Fundamental Frequency Informed Speech Enhancement in a Flexible Statistical Framework

Martin Krawczyk-Becker, *Student Member, IEEE*, and Timo Gerkmann, *Senior Member, IEEE*

***Abstract*—Conventional statistical clean speech estimators, like the Wiener filter, are frequently used for the spectro-temporal enhancement of noise corrupted speech. Most of these approaches estimate the clean speech independently for each time-frequency point, neglecting the structure of the underlying speech sound. In this work, we derive a statistical estimator that explicitly takes into account information about the characteristic structure of voiced speech by means of a harmonic signal model. To this end, we also present a way to estimate a harmonic model-based clean speech representation and the corresponding error variance directly in the short-time Fourier transform domain. The resulting estimator is optimal in the minimum-mean-squared error sense and can conveniently be formulated in terms of a multichannel Wiener filter. The proposed estimator outperforms several reference algorithms in terms of speech quality and intelligibility as predicted by instrumental measures.**

***Index Terms*—Speech enhancement, noise reduction, signal reconstruction.**

I. Introduction

REAL-TIME capable algorithms that mitigate the detrimental effect of acoustic noise are a key component to make communication devices like hearing aids or mobile phones work reliably in adverse conditions. Over the years, various approaches for the reduction of acoustic noise have been proposed. Besides spatial methods that use multiple microphone signals, single-channel noise reduction schemes that utilize spectro-temporal cues are commonly employed either in isolation if only a single microphone is available, or to further enhance the output of a spatial preprocessing stage. In this contribution we focus on single-microphone algorithms that are formulated in the short-time discrete Fourier transform (STFT) domain, which is commonly used due to the low complexity and intuitive interpretation of this transform. Among the most successful proposals are those based on statistical assumptions of the speech and the noise, like the Wiener filter or Ephraim and Malah's short-time spectral amplitude estimator (STSA) [1], which can be derived in a Bayesian framework. Both approaches assume that the spectral coefficients of the speech and the noise are circularly complex Gaussian distributed, mutually independent, and also independent from neighboring time-frequency points. Inspired by the seminal work in [1], numerous extensions and alternatives have been proposed over the past three decades. Improvements have for example been achieved by using more elaborate models for the distribution of the speech and by taking into account the compressive character of the human ear in the optimization function, see e.g. [2] and the references therein.

The achievable performance of these techniques is however limited, since the assumption that neighboring time-frequency points are mutually independent is not fulfilled in practice. For example, correlations between neighboring time-frequency points are inevitable for overlapping STFT segments of finite length. Alternative estimators that explicitly consider the correlations of neighboring spectral coefficients have been presented e.g. in [3]–[5], where the estimation problem is formulated in terms of a single-microphone minimum variance distortionless response (MVDR) beamformer. These algorithms benefit from incorporating more information at the price of a more challenging parameter estimation and a higher complexity. Besides the correlations due to the STFT analysis, the speech signal itself may show characteristic spectro-temporal structures, e.g. for voiced speech. A different approach to exploiting such structures has for example been proposed in [6], where the parameters of a sinusoidal model are iteratively estimated from a noisy observation to recover the clean speech signal. Furthermore, in [7], signal-adaptive filters (and filterbanks) for the time-domain enhancement of noise corrupted periodic signals, like voiced speech, are derived. More specifically, a sinusoidal model is used as the target signal in the error criterion of the filter design. In contrast to traditional statistical estimators like the Wiener filter, such sinusoidal model-based approaches explicitly take into account the structure of the underlying speech sound. This additional a priori knowledge about the observed signals can lead to improvements over conventional statistical estimators as long as the employed signal model holds. For example, a harmonic model, i.e. a sinusoidal model for which the frequencies of the sinusoids are integer multiples of the fundamental frequency, is well suited to represent voiced speech. Its applicability to fricatives and transients is however limited, which may lead to suboptimal speech enhancement results.

There are several approaches that use a harmonic signal model in a more robust way, alleviating its limitations in unvoiced sounds. For example, in [8], the output of [7] is

Manuscript received September 24, 2015; revised January 08, 2016 and February 05, 2016; accepted February 05, 2016. Date of publication February 23, 2016; date of current version March 23, 2016. This work was supported by the DFG Project GE2538/2-1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Graesbøll Christensen.

M. Krawczyk-Becker is with the Speech Signal Processing Group, Department of Medical Physics and Acoustics, Cluster of Excellence "Hearing4all," Universität Oldenburg, Oldenburg 26111, Germany (e-mail: martin.krawczyk-becker@uni-oldenburg.de).

T. Gerkmann is with Technicolor Research and Innovation, 30625 Hanover, Germany (e-mail: timo.gerkmann@uni-oldenburg.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASLP.2016.2533867

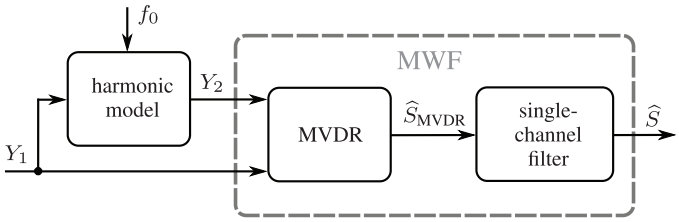

Fig. 1. Block diagram illustrating the proposed STFT-domain estimator for a single time-frequency point. In addition to the noisy observation Y_1, a harmonic model is used to create a second speech representation Y_2, which incorporates information about the fundamental frequency. Both signals are combined using an MWF, i.e. an MVDR filter followed by a single-channel post processing.

not directly used as the enhanced signal, but rather subtracted from the noisy input signal to facilitate the estimation of the noise statistics. This enhanced estimate is then employed in a more robust signal-independent filtering method [9]. In [10], for voiced speech a harmonic signal model is used to design a frequency dependent residual noise floor as well as a spectral gain that protects harmonic components. This is achieved by means of an adaptive comb-filter, which applies less attenuation in the vicinity of a speech component. In unvoiced regions and silence, [10] falls back to a Wiener-like enhancement scheme. A general method to combine two or more speech enhancement algorithms to accumulate their individual strengths has been proposed in [11], where ensemble learning techniques are used to merge the separate enhancement results. For example, a support vector machine is trained to estimate an improved spectro-temporal binary mask based on the spectral weighting of the individual enhancement schemes, which are all applied in parallel.

Two approaches that explicitly combine a harmonic signal model with a statistical estimator for STFT domain speech enhancement have been proposed in [12], [13], where a harmonic model is used to estimate a deterministic speech component. In [12] a statistical estimator and a harmonic model based estimator are employed in parallel and mixed according to the time-frequency dependent probability that the current speech sound is either stochastic, deterministic, or absent. In [13] the speech is modeled by means of a harmonic plus noise model, where the noise represents unvoiced speech, leading to non-zero-mean spectral coefficients of voiced speech. A minimum mean squared error (MMSE) optimal estimator of the amplitudes of the non-zero-mean spectral speech coefficients is derived and combined with a maximum likelihood estimate of the spectral phase.

Also in this paper, extending our work in [14], we aim at getting the best of harmonic model based estimators and statistical estimators. The main novelty of this paper is that we propose to use the harmonic model based signal as an additional input when formulating a multichannel Wiener filter (MWF). Here, one channel is the noisy observation and the other channel is the harmonic model based signal reconstruction. The MWF can then be decomposed into an MVDR filtering that optimally combines the input signals and a single-channel post-filter (see Fig. 1). This general formulation also enables the integration of multiple microphones and state-of-the-art single-channel filters that incorporate super-Gaussian speech models and the compressive character of the human auditory system. See e.g. [2] for an overview of single-channel filters. Further novelties are the presentation of an intuitive and computationally cheap way to estimate the harmonic model directly in the STFT domain, and the way we obtain the error variance of the harmonic model.

In the following section, we introduce the theoretical basics together with the general enhancement framework. The derivation of the STFT domain harmonic model is presented in Sec. III, followed by the estimation of the noise covariance matrix in Sec. IV. The proposed estimator is analyzed in detail in Sec. V and is evaluated and compared to other approaches in Sec. VI before this work is concluded in Sec. VII.

II. Signal Model and Proposed Framework

In the STFT domain, i.e. after chopping the time-domain microphone signal into overlapping segments, applying a spectral analysis window, and computing the discrete Fourier transform (DFT), we define the observed noisy microphone signal as

$$Y_1(k,\ell) = S(k,\ell) + V_1(k,\ell). \tag{1}$$

In each time-frequency point (k,ℓ) the clean speech signal $S(k,\ell)$ is corrupted by additive noise $V_1(k,\ell)$, with frequency index k and segment index ℓ.

Single-channel MMSE optimal clean speech estimators are typically derived by finding the expected value of the spectral speech coefficients $S(k,\ell)$, or a function $f(S(k,\ell))$, given the noisy microphone signal. Implicitly, also the speech variance $\sigma_{\text{S}}^2(k,\ell) = \text{E}\big(|S(k,\ell)|^2\big)$, and the noise variance $\sigma_{\text{V}_1}^2(k,\ell) = \text{E}\big(|V_1(k,\ell)|^2\big)$ are given, where $\text{E}(\cdot)$ denotes statistical expectation. Note that the speech and noise variances can also be interpreted as power spectral densities [15]. Assuming that neighboring time-frequency points are independent, this can be formulated as

$$\widehat{f(S)} = \text{E}\big(f(S)|Y_1, \sigma_{\text{S}}^2, \sigma_{\text{V}_1}^2\big), \tag{2}$$

where we dropped k and ℓ for brevity, as the processing is carried out separately for each time-frequency point. This notational convenience is used for the remainder of this work wherever appropriate. Further, the hat symbol is used to denote estimates, i.e. $\widehat{S}$ is an estimate of S.

For circularly complex Gaussian distributed and mutually independent speech and noise coefficients, the posterior probability of the speech component is (e.g. [16, Chap. 3.11])

$$\begin{aligned} p\left(S|Y_1, \sigma_{\text{S}}^2, \sigma_{\text{V}_1}^2\right) &= \mathcal{N}\left(\frac{\sigma_{\text{S}}^2}{\sigma_{\text{S}}^2 + \sigma_{\text{V}_1}^2} Y_1, \frac{\sigma_{\text{S}}^2 \sigma_{\text{V}_1}^2}{\sigma_{\text{S}}^2 + \sigma_{\text{V}_1}^2}\right) \\ &= \mathcal{N}\left(\widehat{S}_W, \sigma_W^2\right), \end{aligned} \tag{3}$$

where $\mathcal{N}\left(\widehat{S}_W, \sigma_W^2\right)$ denotes a Gaussian distribution with the Wiener filter estimate as its mean

$$\widehat{S}_W = \frac{\sigma_{\text{S}}^2}{\sigma_{\text{S}}^2 + \sigma_{\text{V}_1}^2} Y_1 \tag{4}$$

and the variance

$$\sigma_W^2 = \sigma_{\mathrm{S}}^2 \sigma_{\mathrm{V}_1}^2 / \left(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}_1}^2\right). \tag{5}$$

Accordingly, the MMSE optimal estimator (2) for $f(S) = S$ is the Wiener filter in this case. Under the same distributional assumptions, the STSA [1] can be obtained from (2) for $f(S) = |S|$.

The assumption of independent time-frequency points greatly simplifies the derivations and computations, but at the same time limits the achievable improvement as neither spectral nor temporal structures of the underlying speech sound, such as the characteristic harmonic structure of voiced speech, are utilized. In voiced speech most energy is concentrated around the fundamental frequency f_0 and integer multiples of it, i.e. the harmonic frequencies. The temporal continuity of voiced speech with respect to typical STFT segment lengths as well as the spectral sparsity are strong cues that yield valuable information for the enhancement of noise corrupted speech. In this work, we utilize this information to derive an improved MMSE optimal clean speech estimator. For this, we first create an STFT domain representation of the clean speech signal based on a harmonic-model, i.e. a weighted superposition of sinusoids at the harmonic frequencies, which we denote as

$$Y_2 = S + V_2. \tag{6}$$

Here, $V_2 = Y_2 - S$ is the difference between the speech signal Y_2, reconstructed using the harmonic model, and the true clean speech S. With the use of the harmonic model the spectro-temporal structure of voiced speech is explicitly taken into account in Y_2. Y_2 is then considered as a second, again noisy, observation in addition to the microphone signal Y_1. This interpretation gives us the possibility to formulate novel single-microphone clean speech estimators in a multichannel framework, using Y_1 and Y_2 as input channels as depicted in Fig. 1. In this way, we can incorporate information about the speech's spectro-temporal structure inherent in Y_2 into an MMSE optimal clean speech estimator.

In general, the proposed principle can be extended to multiple microphone signals and also more signal models, e.g. for other speech sounds like stop-consonants. Thus, to retain and to highlight the generality of the proposed approach, we first present a general multichannel framework for M input signals. After the general formulation we focus on the specific case of only a single microphone and a statistical and a harmonic model based representation of the clean speech.

A. General Multichannel Framework

In the STFT domain $\boldsymbol{Y} = [Y_1\ Y_2\ \ldots\ Y_M]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ denotes the column vector of M noisy observations of a single desired speech signal $S \in \mathbb{C}$ at time-frequency bin (k, ℓ), with transposition operator $(\cdot)^{\mathrm{T}}$. Here we use bold letters and symbols to distinguish vectors and matrices from scalar quantities. Each element of $\boldsymbol{Y}$ can either be a microphone signal or any model-based representation of the desired speech signal. Introducing the vector $\boldsymbol{a} = [a_1\ a_2\ \ldots\ a_M]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ and the noise vector $\boldsymbol{V} = [V_1\ V_2\ \ldots\ V_M]^{\mathrm{T}} \in \mathbb{C}^{M\times 1}$ we can write the observation vector as

$$\boldsymbol{Y} = \boldsymbol{a}S + \boldsymbol{V}, \tag{7}$$

which corresponds to an additive superposition of the weighted speech signal and the noise in each observation. We denote the covariance matrix of the noise as

$$\boldsymbol{\Phi}_V = \mathrm{E}\left(\boldsymbol{V}\boldsymbol{V}^{\mathrm{H}}\right) \in \mathbb{C}^{M\times M}, \tag{8}$$

where $\boldsymbol{V}^{\mathrm{H}}$ is the conjugate transpose of $\boldsymbol{V}$.

For mutually independent circularly complex Gaussian distributed speech and noise, the MMSE optimal estimator of the complex clean speech coefficients S is known to be the MWF, e.g. [17], [18]

$$\widehat{S} = \mathrm{E}\left(S|\boldsymbol{Y}, \boldsymbol{\Phi}_V, \sigma_{\mathrm{S}}^2, \boldsymbol{a}\right) = \boldsymbol{H}_{\mathrm{MWF}}^{\mathrm{H}}\boldsymbol{Y} \tag{9}$$

$$= \underbrace{\frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \left(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{\Phi}_V^{-1}\boldsymbol{a}\right)^{-1}}}_{H_{\mathrm{WF}}} \underbrace{\frac{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{\Phi}_V^{-1}}{\boldsymbol{a}^{\mathrm{H}}\boldsymbol{\Phi}_V^{-1}\boldsymbol{a}}}_{\boldsymbol{H}_{\mathrm{MVDR}}^{\mathrm{H}}} \boldsymbol{Y} \tag{10}$$

$$= H_{\mathrm{WF}}\ \widehat{S}_{\mathrm{MVDR}}. \tag{11}$$

The MWF $\boldsymbol{H}_{\mathrm{MWF}}$ can be factorized into a multichannel MVDR filtering $\boldsymbol{H}_{\mathrm{MVDR}}$ and a spectro-temporal post processing H_{WF} on the scalar MVDR output $\widehat{S}_{\mathrm{MVDR}} = \boldsymbol{H}_{\mathrm{MVDR}}^{\mathrm{H}}\boldsymbol{Y}$ [17]. Here, H_{WF} resembles the traditional single-channel Wiener filter,

$$H_{\mathrm{WF}} = \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \Sigma}, \tag{12}$$

with the noise variance

$$\Sigma = \left(\boldsymbol{a}^{\mathrm{H}}\boldsymbol{\Phi}_V^{-1}\boldsymbol{a}\right)^{-1} \tag{13}$$

and the unaltered speech variance σ_{S}^2 at the output of the MVDR filter $\widehat{S}_{\mathrm{MVDR}}$.

If the speech is not Gaussian distributed or we are interested e.g. in the speech spectral amplitude rather than the complex speech coefficients, the MWF is not MMSE optimal anymore. However, Balan and Rosca [19] showed that for the signal model in (7) and Gaussian noise the output of an MVDR beamformer supplies sufficient statistics for S and functions $f(S)$. From this statement it follows that an estimator of S (or any function $f(S)$) that is optimal in the MMSE sense in the single-channel case is also optimal in the multichannel case when it is applied to the MVDR output. This also holds for non-Gaussian speech priors $p(S)$ [18] and allows us to incorporate more advanced post-processing schemes than the traditional Wiener filter in (11). To focus on the general idea and principles of the proposed estimator, in this paper we however only discuss the well known MWF.

B. Combination of a Single Microphone Signal and a Harmonic Model

Now let us consider the case of only a single microphone signal Y_1 and a harmonic model based speech signal representation

Y_2, i.e. we set $M = 2$. In the multi-microphone literature, the vector $\boldsymbol{a}$ is typically referred to as the *propagation* vector and corresponds to the (relative) acoustical transfer function from the source to the microphones. Such acoustic transfer functions can reach from a simple delay according to the direction of arrival of the source (far field in free field condition) to an individualized head related transfer function in a reverberant room. Here, however, the context is different. In our case, from (1) and (6) it follows that $\boldsymbol{a} = [1\ 1]^{\mathrm{T}}$.

To obtain the new estimator, we first need to compute the harmonic model Y_2 as well as estimates of the speech variance σ_{S}^2 and of the noise covariance matrix $\boldsymbol{\Phi}_V$. For the estimation of the speech variance σ_{S}^2 and the acoustic noise variance $\sigma_{\mathrm{V}_1}^2$ we can apply common single-channel methods to Y_1, e.g. the decision-directed approach [1] and the speech presence probability based noise variance estimator [20], respectively. For the computation of the harmonic model Y_2, different methods can be employed, see e.g. [6], [12], [13]. In the following sections we present a new and computationally inexpensive way to estimate Y_2 and the complete noise covariance matrix $\boldsymbol{\Phi}_V$ directly in the STFT domain. The proposed estimation scheme reduces the computational overhead as compared to time-domain formulations by avoiding additional DFTs and allows for a detailed, frequency dependent analysis.

III. STFT-Domain Harmonic Model

A harmonic model is a common tool to accurately describe a voiced speech signal in the time-domain, e.g. [6], [21], [22]. The clean speech signal is modeled as a sum of H harmonic components at the fundamental frequency f_0 and integer multiples of it, the harmonic frequencies $f_h = (h+1)f_0$. The ℓ-th time-domain segment after applying the STFT analysis window $q(n)$ is given by

$$y_{2,\ell}(n) = q(n) \sum_{h=0}^{H-1} 2A_{h,\ell} \cos\left(2\pi \frac{f_h}{f_{\mathrm{s}}} n + \varphi_{h,\ell}\right), \qquad (14)$$

with sampling rate f_{s}, the initial phase $\varphi_{h,\ell}$ of component h at the beginning of segment ℓ, and sample index $n \in [0, \ldots, N-1]$. The fundamental frequency can for example be estimated using [23]. For simplicity we assume that the harmonic model spans the whole frequency range up to $f_{\mathrm{s}}/2$, i.e. we set $H = \mathrm{floor}\left(\frac{f_{\mathrm{s}}}{2f_0}\right)$. We further assume that f_0 and the real-valued harmonic amplitudes $A_{h,\ell}$ are constant over the length N of one signal segment ℓ.

A signal segment $y_{2,\ell}(n)$ can be interpreted as the result of multiplying a stationary, infinitely long and continuous harmonic signal with a discrete and finite length spectral analysis window $q(n)$. With $\cos(x) = 0.5\left(\mathrm{e}^{\mathrm{j}x} + \mathrm{e}^{-\mathrm{j}x}\right)$, the Fourier transform of such a continuous harmonic signal is a weighted pulse train at the harmonic frequencies and their negative counterparts. The time-domain multiplication with $q(n)$ corresponds to a cyclic convolution of this pulse train and the frequency response Q of the analysis window, sampled at the center frequencies of the STFT bands, giving [22]

$$\begin{aligned} Y_2(k) &= \sum_{h=0}^{H-1} A_h \mathrm{e}^{\mathrm{j}\phi_h} Q_{k-\kappa_h} + A_h \mathrm{e}^{-\mathrm{j}\phi_h} Q_{k+\kappa_h} \\ &\approx A_h^k \mathrm{e}^{\mathrm{j}\phi_h^k} Q_{k-\kappa_h^k}, \end{aligned} \qquad (15)$$

where we again drop the segment index ℓ and denote the spectral phase of the h-th harmonic component as ϕ_h. Further, the real-valued $\kappa_h = \frac{N f_h}{f_{\mathrm{s}}}$ maps the harmonic frequency f_h to our index notation, i.e. $Q_{k-\kappa_h}$ denotes the frequency response of $q(n)$ shifted by $2\pi f_h/f_{\mathrm{s}}$ in band k. The frequency responses Q at the desired frequencies can either be obtained analytically for specific analysis windows or by interpolating the discrete Fourier transform of $q(n)$ via zero padding, see e.g. [22]. For the approximation in (15) we assume that the segment length N is large enough to resolve neighboring harmonic components in the spectral domain. We further assume that in each band k only the closest harmonic component is dominant and neglect the influence of all other components, which leads to the simplification in (15). The harmonic component that is closest to band k is found via $h = \max\left(\mathrm{round}\left(\frac{k}{\kappa_0}\right) - 1, 0\right)$. For notational convenience, we introduce the spectral amplitude A_h^k, phase ϕ_h^k, and index κ_h^k of the harmonic component that is closest to band k.

With the simplification in (15), the complex coefficients in bands that are dominated by the same harmonic are directly related by means of the frequency response of the spectral analysis window Q. Starting from bands $k' = \mathrm{round}(\kappa_h)$ that directly contain a harmonic component, we can infer the speech component in all other bands associated to the same harmonic via

$$Y_2(k) = Y_2(k') \frac{Q_{k-\kappa_h^k}}{Q_{k'-\kappa_h^k}}. \qquad (16)$$

With (16) we have a simple and computationally inexpensive way of describing and computing a harmonic model directly in the STFT domain.

In the context of noise reduction, the harmonic signal is typically obtained from a noise corrupted microphone signal. Since for a clean harmonic signal the energy is concentrated around the harmonic frequencies, we assume that in the presence of acoustic noise V_1 the local signal-to-noise ratio (SNR) is the largest in bins k'. Between the spectral harmonics, $k \neq k'$, the local SNR is typically much lower. Accordingly, with (16), we can estimate the speech component in low SNR regions between the harmonics based on the higher SNR bins that directly contain harmonics. Therefore, we first estimate the speech component in bands k' with the help of a Wiener filter with a lower bound $G_{\min}$,

$$\widetilde{G}_{k'} = \max\left(G_{\min}, \frac{\sigma_{\mathrm{S}}^2(k')}{\sigma_{\mathrm{S}}^2(k') + \sigma_{\mathrm{V}_1}^2(k')}\right), \qquad (17)$$

giving

$$\begin{aligned} Y_2(k') &= \widetilde{G}_{k'}\, Y_1(k') \\ &= S(k') + \underbrace{\left(\widetilde{G}_{k'} - 1\right) S(k') + \widetilde{G}_{k'} V_1(k')}_{V_2(k')}. \end{aligned} \qquad (18)$$

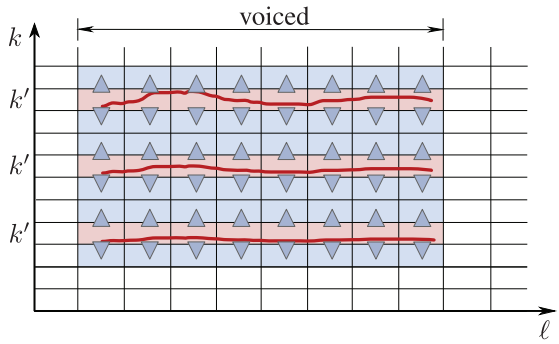

Fig. 2. Illustration of the harmonic model based speech reconstruction. In bands k' (red) the clean speech is estimated using (18). Starting from this estimate, the speech in surrounding bands (blue) is obtained via (16), indicated by the blue arrowheads.

To protect harmonic components at low SNRs, which would be suppressed by the Wiener filter, we set $G_{\min} > 0$, limiting its maximal attenuation. Applying a lower limit to bands k' of the harmonic model (18) utilizes the additional information about the fundamental frequency f_0, in the sense that it determines and protects bands which are more likely to contain relevant speech energy. The harmonic model Y_2 in all other bands $k \neq k'$ is inferred from the estimate in bands k' (18) using (16). This concept is illustrated in Fig. 2. Since a harmonic model per definition has only little energy between the harmonics, Y_2 shows a substantial reduction of acoustic noise between the harmonics as compared to the noisy observation Y_1. An example of a harmonic model based signal representation Y_2 is presented in Fig. 5 in Sec. V.

IV. Computing the Noise Covariance Matrix $\mathbf{\Phi}_V$

For the computation of the final estimator $\widehat{S}$ via (11), an estimate of the noise covariance matrix $\mathbf{\Phi}_V$ is needed. In this section, we first detail the estimation of the error variance of the harmonic model $\sigma^2_{\mathrm{V}_2}$, before discussing the cross-covariances between V_1 and V_2 and constructing the final covariance matrix $\mathbf{\Phi}_V$.

A. Harmonic Model Error Variance $\sigma^2_{\mathrm{V}_2}$

In practice, the harmonic model based representation of the clean speech for a known fundamental frequency is degraded by two conceptually different sources of error. On the one hand, the rather simple harmonic model is not capable of perfectly describing every voiced speech sound S, such as sounds with mixed excitation, like 'v' in 'victory' or 'th' in 'the'. On the other hand, acoustic noise V_1 in bands k' impedes the estimation of speech parameters.

To take into account the former, we define the modeling error variance σ^2_M as the error variance when Y_2 is estimated on *clean* voiced speech. To estimate σ^2_M, we first estimate the normalized modeling error variance

$$\overline{\sigma^2_M}(k) = \frac{\sigma^2_M(k)}{\sigma^2_{\mathrm{S}}(k)} \approx \frac{\sum_{\ell\in\mathbb{L}} |S(k,\ell) - Y_2(k,\ell)|^2}{\sum_{\ell\in\mathbb{L}} |S(k,\ell)|^2} \tag{19}$$

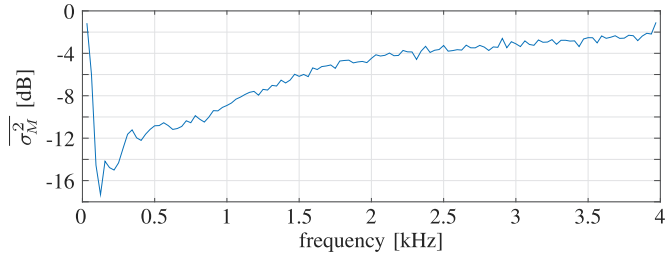

Fig. 3. Normalized modeling error variance $\overline{\sigma^2_M}$ over frequency, i.e. the error variance when the harmonic model is obtained from 500 clean speech TIMIT sentences normalized by the speech variance in each frequency band.

off-line by applying (16) to clean voiced speech taken from 500 gender balanced sentences of the TIMIT [24] training set. Here, $\mathbb{L}$ denotes the set of all voiced speech segments. At runtime, we estimate the actual model error variance via

$$\sigma^2_M(k,\ell) = \overline{\sigma^2_M}(k)\,\sigma^2_{\mathrm{S}}(k,\ell), \tag{20}$$

effectively taking into account the current speech variance σ^2_{S}.

The normalized modeling error variance $\overline{\sigma^2_M}$ is depicted over frequency in Fig. 3. We use a segment length of 32 ms, a segment shift of 16 ms, and a square-root Hann window for analysis and synthesis to compute the STFT. The same STFT setup is employed throughout this work. The fundamental frequency is estimated from the clean speech signal. For this, any state-of-the-art fundamental frequency estimator can be used. Throughout this work, we employ PEFAC [23] to estimate the fundamental frequency f_0 and to classify speech segments as voiced or unvoiced. At low frequencies, where typically most of the energy of voiced sounds is concentrated, the harmonic model yields the most accurate estimates. Towards higher frequencies, $\overline{\sigma^2_M}$ increases, reflecting the increasing inaccuracies of the harmonic model, including fundamental frequency estimation errors which accumulate towards higher harmonics.

Besides the modeling error, the estimate Y_2 is also deteriorated by the acoustic noise V_1 in bands k'. In these bands a limited Wiener filter is applied to the noisy microphone signal $Y_1(k')$ to obtain $Y_2(k')$ (18). The error variance of $Y_2(k')$ is thus given by the error variance of the Wiener filter $\sigma^2_W(k')$ (5), i.e. $\sigma^2_{\mathrm{V}_2}(k') = \sigma^2_W(k')$, where we neglect the influence of the lower limit $G_{\min}$ for simplicity. Since we do not use (16), unlike in the remaining bands, the modeling error variance $\sigma^2_M(k')$ is zero. In STFT bands between spectral harmonics, i.e. $k \neq k'$, the estimate in the closest harmonic band $Y_2(k')$ is then scaled with the frequency response of the analysis window according to (16). Hence, also the error variance is scaled, and we finally obtain

$$\sigma^2_{\mathrm{V}_2}(k) = \frac{1}{B_{\mathrm{V}}} \begin{cases} \sigma^2_W(k), & \text{for } k = k' \\ \sigma^2_{\mathrm{V}_2}(k') \frac{|Q_{k-\kappa^k_h}|^2}{|Q_{k'-\kappa^k_h}|^2} + \sigma^2_M(k), & \text{for } k \neq k', \end{cases} \tag{21}$$

where between harmonics the modeling error variance σ^2_M and the scaled error variance on the harmonics add up. For typical analysis windows the scaling reduces $\sigma^2_{\mathrm{V}_2}$ between harmonics compared to the variance on harmonics, reflecting an increased

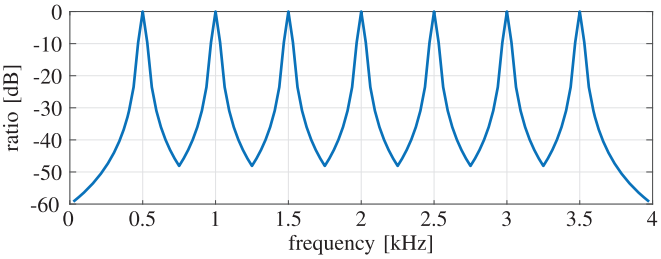

Fig. 4. An example of the ratio of analysis window transfer functions $|Q_{k-\kappa_h^k}|^2/|Q_{k'-\kappa_h^k}|^2$ used in (21) to estimate $\sigma^2_{\mathrm{V}_2}$ for a fundamental frequency of 500 Hz.

certainty about the harmonic model based estimate. As an example, the ratio $\frac{|Q_{k-\kappa_h^k}|^2}{|Q_{k'-\kappa_h^k}|^2}$ for a square-root Hann window of 32 ms and a harmonic signal with a fundamental frequency of 500 Hz at a sampling rate of 8 kHz is presented in Fig. 4. On the other hand, σ_M^2 introduces some residual uncertainty in Y_2 to account for model inaccuracies. In (21), B_{V} denotes the binary decision if the current signal segment ℓ contains voiced speech ($B_{\mathrm{V}}(\ell)=1$) or not ($B_{\mathrm{V}}(\ell)=0$), estimated using PEFAC [23]. As the capability of the harmonic model to describe speech sounds that are not voiced is limited, the harmonic signal representation Y_2 is less accurate and may show strong errors in such segments. This increased error is taken into account in (21) by the multiplication with $\frac{1}{B_{\mathrm{V}}}$, yielding an error variance of $\sigma^2_{\mathrm{V}_2}=\infty$ in frames that do not contain any voiced speech.

B. Cross-Covariance of V_1 and V_2 and building $\mathbf{\Phi}_V$

Now that we have estimates of $\sigma^2_{\mathrm{V}_2}$ and the acoustic noise variance $\sigma^2_{\mathrm{V}_1}$, e.g. obtained via [20], only the cross-covariance $\mathrm{E}\big(V_1 V_2^{\mathrm{H}}\big)$ is needed to complete the noise covariance matrix $\mathbf{\Phi}_V$. In harmonic frequency bands k', the harmonic model Y_2 is obtained by applying a limited Wiener filter to Y_1 (18). Following the formulation of V_2 in (18), the cross-covariance in bands k' is $\mathrm{E}\big(V_1(k')\,V_2^{\mathrm{H}}(k')\big)=\widetilde{G}_{k'}\sigma^2_{\mathrm{V}_1}(k')$. To simplify the estimation of the cross-covariance between the harmonics, $k\neq k'$, we make the common assumption that the acoustic noise V_1 is uncorrelated over frequency bands, i.e. $\mathrm{E}(V_1(k+i)\,V_1(k))=0$ for $i\neq 0$. Due to the way the harmonic model is obtained (16), between the harmonics $V_2(k)$ depends only on the noisy observation $Y_1(k')$ at the respective closest harmonic band k'. Accordingly, the cross-covariance between harmonics is zero,

$$\mathrm{E}(V_1(k)V_2(k))=\mathrm{E}\left(V_1(k)\,\widetilde{G}_{k'}\frac{Q_{k-\kappa_h^k}}{Q_{k'-\kappa_h^k}}V_1(k')\right)=0,\quad \forall k\neq k', \tag{22}$$

where we combined (18) and (16) to reformulate V_2 and then used the independence of speech and acoustic noise, $\mathrm{E}\big(SV_1^{\mathrm{H}}\big)=0$. With these considerations made, we can finally formulate the noise covariance matrix as

$$\mathbf{\Phi}_V(k)=\begin{pmatrix}\sigma^2_{\mathrm{V}_1}(k) & \widetilde{G}_k\sigma^2_{\mathrm{V}_1}(k)\,\delta_{k-k'}\\ \widetilde{G}_k\sigma^2_{\mathrm{V}_1}(k)\,\delta_{k-k'} & \sigma^2_{\mathrm{V}_2}(k)\end{pmatrix}, \tag{23}$$

where $\delta_{k-k'}$ is the Kronecker-delta, which is zero for $k\neq k'$ and one only for $k=k'$. With the noise covariance matrix and the corresponding harmonic model at hand, we can now compute the proposed estimator (11) to take into account the spectral structure of voiced speech in an MMSE optimal framework.

V. Analysis

In this section, we analyze the proposed clean speech estimator in more detail. In a first step we consider only frequency bands between harmonic components before extending the discussion to harmonic bands k'. We then provide a representative example highlighting the potential of the proposed estimation scheme. Finally, the extension to multiple microphones is examined.

A. Between Harmonics

Under the assumption that the acoustic noise V_1 is uncorrelated over frequency bands, between spectral harmonics, the two noise components V_1 and V_2 are mutually uncorrelated and $\mathbf{\Phi}_V$ is a diagonal matrix (23). For this specific case, the MWF formulation in (11) can be simplified, giving more insight into the way the final estimate $\widehat{S}$ is computed. It further results in a practical and robust implementation, also for correlated noises. The inverse covariance matrix $\mathbf{\Phi}_V^{-1}$ is computed by inverting the noise variances on the main diagonal and with $\boldsymbol{a}=[1\;1]^{\mathrm{T}}$ the MVDR filter weight reduces to

$$\boldsymbol{H}_{\mathrm{MVDR}}=\left[\frac{\sigma^2_{\mathrm{V}_2}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}\;,\;\frac{\sigma^2_{\mathrm{V}_1}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}\right]^{\mathrm{H}}. \tag{24}$$

Accordingly, the MVDR output can be written as

$$\widehat{S}_{\mathrm{MVDR}}=\boldsymbol{H}_{\mathrm{MVDR}}^{\mathrm{H}}\boldsymbol{Y}=\frac{\sigma^2_{\mathrm{V}_2}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}Y_1+\left(1-\frac{\sigma^2_{\mathrm{V}_2}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}\right)Y_2, \tag{25}$$

i.e. a weighted combination of the noisy microphone signal Y_1 and the harmonic model Y_2. The mixing factor $\frac{\sigma^2_{\mathrm{V}_2}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}$ approaches one if the variance of the acoustic noise $\sigma^2_{\mathrm{V}_1}$ is much lower than the error variance of the harmonic model $\sigma^2_{\mathrm{V}_2}$, while it approaches zero for $\sigma^2_{\mathrm{V}_2}\ll\sigma^2_{\mathrm{V}_1}$. Considering the variances as measures of uncertainty, the MVDR weighting thus favors the more reliable of the two observations for the computation of $\widehat{S}_{\mathrm{MVDR}}$ (25) in each time-frequency point.

After the MVDR processing, the post-filter H_{WF} (12) is applied to the output $\widehat{S}_{\mathrm{MVDR}}$, with the noise variance (13)

$$\Sigma=\left(\boldsymbol{a}^{\mathrm{H}}\mathbf{\Phi}_V^{-1}\boldsymbol{a}\right)^{-1}=\frac{\sigma^2_{\mathrm{V}_1}\sigma^2_{\mathrm{V}_2}}{\sigma^2_{\mathrm{V}_1}+\sigma^2_{\mathrm{V}_2}}, \tag{26}$$

to obtain the final estimate via the MWF (11). For a closer look at the MWF for mutually independent complex Gaussian speech and noises, we formulate the posterior of the clean speech spectral coefficients analytically:

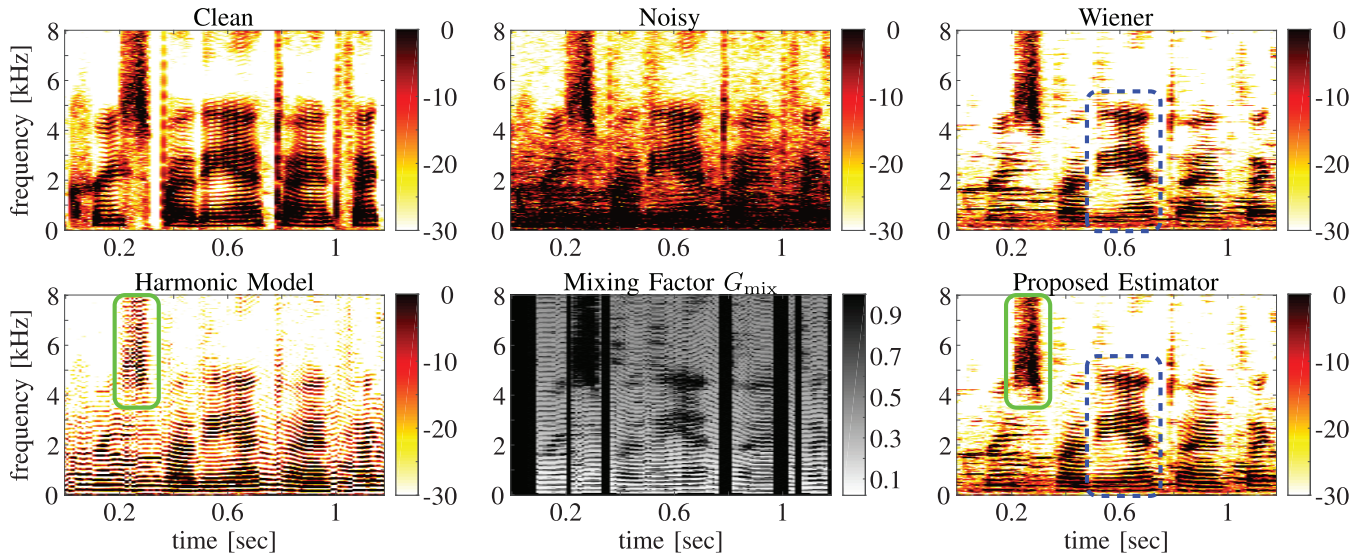

Fig. 5. Spectrograms of a clean speech sentence, the noisy observation degraded by babble noise at 5 dB SNR, the harmonic model based representation Y_2, the Wiener filter estimate $\widehat{S}_W$, the mixing factor G_{mix}, and the proposed estimate $\widehat{S}$. The proposed estimator protects weak harmonic components in voiced speech (e.g. dashed block at 0.6 seconds) while falling back to the Wiener filter in unvoiced sounds (e.g. solid block at 0.25 seconds), where $G_{\mathrm{mix}} \to 1$.

$$p\left(S|\boldsymbol{Y},\boldsymbol{\Phi}_V,\sigma_{\mathrm{S}}^2\right) = \mathcal{N}\left(\frac{\sigma_{\mathrm{V}_2}^2}{\sigma_{\mathrm{V}_2}^2+\sigma_W^2}\,\widehat{S}_W + \left(1-\frac{\sigma_{\mathrm{V}_2}^2}{\sigma_{\mathrm{V}_2}^2+\sigma_W^2}\right)\,Y_2, \frac{\sigma_W^2\sigma_{\mathrm{V}_2}^2}{\sigma_W^2+\sigma_{\mathrm{V}_2}^2}\right). \tag{27}$$

An outline of how (27) can be obtained is presented in the appendix. The posterior is again Gaussian and its mean, the MMSE optimal clean speech estimate $\widehat{S}$ given both, the noisy observation Y_1 and the harmonic model Y_2, is a weighted mixture of Y_2 and the single-channel Wiener filter estimate $\widehat{S}_W$ obtained from Y_1 (4). Accordingly, we can rewrite the MWF (11) for $\boldsymbol{a} = [1\ 1]^{\mathrm{T}}$ as

$$\widehat{S} = \mathrm{E}\big(S|\boldsymbol{Y},\boldsymbol{\Phi}_V,\sigma_{\mathrm{S}}^2\big) = G_{\mathrm{mix}}\,\widehat{S}_W + (1-G_{\mathrm{mix}})\;Y_2, \tag{28}$$

with mixing factor

$$G_{\mathrm{mix}} = \frac{\sigma_{\mathrm{V}_2}^2}{\sigma_{\mathrm{V}_2}^2+\sigma_W^2}. \tag{29}$$

Let us now have a closer look at the mixing factor G_{mix}. To ease the discussion, we assume that the acoustic noise is spectrally white in the vicinity of the dominant harmonic, i.e. $\sigma_{\mathrm{V}_1}^2(k') = \sigma_{\mathrm{V}_1}^2(k)$. At high SNRs, i.e. $\sigma_{\mathrm{S}}^2 \gg \sigma_{\mathrm{V}_1}^2$, we can approximate the error variances as $\sigma_W^2 \approx \sigma_{\mathrm{V}_1}^2$ and $\sigma_{\mathrm{V}_2}^2 \approx \overline{\sigma_M^2}\sigma_{\mathrm{S}}^2$. The mixing factor thus becomes $G_{\mathrm{mix}} \approx \frac{\overline{\sigma_M^2}\sigma_{\mathrm{S}}^2}{\overline{\sigma_M^2}\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{V}_1}^2} \approx 1$ and the normalized modeling error variance $\overline{\sigma_M^2}$ influences how fast G_{mix} approaches 1. Accordingly, the proposed estimator falls back to the single-channel estimate $\widehat{S}_W$ in high SNRs. Only towards lower SNRs, the harmonic model Y_2 has a significant influence on the final estimate $\widehat{S}$.

The main advantage between harmonic components of $\widehat{S}$ (28) over the Wiener filter alone becomes apparent when we consider the presence of highly non-stationary noise. In practice, σ_{S}^2 and $\sigma_{\mathrm{V}_1}^2$ are not known and need to be estimated. Common noise variance estimators assume that the noise is less stationary than the speech component. Thus, fast changes in the noise variance $\sigma_{\mathrm{V}_1}^2$, like babble-bursts, are not tracked correctly. This results in an underestimation of $\sigma_{\mathrm{V}_1}^2$, which in turn leads to an overestimation of σ_{S}^2. Conventional statistical estimators hence erroneously apply too little attenuation to the specific region and noise leaks into the speech estimate. This is where the harmonic signal representation Y_2 comes into play and serves as additional information that allows us to distinguish noise outliers from voiced speech, even if the noise variance is not adequately estimated: for a harmonic speech sound, the true σ_{S}^2 is expected to reduce between the harmonics relative to the harmonic bands k'. If the estimated speech variance deviates from this signal model, e.g. due to a noise burst between the harmonics, for medium to low SNRs, the error variance of the harmonic model will be smaller than that of the Wiener filter, $\sigma_{\mathrm{V}_2}^2 < \sigma_W^2$, due to the scaling in (21). This reduces G_{mix} (29) and puts more weight on the harmonic model Y_2, which does not suffer from the observed noise burst. With the help of the harmonic model, we now have more knowledge about the underlying signal, which allows us to achieve a higher noise reduction between the harmonics while maintaining the speech component on the harmonics. In this sense, the proposed estimator shows some similarities to the phase-aware estimators in [25], [26]. There, the attenuation applied to the noisy observation Y_1 is controlled by the deviation of the observed spectral phase from an estimated clean speech phase. A harmonic model based phase estimate can for instance be obtained by taking the phase of Y_2 or the estimator in [22]. In contrast to [25], [26], however, here we do not only take the phase, but also the amplitude of the harmonic model Y_2 into account.

It is worth noting that the uncertainty of the MWF (27), i.e. its estimation error variance, is always lower than or equal to that of the harmonic model and the single-channel Wiener filter alone, e.g.

$$\frac{\sigma_W^2 \sigma_{\mathrm{V}_2}^2}{\sigma_W^2 + \sigma_{\mathrm{V}_2}^2} \leq \sigma_W^2, \tag{30}$$

where equality is asymptotically reached for $\sigma_W^2 \ll \sigma_{\mathrm{V}_2}^2$. The same relation analogously holds for $\sigma_{\mathrm{V}_2}^2$. Furthermore, as $\sigma_W^2 \leq \sigma_{\mathrm{V}_1}^2$ (5), the uncertainty of the MWF is also lower than or equal to that of the MVDR (26).

The weighted mixture (28) has also been derived in [14] with the motivation to optimally combine a single-channel Wiener filter estimate with a model based estimate. The current work extends and generalizes our approach in [14] by formulating the estimator in terms of the well established MWF. Besides gaining a deeper understanding about the underlying problem, formulating the estimator in this multichannel framework allows us to easily consider also correlated noises, different spectral post-filters, and more than two observations.

Algorithm 1. Step-by-step outline of the proposed FFIWI

for $\forall \ell$ **do**
  Estimate the fundamental frequency f_0 and voicing decision B_{V} e.g. using [23]
  for $\forall k$ **do**
    Estimate the acoustic noise variance $\sigma_{\mathrm{V}_1}^2$ [20]
    Estimate the speech variance σ_{S}^2 [1]
    If $k = k'$ **then**
      Compute the harmonic signal Y_2 via (18)
    else
      Compute the harmonic signal Y_2 via (16)
    end if
    Estimate the error variance $\sigma_{\mathrm{V}_2}^2$ (21)
    Obtain the final estimate $\widehat{S}$ using (28), (29)
  end for
end for

B. On Harmonics

In frequency bands k' that contain spectral harmonics, V_1 and V_2 are *not* uncorrelated, i.e. $\mathbf{\Phi}_V(k')$ is not diagonal (23), and, strictly speaking, the simplification of the MWF presented in (28) is not applicable. Thus, the possibly dense matrix $\mathbf{\Phi}_V(k')$ can be directly applied in (10). However, the resulting estimator looses some of its capability to preserve weak harmonic components as the single-channel post-filter H_{WF} in (10) counteracts the protective lower limit on harmonic bands (17). To increase the preservation of weak harmonics, we propose to neglect the cross-covariance in bins k' in (23). If we do so, equation (28) is applied in bands k' with $\sigma_{\mathrm{V}_2}^2(k') = \sigma_W^2(k')$, which leads to $G_{\mathrm{mix}} = 0.5$. The final estimate $\widehat{S}$ on the harmonics is hence given as the arithmetic mean of $\widehat{S}_W$ and Y_2. In this way, the beneficial effect of the minimum gain $G_{\min}$ in (17) on the harmonics is maintained in the final estimator, resulting in an improved protection of weak harmonic components. In Sec. VI, we show that this modification actually improves the speech enhancement results. We denote the modified estimator as the Fundamental Frequency Informed WIener filter (FFIWI). An outline of FFIWI is presented by means of the pseudo-code in Algorithm 1.

C. A Representative Example

To illustrate the advantage of the proposed estimator over the conventional Wiener filter and the harmonic model alone, we present the enhancement results for a speech excerpt that is degraded by babble noise at 5 dB SNR in Fig. 5. The purely harmonic signal Y_2 protects the harmonic structure of voiced speech sounds while achieving a strong noise suppression between the harmonic components. It however also enforces a harmonic structure onto unvoiced sounds, e.g. the high frequency sound at 0.2-0.3 sec, which leads to annoying speech distortions. The conventional Wiener filter on the other hand, does not introduce such artifacts in unvoiced sounds, but also achieves less noise reduction between the harmonics while at the same time suppressing weak harmonic components. The proposed estimator now combines these two signals according to (28), achieving a protection of low-SNR harmonic components of voiced speech, e.g. at 0.6 sec, and an increased noise reduction while avoiding artifacts in unvoiced speech. In this way, the proposed approach combines the strengths of the individual signals $\widehat{S}_W$ and Y_2 for an improved clean speech estimate.

How $\widehat{S}_W$ and Y_2 are actually combined becomes apparent from the mixing factor G_{mix}. In unvoiced sounds, for which the harmonic model is not well suited, G_{mix} increases and the Wiener filter dominates the final estimate. In voiced speech G_{mix} is lower, giving more weight to the harmonic model. Due to the modeling error variance σ_M^2, we further observe a general increase of G_{mix} towards higher frequencies. This reduces the influence of the harmonic model in higher frequencies, effectively accounting for its increasing inaccuracies e.g. caused by small fundamental frequency estimation errors that accumulate over harmonics. If segments without any harmonic structure are detected by PEFAC [23], we have $B_{\mathrm{V}} = 0$ in (21), leading to $\sigma_{\mathrm{V}_2}^2 = \infty$. Accordingly, the proposed estimator falls back to the conventional single-channel Wiener filter ($G_{\mathrm{mix}} = 1$). The observations made for this example are also reflected in instrumental measures that we evaluate for multiple speakers and various noise conditions in the next section.

D. Extension to Multiple Microphones

If more than one microphone, i.e. $M - 1$ microphones, with $M > 2$, are available, FFIWI can also be computed by applying the MWF-like formulation in (10). The harmonic model then serves as the M^{th} observation $Y_M = S + V_M$. Analogous to the previous sections, we make the assumption that the error of the harmonic model is uncorrelated to the errors in the microphone signals. Under this assumption, the error variance matrix reads

$$\mathbf{\Phi}_V = \begin{pmatrix} & & 0 \\ & \overline{\mathbf{\Phi}_V} & \vdots \\ & & 0 \\ 0 & \dots & 0 \;\; \sigma_{\mathrm{V_M}}^2 \end{pmatrix}, \tag{31}$$

where $\overline{\mathbf{\Phi}_V}$ is the possibly dense $M-1 \times M-1$ covariance matrix of all microphone signals. Due to the block-diagonal structure of $\mathbf{\Phi}_V$, its inverse $\mathbf{\Phi}_V^{-1}$ is also block diagonal and is obtained by simply inverting $\overline{\mathbf{\Phi}_V}$ and $\sigma^2_{\mathrm{V_M}}$. For this specific structure of $\mathbf{\Phi}_V^{-1}$, the MVDR filtering in (10) can be split up into two stages: first, the $M-1$ microphone signals are combined using an MVDR beamformer with $\overline{\mathbf{\Phi}_V^{-1}}$ and the corresponding $M-1$ element propagation vector. Then the output signal of this stage is combined with the harmonic model in a second MVDR, with $a = [1\ 1]^{\mathrm{T}}$ in (7). Consequently, the multi-microphone FFIWI can be implemented as a single-microphone FFIWI applied to the output of an MVDR filter on all microphone signals. The advantage of this two step realization is that we can decouple the multi-microphone processing from the concept of FFIWI. For any microphone setup, we can seamlessly use any of-the-shelf algorithm for the estimation of $\overline{\mathbf{\Phi}_V}$ and the corresponding propagation vector, without the need of adapting the existing FFIWI implementation.

VI. Evaluation

The evaluation is performed on 128 gender balanced sentences taken from the test set of the TIMIT database [24] sampled at 8 kHz. The signals are degraded by various noise types at SNRs ranging from -5 dB to 15 dB. We use segments with a length of 32 ms, overlapping by 50%, and a square-root Hann window for analysis and synthesis. The noise variance is estimated using the speech presence probability based noise variance estimator [20] while the speech variance is estimated with the decision-directed approach [1] with a smoothing factor of 0.98. The fundamental frequency is blindly estimated on the noisy observation using the noise robust fundamental frequency estimator PEFAC [23]. In segments that, according to PEFAC, do not contain voiced speech, we set $B_{\mathrm{V}} = 0$ and the error variance of the harmonic model becomes $\sigma^2_{\mathrm{V}_2} = \infty$ (21). In these segments the proposed estimator hence reduces to a single-channel Wiener filter, see e.g. (28) and (29). In (17) we set $G_{\min} = 0.5$. Also, to increase the perceptual quality, we impose a lower limit of -20 dB relative to the noisy observation on all estimators before synthesizing the time-domain signals via overlap-add. We further only consider the single-microphone case in the evaluation. The extension to multiple microphones is discussed in Sec. V-D, showing that the multi-microphone FFIWI can simply be interpreted as the single-microphone FFIWI applied on the output of an MVDR filter on the microphone signals.

As instrumental measures we employ Perceptual Evaluation of Speech Quality (PESQ) [27], Short-Time Objective Intelligibility Measure (STOI) [28], and the Log Spectral Distance (LSD), e.g. [15, Chap. 3.7]. Even though PESQ has originally been developed for the evaluation of coded speech, it has been shown to correlate also with the *quality* of enhanced speech [29]. The LSD yields the root-mean-square deviation of a modified magnitude squared spectrum in dB from the one of the clean speech, representing a measure of spectral similarity. STOI on the other hand predicts the speech *intelligibility* of a time-frequency weighted noisy speech utterance. In [28], non-linear mapping functions of the raw STOI output values (between zero and one) to actual intelligibility scores have been presented for two databases. Since for the TIMIT database no such mapping is available, here we apply the mapping that has been proposed for the IEEE database used in [28] instead. The absolute numbers hence have to be treated with caution, but the general trends remain. Therefore, and to improve the visual comparability of the results we present improvements relative to the noisy input instead of absolute values. While for PESQ and STOI large values are desirable, for LSD smaller values reflect less deviations from the clean speech spectrogram.

A. Proposed Estimator vs. Wiener Filter and Harmonic Model

In Fig. 6, we compare two versions of the proposed estimator $\widehat{S}$ to the conventional Wiener filter $\widehat{S}_W$ and the harmonic model based signal representation Y_2 in terms of PESQ, LSD, and STOI improvements over the noisy input signal Y_1. The two versions of the proposed estimator FFIWI only differ in the way they process the noisy signal on the harmonics. While FFIWI-C takes the cross-covariance between Y_1 and Y_2 in bands k' into account, FFIWI neglects it and applies (28) in all bands. As discussed in Sec. V-B, neglecting the cross-covariance results in an increased preservation of the speech in harmonic bands k'. In our experiments, i.e. Fig. 6, this leads to improvements in PESQ and STOI, as well as a reduction in LSD for FFIWI over FFIWI-C. For the remaining discussions, we therefore focus on FFIWI.

Both, the harmonic model and the proposed estimator, achieve an improvement in STOI for low SNRs relative to the Wiener filter, but also relative to the noisy input. For the rather stationary noise inside of a driving car [30], the proposed FFIWI outperforms the separate estimators in PESQ and STOI in all conditions, with larger improvements for low SNRs. The proposed combination benefits from the information inherent in Y_2 and leads to a consistent improvement over the conventional Wiener filter alone, e.g. about 0.1 MOS in PESQ and 6% in STOI at 0 dB SNR. Similar results are achieved also in non-stationary pink noise, which has been modulated with a modulation frequency of 0.5 Hz.

The influence of the harmonic model Y_2 on the proposed estimator is even more prominent for the non-stationary acoustic scenario of rain on a roof[1]. For this noise type, also isolated raindrops are audible, which can be considered as impulsive disturbances. As discussed in Sec. V-A, in practice, the conventional Wiener filter is not capable of adequately suppressing such impulsive sounds, since the sudden rise in the noise variance σ^2_{V} is not tracked by the estimator [20]. This is different from the modulated pink noise, for which the noise variance can still be adequately tracked [20]. In low SNRs, the harmonic model benefits from enforcing a perfectly harmonic structure onto the noisy observation, which effectively suppresses the rather broadband, impulsive raindrops. For increasing SNRs, however, the accompanying speech distortions, especially in unvoiced sounds, outweigh the noise reduction and lead to a decreased speech quality, reflected in the steep drop in PESQ.

[1] http://www.freesound.org/people/mmorast/sounds/192149/

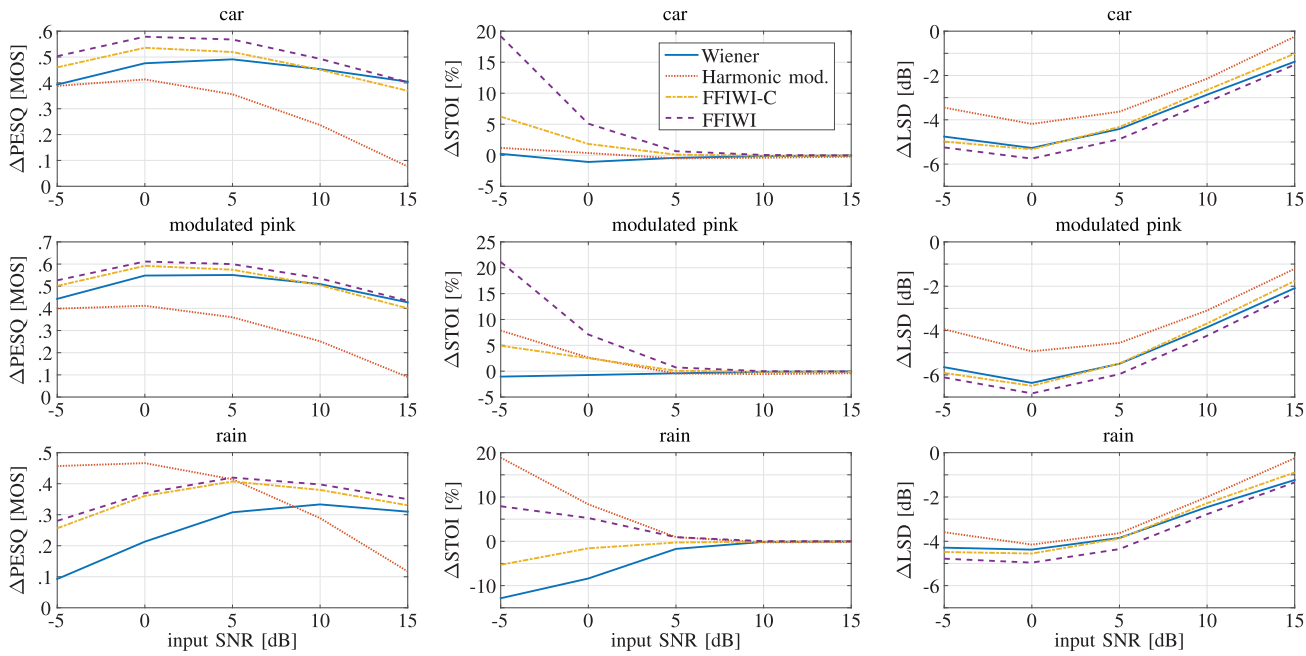

Fig. 6. PESQ, STOI, and LSD improvements over the noisy microphone signal for the conventional Wiener filter, the harmonic model, and two versions of the proposed estimator. 'FFIWI-C' considers the cross-covariance of Y_1 and Y_2 in harmonic bands k', while 'FFIWI' applies (28) in all bands.

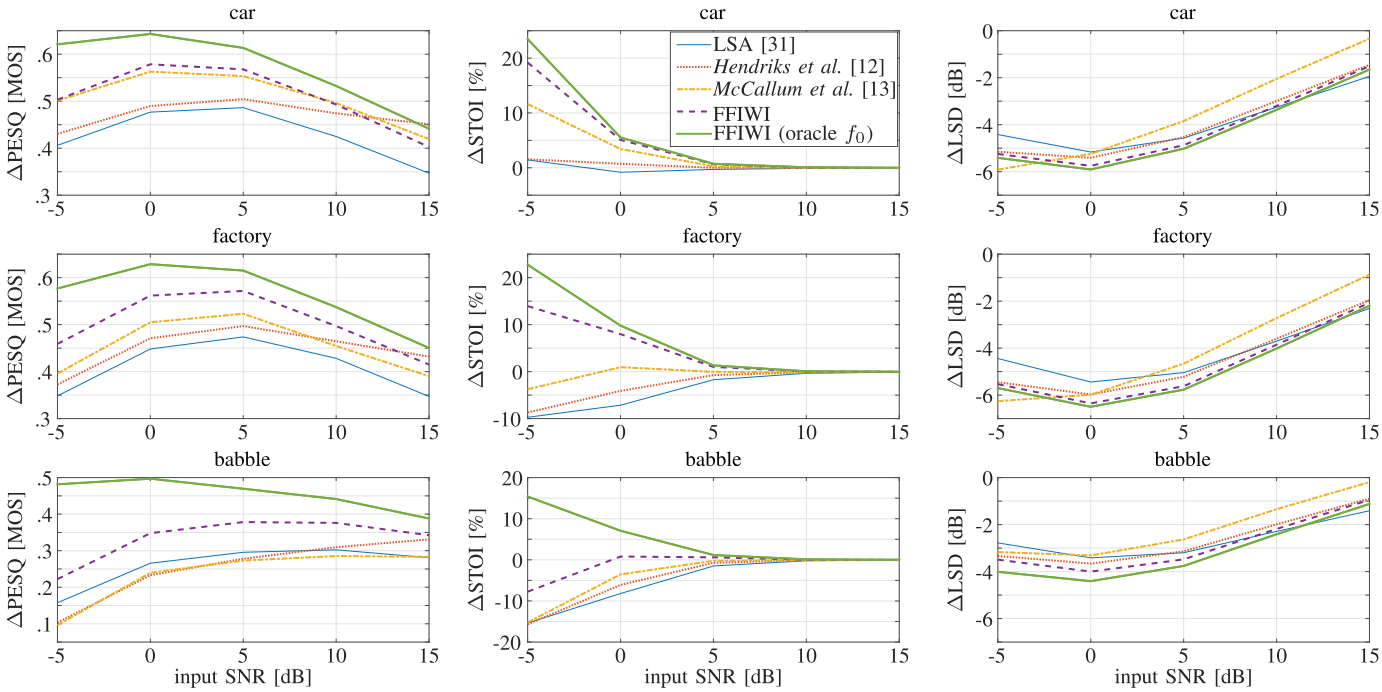

Fig. 7. PESQ, STOI, and LSD improvements over the noisy microphone signal for the LSA [31], *Hendriks et al.* [12], *McCallum et al.* [13], and the proposed estimator FFIWI. As a reference, we also include the results for FFIWI using the fundamental frequency from the annotation in [33].

In the proposed estimator, this is encountered for by differentiating between voiced and unvoiced speech and by the behavior of the mixing factor G_{mix} (28), which favors the Wiener filter in high local SNRs. In this scenario, the proposed estimator achieves an improvement of up to 0.2 MOS in PESQ and 20% in STOI over the conventional Wiener filter. The proposed FFIWI also achieves a lower LSD, meaning that the estimated spectrum is closer to the desired clean speech spectrum. This is the case for all noise types and SNRs considered here, achieving improvements of around 0.5 dB at lower SNRs with respect to the Wiener filter.

B. Comparison to Other State-of-the-Art Algorithms

Now we compare the proposed FFIWI against the two statistical estimators [12], [13], which also utilize a harmonic signal model. To denote the reference algorithms, here we use the names of the respective authors: *McCallum et al.* for [13] and

Hendriks et al. for [12]. More specifically, *Hendriks et al.* corresponds to the algorithm SOFT-SD-U with a Gaussian speech prior in [12]. As an example for a commonly used statistical approach, we further include the log-spectral amplitude estimator (LSA) [31] as a baseline. All estimators employ the STFT setup as described above, except for [12], for which a 32 ms Hamming window for analysis and 50% zero padding is used, which is very close to the original proposal and showed to be beneficial for the general performance of this algorithm. In Fig. 7, we present the results for car noise [30], factory noise [32], and babble noise [32].

The proposed estimator outperforms all three reference algorithms in PESQ for most scenarios and in STOI for all scenarios. Only for high SNRs, *Hendriks et al.* may achieve slightly larger PESQ scores and for car noise *McCallum et al.* produces very similar PESQ results. At 0 dB SNR, improvements of about 0.1 MOS in PESQ relative to the LSA are achieved for all three noise types. Particularly for babble noise this improvement is substantial, considering that the improvement of the LSA over the noisy signal in this condition is only about 0.25 MOS. For babble noise, FFIWI also achieves the lowest LSD for SNRs below 10 dB. The same trend can be observed for the other two noise types as well, except for very low SNRs, where *McCallum et al.* achieves the lowest LSD.

To evaluate the influence of fundamental frequency estimation errors, we also present the results for the case that the fundamental frequency is not blindly estimated, but taken from the annotation provided in [33]. Given this oracle information, the gain in PESQ and STOI increases, especially in low SNRs and challenging noise scenarios. For babble noise, the improvement with respect to the LSA increases to about 0.3 MOS in PESQ and almost 30% in STOI at -5 dB SNR. Also *Hendriks et al.* and *McCallum et al.* benefit from an oracle fundamental frequency, but experiments showed that the increase in performance is not as prominent as for the proposed approach. Accordingly, with a better estimate of the fundamental frequency, the benefit of the proposed estimator over *Hendriks et al.* and *McCallum et al.* can be expected to increase even further.

VII. Conclusion and Outlook

In this contribution, we presented a novel STFT domain clean speech estimator that incorporates information about the structure of voiced speech by means of a harmonic model into an MMSE optimal statistical estimator. To this end, we proposed a way to estimate a harmonic signal representation directly in the STFT domain and combined it with the noisy microphone signal using a multichannel Wiener filter. The resulting estimator yields an increased noise reduction between harmonics while preserving weak harmonic components, leading to improvements in speech quality and intelligibility as predicted by PESQ and STOI over several reference algorithms. Thanks to the formulation in terms of a general, well understood multichannel framework, the proposed estimator can seamlessly be extended to use multiple microphones, alternative post-filtering techniques, or different signal models, e.g. for transient sounds.

Appendix

To obtain the posterior of the MWF for mutually uncorrelated Gaussian noise and $\boldsymbol{a} = [1\ 1]^{\mathrm{T}}$ (27), we apply Bayes rule, i.e.

$$p\left(S|Y_1, Y_2\right) = \frac{p\left(Y_1|S\right) p\left(Y_2|S\right) p\left(S\right)}{p\left(Y_2|Y_1\right) p\left(Y_1\right)}, \tag{32}$$

with

$$p\left(S\right) = \mathcal{N}\left(0, \sigma_{\mathrm{S}}^2\right) \tag{33}$$

$$p\left(Y_1\right) = \mathcal{N}\left(0, \sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}_1}^2\right) \tag{34}$$

$$p\left(Y_2|S\right) = p\left(S + V_2|S\right) = \mathcal{N}\left(S, \sigma_{\mathrm{V}_2}^2\right) \tag{35}$$

$$p\left(Y_1|S\right) = p\left(S + V_1|S\right) = \mathcal{N}\left(S, \sigma_{\mathrm{V}_1}^2\right) \tag{36}$$

$$p\left(Y_2|Y_1\right) = p\left(S + V_2|Y_1\right) = \mathcal{N}\left(\widehat{S}_W, \sigma_W^2 + \sigma_{\mathrm{V}_2}^2\right), \tag{37}$$

where all probability density functions are given $\boldsymbol{\Phi}_V$ and σ_{S}^2, which we drop here for notational convenience. Since S, V_1, and V_2 are assumed to be mutually uncorrelated and Gaussian, we have $p\left(V_2|Y_1\right) = p\left(V_2\right) = \mathcal{N}\left(0, \sigma_{\mathrm{V}_2}^2\right)$ and in the last line the means and variances of V_2 and S given Y_1 add up. Plugging all of the distributions into (32), after some algebraic computations, we obtain the posterior (27).

Acknowledgment

The authors would like to thank M. McCallum for his valuable comments and for providing code for the implementation of [13].

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. SSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State-of-the-Art*. Fort Collins, CO, USA: Morgan & Claypool, Feb. 2013.

[3] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 273–276.

[4] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 9, pp. 1355–1365, Sep. 2014.

[5] D. Fischer and T. Gerkmann, "Single-microphone speech enhancement using MVDR filtering and Wiener post-filtering," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016.

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

[7] 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.

[8] 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 Lang. Process.*, vol. 20, no. 7, pp. 1948–1963, Sep. 2012.

[9] J. Benesty and J. Chen, *Optimal Time-Domain Noise Reduction Filters—A Theoretical Study*. New York, NY, USA: Springer, 2011.

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

[11] J. Le Roux, S. Watanabe, and J. R. Hershey, "Ensemble learning for speech enhancement," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[12] R. C. Hendriks, R. Heusdens, and J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 406–415, Feb. 2007.
[13] M. McCallum and B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1445–1457, Jul. 2013.
[14] M. Krawczyk-Becker and T. Gerkmann, "MMSE-optimal combination of Wiener filtering and harmonic model based speech enhancement in a general framework," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2015, pp. 1–5.
[15] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. Hoboken, NJ, USA: Wiley, 2006.
[16] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Amsterdam, The Netherlands: Elsevier, 2015.
[17] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds. Berlin, Germany: Springer-Verlag, 2001, pp. 39–60.
[18] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 10, pp. 885–888, Oct. 2009.
[19] R. Balan and J. Rosca, "Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase," in *Proc. Sensor Array Multichannel Signal Process. Workshop*, Aug. 2002, pp. 209–213.
[20] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.
[21] T. Quatieri and R. McAulay, "Noise reduction using a soft-decision sine-wave vector quantizer," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Albuquerque, NM, USA, Apr. 1990, vol. 2, pp. 821–824.
[22] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.
[23] S. Gonzalez and M. Brookes, "PEFAC—A pitch estimation algorithm robust to high levels of noise," *IEEE Trans. Audio Speech Lang. Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.
[24] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," *National Institute of Standards and Technology (NIST)*, Gaithersburg, MD, USA, 1993.
[25] T. Gerkmann and M. Krawczyk, "MMSE-optimal spectral amplitude estimation given the STFT-phase," *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 129–132, Feb. 2013.
[26] T. Gerkmann, "Bayesian estimation of clean speech spectral coefficients given a priori knowledge of the phase," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4199–4208, Aug. 2014.
[27] ITU-T, "Perceptual evaluation of speech quality (PESQ)," ITU-T Recommendation P.862, 2001.
[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[29] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[30] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Automat. Speech Recog. (ASR) Challenges New Millenium ISCA Tut. Res. Workshop (ITRW)*, Paris, France, Sep. 2000, pp. 29–32.
[31] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. SSP-33, no. 2, pp. 443–445, Apr. 1985.
[32] A. Varga and H. Steeneken, "Assessment for automatic speech recognition—II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, Jul. 1993.
[33] S. Gonzalez. (2014, Feb.). *Pitch of the Core TIMIT Database Set* [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/data/TIMITfxv.zip

Martin Krawczyk-Becker (S'15) received the Dipl.-Ing. degree in electrical engineering and information technology from the Ruhr-Universität Bochum, Bochum, Germany in 2011. From January 2010 to July 2010, he was with Siemens Corporate Research, Princeton, NJ, USA. He is currently pursuing the Ph.D. degree in speech enhancement and noise reduction at the Universität Oldenburg, Oldenburg, Germany.

Timo Gerkmann (S'08–M'10–SM'15) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering and information technology from the Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, Bochum, Germany, in 2004 and 2010, respectively. From January 2005 to July 2005, he was with Siemens Corporate Research, Princeton, NJ, USA. Between 2010 and 2011, he was a Postdoctoral Researcher with the Sound and Image Processing Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015, he headed the Speech Signal Processing Group, Universität Oldenburg, Oldenburg, Germany. Since 2015, he has been a Principal Scientist for Audio and Acoustics at Technicolor Research and Innovation, Hanover, Germany. His research interests include signal processing for acoustic communication devices, human-machine interfaces, and audiovisual media.