HARMONIC MINIMUM MEAN SQUARED ERROR FILTERS FOR MULTICHANNEL SPEECH ENHANCEMENT

Jesper Rindom Jensen[†]*, Mads Græsbøll Christensen*[†]*, and Andreas Jakobsson*[‡]

[†]Audio Analysis Lab, AD:MT, Aalborg University, Denmark
[‡]Centre for Mathematical Sciences, Lund University, Sweden
Emails: {jrj,mgc}@create.aau.dk, aj@maths.lth.se

ABSTRACT

Many state-of-the-art multichannel speech enhancement methods rely on second-order statistics of the desired speech signal, the noise signal, or both. Estimation of those are difficult in practice, resulting in a practical performance that is typically much lower than their potential theoretical performance. We propose two multichannel enhancement techniques that instead rely on a model for voiced speech. That is, the proposed methods are driven by the signals' fundamental frequencies, which may be accurately estimated even in noisy scenarios. The first method is designed independently of the microphone array geometry and source position, whereas these are utilized in the second approach. Thereby, we can investigate when to exploit such information in the case of localization errors and violations of the spatial assumptions. Numerical results show that the proposed method is able to outperform competing methods in terms of both output SNRs and PESQ scores.

Index Terms— multichannel speech enhancement, voiced speech, MMSE filtering, harmonic filters, DOA mismatch.

1. INTRODUCTION

During recent decades, much effort has gone into removing noise from recordings of speech; this problem is referred to as speech enhancement. The problem is of uttermost importance in numerous applications, such as human-machine interaction, hearing-aids, and hands-free communication. Recently, solutions to the enhancement problem have been sought through employment of multiple microphones, adding extra dimensionality to the problem, which in turn allow for higher degrees of noise reduction. Such hardware setups are becoming common in modern audio equipment, such as in hearing aids, wireless headsets, and loudspeakers.

Typically, many of the recent proposals for multichannel enhancement are driven by knowledge about the noise statistics. Examples of such are linear filtering, subspace, statistical, and spectral subtractive methods (see, e.g., [1–4]). Some of these methods have been shown to achieve excellent results in theory, when accurate estimates of the necessary statistics are available [1]. However, in practice, the statistics are often difficult to estimate, e.g., due to simultaneous speech and noise presence, having a detrimental impact on the noise reduction performance. This has spurred interest in parametric model-based speech enhancement, in which estimates of speech parameters, such as the fundamental frequency, drives the noise reduction rather than the noise statistics. It has been shown in earlier work that these parameters can be estimated very accurately even at low signal-to-noise ratios (SNRs) [5, 6]. The model-based approach has mainly been considered for single-channel enhancement; some examples of this development be found in [7–10]. A purely parametric model-based approach will not suffer from the difficulties of estimating noise statistics, but it is targeted only towards some parts of speech signals like voiced speech. In practice, it should therefore be applied only to these parts of the speech, e.g., by using voiced-unvoiced speech detection [11–13] or a parallel approach as in [14, 15].

In this paper, we propose two parametric model-based speech enhancement methods based on linear filtering. The methods are based on minimization of the mean square error between the filter output and the desired speech signal, which herein is assumed to be periodic voiced speech. This approach has previously been used for spectral estimation [16], and single-channel speech enhancement [9]. Compared to the state-of-the art method in [15], the proposed methods are formulated in the time-domain instead of the frequency domain. By designing the filters in the time-domain, we can ensure that the desired signal is undistorted by satisfying only a few linear constraints. This is much more complicated to achieve in the frequency domain due to spectral leakage. Moreover, the proposed approach tackles the multichannel speech enhancement problem as opposed to the single-channel methods in [7–10, 15]. The first of the proposed methods is derived independently of the array geometry and of the relative positions of the array and the source, whereas the second method takes these aspects into account. Another contribution of the paper is that this enables us to shed light on which of these approaches to utilize in case of localization errors and model violations.

These topics are covered by first introducing the signal model and problem formulation in Section 2. Then, in Section 3, we present the proposed filtering methods, followed by experimental results in Section 4. Finally, a discussion of the obtained results is found in Section 5.

2. PROBLEM FORMULATION

We consider a scenario where K microphones are used for recording a desired speech signal corrupted by additive noise. Mathematically, the recording by microphone k may be formulated as

$$y_k(n) = x_k(n) + v_k(n), \tag{1}$$

for time instances $n = 0, \dots, N-1$ and $k = 0, \dots, K-1$, where $x_k(n)$ is the desired speech signal and $v_k(n)$ the additive noise, which could be constituted by, e.g., late reverberation, stationary background noise, or interfering speech sources or a combination thereof. The multichannel enhancement problem aims at extracting $x_0(n)$ from the recorded data, with little or no speech distortion, and

This work was supported in part by the Swedish Research Council.

with as much noise reduction as possible. To facilitate this task, we consider N time-consecutive and time-reversed samples from each microphone, such that

$$\mathbf{y}_k = \begin{bmatrix} y_k(0) & y_k(-1) & \cdots & y_k(-N+1) \end{bmatrix}^T = \mathbf{x}_k + \mathbf{v}_k, \quad (2)$$

with $(\cdot)^T$ denoting the transpose, and where the signal and noise vectors, $\mathbf{x}_k$ and $\mathbf{v}_k$, are defined similarly to $\mathbf{y}_k$. Our aim is thus to extract $x_0(0)$ from the recordings in $\mathbf{y}_0, \ldots, \mathbf{y}_{K-1}$. The enhancement methods proposed herein are targeted towards voiced parts of the speech, which typically is the most dominant part of a general speech recording. For such parts, we can accurately model frames shorter than 20 ms as [17]

$$y_k(n) = \sum_{l=-L}^{L} \alpha_{l,k} e^{jl\omega_0 n} + v_k(n), \quad (3)$$

where ω_0 denotes the fundamental frequency, L the number of harmonic components, and $\alpha_{l,k}$ the complex harmonic amplitudes. It is assumed that the sought speech signal is zero mean, such that $\alpha_{0,k} = 0$. At this point, it should be noted that the model in (3) and the filters derived in Sec. 3 may easily be extended for non-stationary speech by allowing the fundamental frequency to vary linearly over time as shown in [17]. While the model in (3) allows for an accurate modeling of voiced speech, we envision that the harmonicity-based filters proposed in the following section also to be used in a hybrid scheme, like that of [14, 15]. That is, they should only be applied to the relevant parts of the speech part of the signal, whereas traditional noise-driven filters should be used for the unvoiced parts of the speech.

3. HARMONIC MULTICHANNEL ENHANCEMENT

To tackle the speech enhancement problem, we consider two different approaches both based on linear filtering of the observed data. That is, we extract the desired speech signal, $x_0(n)$, from the microphone k recordings, $\mathbf{y}_k(n)$, as

$$\widehat{x}_{0,k}(n) = \mathbf{h}_k^T \mathbf{y}_k(n), \quad (4)$$

with

$$\mathbf{h}_k = \begin{bmatrix} h_k(0) & \cdots & h(M-1) \end{bmatrix}^T, \quad (5)$$

$$\mathbf{y}_k(n) = \begin{bmatrix} y_k(n) & \cdots & y_k(n-M+1) \end{bmatrix}^T, \quad (6)$$

where M denotes the filter length. In the first approach, we design the filters independently of the array geometry and of the relative positioning of the array and the source, whereas we take this information into account in the second approach. We do this in order to investigate which approach is better in case of, e.g., localization errors and different degrees of violations to the model assumptions. In the following subsections, we refer to these filtering approaches as the geometry-based and geometry independent filters.

3.1. Geometry Independent Filters

When the geometry of the sensor array is not known or utilized, we may model the M observations in each microphone as

$$\mathbf{y}_k(n) = \mathbf{Z}\boldsymbol{\alpha}_k(n) + \mathbf{v}_k(n), \quad (7)$$

with

$$\boldsymbol{\alpha}_k(n) = \mathbf{w}^*(n) \odot \boldsymbol{\alpha}_k, \quad (8)$$

$$\mathbf{w}(n) = \begin{bmatrix} e^{-jL\omega_0 n} \cdots e^{-j\omega_0 n} \; e^{j\omega_0 n} \cdots e^{jL\omega_0 n} \end{bmatrix}^H,$$

$$\boldsymbol{\alpha}_k = \begin{bmatrix} \alpha_{k,L}^* \cdots \alpha_{k,1}^* \; \alpha_{k,1} \cdots \alpha_{k,L} \end{bmatrix}^T,$$

where $(\cdot)^*$ and $\odot$ denote the elementwise complex conjugate and product operators, respectively. Furthermore, let

$$\mathbf{Z} = [\mathbf{z}_{-L} \cdots \mathbf{z}_{-1} \, \mathbf{z}_1 \cdots \mathbf{z}_L], \quad (9)$$

$$\mathbf{z}_l = \begin{bmatrix} 1 \; e^{-jl\omega_0} \cdots e^{-jl\omega_0(M-1)} \end{bmatrix}^T. \quad (10)$$

Using this model, one may introduce a multichannel speech enhancement filter without information about the array geometry and source positioning. The proposed method consists of two stages: first, one extracts the desired speech individually from each channel and, secondly, one optimally weighs the different speech estimates to achieve the multichannel enhancement.

The filters are based on minimization of a mean squared error (MSE), reminiscent of the technique used in the amplitude and phase estimation (APES) method introduced in [16], and in the filters for single channel enhancement and separation of periodic signals introduced in [9]. The idea is to formulate a criterion that enables us to make the output of the filter at microphone k resemble a periodic signal at microphone 0 as much as possible. This may be achieved through minimization of the following MSE:

$$P_k = \frac{1}{N-M+1} \sum_{n=M-1}^{N-1} \left| \mathbf{h}_k^T \mathbf{y}_k(n) - \boldsymbol{\alpha}_0^H \mathbf{w}(n) \right|^2. \quad (11)$$

In order to do so, we first solve this problem for microphone 0, noting that the MSE may be expressed as

$$P_0 = \mathbf{h}_0^T \mathbf{R}_0 \mathbf{h}_0 - \boldsymbol{\alpha}_0^H \mathbf{G}_0 \mathbf{h}_0 - \mathbf{h}_0^T \mathbf{G}_0^H \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_0^H \mathbf{W} \boldsymbol{\alpha}_0 \quad (12)$$

with

$$\mathbf{R}_k = \sum_{n=M-1}^{N-1} \frac{\mathbf{y}_k(n)\mathbf{y}_k^T(n)}{N-M+1}, \quad \mathbf{G}_k = \sum_{n=M-1}^{N-1} \frac{\mathbf{w}(n)\mathbf{y}_k^T(n)}{N-M+1},$$

$$\mathbf{W} = \sum_{n=M-1}^{N-1} \frac{\mathbf{w}(n)\mathbf{w}^H(n)}{N-M+1}.$$

Minimizing the MSE with respect to the amplitudes, $\boldsymbol{\alpha}_0$, typically being unknown in practice, yields

$$\widehat{\boldsymbol{\alpha}}_0 = \mathbf{W}^{-1} \mathbf{G}_0 \mathbf{h}_0. \quad (13)$$

These estimates can be used to rewrite the MSE as

$$P_0 = \mathbf{h}_0^T \mathbf{Q}_0 \mathbf{h}_0, \quad (14)$$

where $\mathbf{Q}_k = \mathbf{R}_0 - \mathbf{G}_0^H \mathbf{W}^{-1} \mathbf{G}_0$. We note that $\mathbf{Q}_k$ might be interpreted as a particular noise covariance matrix estimate for microphone k (see also [18]), i.e., the MSE in (14) can be seen as the residual noise power after filtering. That is, to design our noise reduction filter for microphone 0, we minimize (14) with respect to the residual noise power, with the constraint that the desired signal should be undistorted in the process. From the model in (7), we deduct that this is achieved by solving

$$\min_{\mathbf{h}_0} \mathbf{h}_0^T \mathbf{Q}_0 \mathbf{h}_0 \quad \text{s.t.} \quad \mathbf{h}_0^T \mathbf{Z} = \mathbf{1}^T, \quad (15)$$

with $\mathbf{1}$ being a $L \times 1$ vector of ones. The solution to this quadratic optimization problem is given in closed-form as

$$\mathbf{h}_0 = \mathbf{Q}_0^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Q}_0^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \quad (16)$$

That is, the MSE of the enhancement filter for microphone 0 is

$$P_0 = \mathbf{1}^T\left(\mathbf{Z}^H\mathbf{Q}_0^{-1}\mathbf{Z}\right)^{-1}\mathbf{1}. \quad (17)$$

Next is to extract the desired signal from the other microphones. We achieve this by minimizing P_k with respect to $\mathbf{h}_k$, for $k = 1, \dots, K-1$, using the amplitude estimates in (13) for microphone 0 in place of $\boldsymbol{\alpha}_0$ in (11). The resulting filter designs are given by

$$\mathbf{h}_k = \mathbf{R}_k^{-1}\mathbf{G}_k^H\mathbf{W}^{-1}\mathbf{G}_0\mathbf{h}_0. \quad (18)$$

The MSEs of these filters are found by inserting (18) into (11), yielding

$$P_k = \mathbf{h}_0^T\mathbf{G}_0^H\mathbf{W}^{-1}\left(\mathbf{W} - \mathbf{G}_k\mathbf{R}_k^{-1}\mathbf{G}_k^H\right)\mathbf{W}^{-1}\mathbf{G}_0\mathbf{h}_0. \quad (19)$$

Since one may easily obtain the MSEs of the different filters, we can use these to combine the extracted speech from each channel into a single enhanced speech signal. This is achieved by taking a weighted mean of the estimates as [19]

$$\widehat{x}_0(n) = \frac{\mathbf{1}^T\mathbf{P}^{-1}\widehat{\mathbf{x}}_0(n)}{\mathbf{1}^T\mathbf{P}^{-1}\mathbf{1}}, \quad (20)$$

where

$$\mathbf{P} = \text{diag}\left(\begin{bmatrix} P_0 & \cdots & P_{K-1} \end{bmatrix}^T\right), \quad (21)$$

$$\widehat{\mathbf{x}}_0(n) = \begin{bmatrix} \widehat{x}_{0,0}(n) & \cdots & \widehat{x}_{0,K-1}(n) \end{bmatrix}^T. \quad (22)$$

In the remainder of the paper, we refer to this method as the harmonic multichannel MMSE method (M-MMSE).

3.2. Geometry-based Filters

With knowledge of the array geometry and the location of the speech source, one may beneficially take this information into account in the filter design. This is here illustrated for the case of a uniform linear array (ULA) under a far-field assumption. In such scenarios, the signal model can be further specified as [20]

$$y_k(n) = \sum_{l=-L}^{L} \alpha_l e^{jl\omega_0 n} e^{-j\tau_k l \omega_0 f_s} + v_k(n) \quad (23)$$

with $\tau_k = k\frac{d \sin\theta}{c}$ being the time difference of arrival (TDOA) of the speech source between microphones 0 and k, d the microphone spacing, θ the source direction of arrival (DOA), c the sound propagation speed, and f_s the sampling frequency. Thus, one may model the M samples recorded by each microphone as

$$\mathbf{y}_k(n) = \mathbf{Z}\mathbf{D}_k\boldsymbol{\alpha}_0(n) + \mathbf{v}_k(n), \quad (24)$$

where

$$\mathbf{D}_k = \text{diag}\left(\begin{bmatrix} e^{-jLk\eta_\theta} & \cdots & e^{-jk\eta_\theta} & e^{jk\eta_\theta} & \cdots & e^{jLk\eta_\theta} \end{bmatrix}\right),$$

with $\eta_\theta = \omega_0 f_s \frac{d \sin\theta}{c}$. We then design the filters for each microphone by minimizing (11) under the constraint that the desired signal from the DOA, θ, should be passed undistorted, i.e.,

$$\min_{\mathbf{h}_k} \mathbf{h}_k^T\mathbf{Q}_k\mathbf{h}_k \quad \text{s.t.} \quad \mathbf{h}_k^T\mathbf{Z}\mathbf{D}_k = \mathbf{1}^T. \quad (25)$$

Since the assumed DOA is included in the constraints, the outputs of microphone filters will be time aligned. The solution to the above optimization problem is given by

$$\mathbf{h}_k = \mathbf{Q}_k^{-1}\mathbf{Z}\left(\mathbf{Z}^H\mathbf{Q}_k^{-1}\mathbf{Z}\right)^{-1}\mathbf{D}_k\mathbf{1}. \quad (26)$$

Thus, the MSE of filter k is

$$P_k = \mathbf{1}^T\mathbf{D}_k^H\left(\mathbf{Z}^H\mathbf{Q}_k^{-1}\mathbf{Z}\right)^{-1}\mathbf{D}_k\mathbf{1}. \quad (27)$$

To obtain the final signal estimate, we combine the signal estimates from each channel as in (20), with the weights as given by (27). This method is termed the geometry-based harmonic multichannel MMSE (GM-MMSE) method.

4. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed methods, we applied them on different scenarios where speech signals of interest sampled at 8 kHz are corrupted by reverberation and diffuse babble noise [21] at an input SNR of 10 dB. Also, we included two comparison methods in the evaluation: a single-channel harmonic MMSE (S-MMSE) filter [9] applied on microphone 0, and delay-and-sum beamforming followed by the S-MMSE filter as a postfilter (DSB+MMSE). The proposed methods were only compared to competing signal-driven enhancement methods based on the harmonic model, since in practice they would have to be used in a hybrid scheme where they are only applied to the voiced speech signal parts as mentioned in the introduction. This may though be achieved by using a voiced-unvoiced speech detector [11–13] or a parallel approach as in [14, 15], but this is out of the scope of this paper. The multichannel microphone data were generated from single-channel recordings speech recordings (three male and one female speech sentence [22]) by convolving with room impulse responses (RIRs) obtained using an RIR generator [23], with the following setup: the center of a uniform linear two-microphone array was placed at $[3.5, 1, 1]$ m in parallel with the x-axis in a room with dimensions $8 \times 6 \times 4$ m. Moreover, the microphones were omnidirectional and spaced by 0.2 m. The source was placed 2 m from the array at an angle of $-40°$, emitting sound with a speed of 343 m/s. Finally, the T_{60} was 0.2 s and the simulated RIRs were highpass filtered. Before applying the enhancement, the fundamental frequency was estimated from microphone 0 for each time instance using a fast implementation [24] of the nonlinear least squares estimator in [5, 6] and the model order was obtained using the method in [25]. Then, the harmonic MMSE filters were all implemented with $M = 40$, $N = 160$, and $L = 12$. We could have used the model order estimates from the fundamental frequency estimation step, but we found these to be perceptually too low. Moreover, a small diagonal loading of $\epsilon = 1e-8$ was added to the matrix inverses in the MMSE filters, since the harmonics can be linearly dependent for low fundamental frequencies. The DSB beamformer used in one of the comparison methods was applied on 50 % time-overlapping blocks of 200 samples, using overlap-add with Hanning windows.

With this setup, we first investigated the enhancement performance versus errors in the assumed DOA of the speech source. The

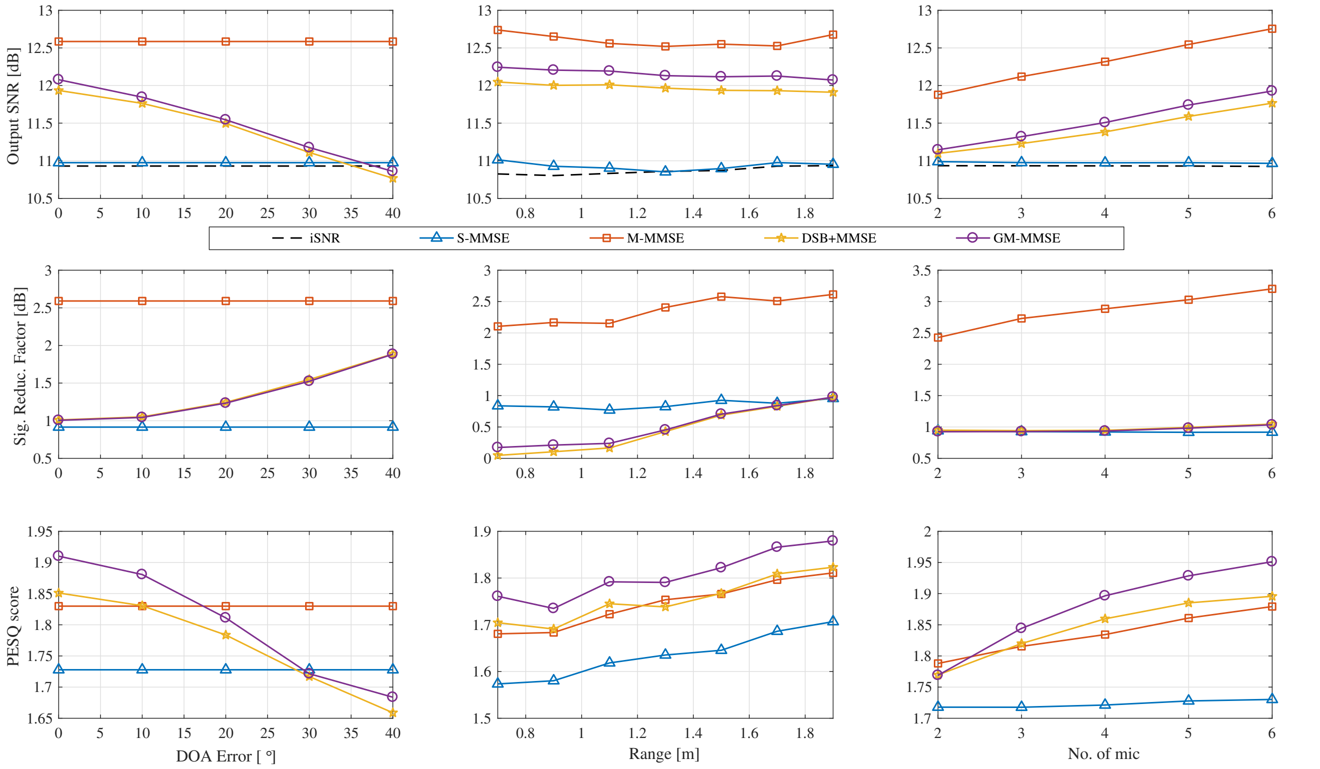

Fig. 1. Measured output SNRs, signal reduction factors, and PESQ scores for the proposed and comparison methods in speech with diffuse babble noise scenarios.

results from this and the following evaluations are shown in Fig. 1. As can be seen in the figure, the M-MMSE constantly delivers the highest output SNR for all errors, which is expected as it does not make use of the geometry. The DSB+MMSE and GM-MMSE methods have a decreasing output SNR for increasing errors, and for high errors, they may even give lower output SNR than the single-channel S-MMSE method. However, the geometry based methods have less distortion than the M-MMSE method for all errors. The PESQ scores [26], which are objective measures reflecting the perceptual quality, show that the GM-MMSE is preferred for errors below 17°, and the DSB+MMSE below 10°, when compared with the M-MMSE method. In the second evaluation, we looked at the performance versus the source range. The purpose of this is to see what the effects are of violating the far-field assumption, being assumed by the GM-MMSE and DSB+MMSE methods. Here, the filters show approximately the same trend, i.e., almost constant output SNR, slightly increasing signal distortion, and increasing PESQ scores for an increasing range. Moreover, in terms of PESQ scores, the best performance is obtained by the GM-MMSE method followed by the M-MMSE and DSB+MMSE methods having similar performance and then the S-MMSE method. It should be noted that perfect DOA information is assumed here, so the M-MMSE method will be closer to or better than the GM-MMSE method in practice. Finally, we measured the performance versus the number of microphones. In this experiment, the microphone spacing was 0.05 m. We see that all the multichannel methods benefit from having further microphones in terms of output SNR and PESQ scores. However, the gain in PESQ scores is higher with the GM-MMSE and DSB+MMSE methods than with the M-MMSE approach. Again, the M-MMSE method is expected to perform better than the others in practice according to the first evaluation, since we will have DOA estimation errors due to reverberation, interferers, and background noise.

5. DISCUSSION

The topic considered in the paper is multichannel enhancement of speech, which is a classical signal processing problem within the audio and speech processing community. Most existing methods for this problem rely on second-order statistics of the desired speech signal, the noise signal, or both. However, obtaining these statistics in practice is difficult due to the nonstationarity of the speech and noise. An alternative approach is to use model-based approach, e.g., to exploit the periodicity of voiced speech through a harmonic model. With this in mind, we proposed two multichannel speech enhancement methods: the first is derived independent of the microphone array geometry and the speech location, whereas this information is utilized in the other method. Compared to existing harmonic multichannel enhancement methods, the proposed methods are formulated in the time-domain, which means that we can ensure a distortionless filter design by meeting a few linear constraints. By introducing two different methods, we investigate which approach is preferable to use in scenarios with, e.g., DOA errors and violation of far-field assumptions. Our evaluation in terms PESQ scores showed that if the DOA is known, we should apply the method utilizing the geometry even if the far-field assumptions are violated. However, if we introduce DOA errors, which are always present and significant in practice, the geometry independent approach is preferable in terms of both output SNR and PESQ scores. Finally, the proposed methods also show output SNR and PESQ improvements over the comparison methods.

6. REFERENCES

[1] J. R. Jensen, J. Benesty, and M. G. Christensen, "Noise reduction with optimal variable span linear filters," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 4, pp. 631–644, Apr. 2016.

[2] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sep. 2002.

[3] J. Benesty, M. Souden, and J. Chen, "A perspective on multichannel noise reduction in the time domain," *Applied Acoustics*, vol. 74, no. 3, pp. 343–355, Mar. 2013.

[4] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel signal enhancement algorithms for assisted listening devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Proc. Mag.*, vol. 32, no. 2, pp. 18–30, Mar. 2015.

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

[6] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Elsevier Signal Process.*, vol. 88, no. 4, pp. 972–983, Apr. 2008.

[7] J. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 7, pp. 731–740, Oct. 2001.

[8] W. Jin, X. Liu, M. S. Scordilis, and L. Han, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 2, pp. 356–368, Feb. 2010.

[9] M. G. Christensen and A. Jakobsson, "Optimal filter designs for separating and enhancing periodic signals," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 5969–5983, Dec. 2010.

[10] J. R. Jensen, J. Benesty, M. G. Christensen, and S. H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[11] Y. Qi and B. R. Hunt, "Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 250–255, Apr. 1993.

[12] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 2, pp. 502–510, Mar. 2006.

[13] M. K. I. Molla, K. Hirose, and N. Minematsu, "Robust voiced/unvoiced speech classification using empirical mode decomposition and periodic correlation model," in *Proc. Interspeech*, Sep. 2008, pp. 2530–2533.

[14] J. Le Roux, S. Watanabe, and J. R. Hershey, "Ensemble learning for speech enhancement," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.*, Oct. 2013, pp. 1–4.

[15] M. Krawczyk-Becker and T. Gerkmann, "Fundamental frequency informed speech enhancement in a flexible statistical framework," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 24, no. 5, pp. 940–951, May 2016.

[16] J. Li and P. Stoica, "An adaptive filtering approach to spectral estimation and SAR imaging," *IEEE Trans. Signal Process.*, vol. 44, no. 6, pp. 1469–1484, Jun. 1996.

[17] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Parameter estimation and optimal segmentation of non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, pp. 1–11, 2016, accepted.

[18] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, Inc., 1993.

[20] 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.

[21] E. A. P. Habets, I. Cohen, and S. Gannot, "Generating nonstationary multisensor signals under a spatial coherence constraint," *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 2911–2917, Nov. 2008.

[22] M. P. Cooke, *Modelling auditory processing and organisation*, Ph.D. thesis, Published by Cambridge University Press, 1993, Data: ftp://ftp.dcs.shef.ac.uk/share/spandh/cookephd/.

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

[24] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast and statistically efficient fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 86–90.

[25] J. K. Nielsen, M. G. Christensen, and S. H. Jensen, "Default bayesian estimation of the fundamental frequency," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 3, pp. 598–610, Mar. 2013.

[26] ITU-T, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," , no. P.862, pp. 1–30, Feb. 2001.