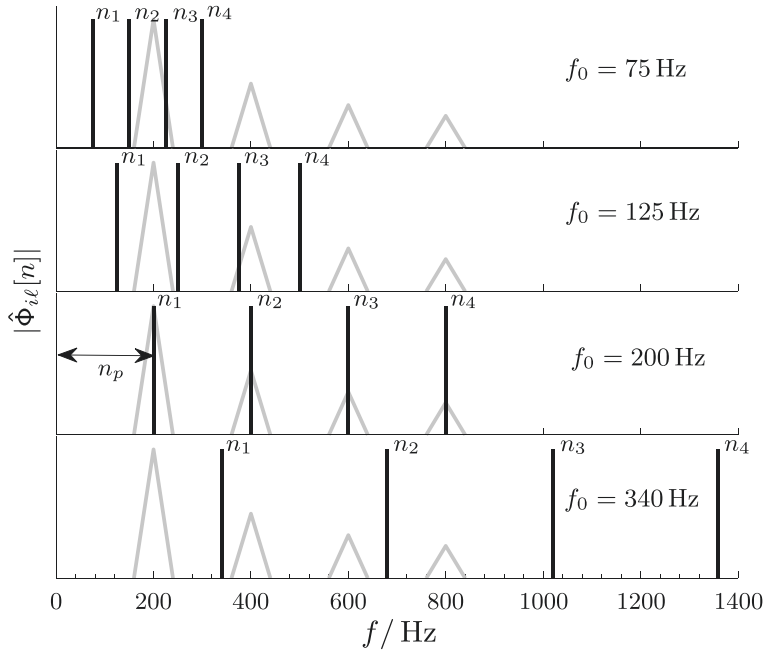

Figure 4 Harmonic sieve with four different pitch values f_0 up to the fourth harmonic ($P = 4$; black lines). The exemplary harmonic signal consists of equally spaced triangles. Best estimation results would be achieved with example 3 ($f_0 = 200$ Hz) where the pitch of the signal and harmonic sieve match.

In addition to pitch estimation, the DOA estimation is performed by analyzing the phase $\psi_{i\ell}[n]$ of the CPSD. To this end, the harmonic sieve is applied to the recombined CPSD in Equation 8, where the amplitude $|\hat{\Phi}_{i\ell}[n]|$ and the phase $\psi_{i\ell}[n]$ are treated differently to obtain the 2D DOA/pitch pattern $\rho_{i\ell}(\varphi, f_0)$ as follows:

$$\rho_{i\ell}(\varphi, f_0) = \sum_{p=1}^{P} \left|\hat{\Phi}_{i\ell}\left[n_p\right]\right| \cdot \mathrm{T}\left\{\hat{\psi}_{i\ell}\left[n_p\right] - \psi_{i\ell}^{0}\left[n_p\right]\right\} \tag{13}$$

$$\psi_{i\ell}^{0}\left[n_p\right] = p \cdot 2\pi f_0 \frac{d_{i\ell} \cdot \cos(\varphi)}{c} \tag{14}$$

$$\hat{\psi}_{i\ell}\left[n_p\right] = \arg\left\{\hat{\Phi}_{i\ell}\left[n_p\right]\right\}, \tag{15}$$

where $\hat{\psi}_{i\ell}\left[n_p\right]$ denotes the phase of the CPSD, and $\psi_{i\ell}^{0}\left[n_p\right]$ denotes the expected phase for a combination of pitch f_0 and DOA φ. The sum is taken over the P discrete frequency bins n_p belonging to the harmonic sieve. The amplitude $|\hat{\Phi}_{i\ell}\left[n_p\right]|$ encodes pitch information due to the harmonic multiples of f_0, whereas DOA information is encoded in the phase $\hat{\psi}_{i\ell}\left[n_p\right]$. The result for all considered combinations of pitch values f_0 and DOA values φ are stored in the 2D pattern $\rho_{i\ell}(\varphi, f_0)$. For computational efficiency, the values n_p and $\psi_{i\ell}^{0}\left[n_p\right]$ can be calculated beforehand and stored in look-up tables. Figure 5 shows the magnitude and phase spectrum of the harmonic sieve filter for a speech signal. The example depicts the case in which the harmonic sieve fits to the pitch of the speaker.

The operator $\mathrm{T}\{\cdot\}$ in (13) can be considered as an additional phase transform. Different phase transforms $\mathrm{T}\{\cdot\}$ are possible in order to enhance the 2D pattern $\rho_{i\ell}(\varphi, f_0)$, which are all real-valued, even, and 2π periodic functions [22]. These transforms increase the impact of the phase weighting on the harmonic sieve (cf. Equation 13). The transform used in this contribution is the one proposed in [22], i.e.,

$$\mathrm{T}\{\chi\} = \frac{1}{1 + \beta - \cos(\chi)}. \tag{16}$$

For χ, we use the mismatch between $\psi_{i\ell}^{0}\left[n_p\right]$ and $\hat{\psi}_{i\ell}\left[n_p\right]$ as stated in Equation 13, where the parameter $0 < \beta \leq 1$ affects the width of the preferred direction. A small mismatch from 0 or a multiple of 2π causes a large weighting factor. Accordingly, a large mismatch in χ leads to a small weighting factor. Hence, if the pair φ, f_0 corresponds to a source, the amplitude $|\hat{\Phi}_{i\ell}\left[n_p\right]|$ is weighted more. Figure 6 depicts the case in which $\psi_{i\ell}^{0}\left[n_p\right]$ corresponds to the measured phase $\hat{\psi}_{i\ell}\left[n_p\right]$, and the transform $\mathrm{T}\{\cdot\}$ is large for the analyzed frequency bins (marked by vertical dashed lines).

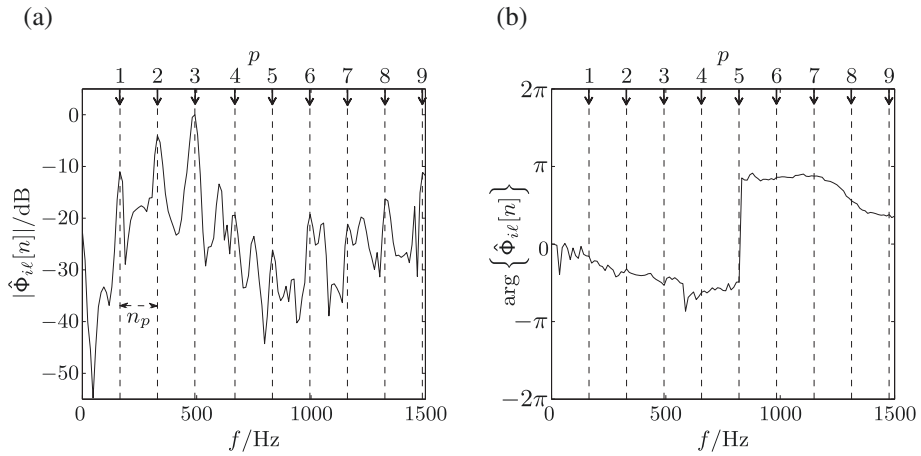

Figure 5 CPSD of a speech signal with a pitch of $f_0 = 164$ Hz. (a) Amplitude. **(b)** Phase. The dotted lines represent a harmonic sieve filtering at the discrete frequency positions n_p with $p = 1 \ldots P$ and $P = 9$. In this case, the pitch of the the signal and the harmonic sieve match.

A cepstral weighting (cf. Figure 1) of the 2D pattern $\rho_{i\ell}(\varphi, f_0)$, based on the cepstrum of the cross-correlation, was proposed in [23] to further increase the pitch estimation for disturbed input signals. The cepstrum is computed on the inverse STFT of the logarithm of the amplitude of the spectrum $\mathsf{y}_i[n]$. This transform leads to an additive representation of the signal components rather than a multiplicative one in the superimposed spectrum [33]. Thus, the so-called quefrency [35] for a dominant peak can be interpreted as a pitch candidate and the pitch relevant part of the cepstrum can be used as a weighting function.

In the post-processing, a particle filter is applied to the 2D pattern $\rho(\varphi, f_0)$, combined of DOA and pitch estimate. The particle filter tries to represent an unknown probability function by using a sequential Monte Carlo simulation with a set of particles and respective probabilities. The particles $\nu^u[\lambda]$ for frame λ incorporate the DOA $\varphi^u[\lambda]$, angular velocity $\omega^u[\lambda]$, and pitch $f_0^u[\lambda]$, i.e.,

$$\nu^u[\lambda] = \left[\varphi^u[\lambda], \omega^u[\lambda], f_0^u[\lambda]\right], \qquad u = 1 \ldots U, \quad (17)$$

where U denotes the total number of particles.

Each particle $\nu^u[\lambda]$ has a weight ξ^u representing its probability. The evolution of the particles can be described in two stages. First, the state of a particle is predicted using the particle $\nu^u[\lambda - 1]$ from the previous

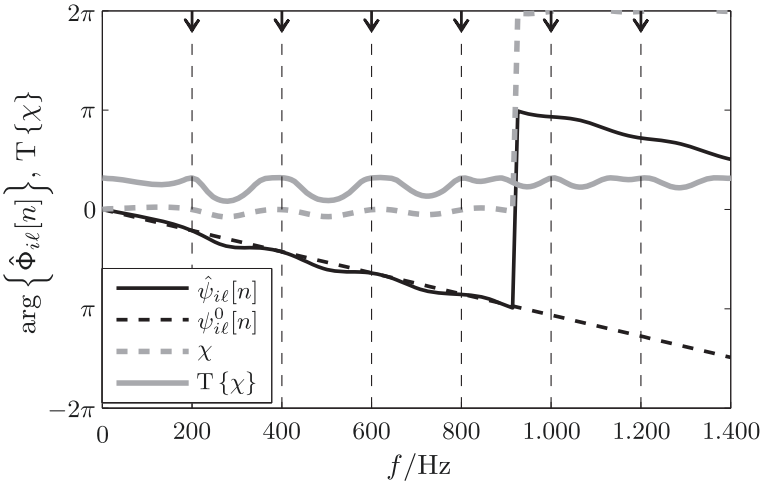

Figure 6 Phase transform $\mathrm{T}\{\chi\}$, in case when measured (solid black) and expected (dashed black) phase are close. The gray dotted line is the difference $\chi = \hat{\psi}_{i\ell}[n] - \psi^0_{i\ell}[n]$ between both phases. The transform $\mathrm{T}\{\chi\}$ (solid gray) produces a high value at frequency bins relevant for the harmonic sieve (assuming matching phases) and furthermore acts as an unwrapping function.

frame, taking into account possible physical restrictions. The pitch change is described using

$$f_0^u[\lambda] = f_0^u[\lambda-1] + \frac{N}{f_s} \cdot \beta_f \cdot r, \tag{18}$$

where r is a Gaussian distributed random variable, and β_f is the pitch shift prediction value. In addition, it is assumed that the pitch only changes within a certain range. The DOA change is described using the so-called *Langevin Model* [37], i.e.,

$$\omega^u[\lambda] = a_\varphi \omega^u[\lambda-1] + b_\varphi r \tag{19}$$

$$\varphi^u[\lambda] = \varphi^u[\lambda-1] + \frac{N}{f_s}\omega^u[\lambda] \tag{20}$$

$$a_\varphi = e^{-\beta_\varphi \frac{N}{f_s}} \tag{21}$$

$$b_\varphi = \bar{\omega}\sqrt{1-a_\varphi^2} \cdot \frac{180^\circ}{\pi}, \tag{22}$$

where $\bar{\omega}$ is the steady-state angular velocity, and β_φ is the DOA shift prediction value. In the second stage, we use the 2D pattern $\rho(\varphi, f_0)$ as a pseudo-likelihood function to determine the weights ξ^u, i.e.,

$$\xi^u = \rho\left(\varphi^u[\lambda], f_0^u[\lambda]\right), \tag{23}$$

where subsequently, the sum of all weights is normalized to unity such that $\sum_{u=1}^{U} \xi^u = 1$. The final DOA and pitch estimate for time frame λ is obtained by summing all weighted particles, i.e.,

$$\tilde{\varphi}[\lambda] = \sum_{u=1}^{U} \xi^u \cdot \varphi^u[\lambda] \tag{24}$$

$$\tilde{f}_0[\lambda] = \sum_{u=1}^{U} \xi^u \cdot f_0^u[\lambda]. \tag{25}$$

To avoid the so-called degeneracy problem, we use the systematic resampling approach, as proposed in [38], and an additional module for removal and addition of particles, as proposed in [26]. The advantages of a particle filter compared to a simple maximum search is based on its inherent tracking capabilities and its robustness against reverberation [37]. This is because in adverse conditions, the 2D pattern $\rho(\varphi, f_0)$ does not exhibit a clear main peak at the source positions, but instead a fuzzy sometimes biased area with multiple peaks is observable. The particle filter can be considered as a self-adapting smoothing function of the estimate due to the predicted source behavior and the imposed physical restrictions of this behavior.

2.3 Methods to increase the robustness

For a single speaker scenario and clean speech recordings, the basic DOA and pitch estimation algorithm in [23] performs quite well. However, its performance decreases in noisy and reverberant conditions as well as in multi-speaker scenarios. Different extensions have been proposed in [23-27] to increase the robustness of the algorithm in various aspects. The above stated subgrouping of the spectra (cf. Equation 5) as well as the already mentioned cepstral weighing [23], both part of the core algorithm and discussed in Section 2.2, are two of these extensions.

In the following sections, we will explain three novel extensions, namely, a spectral comb filter to better cope with concurrent speakers, a generalized cross-correlation (GCC)-phase transform (PHAT) weighting function to improve the DOA estimation accuracy, and a multi-channel cross-correlation approach to improve the robustness against noise and reverberation. The order of the extensions corresponds to their occurrence in the algorithm, cf. Figure 1.

2.3.1 Spectral comb filter

In [25], the authors observed that if more than one source is active simultaneously, a dominant source masks other concurrent sources in the CPSD $\hat{\Phi}_{i\ell}[n]$ and eventually in the 2D pattern $\rho_{i\ell}(\varphi, f_0)$. Assuming the number of sources Q is known, we propose to introduce a comb filter $\gamma[n]$ in order to suppress components of the CPSD $\hat{\Phi}_{i\ell}[n]$ corresponding to already estimated sources, i.e.,

$$\left|\hat{\Phi}'_{i\ell}[n]\right| = \left|\hat{\Phi}_{i\ell}[n]\right| \cdot \gamma[n] \tag{26}$$

$$\gamma[n] = \begin{cases} 0\,, \text{if } n \in \left[n'_p - \beta_c, n'_p + \beta_c\right], \text{ with } p = 1 \ldots P \\ 1\,, \text{else.} \end{cases} \tag{27}$$

The parameter β_c indicates the width of one tooth of the comb filter, and P denotes the number of teeth in the comb filter, which is equal to the number of considered harmonics in Equation 12. The comb filter $\gamma[n]$ only depends on already estimated pitch values $\hat{f}_0$, i.e.,

$$n'_p = \underbrace{\left\lfloor p \cdot \frac{\hat{f}_0}{f_s} \cdot N + 0.5 \right\rfloor}_{\text{round}}, \qquad p = 1 \ldots P. \tag{28}$$

Using estimated pitch values $\hat{f}_0$, the spectral comb filter is build to suppress the influence of the already estimated speech sources in the CPSD; this leads to a more robust estimation of the remaining speech sources. If the concurrent sources are not yet estimated in the current frame,

the pitch estimate from the previous timeframe $\lambda - 1$ is used. Accordingly, if there is no previous estimate available, the very first pitch estimate is determined using the unmodified $|\hat{\Phi}_{i\ell}[n]|$ in Equation 13.

For each time frame, the filtering is applied repeatedly to the original CPSD $\hat{\Phi}_{i\ell}[n]$ as often as sources are estimated. All successive processing steps, including the harmonic sieve, are repeated respectively. Figure 7 illustrates the effect of the comb filter for two concurrent sources. It can be seen that the secondary source is suppressed, whereas the target source is highlighted.

2.3.2 GCC-PHAT weighting

When using the core algorithm discussed in Section 2.2, the 2D pattern $\rho_{i\ell}(\varphi, f_0)$ exhibits a wide spread of the peaks with regard to the DOA φ. Similar to the cepstral weighting, which aims to improve the pitch estimation, we propose a second weighting function $w_{i\ell}(\varphi)$ which aims to improve the DOA estimation accuracy. This extension is derived from the GCC-PHAT algorithm [3], not used as a DOA estimator itself, but only as a weighting function of the 2D pattern $\rho_{i\ell}(\varphi, f_0)$, i.e.,

$$\rho_{i\ell}^{\text{phat}}(\varphi, f_0) = \rho_{i\ell}(\varphi, f_0) \cdot w_{i\ell}(\varphi) \tag{29}$$

$$w_{i\ell}(\varphi) = r_{i\ell}^{\text{phat}}\left[\left\lfloor \frac{d_{i\ell} \cdot \cos(\varphi) \cdot f_s}{c} \right\rfloor\right], \tag{30}$$

where $r_{i\ell}^{\text{phat}}[k]$ denotes the resampled generalized cross-correlation between the microphone signals i and ℓ using the phase transform PHAT weighting [3]. The weighting function $w_{i\ell}(\varphi)$ can be interpreted as a warped extract of the cross-correlation with respect to the DOA φ and the microphone distance. In Figure 8, the upper graph depicts an example of a complete GCC-PHAT, whereas the lower graph only shows the relevant part for the DOA estimation, which is used as a weighting function.

Please note that in [25], a different GCC-PHAT extension was proposed, in which the central part of the

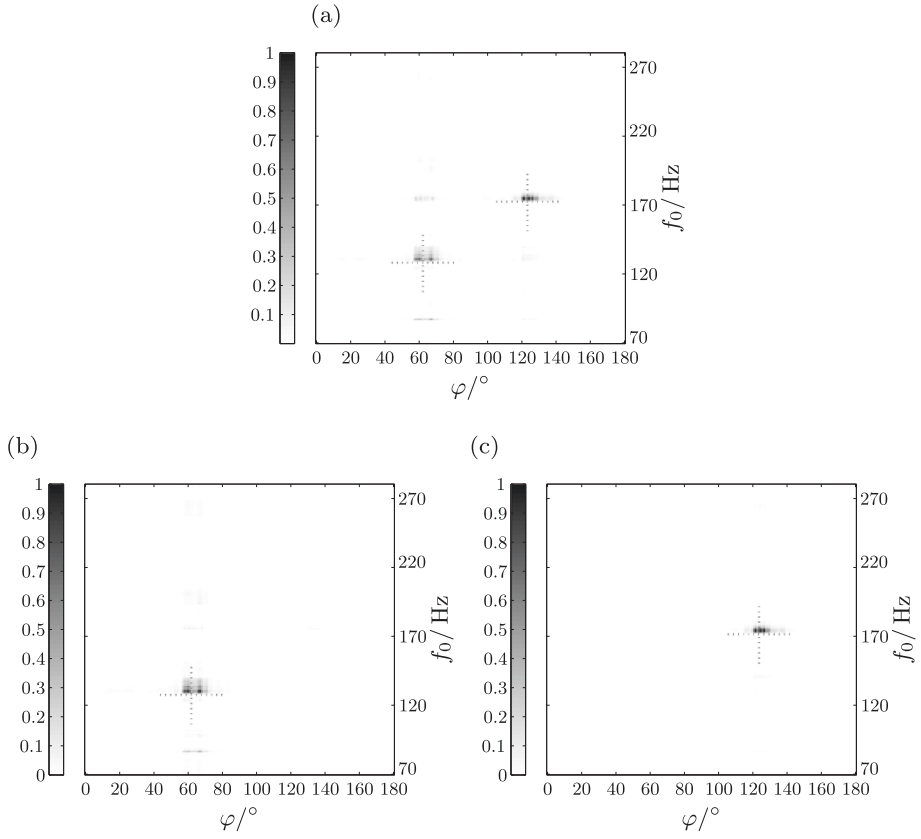

Figure 7 2D pattern $\rho(\varphi, f_0)$ of joint DOA and pitch estimation. (a) Resulting pattern for two concurrent sources with $f_{0,1} = 132$ and $f_{0,2} = 175$ Hz and DOA $\varphi_1 = 63°$ and $\varphi_2 = 124°$ without spectral comb filtering. **(b, c)** The patterns after spectral comb filtering for each of the sources. The suppressing influence of the comb filtering is clearly visible. Real recorded vowel utterances from two different speakers where used as sources. The dotted crosses indicate the true source positions. **(a)** original pattern $\rho(\varphi, f_0)$. **(b)** $\rho(\varphi, f_0)$ after comb filter for source 1. **(c)** $\rho(\varphi, f_0)$ after comb filter for source 2.

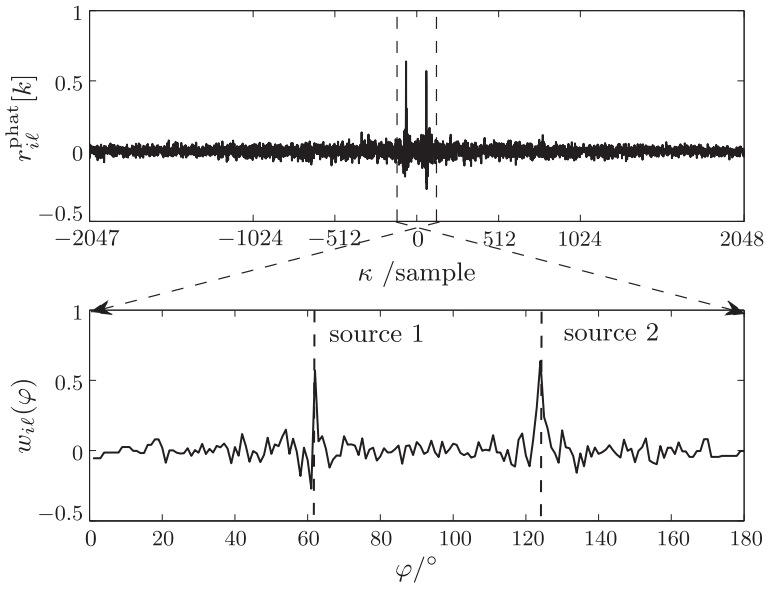

Figure 8 GCC-PHAT weighting. The upper graph depicts a complete GCC-PHAT of two speech signals at different DOA, recorded with a single microphone pair ($d_{i\ell} = 22$ cm). The lower graph depicts only the DOA relevant informations according to Equation 30, whereby the *x*-axis is transformed to DOA values. The speech sources were located at 63° and 124° relative to the microphones.

unweighted cross-correlation is replaced with the GCC-PHAT weighted cross-correlation. Afterwards, in contrast to the GCC-PHAT weighing proposed here, a time-domain-based harmonic sieve is applied to the cross-correlation to obtain the 2D pattern $\rho_{i\ell}(\varphi, f_0)$.

2.3.3 Multi-channel cross-correlation

Up to now, we have discussed methods and extensions to compute the 2D pattern $\rho_{i\ell}^{\mathrm{phat}}(\varphi, f_0)$ using one microphone pair i and ℓ. An intuitive approach to combine multiple microphone pairs is the arithmetic mean of all 2D pattern, as already performed in [22]. However, averaging is sensitive to microphone malfunctions and mutual cancelation of opposite erroneous estimates. Therefore, we introduce a more sophisticated method based on the multi-channel cross-correlation (MCCC) [34], which exploits the redundancy among multiple microphones pairs and can be understood as the generalized multi-channel extension of the cross-correlation. We adapted the MCCC to the joint DOA and pitch estimation problem, in order to generate an overall 2D pattern using multiple microphone pairs. First, a $M \times M$ matrix $\mathbf{P}(\varphi, f_0)$ with the 2D pattern of all microphone pairs is constructed, i.e.,

$$\mathbf{P}(\varphi, f_0) = \begin{pmatrix} \rho_{11}(\varphi, f_0) & \rho_{12}(\varphi, f_0) & \cdots & \rho_{1M}(\varphi, f_0) \\ \rho_{21}(\varphi, f_0) & \rho_{22}(\varphi, f_0) & \cdots & \rho_{2M}(\varphi, f_0) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{M1}(\varphi, f_0) & \rho_{M2}(\varphi, f_0) & \cdots & \rho_{MM}(\varphi, f_0) \end{pmatrix}, \tag{31}$$

which is a symmetry matrix, since $\rho_{il}(\varphi, f_0) = \rho_{li}(\varphi, f_0)$. Similarly to [34], the determinant $\det(\mathbf{P}(\varphi, f_0))$ of this matrix is subsequently used to define the overall 2D pattern, i.e.,

$$\rho(\varphi, f_0) = 1 - \det\big(\mathbf{P}(\varphi, f_0)\big). \tag{32}$$

Although it has been shown in [34] that the MCCC always lies between 0 and 1, this does not hold anymore for $\rho(\varphi, f_0)$ defined in Equation 32.

The adaptation of the MCCC algorithm shows two advantages compared to the arithmetic mean. Firstly, it is robust against malfunctions of single microphones. In case of defective microphones, only the remaining microphones are taken into account for the estimation. Secondly, if two microphone signals, highly or perfectly, match (with regard to the considered pitch and DOA combination), the overall result becomes 1, independent of the remaining microphones and opposite erroneous estimates no longer cancel themselves.

3 Evaluation

We have conducted experimental evaluations for different acoustic conditions and scenarios. Three different scenarios with increasing complexity are evaluated. In Scenario 1, two simultaneous speakers at fixed positions are simulated. In Scenario 2, the two speakers move stepwise while speaking. In the most difficult Scenario 3, the two speakers move stepwise on intersecting pathways. Measured and simulated room impulse responses (RIRs) are used to generate the microphone signals. Reverberation

times RT_{60} ranging from 0 to 560 ms, SNRs from 0 to 20 dB, and noise free simulations (SNR $= \infty$) are used. A performance comparison between the core algorithm discussed in Section 2.2 and the extensions proposed in Section 2.3 will be presented in terms of DOA estimation hit rate A_φ and pitch estimation hit rate A_f, as well as root-mean-square error (RMSE) of the DOA estimates.

A comparison with other state-of-the art joint DOA and pitch estimators (cf. Section 1) was not conducted since those algorithms assume single-source scenarios and do not support estimation of multiple sources without introducing further extensions which is beyond the scope of this paper.

3.1 Setup and performance measures

The evaluation was carried out for a conference room (cf. Figure 9) using a microphone line array with $M = 6$ microphones (inter-microphone distance 0.22 m), resulting in 15 microphone pairs. A loudspeaker was used as signal source at nine different positions with a distance of approximately 3.3 m to the microphones at a similar height of 1.21 m to the microphones. The distance between the loudspeaker positions was 0.5 m. The real RIRs were measured at a sampling rate of 48 kHz using the sine sweep method [39]. The reverberation time of the conference room is approximately 560 ms with a direct-to-reverberant ratio (DRR) of 6.2 dB.

To investigate the performance for different reverberation times, we used simulated RIRs that were generated with the image method [40,41]. The same relative microphone and loudspeaker positions were used, but inside a simulated rectangular room of size $l = 4.6\,\text{m} \times w = 5.1\,\text{m} \times h = 2.5\,\text{m}$. Reverberation times of $RT_{60} = \{0, 100, 250, \text{ and } 560\,\text{ms}\}$ with direct-to-reverberant ratios of DRR $= \{\infty, 9.1, 2.5, \text{ and } -3.6\,\text{dB}\}$ were simulated.

The clean microphone signals were generated by convolving the measured or simulated RIRs with clean speech recordings which consisted of male and female speech in German and English. Uncorrelated speech-shaped noise was used as interference signal, played back from all loudspeakers simultaneously. Speech and noise recordings were mixed at six different broadband SNR values (measured at the first microphone) ranging from −5 to 20 dB.

Only the time frames labeled to contain active speech, determined using a VAD [36], are considered for the joint DOA and pitch estimation. These time frames are not further distinguished in voiced or unvoiced speech, assuming the most dominant part of speech is voiced speech.

For the STFT processing, the frame size was set to 85 ms (4,096 samples at 48 kHz sampling rate) using a von Hann window with an overlap of 50%. The resulting spectrum was subdivided in $G = 64$ partly overlapping gammatone-like weighted subspectra. It should be noted that the choice of the frame size is a trade-off between frequency resolution of the harmonic sieve (cf. Equation 12) and tracking capability of the particle filter (cf. Equations 18 and 20).

The smoothing parameter for the CPSD estimation in Equation 7 was set to $\alpha = 0.1$, which is chosen quite low to better deal with simultaneous speech sources. The number of considered harmonics in Equation 12 was set to $P = 5$. In practice, the number of harmonics P is unknown or changes over time and has to be estimated [18-20,32]. The parameter β for the phase transform in Equation 16

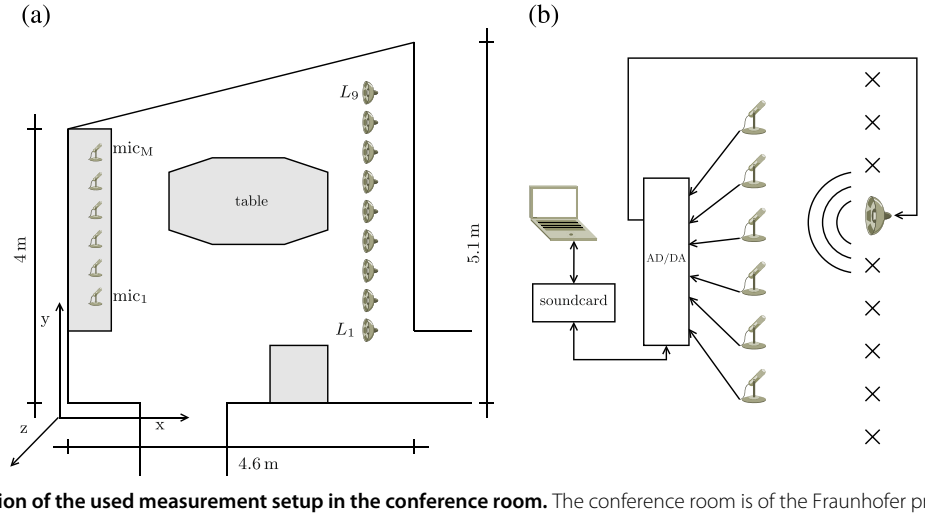

Figure 9 Description of the used measurement setup in the conference room. The conference room is of the Fraunhofer project group HSA in Oldenburg. Room impulse responses were recorded using this setup and convolved with clean speech recordings used for the evaluation. **(a)** Measurement setup. **(b)** Measuring circuit.

was set to $\beta = 0.2$. The teeth width of the proposed spectral comb filter extension in Equation 27 was set to $\beta_c = 1$.

In order to generate the 2D pattern $\rho(\varphi, f_0)$ in Equation 13, the DOA values were analyzed from 0° to 180° with an interval of 1°, where 90° is perpendicular to the microphone axis (cf. Figure 3) , and pitch frequencies were analyzed from 70 to 280 Hz with an interval of 1 Hz. Due to the test setup from Figure 9, only the estimates in front of the array are considered to be valid source positions.

For the particle filter, we used $U = 50$ particles per source to simulate and track the source motion over time, where the steady-state angular velocity was set to $\bar{\omega} = 1$ rad/s; for the DOA shift prediction value, we used $\beta_\varphi = 10\ \mathrm{s}^{-1}$, and for the pitch shift prediction we used $\beta_f = 5$ Hz/s.

The resulting performance was measured in terms of hit rate A_φ (DOA) and A_f (pitch) for all processed time frames with a fault tolerance of $\Delta_\varphi = \pm 10°$ and $\Delta_f = \pm 10$ Hz compared to the true source characteristics. Results with a smaller tolerance interval, i.e., $\Delta_\varphi = \pm 5°$ and $\Delta_f = \pm 5$ Hz, have also been calculated, which lead to an overall reduced hit rate, but showing the same performance comparison between the algorithms under test. Although, it is known that beamformers can be designed to have a more narrow beam width, a tolerance interval of $\Delta_\varphi = \pm 10°$ and $\Delta_f = \pm 10$ Hz allows for more robust beamformers in case of erroneous DOA estimates.

Please note that the exact pitch of real speech signals, required to calculate the hit rate A_f, is unknown and can only be estimated. We used the overall mean of the PEFAC pitch estimate [17] of the clean speech signal as ground-truth pitch.

3.1.1 Scenario 1: two speech sources at fixed positions

In Scenario 1, we investigated the influence of each proposed extension on the hit rates A_φ and A_f separately. We chose a scenario where two persons (male and female) were simultaneously speaking at fixed positions. The signals were about 5 s long, wherein each speaker is pronouncing one sentence. These signals were processed with different extensions enabled, resulting in five different setups shown in Table 1.

We performed simulations for an SNR of 15 dB and noiseless and for reverberation times $\mathrm{RT}_{60} = 0$ and 560 ms. The results shown in Figure 10 are separated into DOA and pitch results for every source separately.

As seen from Figure 10, the core algorithm in Setup I performs moderately at SNR $= \infty$ and in an anechoic environment, but deteriorates fast in adverse conditions for both DOA and pitch estimation. Setup I seems to be particularly susceptible to reverberation.

Table 1 Setup specification of used extensions

Setup	Specification
I	Core algorithm
II	Core algorithm and spectral comb filter
III	Core algorithm and spectral comb filter and GCC-PHAT weighting
IV	Core algorithm and spectral comb filter and MCCC
V	Core algorithm and spectral comb filter and GCC-PHAT weighting and MCCC

With the spectral comb filter activated in Setup II, the two sources are estimated equally good for the scenarios without reverberation (top panel in Figure 10). Unfortunately, there is no improvement observable for the scenarios with $\mathrm{RT}_{60} = 560$ ms (bottom panel in Figure 10). Nevertheless, if the spectral comb filter is missing, as in Setup I, we can see that the algorithm preferably estimates the dominant source. Hence, in our implementation, this filter is beneficial for tracking of two sources simultaneously. Considering that we are exploring multi-source localization, in all following scenarios, the spectral comb filter will be a ctivated using Setups II to V.

Using Setup III, focusing on the GCC-PHAT weighting as our second extension, we can observe that especially the DOA estimate improves considerably in all four SNR and RT_{60} combinations, compared to Setup I. In comparison to Setups I and II, no substantial difference in the pitch estimation can be identified.

In case of low reverberation, the MCCC extension (Setup IV) also improves the DOA estimation compared to Setup I, but deteriorates strongly with larger reverberation times. The simulation results of Setup V, in which all proposed extensions are enabled, show a good overall performance for all different acoustic conditions. It seems that the GCC-PHAT weighting has the largest influence in terms of DOA estimation, in that the DOA hit rates A_φ of Setup III are equal or better than those of Setup V. However, especially for $\mathrm{RT}_{60} = 560$ ms and SNR $= 15$ dB, Setup V shows the best hit rate compared to all other setups.

Figure 11 shows the pitch estimates for all processed time frames. It can be observed that the pitch estimates are less scattered using Setup V in Figure 11b, compared to Setup II used in Figure 11a. This leads to a more robust DOA estimation even if the estimated pitch does not correspond to the true value. Again, the pitch estimation is not very accurate, but it is still beneficial in case of multi-source scenarios to improve the source differentiation.

At an SNR of 15 dB and a reverberation time RT_{60} of 560 ms, we obtained DOA hit rates of $A_\varphi = 72\%$ with Setup V, compared to DOA hit rates of 5% obtained with Setup II.

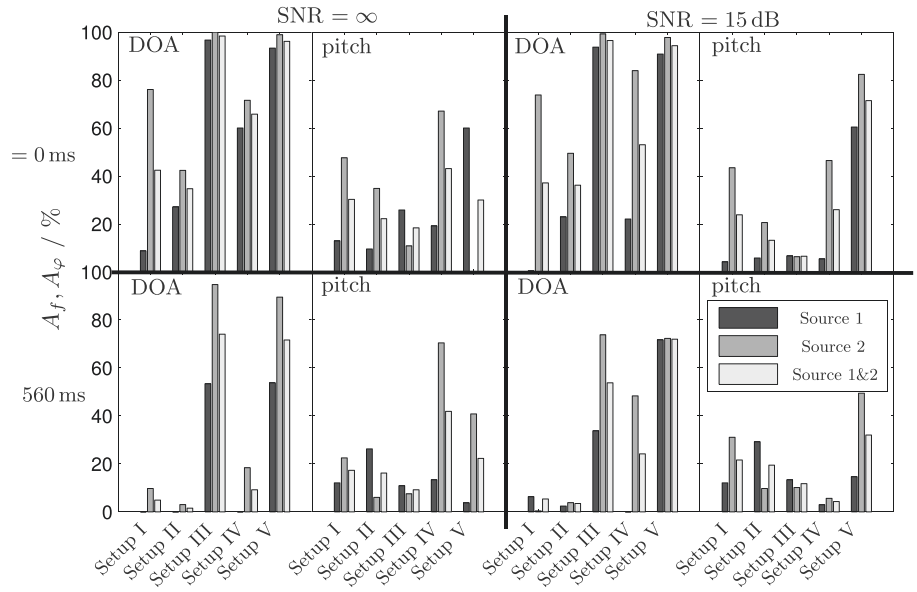

Figure 10 Hit rates in terms of pitch (A_φ) and DOA (A_f). The rates are for two simultaneous speakers at fixed positions. Left panels show the results for SNR = 0 dB; right panels shows the results for SNR = 15 dB. The top panels show the results for anechoic environment (RT_{60} = 0 ms); the bottom panels show the results for reverberant environment (RT_{60} = 560 ms). The scenarios were processed with the core algorithm only (Setup I), with single extensions activated, and up to the proposed algorithm (Setup V) (cf. Table 1).

3.1.2 Scenario 2: two stepwise moving speech sources

Since we aim at real-world scenarios, moving sources are considered in the second scenario. Two concurrent speech sources move towards one another in a stepwise manner, heading to the middle of the monitored area (cf. Figure 9). We simply switched between adjacent loudspeaker positions to simulate the movements of the sources. At every new position, the speaker repeated the same sentence. The results shown in Figure 12 are calculated with Setup V (bottom panel of Figure 12) and with Setup II (top panel of Figure 12).

Figure 12(a) shows the results for several SNR conditions from 20 to −5 dB without reverberation. It is apparent that the proposed algorithm (Setup V) outperforms Setup II for all conditions, resulting in mean hit rates $\bar{A}_\varphi = 87.1\%$ for Setup V over all SNR (bottom panel in column (a)) and 55.3% for Setup II (top panel in column (a)). The proposed algorithm (Setup V) results in high hit rates even at low SNR.

Figure 12(b) shows the result for the measured RIRs with reverberation time RT_{60} = 560 ms and varying SNRs. For both setups, II and V, the hit rate decreases compared to RT_{60} = 560 ms. However, the proposed algorithm (Setup V) still outperforms Setup II by 25.7% on average for all conditions. Figure 12(c) shows the results for different reverberation times without noise

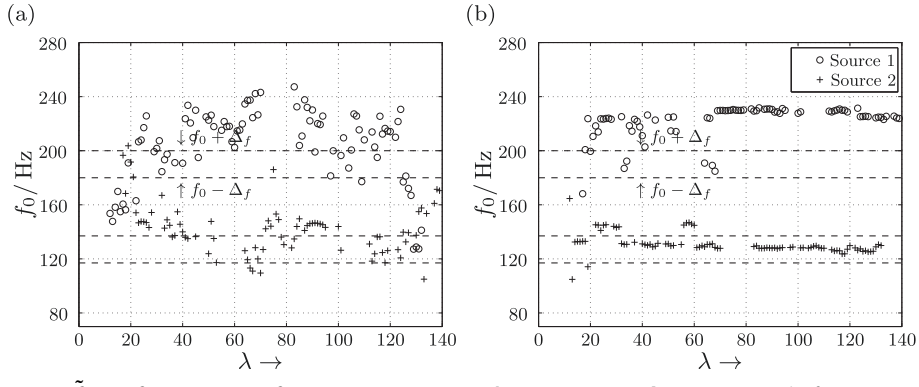

Figure 11 Pitch estimates $\tilde{f}_0[\lambda]$ (cf. Equation 25) for two concurrent speakers. (a) Setup II. **(b)** Setup V. Results for RT_{60} = 560 ms and SNR = 15 dB. The dashed lines indicate the tolerance interval around the true pitch values.

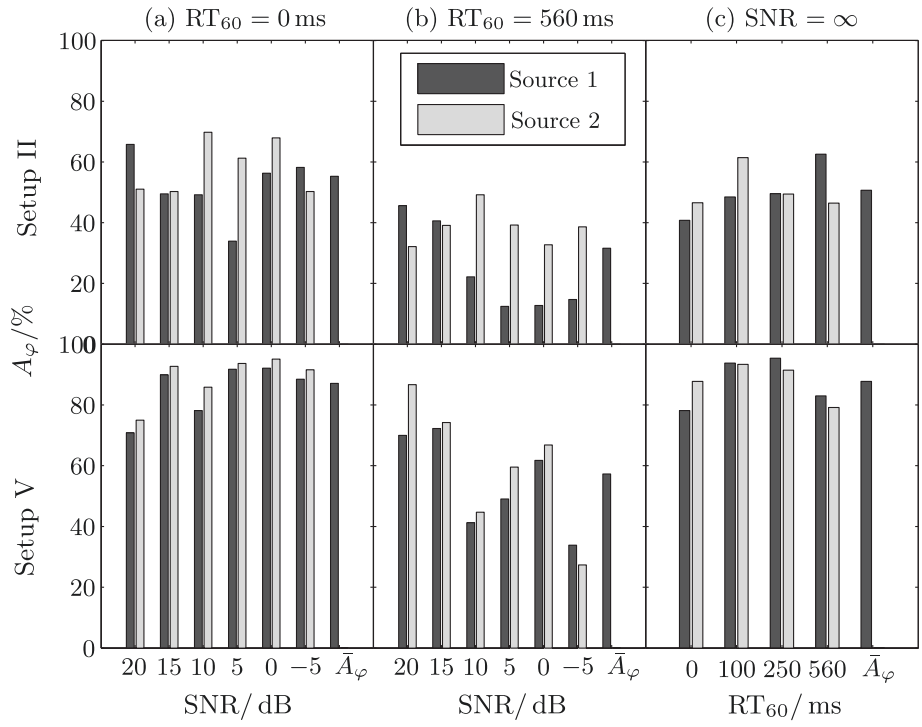

Figure 12 Hit rates A_φ for two concurrent speakers moving towards each other. The top panels show the results for the core algorithm including spectral comb filter (Setup II); the bottom panels show the results for the proposed algorithm (Setup V). Columns **(a)** and **(b)** show the results for $RT_{60} = 0$ ms and $RT_{60} = 560$ ms (measured RIR) and varying SNR values. Column **(c)** shows the results for SNR $= \infty$ and several simulated RIRs. The very right bar in each panel shows the mean hit rate $\bar{A}_\varphi$ for the respective condition.

(SNR $= \infty$). Comparing the mean hit rates $\bar{A}_\varphi$, the proposed algorithm (Setup V) achieves a hit rate of 87.7% and surpasses the core algorithm with the spectral comb filter (Setup II) by 37%. Additionally, Figure 13 shows the RMSEs of the DOA estimates for the same setups (II and V) and conditions as in Figure 12.

For $RT_{60} = 0$ ms (Figure 13(a)), Setup II achieves a mean RMSE of 15.7°, whereas the mean RMSE of Setup V decreases to 8.5°. For $RT_{60} = 560$ ms (Figure 13(b)), the mean RMSE for both setups increases to 19.6° for Setup II and to 15.9° for Setup V. For different reverberation times (Figure 13(c)), Setup V achieves a mean RMSE of 7.9°), which is 10° better than the mean RMSE of Setup II.

Figure 14 shows the DOA estimates for all processed time frames for the condition SNR $= 15$ dB and $RT_{60} = 560$ ms. It can be observed that the proposed algorithm (Setup V) achieves less scattered estimates than Setup II.

The hit rates A_φ in Scenario 2 are considerably higher than in Scenario 1, especially for the core algorithm with the spectral comb filter (Setup II). This is because wrong DOA estimates tend to be located in the frontal direction (around $\varphi = 90°$), as can be seen in Figure 14a. Due to the scenario definition, in which the sources move towards $\varphi = 90°$, it occurred that erroneous estimates are actually counted as correct hits, e.g., for time frames around $\lambda = 400$, which does not necessarily indicate a more reliable estimation but still increases the A_φ value. At an SNR of 15 dB and a reverberation time RT_{60} of 560 ms, we obtained DOA hit rates of $A_\varphi = 73\%$ with Setup V, compares to DOA hit rates of $A_\varphi = 40\%$ obtained with Setup II.

3.1.3 Scenario 3: two intersecting speech sources

The proposed algorithm is intended to be capable of tracking stepwise intersecting sources by using the particle filter. Therefore, in the third scenario, we considered two speakers on crossing paths while speaking. The intersecting source movement can be considered as the most ambitious, but also the most realistic scenario in this evaluation. The movement of the two concurrent speakers was, again, simulated with a stepwise switching between subsequent loudspeaker positions. Hence, at a certain position, the two speech signals were emitted by a single loudspeaker. Similar to Scenario 2, we performed three experiments in which either the reverberation time RT_{60} or the SNR was kept constant and the other value varied over the range of interest.

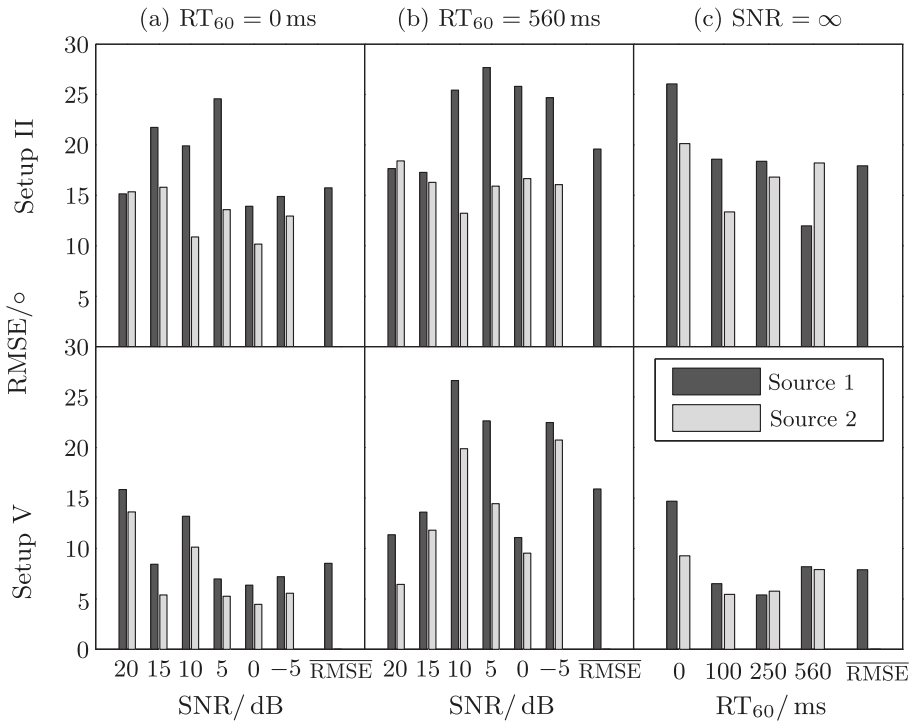

Figure 13 RMSE (in degree) for two concurrent speakers moving towards each other. The top panel shows the results for the core algorithm including spectral comb filter (Setup II); the bottom panel shows the results for the proposed algorithm (Setup V). Columns **(a)** and **(b)** show the results for $RT_{60} = 0$ ms and $RT_{60} = 560$ ms (measured RIR) and varying SNR values. Column **(c)** shows the results for SNR $= \infty$ and several simulated RIRs. The very right bar in each panel shows the mean RMSE for the respective condition.

The results for the third scenario are shown in Figure 15. It can be observed that the mean hit rate decreases slightly in all conditions, compared to the results in Scenario 2. Again, the proposed algorithm (Setup V) outperforms the core algorithm with the spectral comb filter (Setup II) in all conditions, e.g., the mean hit rate $\bar{A}_{\varphi}$ for $RT_{60} = 560$ ms (cf. Figure 15(b)) is 54.6% for Setup V and only 34.5% for Setup II.

The corresponding RMSE is shown in Figure 16. It can be observed that for all conditions, the proposed Setup V achieves a better mean RMSE, i.e., Setup II achieves mean RMSEs of 13.6° ($RT_{60} = 0$ ms), 17° ($RT_{60} = 560$ ms), and 17.2° (SNR $= \infty$), but Setup V achieves better mean RMSE of 11.3°, 15.6°, and 10.4°, respectively.

Figure 17 shows the DOA estimates for all processed time frames λ for an SNR of 15–dB and a reverbera-

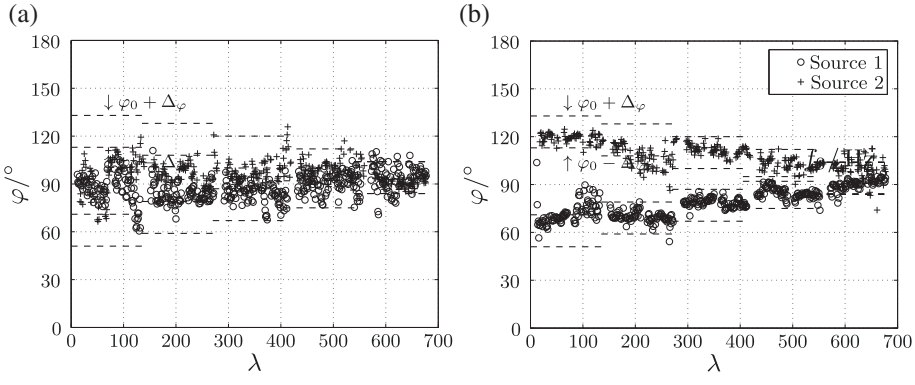

Figure 14 DOA estimation $\tilde{\varphi}[\lambda]$ with two speech sources moving towards each other. (a) Setup II. **(b)** Setup V. Results for SNR = 15 dB and $RT_{60} = 560$ ms (measured RIR). Setup V shows a significantly better performance than Setup II. The dashed lines indicate the tolerance interval around the true DOA value.

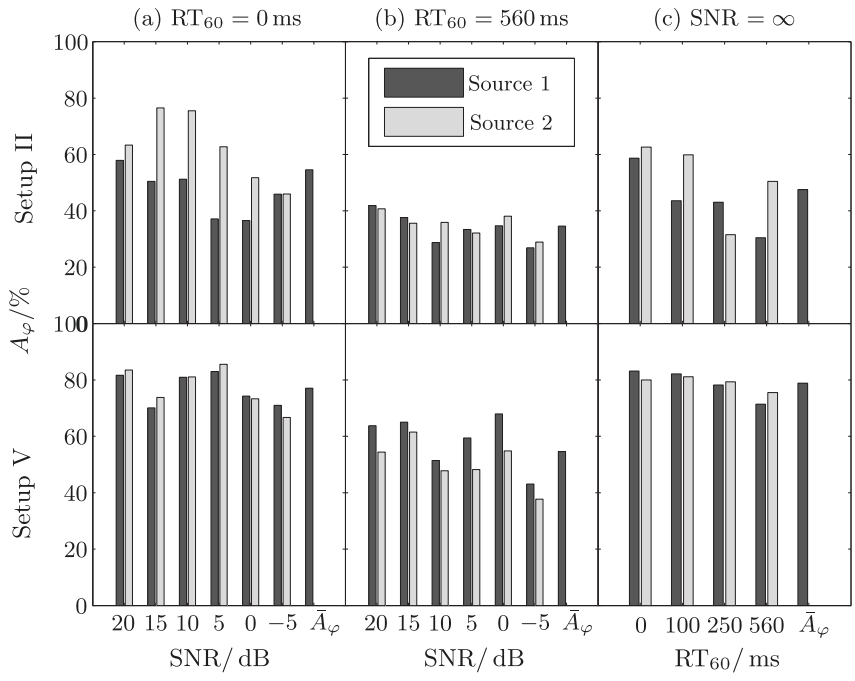

Figure 15 Hit rates A_φ for two concurrent speakers on crossing paths. The top panels show the results for the core algorithm including spectral comb filter (Setup II); the bottom panels show the results for the proposed algorithm (Setup V). Columns **(a)** and **(b)** show the results for $\mathrm{RT}_{60} = 0$ ms and $\mathrm{RT}_{60} = 560$ ms (measured RIR) and varying SNR values. Column **(c)** shows the results for $\mathrm{SNR} = \infty$ and several simulated RIRs. The very right bar in each panel shows the mean hit rate $\bar{A}_\varphi$ for the respective condition.

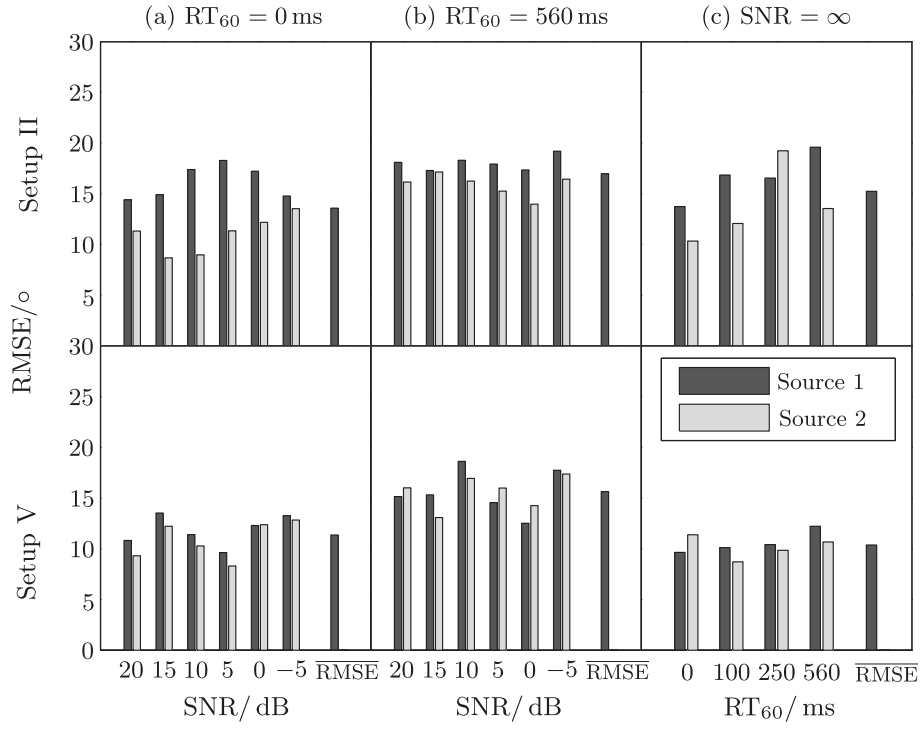

Figure 16 RMSE (in degree) for two concurrent speakers on crossing paths. The top panels show the results for the core algorithm including spectral comb filter (Setup II); the bottom panels show the results for the proposed algorithm (Setup V). Columns **(a)** and **(b)** show the results for $\mathrm{RT}_{60} = 0$ ms and $\mathrm{RT}_{60} = 560$ ms (measured RIR) and varying SNR values. Column **(c)** shows the results for $\mathrm{SNR} = \infty$ and several simulated RIRs. The very right bar in each panel shows the mean RMSE for the respective condition.

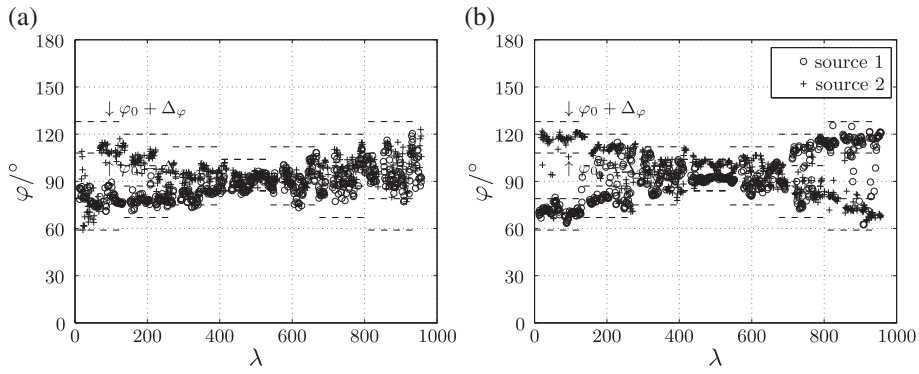

Figure 17 DOA estimate $\tilde{\varphi}[\lambda]$ for two concurrent speakers on crossing paths. (a) Setup II. **(b)** Setup V. Results for SNR = 15 dB and RT_{60} = 560 ms (measured RIR). Cross-over takes place at time frames $400 < \lambda < 550$. The proposed algorithm (Setup V) is able to properly track the movement.

tion time of 560 ms. The corresponding DOA hit rates are $A_\varphi = 63.2\%$ for Setup V and $A_\varphi = 36.6\%$ for Setup II.

Due to the spectral comb filter, it is impossible for the proposed algorithm to exactly estimate two sources coming from a single direction. Nevertheless, Figure 17 shows that the proposed algorithm is still capable to estimate the movement of intersecting speakers. In particular, when the speakers are crossing (at frames $400 < \lambda < 550$), it can be seen that the proposed algorithm estimates the sources to be in close proximity to each other; however, they never overlap.

4 Conclusion

In this paper, several extensions to the core joint DOA and pitch estimation algorithm were proposed, which were shown to increase the robustness and hit rate even for difficult acoustic situations. In particular, the generalized cross-correlation GCC-PHAT weighting achieves a considerable improvement of the DOA estimation accuracy. To cope with multi-speaker situations, the spectral comb filter was proposed, which achieves that the proposed method is less unaffected by dominant sources and more or less estimates the DOA and pitch of all sources to the same extent. Furthermore, the MCCC extension improves the robustness and accuracy and, in addition, makes the algorithm less sensitive to microphone malfunctions. Even intersecting sources can be tracked by usage of a particle filter.

At an SNR of 15 dB and a reverberation time RT_{60} of 560 ms, the proposed algorithm (Setup V) achieved DOA hit rates of $A_\varphi = 72\%, 73\%$, and 63.2% for two fixed, moving, and intersecting speech sources, respectively, compared to $A_\varphi = 5\%, 40\%$, and 36.6% achieved with the core algorithm including the spectral comb filter (Setup II).

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was partly supported by the Research Unit FOR 1732 'Individualized Hearing Acoustics', funded by the German Research Foundation (DFG), and EcoShopping 'Energy efficient & cost competitive retrofitting solutions for shopping buildings' grant no. 609180, funded by the European Commission.

Author details
[1]Project Group Hearing, Speech and Audio Technology (HSA), Fraunhofer Institute for Digital Media Technology (IDMT), Oldenburg 26129, Germany. [2]Jade University of Applied Sciences, Oldenburg 26121, Germany. [3]Department of Medical Physics and Acoustics, University of Oldenburg, Oldenburg 26111, Germany. [4]Cluster of Excellence Hearing for All, Oldenburg 26129, Germany.

Received: 6 September 2013 Accepted: 16 July 2014

References

1. J Chen, J Benesty, Y Huang, Time delay estimation in room acoustic environments: an overview. EURASIP J. Appl. Signal Proces. **26503**(1), 1–19 (2006)
2. N Madhu, R Martin, in Acoustic Source Localization with Microphone Arrays in, in *Advances in Digital Speech Transmission*, ed. by R Martin, U Heute, and Antweiler C (Wiley Chichester, UK, 2008), pp. 135–170. http://dx.doi.org/10.1002/9780470727188.ch6
3. C Knapp, G Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Processing. **24**, 320–327 (1976)
4. D Bechler, K Kroschel, Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array, in *Proceedings of the International Workshop on Acoustic Echo and Noise Cancellation (IWAENC)* (Kyoto, Japan, Sept. 2003), pp. 315–318
5. A Brutti, M Omologo, P Svaizer, Comparison between different sound source localization techniques based on a real data collection, in *Hands-Free Speech Communication and Microphone Arrays, HSCMA* (Trento, Italy, 2008), p. 69–72. doi:10.1109/HSCMA.2008.4538690
6. J Scheuing, B Yang, Correlation-based TDOA-estimation for multiple sources in reverberant environments, in *Signals and Communication Technology: Speech and Audio Processing in Adverse Environments*, ed. by E Hänsler, G Schmidt (Springer Berlin, Germany, 2008), pp. 381–416. http://dx.doi.org/10.1007/978-3-540-70602-1_11
7. B Kwon, Y Park, Y Park, Multiple sound sources localization using the spatially mapped GCC functions, in *ICROS-SICE International Joint Conference* (Fukuoka, Japan, 2009), pp. 1773–1776

8. C Liu, BC Wheeler, WD O'Brien, RC Bilger, CR Lansing, AS Feng, Localization of multiple sound sources with two microphones. J. Acoust. Soc. Am. **108**(4), 1888–1905 (2000)
9. J Benesty, Adaptive eigenvalue decomposition algorithm for passive source localization. J. Acoust. Soc. Am. **107**(1), 384–391 (2000)
10. S Doclo, M Moonen, Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments. EURASIP J. Appl. Signal Proces. **11**, 1110–1124 (2003)
11. R Schmidt, Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propagation. **34**(3), 276–280 (1986). doi:10.1109/TAP.1986.1143830
12. D Ampeliotis, K Berberidis, Low complexity multiple acoustic source localization in sensor networks based on energy measurements. Signal Proces. **90**(4), 1300–1312 (2010). doi:10.1016/j.sigpro.2009.10.015
13. W Cui, Z Cao, J Wei, Dual-microphone source location method in 2-D space, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4 (Toulouse, France, May 2006), pp. 845–848. doi:10.1109/ICASSP.2006.1661101
14. KC Ho, M Sun, Passive source localization using time differences of arrival and gain ratios of arrival. IEEE Trans. Signal Proces. **56**(2), 464–477 (2008). doi:10.1109/TSP.2007.906728
15. DP Morgan, EB George, LT Lee, SM Kay, Co-channel speaker separation, in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1 (Detroit, USA, May 1995), pp. 828–831
16. D Sharma, PA Naylor, *Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment*, (Glasgow, Scotland, Aug. 2009)
17. S Gonzalez, M Brookes, PEFAC - a pitch estimation algorithm robust to high levels of noise. IEEE/ACM Trans. Audio, Speech Lang. Proces. **22**(2), 518–530 (2014). doi:10.1109/TASLP.2013.2295918
18. M Christensen, L HÃÿjvang, A Jakobsson, S Jensen, Joint fundamental frequency and order estimation using optimal filtering. EURASIP J. Adv. Signal Proces. **2011**(1), 1–18 (2011)
19. J Nielsen, M Christensen Jensen, Default Bayesian estimation of the fundamental frequency. IEEE Trans. Audio Speech Lang. Proces. **21**(3), 598–610 (2013). doi:10.1109/TASL.2012.2229979
20. JK Nielsen, MG Christensen, AT Cemgil, SH Jensen, Bayesian model comparison with the g-prior. IEEE Trans. Signal Proces. **62**(1), 225–238 (2014)
21. MG Christensen, A Jakobsson, Multi-pitch estimation, in *Synthesis Lectures on Speech & Audio Processing*, ed. by BH Juang, vol. 5 (Morgan & Claypool San Rafael, 2009). http://dx.doi.org/10.2200/S00178ED1V01Y200903SAP005
22. M Wohlmayr, M Képesi, Joint position-pitch extraction from multichannel audio, in *8th Conference of the International Speech Communication Association, Interspeech* (Antwerp Belgium, Aug. 2007), pp. 1629–1632
23. T Habib, M Képesi, L Ottowitz, Experimental evaluation of the joint position-pitch estimation (POPI) algorithm in noisy environments, in *5th IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)* (Darmstadt, Germany, July 2008), pp. 369–372
24. M Képesi, L Ottowitz, T Habib, Joint position-pitch estimation for multiple speaker scenarios, in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)* (Trento, Italy, May 2008), pp. 85–88. doi:10.1109/HSCMA.2008.4538694
25. T Habib, L Ottowitz, M Képesi, Experimental evaluation of multi-band position-pitch estimation (M-PoPi) algorithm for multi-speaker localization, in *9th Conference of the International Speech Communication Association, Interspeech* (Brisbane, Australia, Sept. 2008), pp. 1317–1320
26. T Habib, H Romsdorfer, Comparison of SRP-PHAT and multiband-Popi algorithms for speaker localization using particle filters, in *13th International Conference on Digital Audio Effects (DAFX)* (Graz, Austria, Sept. 2010)
27. T Habib, H Romsdorfer, Auditory inspired methods for localization of multiple concurrent speakers. Comput. Speech Lang. Spec. Issue Speech Sep. Recognit. Multisource Environ. **27**(3), 634–659 (2013). doi:10.1016/j.csl.2012.09.003
28. SN Wrigley, GJ Brown, Recurrent timing neural networks for joint F0-localization based speech separation, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Hawaii, USA, April 2007)
29. JR Jensen, MG Christensen Jensen, Joint DOA and fundamental frequency estimation methods based on 2-D filtering, in *European Signal Processing Conference, EUSIPCO* (Aalborg, Denmark, Aug. 2010), pp. 2091–2095
30. Z Zhou, MG Christensen, JR Jensen, HC So, Joint DOA and fundamental frequency estimation based on relaxed iterative adaptive approach and optimal filtering, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Vancouver, Canada, May 2013), pp. 6812–6816. doi:10.1109/ICASSP.2013.6638981
31. J Zhang, M Christensen, S Jensen, M Moonen, Joint DOA and multi-pitch estimation based on subspace techniques.EURASIP. J. Adv. Signal Proces. **2012**, 1 (2012)
32. S Karimian-Azari, JR Jensen, MG Christensen, Fast joint DOA and pitch estimation using a broadband MVDR beamformer, in *European Signal Processing Conference EUSIPCO* (Marrakech, Morocco, p. Sept. 2013
33. JR Jensen, MG Christensen, SH Jensen, Nonlinear least squares methods for joint DOA and pitch estimation. IEEE Trans. Audio Speech Lang. Proces. **21**(5), 923–933 (2013). doi:10.1109/TASL.2013.2239290
34. J Benesty, J Chen, Y Huang, Time-delay estimation via linear interpolation and cross correlation. IEEE Trans. Speech Audio Proces. **12**(5), 509–519 (2004)
35. P Vary, R Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Wiley, Chichester, 2006)
36. I Shafran, R Rose, Robust speech detection and segmentation for real-time ASR applications, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1 (Hong Kong, China, April 2003), pp. 432–435
37. DB Ward, EA Lehmann, RC Williamson, Particle filtering algorithms for tracking an acoustic source in a reverberant environment. IEEE Trans. Speech Audio Proces. **11**(6), 826–836 (2003)
38. MS Arulampalam, S Maskell, N Gordon, T Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Proces. **50**(2), 174–188 (2002). doi:10.1109/78.978374
39. S Müller, P Massarani, Transfer-function measurement with sweeps. J. Audio Eng. Soc. (AES). **49**(6), 443–471 (2001)
40. JB Allen, DA Berkley, Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. **65**(4), 943–950 (1979)
41. E Habets, Room impulse response generator. Internal Report (2010). http://home.tiscali.nl/ehabets/rir_generator.html. Accessed 18 March 2014

doi:10.1186/s13636-014-0031-8
Cite this article as: Gerlach *et al.*: **Joint estimation of pitch and direction of arrival: improving robustness and accuracy for multi-speaker scenarios.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:31.