EUSIPCO 2013 1569744275

FAST JOINT DOA AND PITCH ESTIMATION USING A BROADBAND MVDR BEAMFORMER

Sam Karimian-Azari, Jesper Rindom Jensen and Mads Græsbøll Christensen

Audio Analysis Lab, AD:MT, Aalborg University,
email: {ska, jrj, mgc}@create.aau.dk

ABSTRACT

The harmonic model, i.e., a sum of sinusoids having frequencies that are integer multiples of the pitch, has been widely used for modeling of voiced speech. In microphone arrays, the direction-of-arrival (DOA) adds an additional parameter that can help in obtaining a robust procedure for tracking non-stationary speech signals in noisy conditions. In this paper, a joint DOA and pitch estimation (JDPE) method is proposed. The method is based on the minimum variance distortionless response (MVDR) beamformer in the frequency-domain and is much faster than previous joint methods, as it only requires the computation of the optimal filters once per segment. To exploit that both pitch and DOA evolve piece-wise smoothly over time, we also extend a dynamic programming approach to joint smoothing of both parameters. Simulations show the proposed method is much more robust than parallel and cascaded methods combining existing DOA and pitch estimators.

Index Terms— direction-of-arrival estimation, pitch estimation, MVDR, beamforming, microphone arrays.

1. INTRODUCTION

The estimation of the fundamental frequency, or pitch as it is commonly referred to, of voiced speech signals is a challenging problem, as it is a key feature in many solutions for enhancement, separation, classification, compression, coding, etc. Various methods have been investigated to solve this problem for the single-channel case (see, e.g., [1–3]). However, for the multi-channel case, much less work has been done. When using a microphone array, the spatial information, such as the direction-of-arrival (DOA), provides additional information that can be used for such tasks as separation and enhancement. Likewise, a spatial filter, or beamformer, can be used for extracting signals impinging on the array from any particular DOA [4]. Multiple concurrent speech signals, which each are of broadband nature, may have overlapping spectral features with common pitch and harmonics and are, hence, difficult to separate. A beamforming technique can be used for locating and separating such signals by joint estimation of both the DOA and pitch of the desired signal in cases where it would otherwise not be possible.

The estimation of each parameter has traditionally been treated separately [5], and estimates of both can be obtained using cascaded or parallel combinations of pitch and DOA estimators. In the cascaded approach, a broadband beamformer estimates the DOA and extracts a signal from which the pitch can be found using a standard estimator (e.g., one of those in [2]). Using a parallel approach, multi-channel pitch estimation [6] can be run in parallel with a broadband DOA estimator to obtain both pitch and DOA from the multi-channel signal. It is easy to see that, in multi-source scenarios, these estimation procedures may have problems with sources having the same DOA or overlapping harmonics.

As we have argued, joint DOA and pitch estimation (JDPE) methods are of interest as a robust alternative to cascaded or parallel approaches. Some approaches have recently been proposed, including the non-linear least squares (NLS) method [7], the spatio-temporal filtering based on the linearly constrained minimum variance (LCMV) beamformer [8], the correlation based method [9], and the subspace-based method [10]. While these methods perform well, they are computationally intensive, and faster methods may be required for some applications. More specifically, the methods of [7, 8] have cubic complexity for each combination of DOA and pitch candidates. Furthermore, none of these methods exploit that the pitch and DOA evolve in a piece-wise smooth manner.

In this paper, we present a fast algorithm for JDPE. The method is based on the frequency-domain minimum variance distortionless response (MVDR) beamformer, which is used to estimate the 2D spatio-temporal spectrum of the observed signal once per segment. From this 2D spectrum, the DOA and the pitch are estimated jointly by forming sums over the 2D spectrum for combinations of DOAs and pitches. This process essentially estimates the power of the assumed underlying periodic signal. Also, the number of harmonics is determined in the process using the maximum a posteriori method of [11, 12], something that is required to avoid ambiguities in the pitch estimates. Finally, the piece-wise smoothness of the DOA and pitch over time is exploited by the extension of the

This work was funded by the Villum Foundation.

method [13] to include also the DOA.

The rest of this paper is organized as follows: in Section 2, we introduce the signal model and MVDR-based broadband beamforming, from which the proposed method is developed in Section 3. Later on, in Section 4, experimental results are reported. Finally, the paper is concluded in Section 5.

2. PROBLEM FORMULATION

2.1. Signal model

An array of M microphones receives broadband acoustic waves from D desired sound sources in a noisy environment without reverberation. We assume each desired complex signal, $s_d(n)$, is quasi-periodic with L_d number of harmonics, where $d = 1, ..., D$, and it is stationary over the sampling interval N. The signals captured by the mth microphone relating to the dth source are delayed by τ_{md} depending on their distance and the sampling frequency f_s. A linear combination of all sources besides an additive noise $e_m(n)$ constitutes this model:

$$x_m(n) = \sum_{d=1}^{D} s_d(n - f_\mathrm{s}\tau_{md}) + e_m(n), \tag{1}$$

where

$$s_d(n - f_\mathrm{s}\tau_{md}) = \sum_{l=1}^{L_d} a_{dl}\ e^{jl\omega_{0d}n}\, e^{-jl\omega_{0d}f_\mathrm{s}\tau_{md}}, \tag{2}$$

and ω_{0d} is the fundamental frequency of the dth source. While many different array structures can be considered, we will assume a uniform linear array (ULA) structure herein for a proof of our concept. Consider a ULA denoted by M consecutive microphones with the specific inter-distance δ. Supposing a long distance from the ULA to the sources in comparison with δ, a plane wave and a homogeneous magnitude a_{dl} can be assumed across the array. By choosing the first microphone as a reference, the time delay between the other microphones and the reference is $\Delta\tau_{md} = (m-1)\delta\sin(\theta_d)/c$, where θ_d is the direction of desired signal, c is the wave propagation velocity and $\Delta\tau_{md} = \tau_{md} - \tau_{1d}$.

We can also formulate our signal model in the frequency domain, which is useful for deriving the proposed method. Stacking the spectral amplitudes of the observed signals at the M sensors for the frequency bin ω gives

$$\mathbf{X}(\omega) = [\, X_1(\omega)\, X_2(\omega)\, \ldots\, X_M(\omega)\,]^T. \tag{3}$$

By exploiting the relation between the sensor signals described by (2), the model further yields

$$\mathbf{X}(\omega) = \sum_{d=1}^{D} \mathbf{z}(\theta_d, \omega)\, S_d(\omega)\ +\ \mathbf{E}(\omega), \tag{4}$$

where $\mathbf{z}(\theta_d, \omega) = e^{-j\omega f_\mathrm{s}\tau_{1d}}[\,1\ e^{-j\psi_{2d}}\ \ldots\ e^{-j\psi_{Md}}\,]^T$ and $\psi_{md} = \omega f_\mathrm{s}\Delta\tau_{md}$.

2.2. MVDR broadband beamformer

The Capon beamformer, which is also known as the minimum variance distortionless response (MVDR) beamformer, is a type of baseband filter that can be extended to a broadband filter through a filter bank approach (FBA) [3]. To approach the proposed JDPE method, we introduce the broadband frequency-domain MVDR (FMV) algorithm as a quick solution in comparison with other time-domain beamformers [14].

First, a narrowband beamformer $\mathbf{W}(\theta, \omega)$ is designed to minimize the output power of the filter while it has a unit gain at a specific DOA, $\theta \in [-90°, +90°]$, and a sub-band frequency ω. In this way, we design a narrowband beamformer for a wide range of frequencies to get the broadband beamformer, i.e.,

$$\begin{aligned} \min_{\mathbf{W}(\theta,\omega)}\ & \mathbf{W}^H(\theta,\omega)\mathbf{R}_\mathbf{X}(\omega)\mathbf{W}(\theta,\omega) \\ \text{s.t.}\quad & \mathbf{W}^H(\theta,\omega)\mathbf{z}(\theta,\omega) = 1, \end{aligned} \tag{5}$$

where $\mathbf{R}_\mathbf{X}(\omega) \in \mathbb{C}^{M\,\mathrm{x}\,M}$ is the correlation matrix of $\mathbf{X}(\omega)$ i.e., $\mathbf{R}_\mathbf{X}(\omega) = E\{\mathbf{X}(\omega)\mathbf{X}^H(\omega)\}$, $E\{.\}$ represents the expectation operation, and $\{.\}^H$ represents the conjugate transpose of a matrix. The correlation matrix $\mathbf{R}_\mathbf{X}(\omega)$ is not known in most practical scenarios, so we estimate it as

$$\hat{\mathbf{R}}_\mathbf{X}(\omega) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{X}_b(\omega)\, \mathbf{X}_b^H(\omega), \tag{6}$$

where $\mathbf{X}_b(\omega)$ denotes the bth complex spectral amplitude out of the last B estimates, and $\{\hat{.}\}$ denotes the estimate. In practice, blocks of N samples are used to obtain the spectral amplitude estimates with consecutive blocks overlapping by Q samples.

The adaptive weights of the beamformer $\mathbf{W}(\theta, \omega)$ are formed using the Lagrange multiplier method [3] which yield

$$\mathbf{W}(\theta,\omega) = \frac{\hat{\mathbf{R}}_\mathbf{X}^{-1}(\omega)\mathbf{z}(\theta,\omega)}{\mathbf{z}^H(\theta,\omega)\hat{\mathbf{R}}_\mathbf{X}^{-1}(\omega)\mathbf{z}(\theta,\omega)}. \tag{7}$$

Inserting the optimal beamformer in the output power expression $\mathbf{W}^H(\theta,\omega)\hat{\mathbf{R}}_\mathbf{X}(\omega)\mathbf{W}(\theta,\omega)$ results in

$$J(\theta,\omega) = \frac{1}{\mathbf{z}^H(\theta,\omega)\hat{\mathbf{R}}_\mathbf{X}^{-1}(\omega)\mathbf{z}(\theta,\omega)}, \tag{8}$$

which should be an estimate of the spatio-temporal spectral power for θ and ω.

According to the designed filter, the estimated covariance matrix has to be invertible (7-8). This can be ensured by choosing $B \geq M$ in (6). In practice, B and Q should be chosen such that a good trade off between the robustness of the estimate and the validity of the stationary assumption is obtained.

3. PROPOSED METHOD

3.1. Order estimation

To estimate the parameters (θ_d, ω_{0d}) of the desired signal source, we need an estimate of the number of harmonics L_d according to the signal model in (2). Here, we propose a model-order estimator, which is optimal in single source scenarios. The method is inspired by the maximum a posterior (MAP) estimator in [2, 11], where the noise variance is estimated using the directional spectrum obtained using the FMV method. It penalizes a maximum likelihood (ML) estimation method to find the maximum a posteriori probability of Φ and the number of harmonics, $L(\theta, \omega_0)$, given the temporal spectrum at the candidate direction θ. In this way, we can estimate the model order for the relative pair of DOA and fundamental frequency [15]:

$$\hat{L}(\theta, \omega_0) = \arg \min_{L(\theta,\omega_0)} \{-\ln f(\mathbf{J}(\theta)|\Phi, L(\theta, \omega_0)) + \frac{1}{2} \ln |\hat{\mathbf{G}}|\}, \quad (9)$$

where $\mathbf{J}(\theta) = [\, J(\theta, 0)\, J(\theta, \frac{2\pi}{N_\mathrm{f}}) \ldots J(\theta, (\frac{N_\mathrm{f}}{2} - 1)\frac{2\pi}{N_\mathrm{f}})\,]$, N_f is the length of the discrete Fourier transform (DFT), and Φ denotes the vector containing the other estimation parameters: fundamental frequency, amplitudes, and phases. In the following, the notation of $L(\theta, \omega_0)$ is simplified to be exposed by L. The penalty part $\hat{\mathbf{G}}$ of this estimation is an approximation of the Fisher information matrix (FIM) relating to Φ, i.e.,

$$\hat{\mathbf{G}} \approx -E \left\{ \frac{\partial^2 \ln f(\mathbf{J}(\theta)|\Phi, L)}{\partial \Phi \partial \Phi^T} \right\}_{\Phi=\hat{\Phi}}. \quad (10)$$

The determinant of the given Hessian matrix $\hat{\mathbf{G}}$ can be normalized [15] with respect to the number of samples N (see [11]) as:

$$|\hat{\mathbf{G}}| = |\mathbf{K}^{-2}|\, |\mathbf{K}\, \hat{\mathbf{G}}\, \mathbf{K}|, \quad (11)$$

where it can be shown that $\mathbf{K}$ is given by

$$\mathbf{K} = \begin{bmatrix} N^{-\frac{3}{2}} & 0 \\ 0 & N^{-\frac{1}{2}} \mathbf{I}_{2Lx2L} \end{bmatrix}. \quad (12)$$

The estimate of the number of harmonics in (9) can be simplified by assuming that N is large, in which case [12] we obtain

$$\hat{L} \approx \arg \min_L \{N \ln \hat{\sigma}_L^2 + \frac{3}{2} \ln N + L \ln N\}, \quad (13)$$

where $\hat{\sigma}_L^2$ is the noise variance related to every candidates of a number of harmonics, and a fundamental frequency, ω_0. That is

$$\hat{\sigma}_L^2 = 2 \left(\frac{N}{N_\mathrm{f}} \sum_{i=0}^{N_\mathrm{f}/2-1} J(\theta, i\frac{2\pi}{N_\mathrm{f}}) - \sum_{l=1}^{L} J(\theta, l\omega_0) + r \right), \quad (14)$$

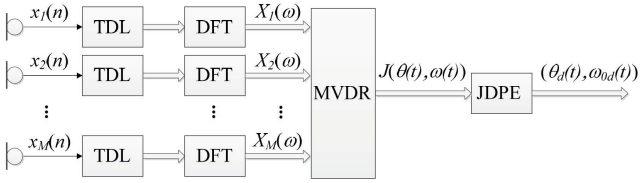

Fig. 1. General diagram of the frequency-domain MVDR beamformer to joint DOA and pitch estimation (JDPE)

where, we have introduced the regularization factor r to account for inaccurate noise variance estimates for relatively small N.

3.2. Joint DOA and pitch estimation and smoothing

A JDPE method for speech signals in a noisy field is proposed in this section based on the FMV method, and the general idea is depicted in Fig. 1. To estimate the fundamental frequency, the initial assumption of stationary temporal samples is introduced for an appropriate number of samples N in a tapped-delay line (TDL). The samples are mapped to the frequency domain using the DFT, and then the MVDR beamformer estimates a time-dependent 2D spectrum $J(\theta(t), \omega(t))$ at each time instance t.

According to the model in (2), the desired signals only have frequency contents at the harmonic frequencies. Hence, to estimate the pitch, we should only consider those bins. Therefore, we introduce the cost function

$$J_0(\theta_d(t), \omega_{0d}(t)) = \sum_{l=1}^{\hat{L}_d} J(\theta_d(t), l\omega_{0d}(t)). \quad (15)$$

Then, the pitch and the DOA are estimated jointly by maximizing $J_0(\theta_d(t), \omega_{0d}(t))$ for one source as

$$(\hat{\theta}_d(t), \hat{\omega}_{0d}(t)) = \arg \max_{(\theta_d(t), \omega_{0d}(t))} J_0(\theta_d(t), \omega_{0d}(t)). \quad (16)$$

The series of estimated parameters $[\hat{\theta}_d, \hat{\omega}_{0d}]$ have to be a continuous and smooth function of time according to the pitch and the position of the actual sound source. Smoothing of one dynamic parameter had been solved using a recursive algorithm in [13]. In this approach, a transition cost function $c(t, t_0)$ at time t is accumulated over the preceding states since t_0. The forward path is then the path that minimize the accumulated cost function $D(t)$ among all previous cost functions;

$$D(t) = \min_{t^*} \{D(t^*) + c(t^*, t_0)\} - B_\mathrm{s}, \quad (17)$$

where $t^* \in [t_0, t)$, B_s is a smoothing factor ($B_\mathrm{s} > 0$), and

$$c(t^*, t_0) = \frac{||[\hat{\theta}_d(t^*), \hat{\omega}_{0d}(t^*)] - [\hat{\theta}_d(t_0), \hat{\omega}_{0d}(t_0)]||_2}{(t^* - t_0)}. \quad (18)$$

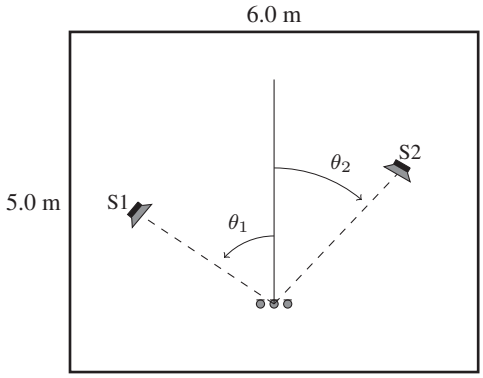

Fig. 2. The room layout and the azimuth angles of sound sources; $\theta_1 \approx$ -53° and $\theta_2 \approx 45°$

Note that the transition cost function proposed here, is a generalization of the function in [13] from 1D to 2D.

4. EXPERIMENTAL RESULTS

To evaluate the proposed JDPE method, we simulated a cocktail party in a room without reverberation with zero reflection order, as shown in Fig. 2. The desired voiced speech, S1 uttering the sentence: Why were you away a year Roy?, and a periodic signal with five harmonics, S2, as an interference are played simultaneously along with diffusive white noise. The sound propagation in a rectangular room is simulated using the image method [16]. This method simulates the room impulse response relating to the dimension, the reflection order and geometric position of acoustic sources and microphones [17]. We used this method to simulate acoustic waves on a ULA with three hyper-cardioid microphones, $M = 3$, at the specified positions with inter-distance of $\delta =$ 0.04 m, and those were oriented respecting to the zero azimuth and elevation angles of the ULA. The distance between the microphones in the ULA should be smaller than half of the wavelength to avoid aliasing [3]. In addition, a real-life acoustic ambiance was simulated by adding diffusive acoustic noise, and the wave propagation speed was assumed $c =$ 343.2 m/s. The spectrograms of the desired voiced speech and the interfered signal are shown in Fig. 3. For the signal measured using the first microphone, the signal-to-interference-ratio (SIR) was 12.8 dB, while the signal-to-noise-ratio (SNR) was 10 dB.

Time-domain input signals were sampled across TDLs with $f_s = 8.0$ kHz of length $N = 256$, refreshed every 2.5 ms (20 time steps), and preserved in a buffer containing the $B = 10$ most recent ones. We calculated the DFTs of these vectors using zero-padded FFTs with a rectangular window of length $N_f = 4096$. The cross-correlation matrices for all frequency bins were then estimated from the B past DFTs, and they were used in three different methods: a parallel method,

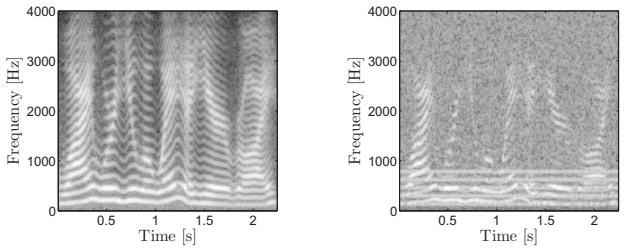

Fig. 3. Spectrograms of the desired voiced speech (left), and the interfered mixture (right)

a cascade method, and the proposed method. The smoothing and the regularization factors in the proposed method were set to $B_s = 0.001\pi$ and $r = 10^{-6}$. In the parallel method, the FMV beamformer [14] runs along with the multi-channel pitch estimator in [6]. Finally, in the cascaded method, the NLS method [2] is applied after beamforming with the aforementioned parameters. Note that, in the cascaded method we used a Hanning window with 50% overlapping to recover the time domain signals.

Fig. 4 depicts the results of the proposed JDPE method in comparison with the results of the cascaded and parallel methods. In order to evaluate the acquired results, we compared the estimates with the true DOA, $\theta_1 \approx$ -53°, and single channel pitch estimates of the clean signal obtained using the NLS method [2]. It shows the continuity and smoothness of both estimated DOA and pitch analytically, and the robust estimations are approved in terms of measured mean-square error (MSE) (see Table 1).

5. CONCLUSION

In this paper, we have proposed the JDPE method. In this method, the DOA and pitch are estimated jointly by integrating the broadband MVDR spectrum over the harmonic frequencies for a set of candidate pitches and DOAs, and later on maximizing. The simplicity of the proposed method is a significant advantage in comparison with other joint estimation methods. The MVDR spectrum which can be implemented efficiently using FFTs, needs to be calculated once. It benefits the method as a fast spectral estimation. Our second contribution, is the spatio-temporal smoothing of the obtained DOA and pitch estimates using dynamic programming, which improves the robustness of the underlying estimator. Our sim-

Table 1. Mean square error (MSE) of estimated DOA and pitch of the different experiments

	MSE(ω_0) [Hz^2]	MSE(θ) [degree^2]
JDPE	122.1	1.5
Parallel	2.3x10^3	0.4
Cascade	2.9x10^3	0.4

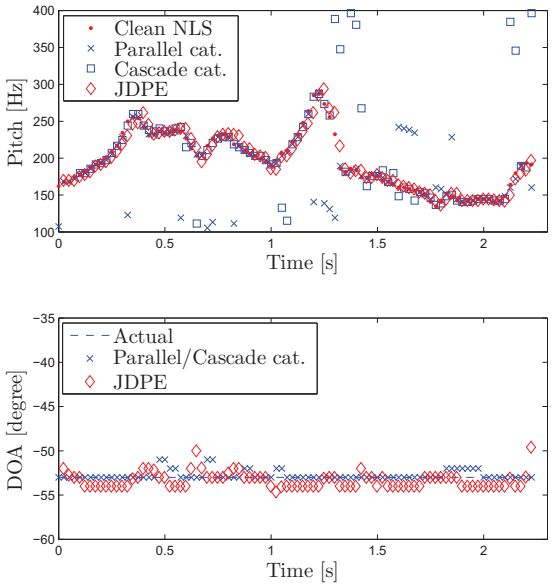

Fig. 4. Estimation of pitch (top), and DOA (bottom) at the different experiments

ulations show that the proposed joint estimator outperforms traditional methods, i.e., a cascaded and a parallel approaches which estimate pitch and the DOA separately in a real-life scenario.

6. REFERENCES

[1] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76 – 87, Jan. 2004.

[2] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech and Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[3] P. Stoica and R. Moses, *Spectral Analysis of Signals*, Pearson Education, Inc., 2005.

[4] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[5] Z. Zhou, M. G. Christensen, and H. C. So, "Two stage DOA and fundamental frequency estimation based on subspace techniques," in *Proc. IEEE Int. Conf. Signal Process.*, Oct. 2012.

[6] M. G. Christensen, "Multi-channel maximum likelihood pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2012, pp. 409–412.

[7] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 923 –933, May 2013.

[8] 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 Conf.*, Aug. 2010, pp. 2091–2095.

[9] M. Wohlmayr and M. Képesi, "Joint position-pitch extraction from multichannel audio," in *Proc. Interspeech*, Aug. 2007, pp. 1629–1632.

[10] L. Y. Ngan, Y. Wu, H. C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. IEEE Int. Symp. Circuits and Systems*, May 2003, vol. 3, pp. 722–725.

[11] P. M. Djuric, "Asymptotic MAP criteria for model selection," *IEEE Trans. Signal Process.*, vol. 46, no. 10, pp. 2726 –2735, Oct. 1998.

[12] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[13] Hermann Ney, "A dynamic programming algorithm for nonlinear smoothing," *Signal Processing*, vol. 5, no. 2, pp. 163–173, March 1983.

[14] M. E. Lockwood, D. L. Jones, R. C. Bilger, C. R. Lansing, Jr. W. D. O'Brien, B. C. Wheeler, and A. S. Feng, "Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms," *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 379–391, Jan. 2004.

[15] M. G. Christensen, J. L. Højvang, A. Jakobsson, and S. H. Jensen, "Joint fundamental frequency and order estimation using optimal filtering," *EURASIP J. on Applied Signal Processing*, vol. 2011, no. 1, pp. 1–18, Jun. 2011.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[17] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010, Ver. 2.0.20100920.