MMSE-OPTIMAL COMBINATION OF WIENER FILTERING AND HARMONIC MODEL BASED SPEECH ENHANCEMENT IN A GENERAL FRAMEWORK

Martin Krawczyk-Becker and Timo Gerkmann

Speech Signal Processing Group, Department of Medical Physics and Acoustics,
Cluster of Excellence ”Hearing4all”, University of Oldenburg, Germany
martin.krawczyk-becker@uni-oldenburg.de, timo.gerkmann@uni-oldenburg.de

ABSTRACT

For the reduction of additive acoustic noise, various methods and clean speech estimators are available, with specific strengths and weaknesses. In order to combine the strengths of two such approaches, we derive a minimum mean squared error (MMSE)-optimal estimator of the clean speech given two independent initial clean speech estimates. As an example we present a specific combination that results in a weighted mixture of the Wiener filter and a simple, low-cost harmonic speech model. The proposed estimator benefits from the additional information provided by the harmonic model, leading to a better protection of harmonic components of voiced speech as compared to the traditional Wiener filter. Instrumental measures predict improvements in speech quality and speech intelligibility for the proposed combination over each individual estimator.

Index Terms— Speech enhancement, noise reduction, signal reconstruction

1. INTRODUCTION

Over the years, numerous approaches for the reduction of undesired acoustic noise have been proposed to increase the robustness of communication devices like hearing aids or mobile phones. 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. A majority of these approaches is formulated in some spectro-temporal domain, most commonly the short time discrete Fourier transform (STFT) domain due to its low complexity and intuitive interpretation. 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 [1]. 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.

Improvements over the original approaches 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, e.g. [2, 3, 4, 5]. Furthermore, dropping the assumption that neighboring spectral coefficients are uncorrelated has been shown to lead to alternative estimators, e.g. [6, 7], which benefit from incorporating more information at the price of a more challenging parameter estimation. Besides such statistical approaches, also methods based on sinusoidal or harmonic speech signal models have been proposed, like e.g. [8], where the parameters of a sinusoidal model are iteratively estimated from a noisy observation to recover the underlying speech signal. These model-based estimators can provide some benefit over simple approaches like the traditional Wiener filter due to the consideration of more a priori knowledge about the observed signals. In particular situations the simple approaches might however still provide more reliable estimates. For example, while a sinusoidal model is well suited to represent voiced speech, its applicability to transients or stop consonants is limited, which may lead to suboptimal speech enhancement results.

In this contribution we therefore propose a framework which allows for a MMSE-optimal combination of two clean speech estimates, of which one is formulated as a multiplication of the noisy observation with a spectral gain function. For this, we derive a MMSE-optimal estimator of the clean speech given two independent initial clean speech estimates. As an example, we present the combination of the traditional Wiener filter with an approach based on a simple, low-cost harmonic model and show that this combination outperforms the two individual estimators. For this specific example, the combined estimator shows some conceptual similarities to [9, 10], where a harmonic signal model is considered as additional deterministic information to facilitate the estimation of clean speech spectral coefficients. The proposed estimator yields a time and frequency dependent weighting of the two individual estimators based on their estimation error variances.

2. MMSE-OPTIMAL COMBINATION

We define the noisy observation in the STFT domain $Y_{k,\ell}$ in each time-frequency point (k,ℓ) as an additive superposition of clean speech $S_{k,\ell}$ and environmental noise $V_{k,\ell}$, i.e

$$Y_{k,\ell} = S_{k,\ell} + V_{k,\ell}. \tag{1}$$

In the remainder of this paper we drop the time and frequency indices ℓ and k for notational convenience where appropriate. We assume that the spectral coefficients of both, speech and noise, follow a zero-mean circular complex Gaussian distribution with variances σ_{S}^2 and σ_{V}^2, respectively, and that S and V are mutually uncorrelated. In this case, the speech posterior is given by, e.g. [11, Chap. 5.3.2] and [12, Chap. 3.4],

$$p\left(S \mid Y\right) = \mathcal{N}\left(\frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{V}}^2}Y, \frac{\sigma_{\mathrm{S}}^2\sigma_{\mathrm{V}}^2}{\sigma_{\mathrm{S}}^2+\sigma_{\mathrm{V}}^2}\right) = \mathcal{N}\left(S_{\mathrm{W}}, \sigma_W^2\right), \tag{2}$$

where $\mathcal{N}\left(S_{\mathrm{W}}, \sigma_W^2\right)$ denotes a Gaussian distribution with mean S_{W} and variance σ_W^2. The Wiener filter estimate S_{W} can also be written

This work was supported by the DFG Project GE2538/2-1.

978-1-4799-7450-4/15/$31.00 ©2015 IEEE

in terms of the true speech signal S and an error term W, as

$$S_{\mathrm{W}} = G_{\mathrm{W}} Y = S + \underbrace{(G_{\mathrm{W}} - 1)\, S + G_{\mathrm{W}} V}_{W}, \tag{3}$$

with the Wiener filter gain $G_{\mathrm{W}} = \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2}$. Under the assumptions that have been made, the Wiener filter estimate is optimal in the MMSE sense. Nevertheless, other estimators, like the spectral amplitude estimators [1, 4], could also be used here to obtain a spectral gain G_{W}.

As discussed in the introduction, there are numerous proposals for alternative clean speech estimators of which some incorporate prior information about the underlying signal and/or the employed spectral analysis. We denote this second clean speech estimate as

$$\widetilde{S} = S + E, \tag{4}$$

where the estimation error E is assumed to follow a zero-mean circularly complex Gaussian distribution with variance σ_E^2 and to be statistically independent of S. Depending on the actual estimator, there will be spectro-temporal regions where $\widetilde{S}$ is more accurate than S_{W} and vice versa. Our goal now is to combine the two estimates S_{W} and $\widetilde{S}$ in a MMSE-optimal fashion such that the combined estimator, denoted as $\widehat{S}$, benefits from the individual strengths of each estimate and eventually outperforms both of them. In this paper, we obtain $\widetilde{S}$ using a clean speech estimator that is based on a harmonic model for voiced speech which is presented in Sec. 3. The model-based estimate might well preserve the harmonics of a voiced speech sound better than a Wiener filter, but at the same time it might also introduce annoying artifacts in unvoiced speech or noise-only regions, which are not present in S_{W}. Combining the two estimates in an optimal fashion potentially results in an increased protection of harmonics relative to the Wiener filter while artifacts of the model-based estimate are strongly reduced.

In order to obtain the combined estimator $\widehat{S}$, we derive the speech posterior, given both estimates S_{W} and $\widetilde{S}$:

$$p\left(S \mid S_{\mathrm{W}}, \widetilde{S}\right) = \frac{p\left(S_{\mathrm{W}} \mid S\right) p\left(\widetilde{S} \mid S\right) p\left(S\right)}{p\left(\widetilde{S} \mid S_{\mathrm{W}}\right) p\left(S_{\mathrm{W}}\right)}, \tag{5}$$

with

$$p\left(S_{\mathrm{W}}\right) = p\left(G_{\mathrm{W}} Y\right) = \mathcal{N}\left(0, G_{\mathrm{W}}^2 \left(\sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{V}}^2\right)\right) \tag{6}$$

$$p\left(\widetilde{S} \mid S\right) = p\left(S + E \mid S\right) = \mathcal{N}\left(S, \sigma_E^2\right) \tag{7}$$

$$p\left(S_{\mathrm{W}} \mid S\right) = p\left(G_{\mathrm{W}}\left(S + V\right) \mid S\right) = \mathcal{N}\left(G_{\mathrm{W}} S, G_{\mathrm{W}}^2 \sigma_{\mathrm{V}}^2\right) \tag{8}$$

$$p\left(\widetilde{S} \mid S_{\mathrm{W}}\right) = p\left(S + E \mid G_{\mathrm{W}} Y\right) = \mathcal{N}\left(S_{\mathrm{W}}, \sigma_W^2 + \sigma_E^2\right), \tag{9}$$

where in (5) and (9) we assume that E and W are independent. Under this assumption, $p\left(E \mid S_{\mathrm{W}}\right) = p\left(E\right) = \mathcal{N}\left(0, \sigma_E^2\right)$. As both, S given S_{W} and E, are Gaussian also their sum is Gaussian with means and variances adding up, leading to (9).

Plugging all of the distributions into (5), after some algebraic computations, we finally obtain the posterior

$$p\left(S \mid S_{\mathrm{W}}, \widetilde{S}\right) = \mathcal{N}\left(S_{\mathrm{W}} \frac{\sigma_E^2}{\sigma_E^2 + \sigma_W^2} + \widetilde{S}\left(1 - \frac{\sigma_E^2}{\sigma_E^2 + \sigma_W^2}\right), \frac{\sigma_W^2 \sigma_E^2}{\sigma_W^2 + \sigma_E^2}\right). \tag{10}$$

The posterior is again Gaussian and its mean, i.e. the MMSE-optimal estimate $\widehat{S}$ of the clean speech S given both, S_{W} and $\widetilde{S}$, is given as a weighted mixture of the two estimates

$$\widehat{S} = \mathrm{E}\left(S \mid S_{\mathrm{W}}, \widetilde{S}\right) = S_{\mathrm{W}} G_{\mathrm{mix}} + \widetilde{S}\left(1 - G_{\mathrm{mix}}\right), \tag{11}$$

with mixing factor

$$G_{\mathrm{mix}} = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_W^2}. \tag{12}$$

Here, G_{mix} approaches one if the variance of the Wiener estimate σ_W^2 is much lower than the variance of the alternative estimate σ_E^2, while G_{mix} approaches zero for $\sigma_E^2 \ll \sigma_W^2$. Considering the variances as measures of reliability of the estimates, the proposed weighting thus favors the more reliable of the two initial estimates S_{W} and $\widetilde{S}$ in the computation of the final estimate $\widehat{S}$ (11). It is worth noting that in (10) the error variance of the new estimator is *lower than or equal to* the error variance of each individual estimator, e.g.

$$\frac{\sigma_W^2 \sigma_E^2}{\sigma_W^2 + \sigma_E^2} \leq \sigma_W^2, \tag{13}$$

where equality is asymptotically reached for $\sigma_W^2 \ll \sigma_E^2$. The same relation analogously holds for σ_E^2.

In the following section, we present how the second estimate $\widetilde{S}$ and the corresponding error variance can be obtained by employing a simple, low-cost harmonic model for voiced speech. For this specific example, the resulting estimator shows certain similarities to [10], where a clean speech amplitude estimator is derived under the assumption that speech is adequately modeled by a harmonic-plus-noise model. This so called *stochastic-deterministic* estimator results in a weighted mixture of the harmonic speech model and the noisy observation. In contrast to [10] however, here $\widehat{S}$ combines a harmonic model with the Wiener filter estimate, potentially avoiding artifacts encountered in [10] which have to be suppressed by an additional speech presence probability estimation stage. Besides the algorithmic differences between the two approaches, also the parameter estimation, e.g. that of σ_E^2, differs significantly.

3. HARMONIC MODEL

Voiced speech is frequently modeled as a sum of H harmonic components at the fundamental frequency f_0 and integer multiples of it, the harmonic frequencies $f_h = (h+1) f_0$, e.g. in [8, 13, 14]. The ℓ-th time domain segment after applying analysis window $q\left(n\right)$ is given by

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

with sampling rate f_{s} and the initial phase $\varphi_{h,\ell}$ of component h at the beginning of segment ℓ. We assume that f_0 and the real-valued harmonic amplitudes $A_{h,\ell}$ are constant over the length N of one segment ℓ. Under this assumption, the STFT of a harmonic signal is given as the cyclic convolution of a pulse train at the harmonic frequencies with the frequency response of the analysis window Q sampled at the center frequencies of the STFT bands, i.e. [14]

$$\widetilde{S}_k = \sum_{h=0}^{H-1} A_h \mathrm{e}^{\mathrm{j}\phi_h} Q_{k-\kappa_h} + A_h \mathrm{e}^{-\mathrm{j}\phi_h} Q_{k+\kappa_h} \tag{15}$$

$$\approx A_h^k \mathrm{e}^{\mathrm{j}\phi_h^k} Q_{k-\kappa_h^k}, \tag{16}$$

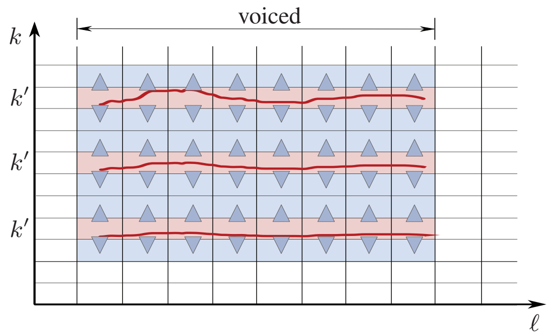

Figure 1: Illustration of the harmonic model based speech estimation. In bands k' (red) the clean speech is estimated using (18). Starting from this estimate, the speech in surrounding bands (blue) is obtained via (17), indicated by the blue arrowheads.

where we again drop segment index ℓ and denote the spectral phase of the h-th harmonic component as ϕ_h. Further, real-valued $\kappa_h = \frac{N f_h}{f_s}$ maps the harmonic frequency f_h to our index notation. For the approximation we assume that segment length N is large enough to well separate neighboring harmonic components in the spectral domain and that in each band k only the closest harmonic component is dominant. 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.

For the approximation in (16), 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 $k' = \text{round}(\kappa_h)$ that directly contain a harmonic component, we can infer the speech component in all other bands associated to the same harmonic via

$$\widetilde{S}_k = \widetilde{S}_{k'} \frac{Q_{k-\kappa_h^k}}{Q_{k'-\kappa_h^k}}. \tag{17}$$

The frequency responses $Q_{k-\kappa_h^k}$ and $Q_{k'-\kappa_h^k}$ 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. [14]. As for a harmonic signal $\widetilde{S}$ the energy is concentrated around the harmonic frequencies, we assume that in the presence of noise V the local signal-to-noise ratio (SNR) is the highest in bins k'. Between the spectral harmonics, $k \neq k'$, the local SNR is typically much lower. Accordingly, with (17), we can estimate the speech component in low SNR regions between the harmonics based on the higher SNR bins that directly contain harmonics. For this, we first estimate the speech component in bands k' with the help of the Wiener filter using a lower limit,

$$\widetilde{S}_{k'} = \widetilde{G}_{k'} Y_{k'} = \max\left(G_{\min}, G_{\mathrm{W},k'}\right)\, Y_{k'}. \tag{18}$$

Then, the signal in all other bands is inferred from this estimate using (17). This concept is illustrated in Figure 1.

For $G_{\min} = 0$, in harmonic bands k' the two estimates S_{W} and $\widetilde{S}$ are identical and $\widehat{S} = S_{\mathrm{W}}$. In between harmonic bands, i.e. $k \neq k'$, the two estimates and also their error variances differ, leading to a weighted mixture of S_{W} and $\widetilde{S}$ according to (11). As $\widetilde{S}$ per definition has only little energy between the harmonics, the final estimator $\widehat{S}$ is capable of reducing noise between harmonics that has not been suppressed by S_{W}. This could for example be the case if a noise burst is not adequately tracked by the noise power estimator.

To protect harmonic components at low SNRs, which would be suppressed by the Wiener filter alone, we set $G_{\min} > 0$, limiting the maximal suppression. Even though with the estimate $\widetilde{S}_{k'}$ as given in (18) in harmonic bands k' the estimation errors W and E are not independent, we keep the independence assumption for simplicity and still compute the final estimate $\widehat{S}_{k'}$ using (11). Applying the lower limit only to (18) utilizes the additional information about the fundamental frequency f_0, in the sense that it determines and protects bins which are more likely to contain relevant speech energy. In this paper we choose $G_{\min} = 0.5$ in (18), resulting in an increased preservation of harmonic components for which $\sigma_{\mathrm{S}}^2 < \sigma_{\mathrm{V}}^2$, i.e. in negative local SNRs.

3.1. Computation of estimation error variance σ_E^2

For the performance of the final estimator $\widehat{S}$ (11), accurate computation of the estimation error variance of the harmonic model σ_E^2 is vital. Only if σ_E^2 is adequately estimated, the mixture in (11) yields the optimal combination of S_{W} and $\widetilde{S}$. In practice, the harmonic model based estimation of 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, environmental noise V in bands k' degrades the estimation performance.

To take into account the former, we define the modeling error variance σ_M^2 as the error variance when $\widetilde{S}$ is applied to *clean* voiced speech. To estimate σ_M^2, we first estimate the inverse modeling SNR

$$\xi_M^{-1}(k) = \frac{\sigma_{M,k}^2}{\sigma_{S,k}^2} \approx \frac{\sum_{\ell\in\mathbb{L}} |S_{k,\ell} - \widetilde{S}_{k,\ell}|^2}{\sum_{\ell\in\mathbb{L}} |S_{k,\ell}|^2} \tag{19}$$

off-line by applying (17) to clean voiced speech taken from 500 gender balanced sentences of the TIMIT [15] training set. Here, $\mathbb{L}$ denotes the set of all voiced speech segments. The so obtained ξ_M^{-1} increases towards higher frequencies, reflecting the increasing inaccuracies of the harmonic model for high frequencies, including fundamental frequency estimation errors which accumulate towards higher harmonics. At runtime, we estimate the actual model error variance via $\sigma_M^2 = \xi_M^{-1}\sigma_{\mathrm{S}}^2$ to consider the current speech power σ_{S}^2.

Besides the modeling error, the estimate is also deteriorated by additive noise V in bands k'. As in these bands a Wiener filter is used to estimate the clean speech (18), the modeling error is zero and the estimation error variance is $\sigma_E^2 = \sigma_W^2$, where for simplicity we neglect the impact of applying a lower limit $G_{\min}$ on the Wiener filter in (18). In STFT bands between spectral harmonics, i.e. $k \neq k'$, the estimate in the closest harmonic band $\widetilde{S}_{k'}$ is scaled with the frequency response of the analysis window according to (17). Hence, also the estimation error variance is scaled, and we finally obtain

$$\sigma_{E,k}^2 = \begin{cases} \sigma_{W,k}^2 & \text{, for } k = k' \\ \sigma_{E,k'}^2 \frac{|Q_{k-\kappa_h^k}|^2}{|Q_{k'-\kappa_h^k}|^2} + \sigma_{M,k}^2 & \text{, for } k \neq k', \end{cases} \tag{20}$$

where between harmonics the modeling error variance σ_M^2 and the scaled error variance on the harmonics add up. The scaling reduces σ_E^2 between harmonics compared to the variance on harmonics, while σ_M^2 introduces some residual uncertainty in the estimate $\widetilde{S}$ due to model inaccuracies.

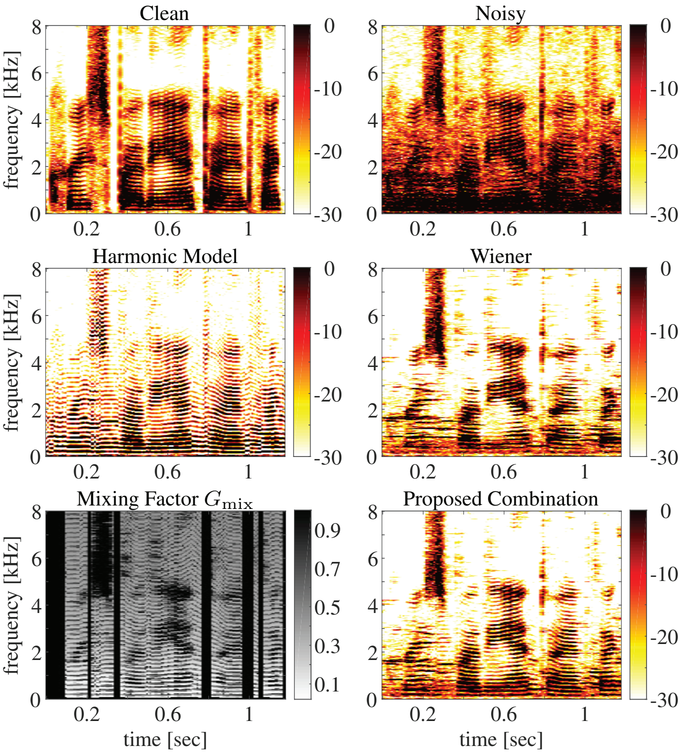

Figure 2: Spectrograms of a clean speech sentence, the noisy observation degraded by babble noise at 5 dB SNR, the harmonic model based estimate $\widetilde{S}$, the Wiener filter estimate S_{W}, the gain G_{mix}, and the proposed estimate $\widehat{S}$. The proposed estimator protects weak harmonic components in voiced speech while falling back to the Wiener filter in unvoiced sounds ($G_{\text{mix}} \to 1$).

4. EVALUATION

We evaluate the proposed estimator on 128 gender balanced sentences taken from the test set of the TIMIT database [15] sampled at 16 kHz. The signals are degraded by babble and traffic noise at SNRs ranging from -5 dB to 15 dB. We use a segment length of 32 ms, a segment shift of 8 ms, and a square-root Hann window for analysis and synthesis. The fundamental frequency is blindly estimated on the noisy observation using PEFAC [16]. The noise power is estimated using [17] while the speech power is estimated by the decision-directed approach [1] with a smoothing parameter of 0.98. To increase the perceptual quality, we impose a lower limit of -20 dB relative to the noisy observation on all three estimators before synthesizing the time domain signals via overlap-add. Further, like e.g. in [8], we combine the amplitude of the harmonic model based speech estimate $|\widetilde{S}|$ with the noisy spectral phase for signal synthesis.

The advantage of the proposed estimator over the traditional Wiener filter and the model based estimate is illustrated in Figure 2 for a clean speech excerpt degraded by babble noise at 5 dB SNR. The proposed estimator protects low-SNR harmonic components of voiced speech, e.g. at 0.6 sec, that are suppressed by the Wiener filter, including the heavily disturbed fundamental component. In unvoiced sounds, for which the harmonic model is not well suited, e.g. the high frequency sound at 0.2-0.3 sec, the traditional Wiener filter dominates the combined estimator (11), i.e. $G_{\text{mix}} \to 1$. If segments without any harmonic structure are detected by PEFAC [16], the Wiener filter is used by setting $G_{\text{mix}} = 1$. The general increase of G_{mix} towards higher frequencies due to the structure of the modeling error variance σ_M^2 takes into account the limited applicability of $\widetilde{S}$ at high frequencies. In this way, the proposed approach combines the strengths of the individual estimators for an improved clean speech estimate.

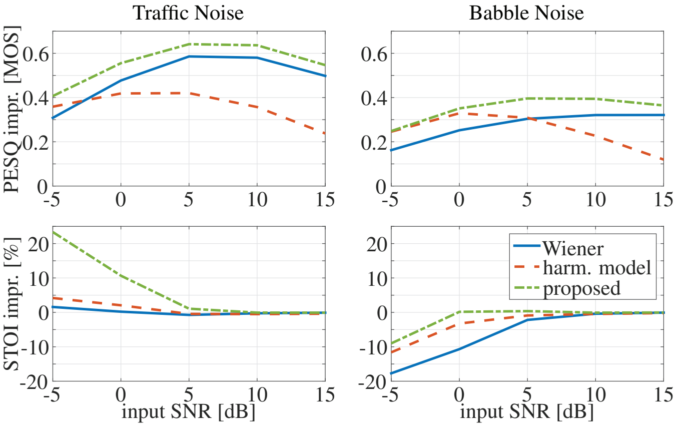

Figure 3: STOI and PESQ improvement over the noisy observation for traffic and babble noise at input SNRs from -5 dB to 15 dB. The proposed method outperforms the traditional Wiener filter and the model-based estimation in terms of both, PESQ and STOI.

In Figure 3 we present the improvements in 'perceptual evaluation of speech quality' (PESQ) [18] and the short-time objective intelligibility measure (STOI) [19] relative to the noisy observation. While STOI predicts the intelligibility of a degraded speech signal, PESQ is used as an instrumental measure of speech quality. Alongside the proposed approach we also present the results for the Wiener filter and the harmonic model. The proposed estimator outperforms the Wiener filter as well as the harmonic model in both, PESQ and STOI in all conditions considered in the evaluation. The performance gain over the traditional Wiener filter increases for decreasing input SNRs, with a relative improvement of around 0.1 PESQ points and 10 % in predicted intelligibility in both noise types at 0 dB input SNR.

In traffic noise, the proposed estimator not only improves STOI with respect to the Wiener filter, but also with respect to the noisy input signal. This improvement is remarkable, as the enhancement of speech intelligibility with single channel techniques is generally a challenging task. Informal listening confirms the general trends, how exactly the improvements in the instrumental measures are reflected in human perception is however still to be evaluated with formal listening test.

5. CONCLUSION

In this contribution we presented a MMSE-optimal clean speech estimator given two independent prior estimates, of which one is formulated as a multiplication with a spectral gain. The proposed estimator results in an intuitive weighting of the individual estimates based on their error variances. For the combination of the Wiener filter with a harmonic model we showed that the proposed method protects weak harmonic components of voiced speech and outperforms the individual estimators in PESQ and STOI over a broad range of SNRs.

6. 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. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] ——, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] R. Martin, “Speech enhancement based on minimum mean-square error estimation and supergaussian priors,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 845–856, Sept. 2005.

[4] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, “Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[5] C. Breithaupt, M. Krawczyk, and R. Martin, “Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4037–4040.

[6] J. Benesty and Y. Huang, “A single-channel noise reduction MVDR filter,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 273–276.

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

[8] 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.

[9] Richard C. Hendriks and Richard Heusdens and Jesper Jensen, “An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 2, Feb. 2007.

[10] M. McCallum and B. Guillemin, “Stochastic-deterministic MMSE STFT speech enhancement with general a priori information,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1445–1457, July 2013.

[11] R. Astudillo, “Integration of short-time Fourier domain speech enhancement and observation uncertainty techniques for robust automatic speech recognition,” Ph.D. dissertation, TU Berlin, Berlin, Germany, 2010.

[12] P. E. H. R. O. Duda and D. G. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[13] T. Quatieri and R. McAulay, “Noise reduction using a soft-decision sine-wave vector quantizer,” in *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Apr 1990, pp. 821–824 vol.2.

[14] M. Krawczyk and T. Gerkmann, “STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1931–1940, Dec 2014.

[15] J. S. Garofolo, “Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,” *National Institute of Standards and Technology (NIST)*, 1988.

[16] S. Gonzalez and M. Brookes, “PEFAC – a pitch estimation algorithm robust to high levels of noise,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 22, no. 2, pp. 518–530, Feb. 2014.

[17] T. Gerkmann and R. C. Hendriks, “Noise power estimation based on the probability of speech presence,” in *IEEE Workshop Appl. Signal Process. Audio, Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 145–148.

[18] ITU-T, “Perceptual evaluation of speech quality (PESQ),” *ITU-T Recommendation P.862*, 2001.

[19] 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, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept. 2011.