INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

MixMax Approximation as a Super-Gaussian Log-Spectral Amplitude Estimator for Speech Enhancement

Robert Rehr, Timo Gerkmann

Signal Processing Group, Department of Informatics, University of Hamburg, Germany
robert.rehr@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

For single-channel speech enhancement, most commonly, the noisy observation is described as the sum of the clean speech signal and the noise signal. For machine learning based enhancement schemes where speech and noise are modeled in the log-spectral domain, however, the log-spectrum of the noisy observation can be described as the maximum of the speech and noise log-spectrum to simplify statistical inference. This approximation is referred to as MixMax model or log-max approximation. In this paper, we show how this approximation can be used in combination with non-trained, blind speech and noise power estimators derived in the spectral domain. Our findings allow to interpret the MixMax based clean speech estimator as a super-Gaussian log-spectral amplitude estimator. This MixMax based estimator is embedded in a pre-trained speech enhancement scheme and compared to a log-spectral amplitude estimator based on an additive mixing model. Instrumental measures indicate that the MixMax based estimator causes less musical tones while it virtually yields the same quality for the enhanced speech signal.

Index Terms: speech enhancement, MixMax model, super-Gaussian PDF, musical noise

1. Introduction

In presence of noise, the intelligibility, as well as, the quality of speech is deteriorated. This may also affect human machine interaction, e.g., applications where a machine is controlled by the user using voice commands. Therefore, speech enhancement algorithms play an important role, e.g., for telecommunication applications, for hearing aids, and for speech recognition. In this paper, enhancement techniques are considered where only a single-channel observation of the noisy speech signal is available.

Single-channel speech enhancement has been a research topic for many years and has led to a variety of different approaches, e.g., [1–7]. Often, these algorithms are based on a framework using the short-time Fourier transform (STFT). Accordingly, the noise is removed in the Fourier domain and the enhanced signal is obtained by taking the inverse of the STFT. The clean speech Fourier coefficients are often estimated within a statistical framework. Commonly, minimum mean-squared error (MMSE) optimal estimators are employed, e.g., [1–3, 8, 9]. These estimators require an estimate of the spectral speech power spectral density (PSD) and noise PSD. Various approaches have been proposed to estimate these quantities blindly from the noisy observation, e.g., [1, 10–12]. This type of enhancement algorithm is referred to as *non-trained* algorithms here. In contrast to non-trained algorithms, *pre-trained* speech enhancement algorithms rely on models of speech and possibly also noise that have been trained prior to the processing. For this, different machine learning algorithms have been used, e.g., generative models such as mixture models and hidden Markov models [6, 13], nonnegative matrix factorization [4], and neural networks [5].

Most non-trained clean speech estimators have been derived based on the assumption that the time domain signals of speech and the noise are additive. Even though this ignores the effect of room characteristics, it reflects the physical properties of sound sufficiently for many applications. For some pre-trained algorithms [6, 7, 14–16], however, an approximation of this model is used which allows to simplify statistical inference if logarithmized spectra are considered. Here, the noisy log-spectral coefficients are modeled as the maximum of the speech and noise coefficients. This is referred to as MixMax model [14] or log-max approximation [15]. In [14], it has been motivated by the empirical finding that the approximation yields spectral representations which are visually similar to the results that are obtained if the additive mixing model is used in the time domain. The validity of this approximation has been further supported in [17] where it has been shown that the MixMax model is the MMSE optimal estimator of the *noisy* log-spectral coefficients if the phase of the complex speech and noise coefficients is uniformly distributed. Further, it is argued in [15] that the error of the MixMax approximation has only a considerable influence if two sources have the same energy in time and frequency. Consequently, as speech has a sparse spectral representation and is uncorrelated to the noise, time-frequency points are often dominated by either speech or noise. From this, the practical expedience is concluded in [15].

The MixMax model is commonly used in combination with pre-trained approaches where speech and noise are modeled using Gaussian distributions in the log-spectral domain [6, 7, 14–16]. In [14], it has been used to adapt the clean speech models to the background noise for robust speech recognition. In the context of speech enhancement, it has been used to infer the log-spectrum of the target speech from noisy observations [6, 7] or mixtures of multiple speakers [15, 16] in an MMSE optimal way. In this paper, we show that the MixMax based clean speech estimator can be interpreted as a super-Gaussian log-spectral amplitude estimator (LSA) estimator similar to [2, 9]. For this, the relationship between spectral and log-spectral coefficients described in [18] is exploited. This relationship also allows to combine pre-trained enhancement schemes based on the MixMax model with non-trained speech and noise PSD estimators such as [1, 10–12]. Following this, we employ the MixMax model in a pre-trained speech enhancement scheme similar to [7] and show that the MixMax based speech estimator [14] causes less artifacts in the background noise compared to super-Gaussian LSA [2, 9] without degrading the speech quality.

First, we recapitulate the MixMax based clean speech estimator in Section 2. Following this, we present the relationship between spectral and log-spectral coefficients [18] and analyze the gain functions that result for the MixMax based clean speech estimator in Section 3. In Section 4, the MixMax based estimator is compared to the super-Gaussian LSA [2, 9] within a pre-trained enhancement scheme and Section 5 concludes the paper.

http://dx.doi.org/10.21437/Interspeech.2017-1243

2. MixMax Based Speech Estimator

In this section, we recapitulate the MMSE optimal estimator of the log-spectral speech coefficients that results from the MixMax model. This estimator operates on the short-time Fourier transformed input signal. For this, the signal is split into overlapping frames which are transformed to the Fourier domain after a tapered analysis window has been applied. This results in the noisy spectra $Y_{k,\ell}$ where ℓ is the frame index and k the frequency index. The MixMax model considers the log-spectra of the noisy input which are defined as

$$y_{k,\ell} = \log\left(|Y_{k,\ell}|^2\right). \tag{1}$$

The speech log-spectrum and the noise log-spectrum are defined accordingly as $s_{k,\ell} = \log(|S_{k,\ell}|^2)$ and $n_{k,\ell} = \log(|N_{k,\ell}|^2)$, respectively. Here, $S_{k,\ell}$ is the complex speech spectrum while $N_{k,\ell}$ denotes the complex noise spectrum. The MixMax signal mixing model [14], also known as log-max approximation [15], is given by

$$y_{k,\ell} = \max(s_{k,\ell}, n_{k,\ell}). \tag{2}$$

Under the model in (2), the distribution of the noisy log-spectral coefficients $y_{k,\ell}$ is given by [14]

$$f_y(y_{k,\ell}) = f_s(y_{k,\ell})F_n(y_{k,\ell}) + f_n(y_{k,\ell})F_s(s_{k,\ell}). \tag{3}$$

Here, $f_s(\cdot)$ and $F_s(\cdot)$ denotes the probability density function (PDF) and the cumulative distribution function (CDF) of the speech log-spectral coefficients $s_{k,\ell}$, respectively. Similarly, $f_n(\cdot)$ and $F_n(\cdot)$ denote the PDF and the CDF of the noise. In [6, 7, 14], $f_s(\cdot)$ is set to a Gaussian distribution

$$f_s(s_{k,\ell}) = \frac{1}{\sqrt{2\pi\lambda^s_{k,\ell}}}\exp\left(-\frac{1}{2}\frac{(s_{k,\ell}-\mu^s_{k,\ell})^2}{\lambda^s_{k,\ell}}\right) \tag{4}$$

$$= \mathcal{N}(s_{k,\ell}|\mu^s_{k,\ell}, \lambda^s_{k,\ell}). \tag{5}$$

The quantities $\mu^s_{k,\ell}$ and $\lambda^s_{k,\ell}$ denote the mean and the variance of the speech log-spectral coefficients $s_{k,\ell}$, respectively. Similarly, also the noise log-spectral coefficients $n_{k,\ell}$ are also assumed to follow a Gaussian distribution as

$$f_n(n_{k,\ell}) = \mathcal{N}(n_{k,\ell}|\mu^n_{k,\ell}, \lambda^n_{k,\ell}). \tag{6}$$

Similar to the speech log-spectrum $s_{k,\ell}$, $\mu^n_{k,\ell}$ and $\lambda^n_{k,\ell}$ denote the mean and the variance of the noise log-spectral coefficients, respectively. Using the mixing model in (2) and the PDFs used for $s_{k,\ell}$ and $n_{k,\ell}$, the MMSE optimal estimator of the speech log-spectral coefficients is considered, i.e., the $\hat{s}_{k,\ell}$ which minimizes $\mathbb{E}\{(s_{k,\ell} - \hat{s}_{k,\ell})^2\}$. Here, $\mathbb{E}\{\cdot\}$ denotes the expectation operator. This can be equivalently expressed as $\hat{s}_{k,\ell} = \mathbb{E}\{s_{k,\ell}|y_{k,\ell}\}$ [19, Chapter 5.2] and the result is given by [7, 14]

$$\hat{s}_{k,\ell} = \rho_{k,\ell}y_{k,\ell} + (1-\rho_{k,\ell})\left(\mu^s_{k,\ell} - \lambda^s_{k,\ell}\frac{f_s(y_{k,\ell})}{F_s(y_{k,\ell})}\right), \tag{7}$$

where $\rho_{k,\ell} = f_s(y_{k,\ell})F_n(y_{k,\ell})/f_y(y_{k,\ell})$ [7, 14]. For obtaining an estimate of the spectral clean speech coefficients $\hat{S}_{k,\ell}$, the log-spectral transformation in (1) is reverted and the result is combined with the noisy phase $\Phi^y_{k,\ell} = \arg\{Y_{k,\ell}\}$ as

$$\hat{S}_{k,\ell} = \sqrt{\exp(\hat{s}_{k,\ell})}\exp\left(j\Phi^y_{k,\ell}\right), \tag{8}$$

where $j = \sqrt{-1}$. For obtaining the time-domain representation of the enhanced signal, the inverse Fourier transform of the estimated clean speech spectra $\hat{S}_{k,\ell}$ is taken. The resulting time-domain frames are weighted by a synthesis window and merged using an overlap-add method.

3. Propagation of Spectral PSD Estimates

In this section, the relationship between spectral PSDs and log-spectral means and variances given in [18] is recapitulated. Previously, this relationship has only been used in combination with an additive mixing model in the time domain, e.g., [18, 20]. Here, it is shown that it allows to employ spectral PSD estimates in the MixMax based speech estimator in (7). Following this, the MixMax based speech estimator can be interpreted as a super-Gaussian LSA.

In [18], the speech spectral coefficients $S_{k,\ell}$ are assumed to follow a super-Gaussian distribution. For this, the speech magnitude $A_{k,\ell} = |S_{k,\ell}|$ is modeled using a χ-distribution as

$$f(A_{k,\ell}) = \frac{2}{\Gamma(\nu)}\left(\frac{\nu}{\Lambda^s_{k,\ell}}\right)^{\nu} A^{2\nu-1}_{k,\ell}\exp\left(-\frac{\nu A^2_{k,\ell}}{\Lambda^s_{k,\ell}}\right). \tag{9}$$

Here, $\nu > 0$ denotes the shape parameter and $\nu < 1$ results in super-Gaussian distributed clean speech coefficients $S_{k,\ell}$. Further, $\Gamma(\cdot)$ is the Gamma function [21, (8.31)] and $\Lambda^s_{k,\ell}$ denotes the speech PSD. The phase of $S_{k,\ell}$ is assumed to be uniformly distributed between $-\pi$ and π. Based on this assumption, the mean of the speech log-spectral coefficients $s_{k,\ell}$ is given by [18]

$$\mu^s_{k,\ell} = \mathbb{E}\{\log(|S_{k,\ell}|^2)\} = \log(\Lambda^s_{k,\ell}) + \psi(\nu) - \log(\nu), \tag{10}$$

where $\psi(\cdot)$ denotes the digamma function [21, (8.360.1)]. Further, the variance is given by [18]

$$\lambda^s_{k,\ell} = \mathbb{E}\left\{\left(\log(|S_{k,\ell}|^2) - \mu^s_{k,\ell}\right)^2\right\} = \psi_1(\nu), \tag{11}$$

where $\psi_1(\nu)$ is the trigamma function [22, (6.4.10)]. We employ the common assumption that the spectral noise coefficients $N_{k,\ell}$ follow a circular symmetric Gaussian distribution. This distribution results if $\nu = 1$ is used in (9) and the same uniform distribution is used for the phase. Hence, the parameters $\mu^n_{k,\ell}$ and $\lambda^n_{k,\ell}$ can be obtained by using $\nu = 1$ and replacing $\Lambda^s_{k,\ell}$ with the noise PSD $\Lambda^n_{k,\ell}$ in (10) and (11).

The relationship between spectral and log-spectral coefficients in (10) and (11), allows to use spectral speech PSDs $\Lambda^s_{k,\ell}$ and noise PSDs $\Lambda^n_{k,\ell}$ in combination with the MixMax based clean speech estimator in (7). As a result, pre-trained models using log-spectral representations can easily be used in combination with speech and noise spectral PSD estimators, e.g., [1, 11, 12]. Hence, the advantages of both domains can be exploited: on the one hand, many different non-trained approaches are available for spectral PSD estimation [1, 10, 11] while, on the other hand, log-spectral representations are better suited for constructing generalizing pre-trained speech models. Further, this relationship gives an interpretation of the means and variances of the pre-trained models. Considering (11), the log-spectral variance depends only on the shape ν. In other words, the log-spectral variance of a pre-trained model can be associated with a specific assumption about the shape of the spectral coefficients. Correspondingly, the log-spectral mean is related to the spectral PSD.

Moreover, the relationship in (10) and (11) allows to interpret the estimator given by (7) and (8) as a real-valued spectral gain function $G_{k,\ell}$ that depends on the spectral *a priori* signal-to-noise ratio (SNR) $\xi_{k,\ell} = \Lambda^s_{k,\ell}/\Lambda^n_{k,\ell}$ and *a posteriori* SNR $\gamma_{k,\ell} = |Y_{k,\ell}|^2/\Lambda^n_{k,\ell}$. The spectral gain function $G_{k,\ell}$ allows the estimated speech spectral coefficients to be represented as $\hat{S}_{k,\ell} = G_{k,\ell}Y_{k,\ell}$. Such an interpretation is usually reserved for MMSE optimal estimators that have been defined in the spectral domain, e.g., [1–3, 8, 9]. In Figure 1, the gain function of the MixMax based estimator is shown that results if the relationship in (10) and (11) is exploited. It is compared to the super-Gaussian LSA [2, 9] which is based on an additive mixing model in the time domain.

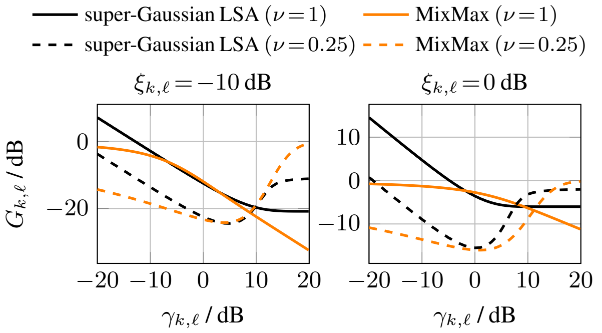

Figure 1: *Gain functions of the super-Gaussian LSA [2, 9] and the MixMax based based clean speech estimator.*

This estimator is implemented using the beta-order estimator proposed in [2] which generalizes [9] if small values are employed for the compression parameter which is set to 0.001 here. This estimator is chosen as it is also an estimator of the log-spectral amplitudes, i.e., the $\hat{S}_{k,\ell}$ is estimated which minimizes $\mathbb{E}\{\log(|S_{k,\ell}|)-\log(|\hat{S}_{k,\ell}|)\}$. Additionally, the same statistical model for the spectral coefficients is used as in the derivation of (10) and (11).

For $\nu=1$, the suppression of both estimators mainly depends on the *a priori* SNR $\xi_{k,\ell}$. With increasing *a priori* SNR, the applied suppression decreases. Differences can be observed for very high and low *a posteriori* SNRs $\gamma_{k,\ell}$ where the MixMax model results in lower gains. Reducing ν, i.e., assuming a super-Gaussian distribution distribution for $S_{k,\ell}$, has a similar effect for both gain functions. In both cases, a higher suppression is applied if the *a posteriori* SNR $\gamma_{k,\ell}$ is close to 0 dB. In [23], it has been shown that this behavior is beneficial if pre-trained speech models are employed that only represent the spectral speech envelope. In this case, it allows to suppress the noise between harmonics which is not represented by the speech models. Little suppression is applied if the *a posteriori* SNR $\gamma_{k,\ell}$ is high, which results in lower speech distortions associated with super-Gaussian estimators.

4. Practical Evaluation

In this section, the MixMax based estimator and the super-Gaussian LSA [9], again realized as in Section 3 using [2], are embedded in a pre-trained enhancement scheme similar to [7]. We show that the MixMax based estimator yields similar results in terms of speech quality which is estimated using Perceptual Evaluation of Speech Quality (PESQ) scores [24] whereas less musical tones are produced as indicated by a modified version of the log-kurtosis ratio [25]. First, the enhancement scheme is described, then the parameters and evaluation setup are considered, and last, the results are presented.

4.1. Pre-Trained Enhancement Scheme

In the employed enhancement scheme, we combine a pre-trained speech PSD estimator with a non-trained noise PSD estimator. Similar to [7], the speech PSD estimator is realized using a phoneme recognizer based on a deep neural network (DNN). The algorithm proposed in [12] is used to estimate the noise PSD.

Figure 2 depicts the architecture of the DNN used for the phoneme recognition. The input is given by a feature vector $\mathbf{v}_\ell=[v_{1,\ell},\ldots,v_{V,\ell}]^T$ where $\cdot^T$ denotes the vector transpose and V the feature dimension. The features are processed by two hidden layers and the output layer returns a score $f(q|\mathbf{v}_\ell)$ for each phoneme q. These scores are interpreted as posterior probability that the phoneme q was spoken given the features $\mathbf{v}_\ell$. The units in the hidden layers are rectified linear units (ReLUs) [7, 26, 27] whose output is given by

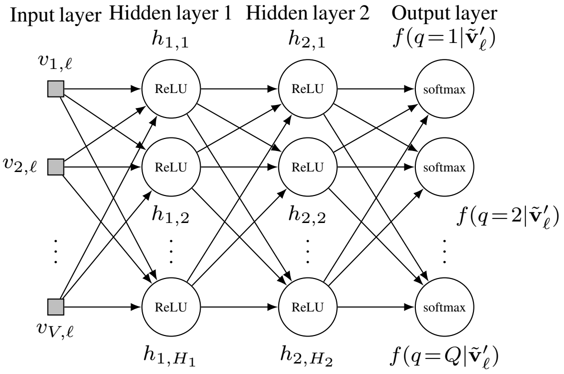

Figure 2: *Architecture of DNN used in the employed phoneme recognizer.*

$$h_{1,j}=\max(0,\mathbf{w}_{1,j}^T\mathbf{v}_\ell+u_{1,j}) \tag{12}$$

$$h_{2,j}=\max(0,\mathbf{w}_{2,j}^T\mathbf{h}_1+u_{2,j}). \tag{13}$$

The vector $\mathbf{w}_{i,j}$ denotes the weights of the jth output of the ith hidden layer and $u_{i,j}$ is the corresponding bias term. The outputs $h_{i,j}$ are pooled in vectors as $\mathbf{h}_i=[h_{i,1},\ldots,h_{i,H_i}]^T$, where H_i denotes the number of units in the ith layer. The transfer function of the final layer is a softmax function which yields the posterior probabilities $f(q|\mathbf{v}_\ell)$ as

$$f(q=j|\mathbf{v}_\ell)=\frac{\exp(\mathbf{w}_{3,j}^T\mathbf{h}_2+u_{3,j})}{\sum_{j'}\exp(\mathbf{w}_{3,j'}^T\mathbf{h}_2+u_{3,j'})}. \tag{14}$$

Before the processing takes place, a speech PSD $\Lambda_k^{s|q}$ is trained for each phoneme q. During enhancement, the following steps are performed for each frame ℓ. First, the noise PSD estimate $\hat{\Lambda}_{k,\ell}^n$ is updated according to the procedure in [12]. Then, using the features $\mathbf{v}_\ell$ extracted from the noisy input spectrum $Y_{k,\ell}$, the posterior probabilities $f(q|\mathbf{v}_\ell)$ are obtained from the DNN. After that, the clean speech coefficients $\hat{S}_{k,\ell}^{(q)}$ are estimated for each phoneme q based on the pre-trained speech PSDs $\Lambda_k^{s|q}$. For this, the super-Gaussian LSA [9] implemented via [2] or the MixMax based estimator in (7) is used. To obtain the final clean speech estimate $\hat{S}_{k,\ell}$, the phoneme dependent estimates $\hat{S}_{k,\ell}^{(q)}$ are combined using the posterior probabilities $f(q|\mathbf{v}_\ell)$ as

$$\hat{S}_{k,\ell}=\sum_{j=1}^{Q}f(q=j|\mathbf{v}_\ell)\hat{S}_{k,\ell}^{(q)}. \tag{15}$$

These steps are repeated until the end of the signal is reached.

4.2. Evaluation Setup

The speech signals processed by the enhancement scheme are sampled at a rate of 16 kHz. For the STFT, 32 ms frames with 50 % overlap are employed and a square-root Hann window is used for spectral analysis and synthesis.

We use 13 Mel-frequency cepstral coefficients (MFCCs) [28] and their Δ and $\Delta\Delta$ derivatives as input features for the DNN based phoneme recognizer. Similar to [7, 26], the features of three previous and three future frames are appended to the feature vector of the current frame to include context information. As 39 MFCCs are extracted per frame, the included context results in a total feature dimension of $V=273$. The features are normalized to unit mean and zero variance [29] before being used in the phoneme recognizer. This normalization is applied per TIMIT utterance. The hidden layers have $H_1 = H_2 = 512$ hidden units. The weights of the DNN are optimized prior to the processing using 1196 sentences taken from the training set of the TIMIT database [30]. It has been ensured that the training sentences are gender and phonetically balanced. As in [7], only clean speech data is used for training to avoid noise specific adaptations of the DNN. The phonemes q are given by the TIMIT annotation which distinguishes between 61 classes. For all frames ℓ in the training data, the annotation is used as training targets which is encoded in 61-dimensional target vectors. For these vectors, all elements are set to zero except the qth element which takes the value 1 to indicate the respective phoneme. For the optimization the cross-entropy is employed as error function which is optimized using scaled conjugate back-propagation [31]. The weights of the first two hidden layers, i.e., $\mathbf{w}_{1,j}$ and $\mathbf{w}_{2,j}$ with $j = 1,\ldots,512$, are initialized using the Glorot method [32] while the weights of output layer $\mathbf{w}_{3,j}$ with $j = 1,\ldots,61$ are initialized using the Nguyen-Widrow method [33]. For each phoneme q, the phoneme dependent speech PSD $\Lambda_k^{s|q}$ is determined by averaging all speech periodograms $|S_{k,\ell}|^2$ labeled as the corresponding phoneme q in the TIMIT annotation.

Due to the averaging of phonemes, only spectral envelopes can be represented by the pre-trained speech PSDs $\Lambda_k^{s|q}$. Hence, similar to [23] a super-Gaussian speech PSD is assumed by using $\nu = 0.25$. This choice allows to suppress noise between the spectral speech harmonics and yields a satisfying compromise for both considered speech estimators. Finally, both gain functions $G_{k,\ell}$ are limited such that a time-frequency bin may not be suppressed by more than 15 dB.

We use PESQ [24] as instrumental measure for the speech quality and a modified version of the log-kurtosis ratio proposed in [25] to evaluate the noise quality in terms of musical tones. Similar to [25], we define the log-kurtosis ratio as

$$\Delta\kappa_{\log} = \log\left(\frac{\kappa_{\tilde{n}}}{\kappa_n}\right), \tag{16}$$

where $\kappa_{\tilde{n}}$ is the empirical kurtosis of the processed noise whereas κ_n denotes the empirical kurtosis of the unprocessed noise. The kurtosis can be considered a measure of outliers and, thus, a positive log-kurtosis ratio $\Delta\kappa_{\log}$ is expected if the processed signals contains musical tones. Instead of estimating the kurtosis for each frame ℓ and using the average along time as κ_n and $\kappa_{\tilde{n}}$, we estimate the kurtosis per frequency band as

$$\kappa_n[k] = \frac{\frac{1}{|\mathbb{I}_k|}\sum_{\ell\in\mathbb{I}_k}\left[|N_{k,\ell}|^2 - \overline{|N|_k^2}\right]^4}{\left(\frac{1}{|\mathbb{I}_k|}\sum_{\ell\in\mathbb{I}_k}\left[|N_{k,\ell}|^2 - \overline{|N|_k^2}\right]^2\right)^2}. \tag{17}$$

In (17), the set $\mathbb{I}_k$ contains only frames in the kth frequency band where the background noise is dominant as

$$\mathbb{I}_k = \{\ell \,|\, |S_{k,\ell}|^2/|N_{k,\ell}|^2 < \theta\}. \tag{18}$$

Here, θ is a threshold value which is set to -10 dB in this evaluation. The cardinality of $\mathbb{I}_k$ is denoted by $|\mathbb{I}_k|$ and $\overline{|N|_k^2}$ is given by $\overline{|N|_k^2} = \sum_{\ell\in\mathbb{I}_k}|N_{k,\ell}|^2/|\mathbb{I}_k|$. Finally, κ_n is given by $\kappa_n = \sum_{k=0}^{K-1}\kappa_n[k]/K$, where K denotes the number Fourier coefficients. Similarly, the kurtosis of the processed noise periodogram $|\tilde{N}_{k,\ell}|^2$ is determined.

For testing, 128 sentences taken from the TIMIT test corpus [30] are used where, again, a gender balanced set is used. The clean speech sentences are corrupted by babble noise, factory 1 noise and pink noise taken from the NOISEX-92 database [34] at SNRs ranging from -5 dB to 20 dB. Additionally, a modulated version of the pink noise similar to [12] and a traffic noise taken from "https://www.freesound.org/s/75375/" is used.

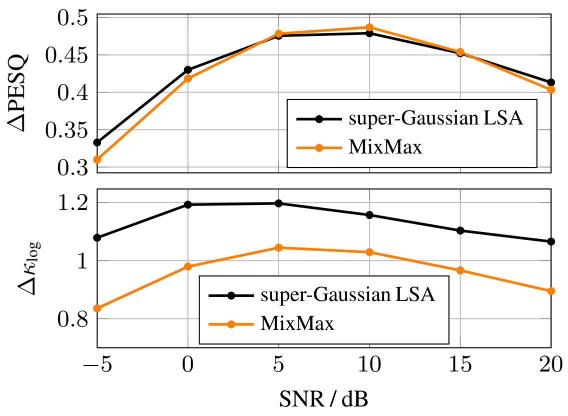

Figure 3: *PESQ improvement score (upper panel) and log-kurtosis ratio (lower panel) of the super-Gaussian LSA [2, 9] and the MixMax based estimator averaged over all noise types.*

4.3. Results

Figure 3 depicts the PESQ improvement scores and log-kurtosis ratio obtained for the used variant of the super-Gaussian LSA [2, 9] and the MixMax based clean speech estimator. The results are averaged over all noise types and the upper panel indicates that the quality of the speech signal is similar for both employed clean speech estimators. The log-kurtosis in the lower panel of Figure 3 shows lower values for the MixMax model, i.e., it indicates less musical tones. We note that if the babble noise and the factory noise are considered separately, the log-kurtosis ratio is higher for the MixMax based estimator. In informal listening tests, however, no disturbing musical tones could be noticed and both clean speech estimators have been found to sound very similar in these highly non-stationary noise types. Part of the reason may be that estimating the fourth-order moments in the kurtosis metric is rather difficult for these noise types. This possibly renders the log-kurtosis ratio unreliable for non-stationary noises. However, for other noise types, such as the pink noise and the traffic noise, it is clearly audible that the MixMax based estimator causes less artifacts. Hence, the overall averaged log-kurtosis ratio in Figure 3 adequately reflects the trend that the MixMax based estimator results in less musical tones which is confirmed in informal listening tests. This is achieved while maintaining the same PESQ scores as the super-Gaussian LSA [2, 9]. Audio examples can be found at "https://www.inf.uni-hamburg.de/en/inst/ab/sp/publications/interspeech2017.html".

5. Conclusions

In this paper, we showed that the MixMax based estimator used in [6, 7] can be interpreted as a super-Gaussian LSA. For this, the relationship described in [18] is exploited. This, additionally, allows to combine pre-trained log-spectral models and spectral speech and noise PSD estimators for speech enhancement. Further, the MixMax based speech estimator is compared to the super-Gaussian LSA proposed in [2, 9] using a pre-trained speech enhancement scheme. The instrumental measures indicate that the speech quality of both estimators is nearly identical while the MixMax based speech estimator causes less musical artifacts in the suppressed background noise.

6. 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, no. 6, pp. 1109–1121, Dec. 1984.

[2] 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, Apr. 4, 2008, pp. 4037–4040.

[3] M. Krawczyk-Becker and T. Gerkmann, "On MMSE-based estimation of amplitude and complex speech spectral coefficients under phase-uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2251–2262, Dec. 2016.

[4] 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, no. 10, pp. 2140–2151, Oct. 2013.

[5] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 577–581.

[6] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, Sep. 2002.

[7] S. E. Chazan, J. Goldberger, and S. Gannot, "A hybrid approach for speech enhancement using MoG model and neural network phoneme classifier," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2516–2530, Dec. 2016.

[8] 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.

[9] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-gaussian densities," in *Interspeech*, Brighton, United Kingdom, 2009, pp. 1319–1322.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[11] 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, Apr. 4, 2008, pp. 4897–4900.

[12] 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, 2011, pp. 145–148.

[13] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[14] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.

[15] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *European Conference on Speech Communication and Technology (Eurospeech/Interspeech)*, Geneva, Switzerland, Sep. 2003.

[16] M. H. Radfar and R. M. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[17] M. H. Radfar, A. H. Banihashemi, R. M. Dansereau, and A. Sayadiyan, "Nonlinear minimum mean square error estimator for mixture-maximisation approximation," *Electronics Letters*, vol. 42, no. 12, pp. 724–725, Jun. 2006.

[18] T. Gerkmann and R. Martin, "On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4165–4174, 2009.

[19] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals*. Cambridge University Press, 2010.

[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, no. 2, pp. 162–176, Mar. 1999.

[21] I. S. Gradshteyn and I. W. Ryzhik, *Table of Integrals, Series, and Products*, 7th, D. Zwillinger and V. Moll, Eds. Academic Press, Feb. 2007.

[22] M. Abramowitz and I. A. Stegun, Eds., *Handbook of mathematical functions : with formulas, graphs, and mathematical tables*, 9th ed., New York: Dover Publ., 1973.

[23] R. Rehr and T. Gerkmann, "On the importance of super-gaussian speech priors for pre-trained speech enhancement," Mar. 15, 2017. arXiv: 1703.05003 [cs.SD].

[24] "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," ITU-T Recommendation, Jan. 2001, [Online]. Available: http://www.itu.int/rec/T-REC-P.862-200102-I/en.

[25] H. Yu and T. Fingscheidt, "Black box measurement of musical tones produced by noise reduction systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 4573–4576.

[26] L. Tóth, "Phone recognition with deep sparse rectifier neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 6985–6989.

[27] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 8609–8613.

[28] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[29] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[30] 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*, 1993.

[31] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[32] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 249–256.

[33] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *International Joint Conference on Neural Networks (IJCNN)*, San Diego, CA, USA, Jun. 1990, 21–26 vol.3.

[34] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG.10 noise database," TNO Institute for perception, Technical Report IZF 1988-3, 1988.