Spatio-temporal filtering methods for enhancement and separation of speech signals

Christensen, Mads Græsbøll; Jensen, Jesper Rindom; Benesty, Jacob; Jakobsson, Andreas

Published in:
Proceedings of the 2013 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP)

DOI (link to publication from Publisher):
10.1109/ChinaSIP.2013.6625349

Publication date:
2013

Document Version
Accepted author manuscript, peer reviewed version

SPATIO-TEMPORAL FILTERING METHODS FOR ENHANCEMENT AND SEPARATION OF SPEECH SIGNALS

Mads Græsbøll Christensen[1], Jesper Rindom Jensen[1], Jacob Benesty[1,2], and Andreas Jakobsson[3]

[1]Audio Analysis Lab, AD:MT
Aalborg University, Denmark
{mgc,jrj}@create.aau.dk

[2]INRS-EMT
University of Quebec, Canada
benesty@emt.inrs.ca

[3]Dept. of Mathematical Statistics
Lund University, Sweden
aj@maths.lth.se

ABSTRACT

In this paper, we give an overview of the background for, the ideas behind, and the challenges to be addressed in the project "Spatio-Temporal Filtering Methods for Enhancement and Separation of Speech Signals," which is funded by the Villum Foundation. The project aims at addressing the problem of enhancing and separating speech signals from noisy mixtures, a problem also known as the cocktail party problem. It aims at exploring new ways of solving this problem by generalizing a new class of optimal temporal filtering methods for periodic signals to multiple microphones, resulting in so-called spatio-temporal filtering methods that are controlled by two parameters, the direction-of-arrival and the fundamental frequency. These filters are optimal in that they let the signal of interest pass undistorted while everything else is attenuated as much as possible. Unlike state-of-the-art methods, they do not require knowledge of the statistics of noise and interfering speech signals, something that is especially important when dealing with non-stationary noise.

Index Terms— Speech enhancement, microphone arrays, beamforming, pitch estimation, DOA estimation

1. INTRODUCTION

Speech enhancement and separation algorithms aim at extracting a speech signal of interest from a signal composed of multiple signals and undesired noise. This should be done in such a way that the influence of interfering sources and noise is minimized while a minimum distortion is incurred on the extracted signal. The quality of the extracted speech signal can be quantified in terms of two aspects, namely perceived quality and intelligibility. The first reflects whether unnatural and annoying distortion has been incurred while the second reflects whether the underlying information has been preserved. In the past few decades, development of methods for speech enhancement from signal recorded with only one microphone has received much attention, and numerous different methods have been proposed. For an overview, we refer the interested reader to [1, 2] and the references therein. Of particular relevance to the present project is the methods based on the quasi-periodicity of voiced speech [3, 4]. There also exist multichannel extensions of some of the single-channels methods, e.g., the multichannel Wiener filter [5, 6]. Using beamforming in microphone arrays, it is possible to steer the focus of the microphone array such that sources impinging on the array from a certain angle, called direction-of-arrival (DOA), are unchanged, whereas noise and interfering sources are suppressed. In recent years, beamforming has found new applications in noise reduction and speech separation in digital hearing aids and entertainment systems, and the use of microphone arrays is becoming ever more widespread. The seminal work [7, 8] laid the foundation of modern beamforming and DOA estimation for narrowband signals and many variations of these designs have since followed (see, e.g., [9, 10]). In the past decade, convex optimization methods have gained widespread use in signal processing and this has also led to advances in beamformer design, notably, in designing robust beamformers [11, 12], wherein model uncertainties, like array calibration errors, are taken into account. Many of these methods have in common that they are based on the narrowband assumption. Furthermore, they rely on only one source impinging on the array from a particular angle, being based only on spatial information. It is well-known in the speech enhancement community that noise reduction algorithms struggle to increase speech intelligibility for signals recorded with a single microphone, at best rendering them unchanged, but rather increase secondary properties like the perceptual quality and reduce listener fatigue. By using multiple microphones, it is, however, in principle possible to increase speech intelligibility. In fact, it is theoretically possibly to achieve super-human performance in tasks such as speech recognition and speaker identification this way.

Filtering is a fundamental tool when separating and enhancing signals, and this project aims at performing filtering in two domains simultaneously, namely in the temporal and spatial domains. Separately, filtering approaches in either domain works well when the contents of the sources are well-separated in either frequency or angle. The fundamental idea

This work was funded by the Villum Foundation.

behind this project is that by combining temporal and spatial filters and operate jointly in these domains, i.e., by forming spatio-temporal filters, it is possible to separate sources under conditions where the other two approaches may fail.

Speech signals do not satisfy the constraints mentioned above, as such signals are well-known to be broadband, and there can be multiple speakers in close vicinity of each other. Generalizations of the methods mentioned above exist in several forms, allowing for multiple sources from the same angle (e.g., [13]), but these are still based on the assumption that the individual sources are narrowband. These methods also do not take the temporal characteristics of the sources into account, and the typical way of extracting such information is to apply a temporal filter after the spatial filter has been applied. Such approaches are therefore, generally, not optimal as they do not exploit the additional knowledge about the signal of interest. Much research has been devoted to the problem of dealing with broadband sources. Many approaches, however, avoid the explicit design of the beamformer, but rather resort to either parametric approaches, where the underlying parameters, or simply time-delays between different microphones, are found [14, 15], or various kinds of heuristics or suboptimal methods for dealing with the problem. These methods, however, generally do not possess the desirable properties of beamformers to adaptively reject interference and noise. In fact, they do not address the question considered here, namely how to extract the signals of interest. The use of explicit speech models in solving the problems associated with array speech processing, which is the underlying idea promoted here, has been strongly advocated by Brandstein [16]. In [17], an attempt to do this was proposed with some success, by modeling the speech signal as an auto-regressive process, but this approach relies on a priori knowledge of the speech signal.

Recently, a new class of filters for enhancing and separating periodic signals was introduced in [18] and these form the theoretical basis of the project. The preliminary, but also very promising, results reported in [18] suggest that the filters can be applied to speech signals, as these are approximately periodic for voiced speech. The filters are temporal filters and operate only on signals recorded by a single microphone. They exploit that periodic signals can be expressed as a sum of sinusoids having frequencies that are integer multiples of a fundamental frequency. The filters can thus be thought of as optimal, signal-adaptive, FIR comb filters. These filters were demonstrated in [18] to have a number of desirable properties compared to existing ones, including IIR comb filters and Fourier-based filters, in fact, they reduce to various well-known designs under certain conditions. In particular, they are signal-adaptive meaning that they automatically adapt to the acoustic environment and they are optimal in that they suppress interfering sources and noise as much a possible while leaving the signal of interest undistorted. A key feature of the filters is that they do not require any knowledge about the noise statistics or the interfering sources, something that is not the case for most state-of-the-art enhancement algorithms. This means that, unlike methods based on the noise statistics, these filters are likely to work for non-stationary noise. An additional feature of the filters is that they also offer a complete, implicit parametrization of periodic signal that can be extracted and processed, if desired.

2. OPTIMAL FILTERS FOR PERIODIC SOURCES

We will now briefly review the generalization of one of the optimal filter designs in [18] to arrays. We will do this by first introducing the signal model. We are concerned with estimating the parameters of and extracting the periodic source $s(n_\mathrm{t})$ which, due to its periodic nature, can be modeled as $s(n_\mathrm{t}) = \sum_{l=1}^{L} \alpha_l e^{j\omega_\mathrm{t} l n_\mathrm{t}}$, where L is the model order and $\alpha_l = A_l e^{j\phi_l}$ is the complex amplitude of the lth sinusoid with $A_l > 0$ and ϕ_l being the amplitude and the phase, respectively. Next, the source $s(n_\mathrm{t})$ impinges on an array containing N_s sensors, which, in our case, are microphones. The signal is corrupted by the noise source $w_{n_\mathrm{s}}(n_\mathrm{t})$ for each sensor. The signal sampled by the n_sth sensor, for $n_\mathrm{t} = 0, \ldots, N_\mathrm{t} - 1$ and $n_\mathrm{s} = 0, \ldots, N_\mathrm{s} - 1$, can then be written as $x_{n_\mathrm{s}}(n_\mathrm{t}) = s(n_\mathrm{t} - \tau_{n_\mathrm{s}}) + w_{n_\mathrm{s}}(n_\mathrm{t})$, where τ_{n_s} is the relative time delay of the signal for sensor n_s. In what follows, we will assume that a uniform linear array (ULA) is used and that the signals of interest are located in the so-called far-field. Curiously, the employed model can be interpreted as a sum of narrowband sources, which means that the principles of narrowband beamforming can be applied to each of these. The problem considered is a) the joint estimation of the DOA θ and the pitch ω_t of the periodic source $s(n_\mathrm{t})$ and b) extraction of the periodic source $s(n_\mathrm{t})$. For this purpose, this project will employ optimal filtering techniques that let the signal $s(n_\mathrm{t})$ pass undistorted while they attenuate $w_{n_\mathrm{s}}(n_\mathrm{t})$ as much as possible to obtain the highest possible quality of the extracted signals.

We will now proceed to state the filter design problem as an optimization problem. To do so, we organize the observed signal $x_{n_\mathrm{s}}(n_\mathrm{t})$ in an $N_\mathrm{s} \times N_\mathrm{t}$ matrix $\mathbf{X}$. Similarly, the impulse response of the $M_\mathrm{s} \times M_\mathrm{t}$ order finite impulse response (FIR) filter is also organized in a matrix $\mathbf{H}_{\omega_\mathrm{t},\omega_\mathrm{s}}$ whose entries are given by $[\mathbf{H}_{\omega_\mathrm{t},\omega_\mathrm{s}}]_{nm} = H_{\omega_\mathrm{t},\omega_\mathrm{s}}(n-1, m-1)$, with n denoting the row, m the column, and $H_{\omega_\mathrm{t},\omega_\mathrm{s}}(\cdot,\cdot)$ is designed for the temporal and spatial frequencies ω_t and ω_s. The filter is then applied to $M_\mathrm{s} \times M_\mathrm{t}$ sub-blocks $\mathbf{X}_{n_\mathrm{s}}(n_\mathrm{t})$ of the data matrix, where the entries are defined as $[\mathbf{X}_{n_\mathrm{s}}(n_\mathrm{t})]_{nm} = x_{n_\mathrm{s}+n-1}(n_\mathrm{t} - m + 1)$. Due to the ULA and far-field assumptions, the spatial frequency is given by $\omega_\mathrm{s} = \omega_\mathrm{t} f_\mathrm{s} \frac{d \sin\theta}{c}$, where f_s is the sampling frequency, d is the inter-sensor spacing, θ is the DOA in radians, and c is the wave propagation velocity. To simplify the notation, the filter response $\mathbf{H}_{\omega_\mathrm{t},\omega_\mathrm{s}}$ and the sub-blocks $\mathbf{X}_{n_\mathrm{s}}(n_\mathrm{t})$ are stacked to form vectors of

length $M_{\mathrm{t}}M_{\mathrm{s}}$, i.e.,

$$\mathbf{h}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} = \mathrm{vec}\{\mathbf{H}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\} \quad \text{and} \quad \mathbf{x}_{n_{\mathrm{s}}}(n_{\mathrm{t}}) = \mathrm{vec}\{\mathbf{X}_{n_{\mathrm{s}}}(n_{\mathrm{t}})\},$$

with $\mathrm{vec}\{\cdot\}$ denoting the column-wise stacking operator. With this notation, the various filter designs in [18] can be derived directly. We will here demonstrate how to do this with the so-called Capon filter design (which is sometimes also referred to as MVDR or LCMV, depending on the context). First, we obtain an expression for the output power as

$$\mathrm{E}\{|y_{n_{\mathrm{s}}}(n_{\mathrm{t}})|^2\} = \mathbf{h}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{R}\mathbf{h}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}, \tag{1}$$

where $\mathbf{R} = \mathrm{E}\{\mathbf{x}_{n_{\mathrm{s}}}(n_{\mathrm{t}})\mathbf{x}^H_{n_{\mathrm{s}}}(n_{\mathrm{t}})\}$ is the $M_{\mathrm{s}}M_{\mathrm{t}} \times M_{\mathrm{s}}M_{\mathrm{t}}$ covariance matrix. Note that $\mathrm{E}\{\cdot\}$ and $(\cdot)^H$ denote the expectation operator and the complex transpose, respectively. In practice, we do not have access to the true covariance matrix and an estimate must be used, like

$$\widehat{\mathbf{R}} = \frac{1}{\kappa_{\mathrm{s}}\kappa_{\mathrm{t}}} \sum_{p=0}^{\kappa_{\mathrm{s}}-1} \sum_{q=0}^{\kappa_{\mathrm{t}}-1} \mathbf{x}_p(n_{\mathrm{t}}-q)\mathbf{x}^H_p(n_{\mathrm{t}}-q)\,, \tag{2}$$

where $\kappa_{\mathrm{s}} = N_{\mathrm{s}} - M_{\mathrm{s}} + 1$ and similarly for κ_{t}. The next task is to design the filter such that the output power is minimized subject to a distortionless constraint at desired angles and frequencies, which are here harmonically related due to the periodic nature of the sources. The presented filter design that follows next can be seen as a generalization of the Capon method [13]. It is obtained by introducing multiple harmonic constraints in the filter design and minimizing the output power:

$$\min_{\mathbf{h}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}} \mathbf{h}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{R}\mathbf{h}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} \;\text{ s.t. }\; \mathbf{h}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{a}_{l\omega_{\mathrm{t}},l\omega_{\mathrm{s}}} = 1, \quad \text{for } l = 1,\ldots,L\,, \tag{3}$$

where $\mathbf{a}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} = \mathbf{a}_{\omega_{\mathrm{t}}} \otimes \mathbf{a}_{\omega_{\mathrm{s}}}$ with $\otimes$ denoting the Kronecker product and $\mathbf{a}_{\omega_{\mathrm{t}}} = \begin{bmatrix}1 & e^{-j\omega_{\mathrm{t}}} & \cdots & e^{-j(M_{\mathrm{t}}-1)\omega_{\mathrm{t}}}\end{bmatrix}^T$ and similarly for $\mathbf{a}_{\omega_{\mathrm{s}}}$. The quantity $\mathbf{a}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}$ can be thought of as the combination of the time-domain Fourier vector accounting for the temporal frequency ω_{t} and the steering vector corresponding to the spatial frequency ω_{s}. For the above problem in (3) to have a non-trivial solution, we require that $L < M_{\mathrm{t}}M_{\mathrm{s}}$. It is a quadratic optimization problem with linear constraints (given ω_{t} and ω_{s}), and its solution can readily be found using the Lagrange multiplier method and is given by

$$\hat{\mathbf{h}}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} = \mathbf{R}^{-1}\mathbf{A}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}(\mathbf{A}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{R}^{-1}\mathbf{A}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}})^{-1}\mathbf{1}, \tag{4}$$

where $\mathbf{A}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} = \begin{bmatrix}\mathbf{a}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}} & \cdots & \mathbf{a}_{L\omega_{\mathrm{t}},L\omega_{\mathrm{s}}}\end{bmatrix}$ and with $\mathbf{1}$ being a column vector containing L ones. From (4), it can be seen that it is required that $\mathbf{R}$ is invertible, and hence that $\kappa_{\mathrm{t}}\kappa_{\mathrm{s}} \geq M_{\mathrm{t}}M_{\mathrm{s}}$. By inserting (4) into (1), we obtain an expression for the output power of the filter for an angle of θ and a fundamental frequency of ω_{t}:

$$\mathrm{E}\{|y_{n_{\mathrm{s}}}(n_{\mathrm{t}})|^2\} = \mathbf{1}^H(\mathbf{A}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{R}^{-1}\mathbf{A}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}})^{-1}\mathbf{1}. \tag{5}$$

From this expression, we can jointly estimate the DOA and the pitch by treating these quantities as unknowns and maximizing the output power of the filter over a set of candidate DOAs Θ and fundamental frequencies Ω as [19]

$$(\hat{\theta},\hat{\omega}_{\mathrm{t}}) = \underset{(\theta,\omega_{\mathrm{t}})\in\Theta\times\Omega}{\mathrm{argmax}}\; \mathbf{1}^H(\mathbf{A}^H_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}}\mathbf{R}^{-1}\mathbf{A}_{\omega_{\mathrm{t}},\omega_{\mathrm{s}}})^{-1}\mathbf{1}, \tag{6}$$

i.e., an optimal filter is essentially designed for each combination of $\theta \in \Theta$ and $\omega_{\mathrm{t}} \in \Omega$ and the output power is measured. Other approaches that are capable of this are [20, 21]. The optimal filter for extracting the periodic source with fundamental frequency $\hat{\omega}_{\mathrm{t}}$ and DOA $\hat{\theta}$ is then obtained by inserting these quantities into the optimal filter design (4). In [18], it was shown how various filter designs can be obtained by replacing $\mathbf{R}$ with, for example, a noise covariance matrix estimate (see also [22]), or an identity matrix, corresponding to the assumption that the noise or observed signal is white. Moreover, by using asymptotic approximations, some simplified filter designs can be obtained that reduce to simply using the Fourier transform to extract the signal of interest, an approach which is often used in the speech enhancement and separation community.

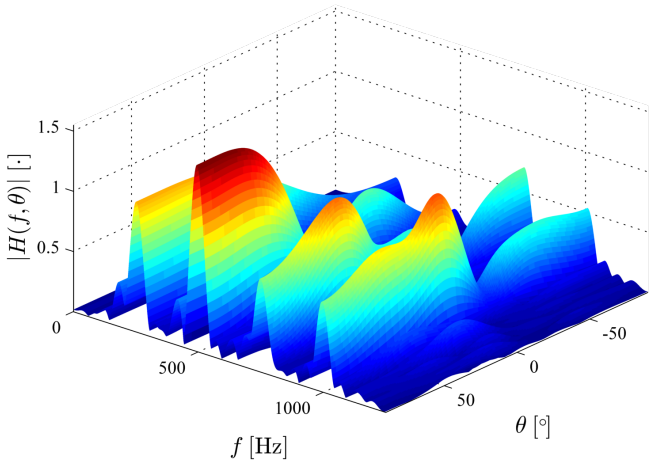

Fig. 1. Example of the response of an optimal filter in (4) designed for an angle of 30° and a pitch of 250 Hz with 4 harmonics for a white noise signal.

Next, we will briefly illustrate the properties of the optimal spatio-temporal filters obtained using (4). In Fig. 1, the magnitude response of an optimal filter is shown for a complex signal comprised of four harmonics with a fundamental frequency of 250 Hz impinging on the array from an angle of 30° and diffuse, white complex Gaussian noise at an SNR of −10 dB. The signal was sampled at a frequency of 2.5 kHz and 8 microphones were used with 300 temporal samples and a temporal filter length of 50. From the figure, it can be seen that the gain of the filter is one at the location and angle of the harmonics of the signal of interest. The output power of the optimal filter in (4) is depicted as a function of the angle

and the fundamental frequency in Fig. 2. As can be seen, despite the poor conditions, it is possible to easily identify both the angle and the pitch at the maximum, as described in (6). A remarkable feature of the adaptive, optimal filters is that this is also the case when strong, periodic interferences are present, even without any knowledge of these.

3. CHALLENGES AND HYPOTHESES

As mentioned earlier, the project aims at generalizing the filters of [18] to the spatio-temporal domain and to address the challenges in using them on real speech signals. Here, we will briefly discuss these challenges and then proceed to discuss some potential solutions. The challenges are the following. 1) Voiced speech is only approximately periodic, as the individual harmonics may deviate from being integer multiples of a fundamental frequency. Moreover, speech signals are nonstationary. These problems must be addressed or the perceived quality and intelligibility of the extracted signals may be compromised. 2) Some parts of speech signals, namely unvoiced speech, are not periodic at all but are still important with respect to both perceived quality and intelligibility. Hence, the filters must be able to handle both unvoiced and voiced speech in a compatible manner. 3) The filter designs of [18] are computationally expensive and their generalization to spatio-temporal filters will lead to even higher dimensionality resulting in a prohibitive complexity for many real-time applications, and fast implementations must be devised to mitigate this. 4) Microphones often behave in a non-ideal way and timing issues between microphones may occur due to calibration errors and this may render the estimated signals and parameters useless. It is therefore important that this problem be addressed in a tractable manner. These challenges can be addressed in the following way. There are several possible solutions to the first one. A promising approach is to use so-called perturbed signal models allowing for small deviations of the frequencies of the individual harmonics [23], although their generalization to the present case is non-trivial. To deal with the nonstationarity, sample-by-sample updates of parameters and filters and by time-recursive implementations, both based on exploiting that the statistics and parameters evolve smoothly most of the time (e.g., fundamental frequency and DOA), are possible solutions. A potential solution to the second challenge is to generalize the filter designs of [18] and incorporate the ideas of [17], in which the speech process is modeled as an auto-regressive process, something that works well for unvoiced speech. The third challenge can be addressed by exploiting the structure of the involved matrices and vectors, using rooting algorithms, time- and order-recursive implementations, LMS-like algorithms and asymptotic approximations [23, 24]. Finally, solutions to the fourth challenge can draw upon inspiration from recent theoretical advances in robust beamforming [12].

The underlying hypothesis of the project is that it is possible to address separation and enhancement of speech signals by incorporating adaptive signal models in beamformers, i.e., taking the properties of speech signals into account, thus resulting in spatio-temporal filters. This should lead to a significant improvement in speech intelligibility and speaker identification as compared to state-of-the-art methods, two factors that are of the utmost importance in many applications. There is reason to believe that the spatio-temporal filters have the potential to do this for the following reasons: 1) they incorporate speech models, such that the design is optimized for the signal of interest; 2) they are capable of adapting to the environment and the speaker of interest, i.e., the design is adaptive; 3) they can reject interfering speakers and noise by suppressing everything as much as possible, while leaving the desired signal unchanged without prior knowledge of the statistics of interfering sources and noise; 4) they do all this jointly and optimally in both time and space.

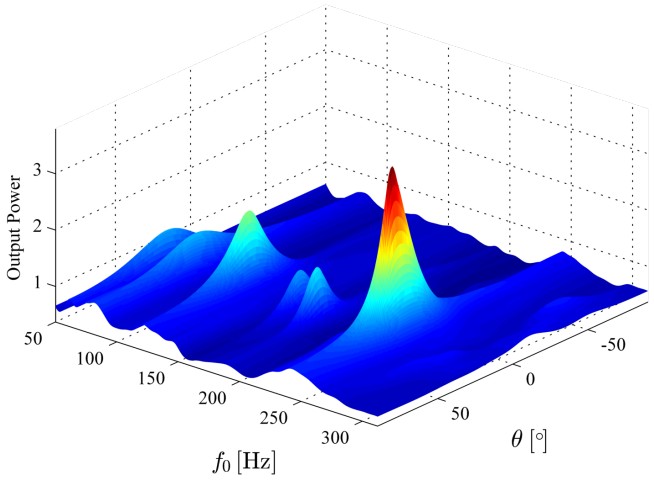

Fig. 2. Example of the output power in (5) as a function of the pitch and the DOA for a signal in white noise having a pitch of 250 Hz and DOA of 30°.

4. CONCLUSION

In this paper, we have given an overview of the motivation and ideas behind an ongoing research project. The project aims at addressing the problems of speech enhancement and separation using microphone arrays based on optimal filtering methods. Using a new class of optimal filtering techniques derived specifically for periodic sources, adaptive spatio-temporal filters are obtained that let the signal of interest pass undistorted while interferences are cancelled and noise is suppressed. The project seeks to address a number of challenges associated with making these filters applicable to speech signals, namely that speech is not perfectly periodic, that both unvoiced and voiced speech should be handled in a consistent manner, that the computational complexity is reduced, and that robustness towards imperfections in the hardware is achieved.

5. REFERENCES

[1] J. Benesty, J. Chen, Y. Huang, and I. Cohen, *Noise Reduction in Speech Processing*. Springer-Verlag, 2009.

[2] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons Ltd, 2006.

[3] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," vol. 34(5), pp. 1124–1138, Oct. 1986.

[4] M.-Y. Zou, C. Zhenming, and R. Unbehauen, "Separation of periodic signals by using an algebraic method," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 5, 1991, pp. 2427–2430.

[5] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14(4), pp. 1218–1234, 2006.

[6] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50(9), pp. 2230–2244, 2002.

[7] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57(8), pp. 1408–1418, 1969.

[8] O. L. Frost III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60(8), pp. 926–935, 1972.

[9] M. Brandstein and D. Ward, *Microphone Arrays - Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[10] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Springer-Verlag, 2010.

[11] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Process.*, vol. 53(5), pp. 1684–1696, 2005.

[12] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust Capon beamformer," *IEEE Trans. Signal Process.*, vol. 52(9), pp. 2407–2423, 2004.

[13] A. Jakobsson, S. L. Marple, and P. Stoica, "Computationally efficient two-dimensional Capon spectrum analysis," *IEEE Trans. Signal Process.*, vol. 48(9), pp. 2651–2661, 2000.

[14] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1997, pp. 375–378.

[15] M. Képesi, L. Ottowitz, and T. Habib, "Joint position-pitch estimation for multiple speaker scenarios," in *Proc. Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 85–88.

[16] M. S. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1998, pp. 3613–3616.

[17] J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16(8), no. 8, pp. 1490–1502, 2008.

[18] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58(12), pp. 5969–5983, 2010.

[19] 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.*, 2010.

[20] ——, "Nonlinear least squares methods for joint DOA and pitch estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21(5), pp. 923–933, 2013.

[21] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. on Advances in Signal Processing*, vol. 2012(1), pp. 1–11, 2012.

[22] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "An optimal spatio-temporal filter for extraction and enhancement of multi-channel periodic signals," in *Rec. Asilomar Conf. Signals, Systems, and Computers*, 2010, pp. 1846–1850.

[23] M. G. Christensen and A. Jakobsson, *Multi-Pitch Estimation*, ser. Synthesis Lectures on Speech & Audio Processing. Morgan & Claypool Publishers, 2009, vol. 5.

[24] G. Glentis and A. Jakobsson, "Efficient implementation of iterative adaptive approach spectral estimation techniques," *IEEE Trans. Signal Process.*, vol. 59(9), pp. 4154–4167, 2011.