A Combination of Pre-Trained Approaches and Generic Methods for an Improved Speech Enhancement

Robert Rehr[1], Timo Gerkmann[2]

[1] Speech Signal Processing Group, Department of Medical Physics and Acoustics, Cluster of Excellence "Hearing4all", University of Oldenburg, Germany

[2] Signal Processing, Vogt-Kölln-Straße 30, Universität Hamburg, Germany

Email: [1] robert.rehr@uni-oldenburg.de [2] timo.gerkmann@uni-hamburg.de

Web: [1] www.speech.uni-oldenburg.de [2] www.inf.uni-hamburg.de/en/inst/ab/sp

Abstract

To improve the quality of single-channel speech enhancement algorithms, various approaches include additional prior knowledge about speech, e.g., in the form of pre-trained speech models. In this paper, we consider a vector Taylor series based approach with a low-rank speech model. While employing a low-rank speech model keeps the complexity feasible, only speech spectral envelopes are represented and noise reduction between spectral harmonics is not possible. To counteract this issue, we propose a combination of generic, single-channel enhancement methods and the pre-trained vector Taylor series approach. Compared to a competing harmonic post-filter approach, the proposed combination is derived within a statistical framework and yields a better quality for the enhanced signal. This is verified using instrumental quality measures.

1 Introduction

Additive background noise is known to reduce speech quality and speech reception in speech communication applications. Hence, speech enhancement algorithms have been a topic of active research for many years. If only a single noisy observation is available, a common approach is to enhance the signal in the short-time Fourier transform (STFT) domain. Due to their quality and efficiency, generic statistical optimal estimators of the clean speech Fourier coefficients are often employed, e.g., [1–3]. These estimators usually depend on the speech power spectral density (PSD) and the noise PSD which have to be estimated from the noisy input signal. Also for this, many different methods and approaches are described in the literature, e.g., [4–7] for estimating the noise PSD and [1, 8] to determine the speech PSD. In this paper, we refer to this type of enhancement algorithms as "generic". This term is motivated by the fact that these methods can potentially be applied to target signals other than speech.

The generic approaches have in common that explicit speech knowledge, e.g., in the form of pre-trained vocal-tract shapes or fundamental frequency estimates, is not exploited. As, in general, the increase of prior knowledge about the target signal can potentially improve the noise reduction, enhancement methods have been considered, which employ pre-trained speech knowledge. Here, the term "pre-trained" is used to distinguish this type of algorithm from the previously considered generic approaches. In [9], a codebook based method has been presented, whereas the methods in [10, 11] are based on hidden Markov models. Nonnegative matrix factorization as in [12, 13] forms another approach to include pre-trained speech information. Similar to the feature enhancement methods presented in [14–17], the methods in [18, 19] employ a pre-trained prior model to represent the log-spectral coefficients of speech which is given by a mixture of Gaussians. Based on a vector Taylor series (VTS) approximation, a minimum mean-squared error (MMSE) optimal estimator of the clean speech log-spectral coefficients can be derived. To obtain a high-resolution estimate of the clean speech, i.e., an estimate which includes the vocal tract shape as well as the pitch, a large amount of mixtures is employed in [18]. This, however, increases the demands with respect to memory and computational complexity. In [19], a reduced amount of mixtures is employed resulting in a low-rank speech model which may only represent the speech spectral envelopes, but typically not the spectral fine structure. As a consequence, noise between the speech spectral harmonics is not reduced. This problem is encountered in [19] by applying a post-filter based on a harmonic model in voiced speech.

Our paper is related to the work in [19]. Here, we propose a novel method to reduce residual noise if a low-rank speech model is employed. For this, we combine a pre-trained method similar to [19] and generic filter-based speech enhancement methods such as the log-spectral amplitude (LSA) estimator [2]. Compared to the approach in [19] where a speech presence probability mask, which is based on a harmonic model, is used as a post-filter, the combination used here is embedded in a statistical framework. It is based on the underlying likelihood models which quantify how well the respective model describes the noisy observation. Experiments indicate that the combination improves the sound quality in comparison to the sole application of the competing generic and pre-trained enhancement method.

This work was funded by the PhD program "Signals and Cognition".

2 Signal Model

First, we start by introducing the signal model which is used throughout this paper. The algorithms considered in this paper operate in the STFT domain. For this, the signal is split into overlapping blocks. After the application of an analysis window function, each block is transformed to the Fourier domain using the discrete Fourier transform (DFT). The complex coefficients of the noisy input signal $Y_{k,\ell}$ are given by the superposition of the speech component $X_{k,\ell}$ and the noise component $N_{k,\ell}$ as

$$Y_{k,\ell} = X_{k,\ell} + N_{k,\ell}. \tag{1}$$

Here, k is the frequency index and ℓ denotes the frame index.

We assume that the complex coefficients of the DFT follow a zero-mean circular-symmetric Gaussian distribution which is commonly used in single-channel speech enhancement literature. The symbols $\Lambda^{y}_{k,\ell}$, $\Lambda^{x}_{k,\ell}$, and $\Lambda^{n}_{k,\ell}$ denote the variance of the noisy, clean and noise coefficients, respectively. Further, we assume that the speech and noise components are uncorrelated.

Often, log-spectral or cepstral representations are used to train models for speech. In this work, similar as in [19], clean speech is modeled in the log-spectral domain. The transformation from the spectral domain to the log-spectrum is given by $y_{k,\ell} = \log\left(|Y_{k,\ell}|^2\right)$. Similarly, the log-spectra of the speech component $x_{k,\ell}$ and the noise component $n_{k,\ell}$ are determined.

The proposed enhancement procedure gives an estimate of the clean speech log-spectrum $\hat{x}_{k,\ell}$. This is used to derive a spectral gain function $G_{k,\ell} = \exp([\hat{x}_{k,\ell} - y_{k,\ell}]/2)$, which is applied to noisy input spectrum to obtain an estimate of the complex clean speech coefficients as $\hat{X}_{k,\ell} = G_{k,\ell} Y_{k,\ell}$. This is equivalent to combining the estimated clean speech magnitude with the noisy phase. The enhanced signal is obtained using the overlap-add method after applying a synthesis window.

3 Statistics of the Log-Spectrum

In the proposed algorithm, we employ estimates of the speech PSD and the noise PSD which are propagated from the spectral

ISBN 978-3-8007-4275-2

domain to the log-spectral domain. This allows the application of state-of-the-art noise PSD and speech PSD estimators, e.g., [6, 8]. The propagation is based on the derivations given in [20, 21]. With the Gaussian assumption for the spectral coefficients, the means in the log-spectral domain of the noisy speech $\mu^y_{k,\ell} = \mathbb{E}\{y_{k,\ell}\}$ are related to the variances in the spectral domain $\Lambda^y_{k,\ell}$. Here, $\mathbb{E}\{\cdot\}$ denotes the expectation operator. The result is given by [20]

$$\mu^y_{k,\ell} = \begin{cases} \log(\Lambda^y_{k,\ell}) - \gamma - \log(2), & \text{for } k = 0, K/2 \\ \log(\Lambda^y_{k,\ell}) - \gamma, & \text{otherwise.} \end{cases} \tag{2}$$

Here, $\gamma \approx 0.5772\ldots$ is the Euler-Mascheroni constant and K is the number of frequency bins, which is assumed to be a multiple of 2. Thus, $K/2$ corresponds to the Nyquist frequency. Similarly, also the means of the clean speech $\mu^x_{k,\ell}$ and the background noise $\mu^n_{k,\ell}$ can be determined.

In [20] it has been shown that for complex Gaussian distributed STFT coefficients, the variance of the log-spectral coefficients $\lambda^y_{k,\ell} = \mathbb{E}\{(y_{k,\ell} - \mu^y_{k,\ell})^2\}$ is constant and results in

$$\lambda^y_{k,\ell} = \begin{cases} \pi^2/2, & \text{if} \quad k = 0, K/2, \\ \pi^2/6, & \text{otherwise.} \end{cases} \tag{3}$$

The same result also holds for the variances of the speech and noise log-spectral coefficients which are denoted by $\lambda^x_{k,\ell}$ and $\lambda^n_{k,\ell}$, respectively.

Finally, the log-spectral cross-covariance $\lambda^{xy}_{k,\ell} = \mathbb{E}\{(x_{k,\ell} - \mu^x_{k,\ell})(y_{k,\ell} - \mu^y_{k,\ell})\}$ is considered. It depends on the squared correlation coefficient $\rho^2_{k,\ell}$ between the spectral coefficients of noisy speech $Y_{k,\ell}$ and clean speech $X_{k,\ell}$, i.e.,

$$\rho^2_{k,\ell} = \frac{|\mathbb{E}\{X_{k,\ell} Y^*_{k,\ell}\}|^2}{\mathbb{E}\{|X_{k,\ell}|^2\}\mathbb{E}\{|Y_{k,\ell}|^2\}}. \tag{4}$$

In [20, 21], it was shown that this quantity is related to the Wiener filter in the spectral domain

$$\rho^2_{k,\ell} = \frac{\Lambda^x_{k,\ell}}{\Lambda^x_{k,\ell} + \Lambda^n_{k,\ell}}. \tag{5}$$

With this, the log-spectral cross-covariance $\lambda^{xy}_{k,\ell}$ can be determined using [20]

$$\lambda^{xy}_{k,\ell} = \begin{cases} \sum\limits_{n=1}^{\infty} \dfrac{n!}{(0.5)_n} \dfrac{\left(\rho^2_{k,\ell}\right)^n}{n^2}, & k = 0, K/2 \\ \sum\limits_{n=1}^{\infty} \dfrac{\left(\rho^2_{k,\ell}\right)^n}{n^2}, & \text{otherwise.} \end{cases} \tag{6}$$

The special function $(a)_n$ is given by $1 \cdot a \cdot (a+1) \cdot (a+2) \cdots (a+n-1)$, which is used for $a = 0.5$ in (6).

4 Pre-Trained Speech Enhancement

In this section, the pre-trained part of the proposed speech enhancement method is presented. It is based on the work in [14, 15, 19]. It is assumed that the joint distribution of the log-spectral speech coefficients can be described by a mixture of Gaussian distributions as

$$p(\boldsymbol{x}_\ell) = \sum_{m=1}^{M} p(m) \left(\prod_{k=0}^{K/2} \mathcal{N}\left(x_{k,\ell} | \mu^{x|m}_k, \lambda^{x|m}_k\right) \right). \tag{7}$$

Here, $\boldsymbol{x}_\ell = \left[x_{0,\ell}, x_{1,\ell}, \ldots, x_{K/2,\ell}\right]^T$ is a vector which comprises the frequency components of the speech log-spectrum at frame ℓ. Each mixture component, which are indexed by m, is a Gaussian denoted by $\mathcal{N}(\cdot)$. Its parameters are the mean $\mu^{x|m}_k$ and variance $\lambda^{x|m}_k$. The log-spectral coefficients are assumed to be independent across frequency allowing each mixture component to be represented by a multiplication over all frequency bins. The probability $p(m)$ is the prior of the mth mixture component and M denotes the number of mixtures. During training, which is performed prior to the application of this algorithm, the parameters $\mu^{x|m}_k$ and $\lambda^{x|m}_k$ and the prior probabilities $p(m)$ are determined. For this, the expectation maximization algorithm [22] is employed.

The linear relationship between speech components and noise components in (1) is in general non-linear in the log-spectral domain. Hence, similar to [14, 15, 19], the relationship in the log-spectral domain (1) is approximated using a first-order VTS. Commonly, the phase information is omitted as originally proposed in [14], so $|Y_{k,\ell}|^2$ can be written as

$$|Y_{k,\ell}|^2 \approx |X_{k,\ell}|^2 + |N_{k,\ell}|^2. \tag{8}$$

This approximation omits the cross-term which additionally depends on the phase difference between speech and noise. While clearly a simplification, it is often used in VTS based enhancement approaches. Under the reasonable assumption that speech and noise are uncorrelated, the cross-term cancels out on average, i.e., at least $\mathbb{E}\{|Y_{k,\ell}|^2\} = \mathbb{E}\{|X_{k,\ell}|^2\} + \mathbb{E}\{|N_{k,\ell}|^2\}$, e.g., [16, 23]. Similar approximations are also used in other pre-trained approaches, e.g., nonnegative matrix factorization [12, 13]. A study on how these approximations affect the quality of enhancement algorithms is given in [24]. While attempts for incorporating the cross-term exist [17, 25, 26], they typically increase the computational complexity. Thus, for simplicity, we stick to the simple model in (8) in this work. In the log-spectral domain, the relationship in (8) can be rewritten as

$$y_{k,\ell} = g(x_{k,\ell}, n_{k,\ell}) = \log\left\{\exp\left(x_{k,\ell}\right) + \exp\left(n_{k,\ell}\right)\right\}. \tag{9}$$

The non-linear mixing function $g\left(x_{k,\ell}, n_{k,\ell}\right)$ is approximated using a first-order VTS with respect to the speech and noise components $x_{k,\ell}$ and $n_{k,\ell}$, as

$$y_{k,\ell} \approx g^{p_0}_x (x_{k,\ell} - x_0) + g^{p_0}_n (n_{k,\ell} - n_0) + g^{p_0}, \tag{10}$$

Here, x_0 and n_0 form the linearization point p_0 as $p_0 = [x_0, n_0]$ and $g^{p_0} = g(x_0, n_0)$. The symbols $g^{p_0}_x$ and $g^{p_0}_n$ denote derivatives with respect to $x_{k,\ell}$ and $n_{k,\ell}$ evaluated at p_0. The linearization point is usually given by $x_0 = \mu^{x|m}_k$ and $n_0 = \mu^n_{k,\ell}$ and therefore depends on the mixture m. As in [19], the approximation in (10) is used to determine the parameters of the likelihood of $x_{k,\ell}$ given the mth mixture $p(y_{k,\ell}|x_{k,\ell}, m)$ which is assumed to follow a Gaussian distribution. Correspondingly, the mean and the variance of $p(y_{k,\ell}|x_{k,\ell}, m)$ are obtained by determining the expected values $\mu^{y|x,m}_{k,\ell} = \mathbb{E}\{y_{k,\ell}\}$ and $\lambda^{y|x,m}_{k,\ell} = \mathbb{E}\{(y_{k,\ell} - \mu^{y|x,m}_{k,\ell})^2\}$ using the simplified $y_{k,\ell}$ in (10). As the speech component $x_{k,\ell}$ is given, the only remaining random variable is the noise $n_{k,\ell}$. Thus, the mean and the variance are given by

$$\mu^{y|x,m}_{k,\ell} = g^{p_0}_x (x_{k,\ell} - x_0) + g^{p_0}_n (\mu^n_{k,\ell} - n_0) + g^{p_0}, \tag{11}$$

$$\lambda^{y|x,m}_{k,\ell} = (g^{p_0}_n)^2 \lambda^n_{k,\ell}. \tag{12}$$

With the model used for $p(y_{k,\ell}|x_{k,\ell}, m)$, also the likelihood of the mth mixture $p(y_{k,\ell}|m)$ and the posterior of $x_{k,\ell}$ given the mth mixture $p(x_{k,\ell}|y_{k,\ell}, m)$ can be determined. Also these probability density functions follow Gaussian distributions due to the Gaussian assumption for $p(y_{k,\ell}|x_{k,\ell}, m)$ and for the speech mixtures in (7). The mean and variance of $p(y_{k,\ell}|m)$ are given by

$$\mu^{y|m}_{k,\ell} = g^{p_0}_x (\mu^{x|m}_k - x_0) + g^{p_0}_n (\mu^n_{k,\ell} - n_0) + g^{p_0}, \tag{13}$$

$$\lambda^{y|m}_{k,\ell} = (g^{p_0}_x)^2 \lambda^{x|m}_k + (g^{p_0}_n)^2 \lambda^n_{k,\ell}, \tag{14}$$

while the mean of the posterior $p(x_{k,\ell}|y_{k,\ell}, m)$ is given by

$$\mu^{x|y,m}_{k,\ell} = \mu^{x|m}_k + \frac{\lambda^{x|m}_k g^{p_0}_x}{\lambda^{y|m}_{k,\ell}} \left(y_{k,\ell} - \mu^{y|m}_{k,\ell}\right). \tag{15}$$

ISBN 978-3-8007-4275-2

With this, the MMSE estimator of the log-spectral clean speech coefficients is determined. The estimator is given by the mean of $p(\boldsymbol{x}_\ell|\boldsymbol{y}_\ell)$, which can be computed for each frequency bin k as

$$\mu_{k,\ell}^{x|y} = \sum_{m=1}^{M} p(m|\boldsymbol{y}_\ell)\mu_{k,\ell}^{x|y,m,z_{\text{PrTr}}}. \tag{16}$$

By setting $\hat{x}_{k,\ell} = \mu_{k,\ell}^{x|y}$, the gain function $G_{k,\ell}$ can be determined, which is then used to to enhance the noisy spectrum $Y_{k,\ell}$. The probability $p(m|\boldsymbol{y}_\ell)$ can be obtained using Bayes' rule as

$$p(m|\boldsymbol{y}_\ell) = \frac{p(\boldsymbol{y}_\ell|m)p(m)}{\sum_{m'=1}^{M} p(\boldsymbol{y}_\ell|m')p(m')}, \tag{17}$$

where $p(\boldsymbol{y}_\ell|m)$ is given by the product over all frequency bins of $p(y_{k,\ell}|m)$. Furthermore, for computing the posterior, an estimate of the log-spectral noise mean $\mu_{k,\ell}^n$ and the log-spectral noise variance $\lambda_{k,\ell}^n$ is required. For obtaining these values, a spectral noise tracking algorithm, e.g., [4, 6, 7], is employed to determine the spectral noise variance $\Lambda_{k,\ell}^n$. In contrast to the static speech model, this estimate is time-variant. Using the method in Section 3, the log-spectral quantities can be obtained.

5 Generic Speech Enhancement

This section gives a brief overview over the generic enhancement methods that we combine with the pre-trained enhancement method described in Section 4. Here, we consider a linear log-spectral estimator related to the linear cepstrum estimator in [20] and the LSA [2]. These generic methods also require an estimate of the speech and noise statistics. Instead of using the static speech model, the spectral speech variance $\Lambda_{k,\ell}^x$ is obtained from the noisy spectrum $Y_{k,\ell}$ and tracked over time, e.g., using [1, 8], i.e., also here, a time-variant estimate is employed. The noise variance is obtained in a similar way as for the pre-trained method. Additionally, in this section, the underlying likelihood models are given as they form the basis of the combination.

5.1 Linear Log-Spectral Filter

The linear log-spectral filter is closely related to the linear cepstrum estimator presented in [20]. In [20], it is shown that the linearly constrained MMSE estimator of the clean speech cepstral coefficients has an equivalent representation in the log-spectral domain. It is given by

$$\mu_{k,\ell}^{x|y,z_{\text{Lin}}} = \mu_{k,\ell}^x + \frac{\lambda_{k,\ell}^{xy}}{\lambda_{k,\ell}^y}\left(y_{k,\ell} - \mu_{k,\ell}^y\right). \tag{18}$$

This filter is also the MMSE optimal estimator of the log-spectral speech coefficients if the log-spectral coefficients of speech $x_{k,\ell}$ and noisy speech $y_{k,\ell}$ are assumed to follow a Gaussian distribution. Hence, we set the likelihood model to

$$p(y_{k,\ell}|z_{\text{Lin}}) = \mathcal{N}(y_{k,\ell}|\mu_{k,\ell}^y, \lambda_{k,\ell}^y). \tag{19}$$

Further, the symbol z_{Lin} is a state indicator for the model assumed in this section and is used for the combination in Section 6. The required means and variances are obtained by propagating the spectral speech and noise variance estimates to the log-spectral domain as described in Section 3.

5.2 Log-Spectral Amplitude Estimator

The second enhancement method that can be used in combination with the considered pre-trained enhancement approach is the LSA estimator [2]. This method is the MMSE optimal estimator of $\log(|X_{k,\ell}|) = 1/2 x_{k,\ell}$. The result of the MMSE estimator in [2] can be rewritten as

$$\mu_{k,\ell}^{x|y,z_{\text{LSA}}} = 2\log\left[\frac{\Lambda_{k,\ell}^x}{\Lambda_{k,\ell}^x + \Lambda_{k,\ell}^n}\right] + y_{k,\ell} + \int_\nu^\infty \frac{e^{-t}}{t}dt, \tag{20}$$

where

$$\nu = \frac{\Lambda_{k,\ell}^x}{\Lambda_{k,\ell}^n + \Lambda_{k,\ell}^x}\frac{\exp(y_{k,\ell})}{\Lambda_{k,\ell}^n}. \tag{21}$$

For this approach, no propagation of the statistics from the spectral domain is required. The likelihood for this estimator can be determined as

$$p(y_{k,\ell}|z_{\text{LSA}}) = \frac{1}{\Lambda_{k,\ell}^y}\exp\left(-\frac{e^{y_{k,\ell}}}{\Lambda_{k,\ell}^y} + y_{k,\ell}\right). \tag{22}$$

Similar to Section 5.1, z_{LSA} is the indicator for the likelihood model employed here.

6 Proposed Combination

In this section, we describe the proposed method for combining the pre-trained approach from Section 4 and the generic enhancement methods given in Section 5.

For the combination, we exploit the fact that each enhancement method exhibits a different underlying likelihood model. Therefore, we define the likelihood of the state $z_{k,\ell}$ given the mth mixture $p(y_{k,\ell}|z_{k,\ell}, m)$ as

$$p(y_{k,\ell}|z_{k,\ell}, m) = \begin{cases} p(y_{k,\ell}|z_{\text{PrTr}}, m), & z_{k,\ell} = z_{\text{PrTr}}, \\ p(y_{k,\ell}|z_{\text{Lin}}), & z_{k,\ell} = z_{\text{Lin}}, \\ p(y_{k,\ell}|z_{\text{LSA}}), & z_{k,\ell} = z_{\text{LSA}}. \end{cases} \tag{23}$$

The different enhancement approaches are distinguished by the discrete state variable $z_{k,\ell}$ which can take the values z_{PrTr}, z_{LSA}, and z_{Lin} for the pre-trained approach, the LSA, and the linear log-spectral estimator, respectively. The state $z_{k,\ell}$ is allowed to be different for each frequency k and frame ℓ. The likelihoods $p(y_{k,\ell}|z_{\text{Lin}})$ and $p(y_{k,\ell}|z_{\text{LSA}})$ are given in (19) and (22). These two likelihoods are independent of the mixture m, such that for all mixtures m the equalities $p(y_{k,\ell}|z_{\text{Lin}}) = p(y_{k,\ell}|z_{\text{Lin}}, m)$ and $p(y_{k,\ell}|z_{\text{LSA}}) = p(y_{k,\ell}|z_{\text{LSA}}, m)$ hold. For the pre-trained approach, $p(y_{k,\ell}|z_{\text{PrTr}}, m)$ is equivalent to the likelihood of the mth mixture $p(y_{k,\ell}|m)$ in Section 4. In other words,

$$p(y_{k,\ell}|z_{\text{PrTr}}, m) = \mathcal{N}(y_{k,\ell}|\mu_{k,\ell}^{y|m}, \lambda_{k,\ell}^{y|m}) \tag{24}$$

with parameters given in (13) and (14). With Bayes' rule, it can be determined which of these states can be considered the most appropriate one for the noisy observation

$$p(z_{k,\ell}|y_{k,\ell}, m) = \frac{p(y_{k,\ell}|z_{k,\ell}, m)p(z_{k,\ell})}{\sum_{z'_{k,\ell}} p(y_{k,\ell}|z'_{k,\ell}, m)p(z'_{k,\ell})}. \tag{25}$$

The state prior probability $p(z_{k,\ell})$ can be used to control the mixing of the combined algorithms such that a specific method may be preferred over the others. The posterior probability in (25) can be included in the calculation of the MMSE estimate of the clean speech log-periodogram. This leads to a weighted combination of all combined enhancement methods as

$$\mu_{k,\ell}^{x|y,m} = \sum_{z_{k,\ell}} p(z_{k,\ell}|y_{k,\ell}, m)\mu_{k,\ell}^{x|y,m,z}. \tag{26}$$

For the pre-trained enhancement method, the mean $\mu_{k,\ell}^{x|y,m,z_{\text{PrTr}}}$ is given by $\mu_{k,\ell}^{x|y,m}$ in (15). For the generic enhancement methods, the means are $\mu_{k,\ell}^{x|y,m,z_{\text{Lin}}} = \mu_{k,\ell}^{x|y,z_{\text{Lin}}}$ and $\mu_{k,\ell}^{x|y,m,z_{\text{LSA}}} = \mu_{k,\ell}^{x|y,z_{\text{LSA}}}$ which are given in (18) and (20), respectively. As the generic enhancement methods are independent of the mixture m, the values have to be computed only once and can be reused for each m in (26). To obtain a final estimate of the clean speech, the obtained $\mu_{k,\ell}^{x|y,m}$ are marginalized over the mixtures m similar to (16). While here we focus on the combination of pre-trained and generic enhancement approaches, it is interesting to note that this method allows for different combinations of algorithms, e.g., it is possible to combine the pre-trained enhancement method with either the linear estimator or the LSA estimator or both generic enhancement algorithms.

ISBN 978-3-8007-4275-2

7 Evaluation

In this section, we evaluate the proposed combination and compare it to the LSA [2] and a pre-trained approach with a harmonic post-filter similar to [19, Section 3] by means of Perceptual Evaluation of Speech Quality (PESQ) [27] improvement scores.

For the evaluation, we use 128 sentences taken from the test set of the TIMIT database [28] where we ensure that the amount of sentences spoken by male and female speakers is balanced. The speech signals are artificially corrupted by different background noises with signal-to-noise ratios (SNRs) ranging from -5 dB to 20 dB. Here, we employ babble noise, pink noise, which are taken from the NOISEX-92 database [29], and a non-stationary traffic noise. Additionally, we also include an amplitude modulated version of pink noise. The modulator is given by $f(n) = 1 + 0.5\sin(2\pi n f_{\text{mod}}/f_s)$ where $f_{\text{mod}} = 0.5$ Hz is the modulation frequency, while n is the sample index and f_s the sampling frequency. In our evaluation, the sampling rate of all signals is $f_s = 16$ kHz.

The corrupted signals are processed in 32 ms blocks with an overlap of 50 %. For spectral analysis and synthesis, a square-root Hann window is used. The speech model used in the pre-trained approach consists of 128 mixtures. These are trained off-line on 784 gender balanced uncorrupted sentences from the TIMIT training corpus using the expectation maximization algorithm [22]. The speech presence probability based harmonic post-filter is implemented according to the description in [19, Section 3]. In [19], the post-filter is only applied to voiced frames. Therefore, we determine the voiced probability for each frame using [30]. If the probability exceeds 50 %, the harmonic post filter is applied. The noise PSD is obtained using the estimator described in [6]. The speech PSD is determined using temporal cepstrum smoothing as described in [8]. For the pre-trained method in Section 4 and the linear log-spectrum estimator in Section 5.1, these estimates are propagated to the log-spectral domain using Section 3. For all enhancement methods, we ensure that the noisy input spectrum is attenuated by a maximum of 12 dB. Further, the VTS approximation may give values larger than the noisy observation such that the input signal may be boosted. We prevent this by setting an upper limit to the amplitude of the estimated clean speech spectrum which is given by the amplitude of the noisy observation. This limit is applied for all algorithms in the comparison.

The results are shown in Figure 1. Here, "LSA" denotes the generic speech enhancement algorithm which uses the gain function from [2] without any pre-trained models. The symbol "PrTr" indicates the pre-trained speech enhancement as in Section 4. The combinations are denoted by PrTr+additional method, where the additional methods are given by the linearly constrained log-spectral filter (Lin), the LSA (LSA), and the harmonic model based speech presence probability mask (H) [19]. Combinations with more than two algorithms are not analyzed in this paper. For the combinations with a generic enhancement method, the prior $p(z_{k,\ell}) = 0.5$ is used in (25), i.e., there is no preference of one algorithm over another.

The results show that the sole application of the pre-trained enhancement, i.e., the algorithm PrTr, is comparable to the LSA in pink noise and traffic noise while lower PESQ scores are obtained for the modulated pink noise and babble noise. Especially babble noise appears to be a challenging situation for the employed pre-trained speech enhancement method. Here, the performance is usually lower compared to the generic LSA. Only in combination with the linear log-spectral filter, the performance of the pre-trained approach is comparable to the LSA in terms of PESQ scores in babble noise. The results for the remaining noise types, however, show an improvement of the proposed combination in contrast to the sole application of either the LSA or the pre-trained approach. Furthermore, the proposed combination also outperforms the competing method PrTr+H that employs a harmonic post-filter [19].

In our experiments, the pre-trained approaches showed a tendency to preserve weak speech components more than generic estimators. This more conservative behaviour, however, has the effect that outliers in the noise sometimes remain unsuppressed.

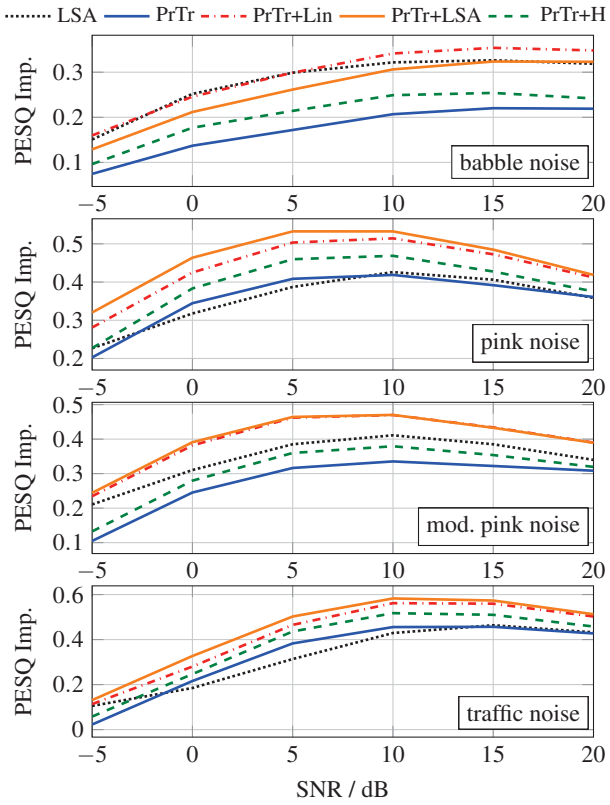

Figure 1: PESQ improvements for evaluated enhancement methods and noise types over different SNRs. LSA: [2], PrTr: pre-trained method, no combination, PrTr+Lin: pre-trained method with linear log-spectral filter, PrTr+LSA: pre-trained method with LSA, PrTr+H: pre-trained method with harmonic filter.

As a result, these enhancement methods generate more audible processing artifacts and noise activations during speech activity. These issues are, on the one hand, related to the speech models which mainly represent the envelope of speech, but, on the other hand, are also linked to the noise PSD estimator which is not able to follow very fast changes in the background noise, e.g., speech bursts in babble. The combination with the generic approaches reduces these artifacts. Informal listening showed that this reduction of artifacts is largest for the PrTr+Lin approach.

8 Conclusions

In this paper, we introduced a combination of a pre-trained speech enhancement method and generic single-channel speech enhancement algorithms. This proposed combination is employed to reduce the processing artifacts of a pre-trained VTS based enhancement method when only a low amount of mixtures is available. In contrast to the post-filter used in [19], the proposed combination rests upon a statistical framework. In non-stationary noise types such as babble noise, the pre-trained approach obtains rather low PESQ scores and the performance is only comparable to the generic LSA in combination with the linear filter. For the other noise types considered in the evaluation, however, the combination with the generic estimators yields a higher quality compared to state-of-the-art single-channel speech enhancement methods. Furthermore, the proposed combination achieves higher PESQ scores than the competing harmonic post-filter.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude

estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[3] C. Breithaupt, M. Krawczyk, and R. Martin, "Parameterized MMSE spectral magnitude estimation for the enhancement of noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Las Vegas, NV, USA), pp. 4037 –4040, Apr. 2008.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 504–512, July 2001.

[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 466–475, Sept. 2003.

[6] T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (New Paltz, NY, USA), pp. 145–148, 2011.

[7] F. Heese and P. Vary, "Noise PSD estimation by logarithmic baseline tracing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brisbane, Queensland, Australia), pp. 4405–4409, Apr. 2015.

[8] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Las Vegas, NV, USA), pp. 4897–4900, Apr. 2008.

[9] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 163–176, Jan. 2006.

[10] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 40, pp. 725–735, Apr. 1992.

[11] D. Y. Zhao and W. B. Kleijn, "HMM-Based Gain Modeling for Enhancement of Speech in Noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 882–892, Mar. 2007.

[12] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1066–1074, Mar. 2007.

[13] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 2140–2151, Oct. 2013.

[14] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series Approach For Environment Independent Speech Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Atlanta, GA, USA), May 1996.

[15] J. C. Segura, A. d. l. Torre, C. Benitez, and A. M. Peinado, "Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks," in *EUROSPEECH*, (Aalborg, Denmark), Sept. 2001.

[16] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero, "ALGONQUIN - Learning Dynamic Noise Models From Noisy Speech for Robust Speech Recognition," in *Advances in Neural Information Processing Systems (NIPS)*, (Vancouver, BC, Canada), pp. 1165–1171, MIT Press, Dec. 2001.

[17] L. Deng, J. Droppo, and A. Acero, "Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 133–143, Mar. 2004.

[18] T. Kristjansson and J. Hershey, "High resolution signal reconstruction," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, (St. Thomas, VI, USA), pp. 291–296, Nov. 2003.

[19] T. Yoshioka and T. Nakatani, "Speech enhancement based on log spectral envelope model and harmonicity-derived spectral mask, and its coupling with feature compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Prague, Czech Republic), pp. 5064–5067, May 2011.

[20] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 162–176, Mar. 1999.

[21] R. F. Astudillo and T. Gerkmann, "On the relation between speech corruption models in the spectral and the cepstral domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Vancouver, BC, Canada), pp. 7044–7048, May 2013.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1 – 38, 1977.

[23] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, (Washington, D.C., USA), pp. 208–211, Apr. 1979.

[24] S. Voran, "Exploration of the additivity approximation for spectral magnitudes," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2015.

[25] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Taipei, Taiwan), pp. 3437–3440, Apr. 2009.

[26] B. J. King and L. Atlas, "Single-Channel Source Separation Using Complex Matrix Factorization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 2591–2597, Nov. 2011.

[27] "P.862 : Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." Jan. 2001.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," tech. rep., Linguistic Data Consortium, Philadelphia, 1993.

[29] A. Varga and H. J. M. 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 Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[30] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 518–530, Feb. 2014.

