MUSICAL NOISE SUPPRESSION FOR SPEECH ENHANCEMENT USING PRE-IMAGE ITERATIONS

Christina Leitner and Franz Pernkopf

Signal Processing and Speech Communication Laboratory
Graz University of Technology
Inffeldgasse 16c, 8010 Graz, Austria

ABSTRACT

In this paper, we propose a method for musical noise suppression in enhanced speech recordings. This method uses the number of iterations needed to compute the so-called pre-images for patches from complex-valued spectral data of the noisy signal. From the number of iterations a continuous mask can be derived which discriminates between speech and non-speech regions. This mask is applied to suppress musical noise that is most disturbing in non-speech regions. Compared to the original enhanced recording the performance in terms of objective quality measures slightly decreases, but the subjectively perceived quality is better, as the musical noise can be significantly reduced.

***Index Terms*—** Speech enhancement, musical noise suppression, pre-image problem

1. INTRODUCTION

In speech enhancement, the occurrence of musical noise is a major problem. Musical noise originates from inaccuracies of the enhancement algorithm at hand and is a random amplification of frequency bins that change quickly over time. This is perceived as "twittering" and degrades the perceptual quality massively. If it is too prominent, it may even be more disturbing than the interference before speech enhancement. Figure 1 shows an example for a speech recording with interfering white Gaussian noise at 10dB SNR that has been enhanced by the generalized subspace method [1]. The "blobs" in the spectrogram are perceived as musical noise.

Much research has been carried out on how to combat musical noise, either by modifying the enhancement method at hand or by postprocessing. The postprocessing method in [3] for spectral subtraction is based on musical noise/speech classification from the spectrum and subsequent processing of the spectral values. In [2], post-filtering with a perceptually inspired filter is applied to the outcome of the used subspace method. The method proposed in [4] can be applied as postprocessing for any speech enhancement method, it performs smoothing of weighting gains using a robust detector for speech pauses and low SNR conditions.

Recently, we presented a new method for speech enhancement that applies kernel principal component analysis (kPCA) on complex spectral data [5]. Kernel PCA is equivalent to principal component analysis (PCA) after a non-linear transformation to a high-dimensional feature space. The transformation to the feature space is performed implicitly via the computation of the kernel matrix. The inverse transformation, however, has to be computed explicitly. Due to the non-linear mapping, the inverse transformation is not unique [6]. Several solutions have been proposed to solve this so-called "pre-image problem" [6, 7, 8, 9].

We gratefully acknowledge funding by the Austrian Science Fund (FWF) under the project number S10610-N13.

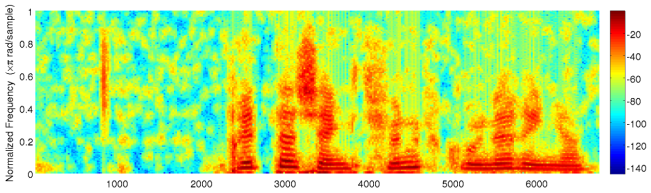


Fig. 1. Spectrogram of the phrase "Britta schenkt fünf grüne Ringe." uttered by a female speaker, corrupted by additive white Gaussian noise at 10dB SNR and enhanced using the generalized subspace method [1]. Note that the "blobs" in the non-speech regions are perceived as musical noise.

In [10], we reported that the used pre-image method can crucially influence the performance of the speech de-noising process. Furthermore, in [11] we observed that the influence of the kernel PCA projection is negligible compared to the computation of the pre-image. During further investigation of the iterative estimation of pre-images in the spectral domain, we discovered a strong correlation between the number of iterations needed and the properties of the underlying signal. To be more precise, we observed that the number of pre-image iterations is an indicator for the presence of speech in the noisy signal. In this paper, we describe a method that uses the number of iterations for musical noise suppression in enhanced speech recordings by deriving a continuous mask. The mask is applied on the magnitude of the enhanced signal and suppresses musical noise in non-speech regions. Listening to the resulting speech recordings confirms the better speech quality while objective quality measures slightly decrease due to inaccuracies of the mask.

The paper is organized as follows: Section 2 explains the motivation for the proposed method and describes the realization. Section 3 presents the experiments, the evaluation, and the results. Section 4 concludes the paper and gives an outlook on future work.

2. PRE-IMAGE ITERATIONS FOR SPEECH ENHANCEMENT AND MUSICAL NOISE SUPPRESSION

Mika et al. [6] proposed an iterative method to compute the pre-image of samples processed by kernel PCA. In [11], we observed that the weights of the kernel PCA only have a minor effect on the outcome. Neglecting these weights, the update equation simplifies

as follows

$$\mathbf{z}_{t+1} = \frac{\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i)}, \quad (1)$$

where $\mathbf{z}_t$ denotes the pre-image or enhanced sample at iteration step t, $\mathbf{x}_i$ is the i^{th} noisy sample, M is the number of noisy samples, and $k(\cdot,\cdot)$ represents the kernel function. We use the Gaussian kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/c), \quad (2)$$

where c is its variance. This kernel serves as similarity measure between the samples.

The method of Mika et al. can be refined by regularization as proposed by Abrahamsen and Hansen in [8]. Applying the observations from above, the corresponding simplified iterative equation is

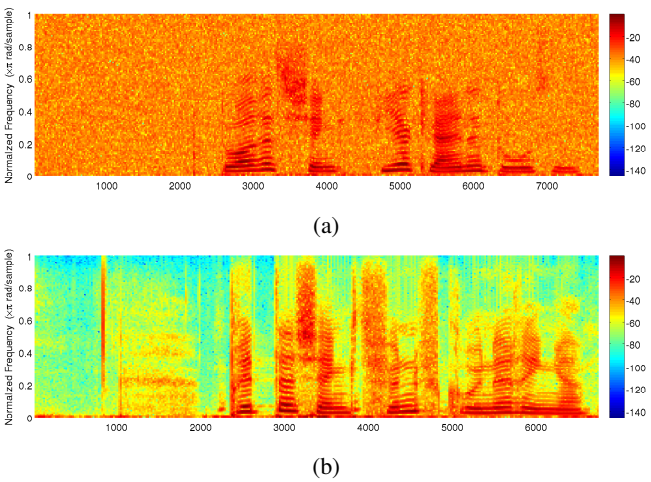
$$\mathbf{z}_{t+1} = \frac{\frac{2}{c}\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i + \lambda\mathbf{x}_0}{\frac{2}{c}\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i) + \lambda} \quad (3)$$

where $\mathbf{x}_0$ is the noisy sample for which the pre-image is computed and λ is the regularization parameter.

The equations (1) and (3) are iterated until convergence of $\mathbf{z}_{t+1}$. The sample vectors for the pre-image iteration are quadratic patches in column-major order which are extracted from the time-frequency representation after application of the short term Fourier transform (for details see section 3).

2.1. Convergence behaviour

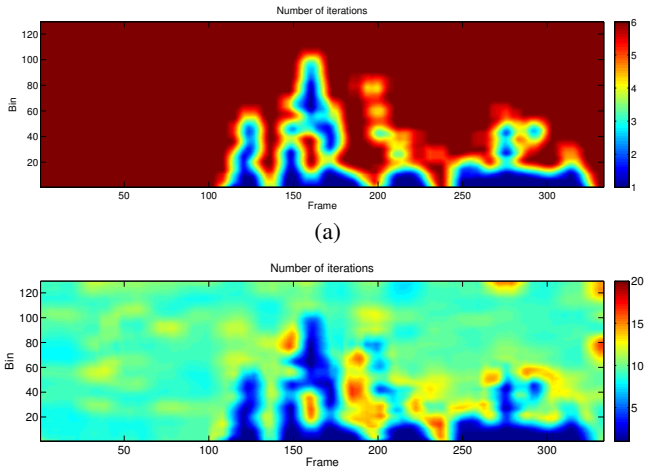
During analysis of the pre-image iterations, we observed a correlation between the convergence rate and the properties of the time-frequency representation of the signal. More precisely, if a patch is extracted from a region with speech only, the number of iterations is very low. If the patch stems from a region with noise only, the number of iterations is higher. Furthermore, if the patch is from a boundary region containing speech and noise at equal amount, the required number of iterations is even higher. Figure 2 (a) shows the spectrogram of a speech utterance by a female speaker corrupted by additive white Gaussian noise at 10 dB SNR, Figure 2 (b) shows the clean recording, and Figure 3 (a) and (b) show the number of iterations computed from the noisy recording, limited to maximally 6 and 20, respectively.



(a)

(b)

Fig. 2. (a) Noisy recording of the phrase visualized in Figure 1, corrupted by additive white Gaussian noise at 10dB SNR. (b) Corresponding clean recording.



(a)

(b)

Fig. 3. (a) Iteration numbers computed with maximally 6 iterations. (b) Iteration numbers computed with maximally 20 iterations.

This observation makes the number of iterations a useful indicator to discriminate between speech and non-speech regions and naturally leads to a continuous mask for time-frequency representations. Musical noise in enhanced speech recordings is most disturbing in non-speech regions, so we apply this mask to attenuate musical noise.

2.2. Musical noise reduction

If the number of iterations for the patches of a speech utterance are computed, speech and non-speech regions can be easily separated by setting a threshold. Instead of using a binary mask, a continuous

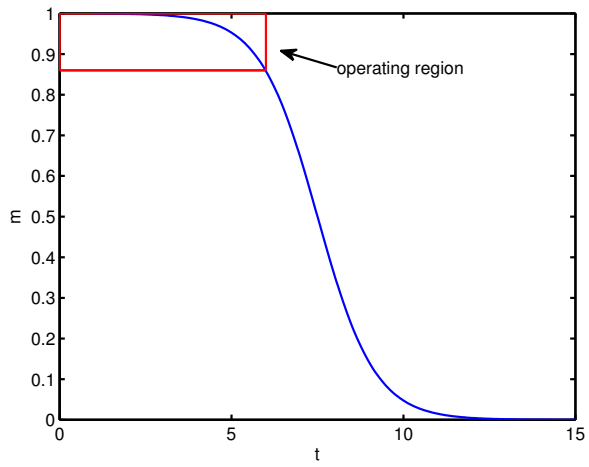
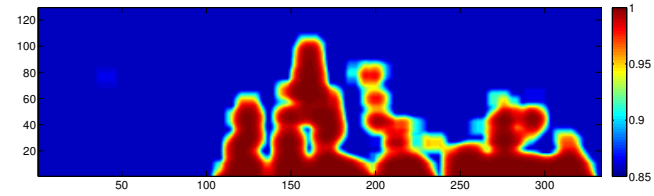


Fig. 4. Sigmoid mapping function from the number of iterations t to the weight of the mask m. The parameters are set to $a = 1.2$ and $b = 9$. The function is used for maximally 6 iterations (see operating region).

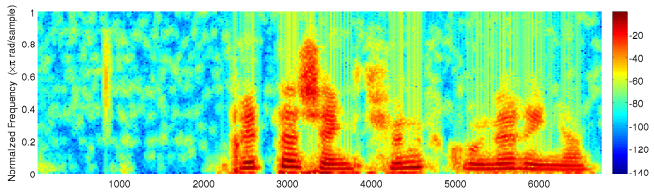
mask that allows smooth transitions between speech and non-speech regions is preferred, as it reduces potential artifacts from inaccuracies of the mask estimation. We compute the mask m by applying the sigmoid function

$$m = \frac{1}{1 + \exp(t * a + b)} \tag{4}$$

to the number of iterations t, where a and b are scaling parameters. Figure 4 shows the mapping function with the parameters set to $a = 1.2$ and $b = 9$, Figure 5 (a) shows the resulting mask. To perform musical noise suppression, the mask is multiplied with the magnitude of the STFT in the logarithmic domain. Figure 5 (b) illustrates the result for the recording plotted in Figure 1. Musical noise is still visible in the spectrogram, however its amplitude is decreased.



(a)



(b)

Fig. 5. (a) Mask computed with the sigmoid function from (4) with $a = 1.2$ and $b = 9$ and maximally 6 iterations. (b) Resulting speech utterance from Figure 1 with suppressed musical noise after application of the mask.

3. EXPERIMENTS

As in previous applications [5], the sample vectors for the pre-image iteration are computed with the following procedure: First, the short term Fourier transform is computed with a frame length of 256 samples, an overlap of 50% and the application of a Hamming window. The resulting time-frequency representation is split along time and frequency axis to reduce computational costs. From the retrieved frequency bands, quadratic patches of size 12×12 are extracted with overlap 11. The frequency range covered by the frequency bands is specified to be 8 patches with an overlap of 4 patches between adjacent bands. The time range of the bands is 20 patches with an overlap of 10 patches. In previous work, windowing of the patches was beneficial, so a 2D Hamming window is applied before rearranging the patches to vectors in column-major order. The pre-image iteration is applied on each frequency band independently, i.e., to the M noisy samples that belong to the current frequency band.

Due to the overlap of the bands, patches at a certain time frequency position belong to different frequency bands at the same time. Consequently, more than one count of iterations is retrieved per patch. Further, we are primarily interested in the number of iterations for each *bin* and not each *patch*. Hence, some more processing has to be done to retrieve a valid mask for musical noise reduction: First, each iteration value is expanded to the size of patches (12×12). Then the patches of iteration numbers corresponding to the same time-frequency position are averaged. Application of function (4) on this iteration map leads to the mask as visualized in Figure 5 (a).

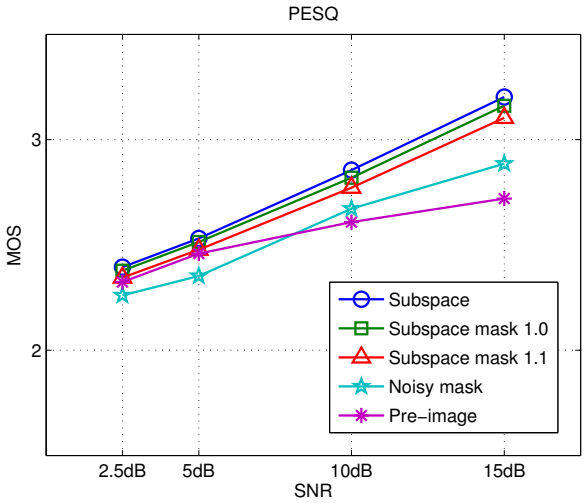


Fig. 6. Perceptual evaluation of speech quality (PESQ) measure for the original subspace method [1] (Subspace), the proposed musical noise suppression applied with two different parameter settings (Subspace mask 1.0 with $a = 1.1$ and $b = 9$ and Subspace mask 1.1 with $a = 1.2$ and $b = 9$), the mask applied on the noisy signal (Noisy mask), and the pre-image iteration method from [11] (Pre-image).

The method was tested on a database with recordings of 6 speakers, 3 male and 3 female. Each of the speakers read a list of 20 sentences. The recordings were performed with a close-talk microphone and 16 kHz sampling frequency. White Gaussian noise was added at 2.5, 5, 10, and 15 dB SNR. Evaluation was done by listening, by visual inspection of the spectrogram and by computation of objective quality measures.

Listening to the files confirmed that musical noise can be reduced with a proper parameter choice without significantly affecting the speech components.[1] The spectrograms confirm this result – one example is shown in 5 (b). If the parameters of the sigmoid mapping function (4) are selected to lead to stronger attenuation, also speech components, mainly fricatives, are affected. This can be explained by the similar properties of noise and the fricatives. Strong attenuation results in a low pass effect, so it was avoided to prevent a degradation of the signal quality.

In addition to the subjective evaluation, the enhanced recordings were evaluated using objective quality measures that have been reported to have a high correlation to subjective listening tests [12]. Figure 6 and 7 show the perceptual evaluation of speech quality (PESQ) measure and the frequency-weighted segmental SNR (fwsegSNR) for the musical noise suppression with two parameter settings compared to the original subspace method [1], the method using pre-image iterations as described in [11], and the mask directly applied on the noisy signal. The PESQ measure shows a minor decrease of performance if the musical noise suppression is

[1] Audio examples can be found on `http://www2.spsc.tugraz.at/people/chrisl/audio/`

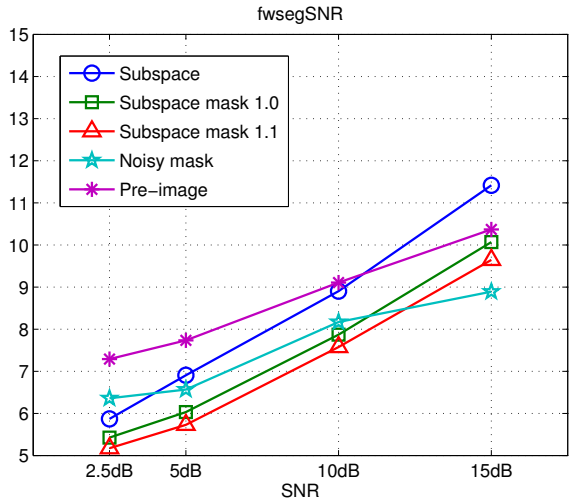


Fig. 7. Frequency-weighted segmental SNR for the original subspace method (Subspace), the proposed musical noise suppression with two parameter settings (Subspace mask 1.0 with $a = 1.0$ and $b = 9$ and Subspace mask 1.1 with $a = 1.1$ and $b = 9$), the mask applied on the noisy signal (Noisy mask), and the pre-image iteration method (Pre-image).

applied, but the performance is better than for the pre-image iteration approach and the application of the mask on the noisy signal. The frequency-weighted SNR of all methods lies in the same range, here the pre-image iteration approach and the approach using only the mask achieve similar performance as the subspace methods with and without musical noise suppression. It is important to note that although the performance measures with musical noise suppression are slightly weaker, the subjective audio quality is higher since there is less disturbing musical noise. The performance measures do not seem to put emphasis on the musical noise, however they rely on changes in the spectrum which might be affected by the application of the mask. Hence subjective listening tests are necessary to reliably evaluate the results.

4. CONCLUSION AND FUTURE WORK

In this paper, we incorporate the knowledge gained from pre-image iterations to suppress musical noise in enhanced speech recordings. If the iterative pre-image method is applied on complex-valued spectral data, the number of iterations until convergence shows a correlation to the properties of the underlying signal and enables a discrimination between speech and non-speech regions. This information is used to create a mask that can be applied on the magnitude of the enhanced signal to suppress the musical noise.

The method was applied on speech recordings corrupted by additive white Gaussian noise at different SNRs which have been enhanced using the generalized subspace method. The empirical comparison of the recordings with suppressed musical noise to the recordings from the subspace method shows a minor decrease in terms of objective quality measures. This can be explained due to attenuations of the speech spectrum caused by inaccuracies of the mask. The subjective quality, however, is better as less musical noise can be perceived.

In future, we plan to optimize the computation of the mask by, e.g., adopting image processing techniques to expand the mask in order to reduce suppression of speech components. Furthermore, we want to extend the experiments to more general scenarios with different noise types such as babble noise. Finally, we plan to do a subjective listening test, as objective quality measures do not always fully reflect the perceived audio quality, as is the case for the presence of musical noise.

5. REFERENCES

[1] Yi Hu and Philipos C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 334–341, 2003.

[2] Mark Klein and Peter Kabal, "Signal subspace speech enhancement with perceptual post-filtering," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 537–540, 2002.

[3] Zenton Goh, Kah-Chye Tan, and T.G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 287 –292, may 1998.

[4] Thomas Esch and Peter Vary, "Efficient musical noise suppression for speech enhancement systems," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4409 –4412, 2009.

[5] Christina Leitner, Franz Pernkopf, and Gernot Kubin, "Kernel PCA for speech enhancement," *12th Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1221–1224, 2011.

[6] Sebastian Mika, Bernhard Schölkopf, Alex Smola, Klaus-Robert Müller, Matthias Scholz, and Gunnar Rätsch, "Kernel PCA and de-noising in feature spaces," *Advances in Neural Information Processing Systems 11*, pp. 536–542, 1999.

[7] James T. Kwok and Ivor W. Tsang, "The pre-image problem in kernel methods," *IEEE Transactions on Neural Networks*, vol. 15, pp. 408–415, 2004.

[8] Trine Julie Abrahamsen and Lars Kai Hansen, "Input space regularization stabilizes pre-images for kernel PCA denoising," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2009.

[9] Paul Honeine and Cédric Richard, "Solving the pre-image problem in kernel machines: A direct method," *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2009.

[10] Christina Leitner and Franz Pernkopf, "The pre-image problem and kernel PCA for speech enhancement," *Nonlinear Speech Processing Workshop (NOLISP) 2011, LNAI (in press)*, 2011.

[11] Christina Leitner and Franz Pernkopf, "Speech enhancement using pre-image iterations," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2012, (submitted)*, 2012.

[12] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229 –238, 2008.