SUPPRESSION OF MUSICAL NOISE IN ENHANCED SPEECH 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 present a post-processing method for musical noise suppression in enhanced speech recordings. The method uses pre-image iterations computed patch-wise on complex-valued spectral data of the enhanced signal to discriminate between speech and non-speech regions. From this knowledge, a binary mask is derived to suppress musical noise in the non-speech regions, where it is most disturbing. The method is evaluated using objective quality measures of the PEASS toolbox. These measures confirm that a suppression of artifacts and an increase in overall quality for several noise conditions is achieved.

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

1. INTRODUCTION

The occurrence of musical noise is a major problem in speech enhancement. Musical noise is caused by inaccuracies of the enhancement algorithm at hand, it originates from a random amplification of frequency bins that change quickly over time. Musical noise is perceived as "twittering" and can severely degrade the perceptual quality of enhanced speech recordings. If it is too prominent, it may even be more disturbing than the interference before enhancement. Figure 1 (a) shows the spectrogram of a speech recording with interfering white Gaussian noise at 10 dB signal-to-noise ration (SNR) that has been enhanced by the generalized subspace method [1]. The "blobs" in the non-speech region of the spectrogram are perceived as musical noise.

Much research has been carried out on how to avoid or suppress musical noise, either by modifying the enhancement method or by post-processing the enhanced utterances. In the context of spectral subtraction, spectral flooring [2] and oversubtraction [3] were introduced. For post-processing musical noise/speech classification of the spectral bins and subsequent manipulation [4], post-filtering [5] and smoothing of the weighting gains [6] were proposed.

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

Recently, we showed how the knowledge gained from the convergence behaviour of pre-image iterations applied for speech de-noising [7] can be employed to suppress musical noise in enhanced speech recordings [8]. In this paper, we perform pre-image iterations on enhanced signals, in contrast to [8], where they were applied on the noisy signal. From the number of iterations until convergence a binary mask is derived that segments the signal into speech and non-speech regions. This mask is used to suppress musical noise in non-speech regions. Experiments were performed on speech data corrupted by additive white Gaussian noise at 0, 5, 10, and 15 dB SNR. The resulting speech recordings were evaluated using the PEASS toolbox [9]. The provided quality measures show an increase in overall quality and a decrease of artifacts - this is consistent with the subjective impression from listening.

This paper is organized as follows: Section 2 introduces pre-image iterations and for musical noise suppression. Section 3 presents the experimental setup, the evaluation and the results. Section 4 concludes the paper.

2. MUSICAL NOISE SUPPRESSION USING PRE-IMAGE ITERATIONS

Pre-image iterations were originally proposed for speech denoising. In [8], we introduced pre-image iterations for suppression of musical noise in enhanced speech recordings, i.e., the pre-image iterations were executed on the noisy signal before enhancement, while in this paper they are computed on the enhanced signal. For speech enhancement any method such as subspace methods or spectral subtraction can be used. The block diagram in Figure 2 illustrates the implementation: First the short-term Fourier transform is computed for the enhanced signal s. Then patches are extracted from the resulting time-frequency representation. The pre-image iterations are used to derive a binary mask for speech/non-speech discrimination, that is further refined by morphological operations from image processing. This mask is used to filter the magnitude of the input signal. The resulting magnitude values are recombined with the phase of the input signal, the inverse

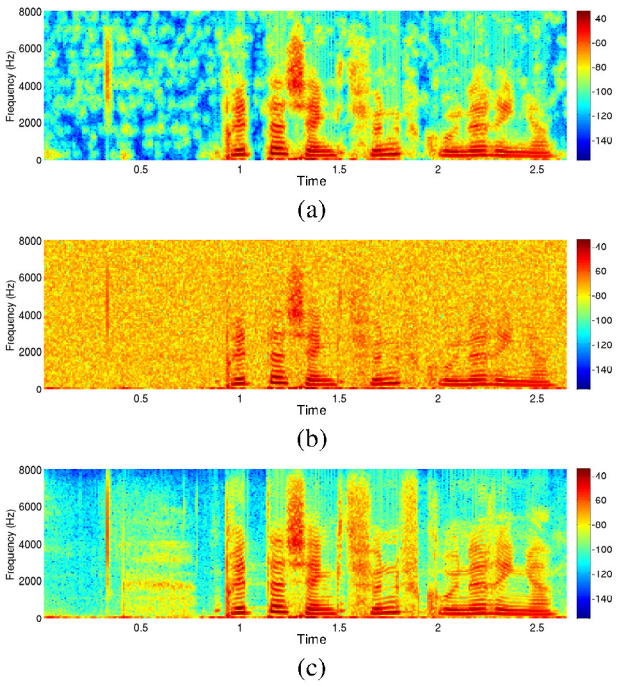


Fig. 1. (a) Spectrogram of the phrase "Britta schenkt fünf grüne Ringe." uttered by a female speaker, corrupted by additive white Gaussian noise at 10 dB SNR and enhanced using the generalized subspace method [1]. Note that the "blobs" in the non-speech regions are perceived as musical noise. (b) Noisy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

Fourier transformation is applied and the signal is synthesized using the overlap-add method. The individual blocks will be explained in more detail in the following sections.

2.1. Extraction of sample vectors

The short-term Fourier transform is computed from frames of 256 samples with 50% overlap and application of a Hamming window. The resulting time-frequency representation is segmented along time and frequency axis into so-called frequency bands to decrease the computational costs (see Figure 3). Each frequency band covers a frequency range of 8 patches and a time range of 5 patches. There is no overlap of bands along the frequency axis, the overlap along the time axis is 2.5 patches to provide smoother transitions. The frequency bands are further split into small quadratic overlapping patches covering 12×12 bins with an overlap of 6 bins both in time and frequency. The bins in each patch are rearranged in column-major order to form the sample vectors for the pre-image iterations.

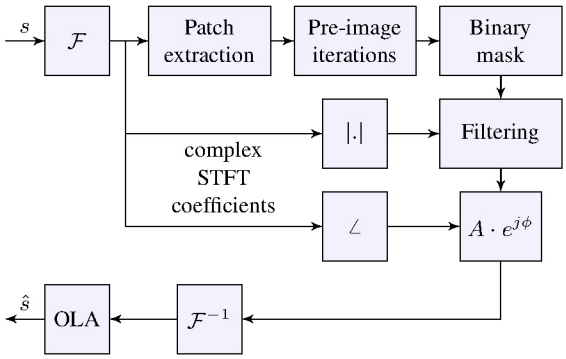


Fig. 2. Block diagram for musical noise suppression.

2.2. Convergence analysis of the pre-image iterations

Pre-image iterations are derived from pre-image estimation in the context of kernel principal component analysis (PCA) [10]. In kernel methods, data is transformed to a high-dimensional feature space, where any further processing is executed. Depending on the application, the resulting samples have to be transformed back to the original input space. These back-transformed samples are called *pre-images*. The pre-image cannot always be computed directly, however, several methods have been proposed to compute an estimate of the pre-image. Our work is based on the iterative method proposed by Mika et al. [10] extended by regularization [11]. This method has proven to be stable [12] and achieved good results for de-noising [7].

When kernel PCA is used for de-noising, the enhanced sample is estimated by the pre-image computed from a linear combination of the noisy samples. In [7], we observed that the weights from the kernel PCA projection in the linear combination have only a minor effect on the de-noising. Therefore, we neglect the weights, and the update equation to estimate the pre-image $\mathbf{z}$ with regularization simplifies to

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

where t denotes the iteration step, $\mathbf{x}_i$ is the i^{th} (noisy) sample

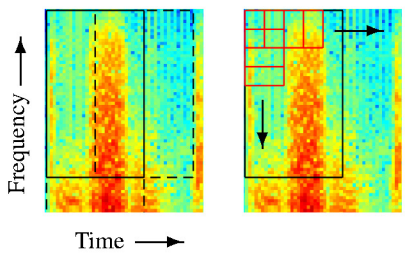
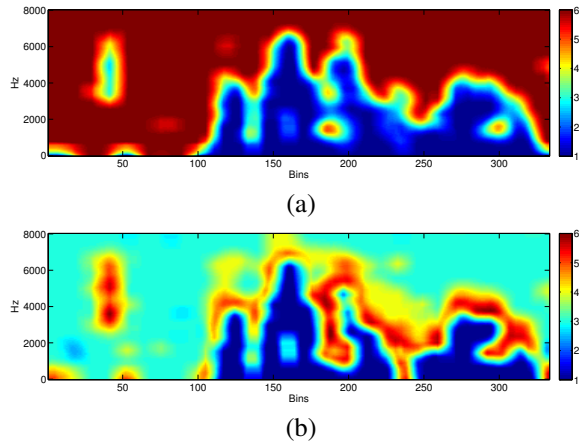


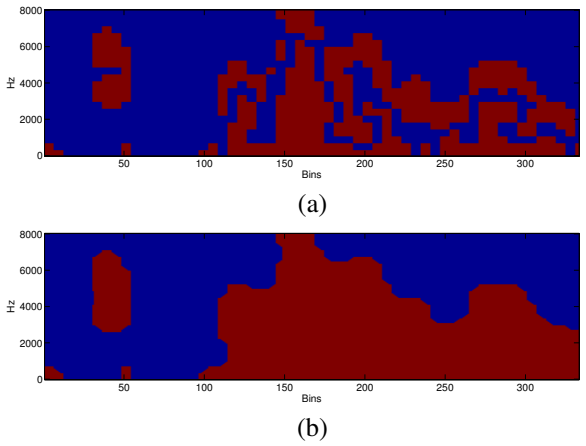
Fig. 3. Spectral detail of the utterance in Figure 1 (c). Left: Extraction of frequency bands with a hopsize of 8 patches. Right: Extraction of 12×12 patches with a hopsize of 6 bins.



(a)

(b)

Fig. 4. Number of iterations for each time-frequency bin, (a) computed from the noisy signal, (b) computed from the enhanced signal.



(a)

(b)

Fig. 5. (a) Binary mask after the threshold operation (2). (b) Smoothed mask after the closing operation.

(i.e. one patch), $\mathbf{x}_0$ denotes the sample for which the pre-image is computed, M is the number of samples, $k(\cdot,\cdot)$ denotes the kernel function and λ is the regularization parameter. We use a Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/c)$, where c is its variance.

For the estimation of the pre-image, equation (1) is iterated until convergence of $\mathbf{z}_{t+1}$. From the number of iterations until convergence, useful information about the properties of the underlying signal can be derived. This is illustrated in Figure 4, where the number of iterations for each patch is visualized according to its position in the time-frequency plane.

Figure 4 (a) shows the number of iterations (encoded as color) computed from the noisy signal as in [8], when maximally 6 iterations are executed. Non-integer values result from averaging between overlapping patches. A comparison to the noisy signal and the clean signal in Figure 1 (b) and (c), respectively, shows that the regions corresponding to speech need fewer iterations until convergence than the regions corresponding to noise. The different convergence behaviour is caused by the similarity measure of the kernel, which returns different values for speech and noise regions. We exploit this observation by setting a threshold to compute a binary mask for suppressing musical noise in non-speech regions.

In this paper, we compute the number of iterations from a signal enhanced by the generalized subspace method [1] (see Figure 4 (b)). Again, we can discriminate between different regions, however the relation between the number of iterations and the content of the signal is not as clear as in the former case. Empirically, we observed that few iterations correspond mainly to speech regions, an intermediate number of iterations corresponds mostly to noise, and more iterations again correspond to speech regions.

2.3. Computation of the binary mask and filtering

For the discrimination between speech and non-speech we set two thresholds, such that the iteration map is segmented as shown in Figure 5 (a): Region 1, in red, covers areas mainly corresponding to speech, while region 2, in blue, covers speech and noise areas. The operation for obtaining the mask m for each bin is

$$m = \begin{cases} 1 & \text{if } n < a \text{ or } n > b \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $m = 1$ if there is speech, n is the number of iterations for a specific bin in the map and a and b are the two threshold values. The values for the thresholds are derived from experiments (see Section 3).

To distinguish between noise and speech, the parts of the blue region within speech areas have to be removed. This is realized with techniques from image processing, namely morphological filtering such as dilation and erosion. The consecutive execution of these operations results in the so-called closing operation [13] that closes the holes in Figure 5 (a). As structural element a disk of radius 10 is used. Figure 5 (b)

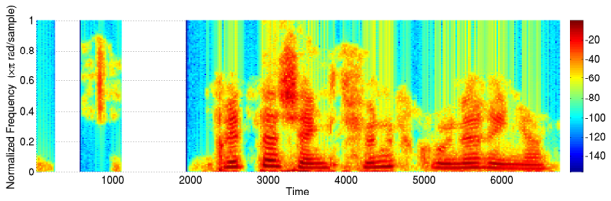


Fig. 6. Speech utterance from Figure 1 after musical noise suppression with the mask from Figure 5 (b). White areas mark regions where no energy is left - this can be avoided by leaving a noise floor instead of setting the spectrogram bins to zero.

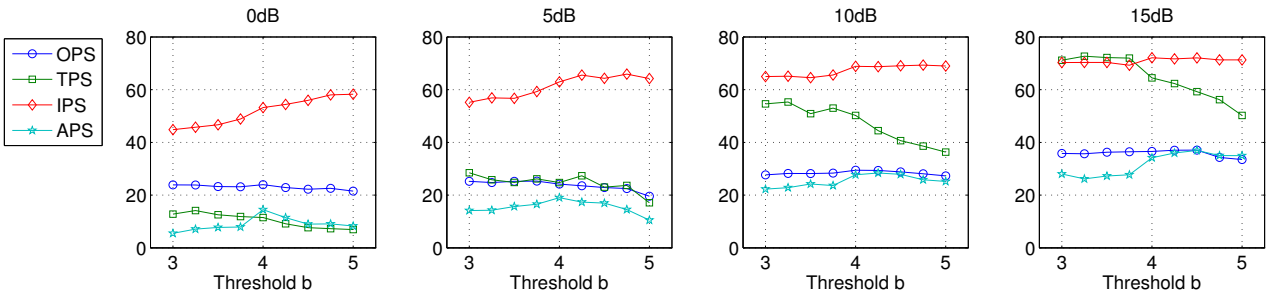


Fig. 7. Overall perceptual score (OPS), target perceptual score (TPS), interference perceptual score (IPS), and artifact perceptual score (APS) computed from the development set for different values of the upper threshold b in different SNR conditions. For the final experiments the threshold maximizing the APS was chosen.

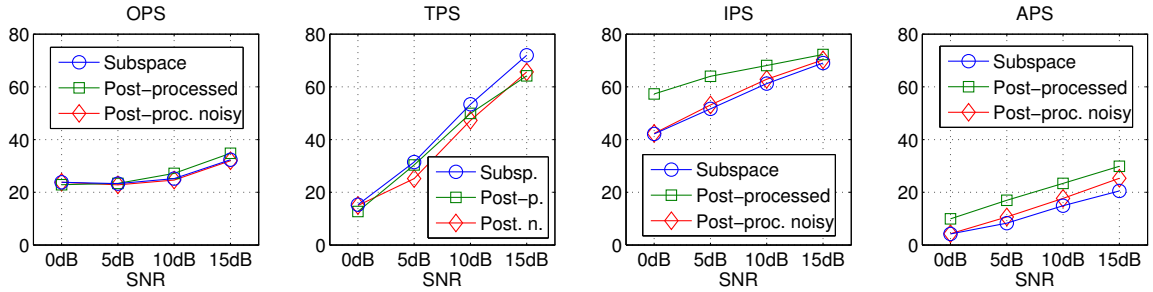


Fig. 8. Evaluation of overall perceptual score (OPS), target perceptual score (TPS), interference perceptual score (IPS), and artifact perceptual score (APS) for the generalized subspace method (Subspace), the proposed post-processing method for musical noise suppression (Post-processed), and the post-processing method from [7] (Post-proc. noisy).

shows the resulting contiguous mask, which is subsequently applied to filter the magnitude of the STFT of the signal. Figure 6 represents the spectrogram after musical noise suppression.

2.4. Resynthesis

After filtering, the inverse Fourier transformation is applied using the phase from the input signal. Finally, the audio signal is synthesized using the weighted overlap-add method [14].

3. EXPERIMENTS

For the evaluation of the approach, we performed experiments on a database with recordings of 6 speakers, 3 male and 3 female. Each speaker read a list of 20 sentences, which makes 120 sentences in total. The recording was performed with a close-talking microphone and 16 kHz sampling frequency. We performed experiments with additive white Gaussian noise[1] at 0, 5, 10, and 15 dB SNR, where the kernel variance for the pre-image iterations was was set to 3, 2, 0.5, and 0.25, respectively, and the regularization parameter λ was set to 0.5. The database was split into a development and a test set, where the development set contains one sentence of each speaker and the test set contains the remaining 114 sentences.

In [9], new quality measures for signals estimated by audio source separation algorithms were proposed. They were designed using the outcome of subjective listening tests. These measures show an improved correlation with subjective scores compared to formerly used measures. The measures evaluate four aspects of the signal: the global quality (OPS - overall perceptual score), the preservation of the target signal (TPS - target perceptual score), the suppression of other signals (IPS - interference perceptual score) and the absence of additional artificial noise (APS - artifact perceptual score). The scores range from 0 to 100, high values denote better quality. We use these measures, because they allow for evaluation of the amount of musical noise by looking at the APS.

To achieve good musical noise suppression, an accurate estimation of the speech regions is needed. The two thresholds in (2) are set as follows: The lower threshold a is fixed to 1.5, as there are few iteration counts in this range and only the interior of the speech region is affected that is properly treated by the closing operation anyway. For the upper threshold b, several values were tested on the development set. The one providing the best tradeoff between OPS and APS was taken, ensuring good quality as well as good musical noise suppres-

[1]For other noise types the generalized subspace method produced less musical noise, therefore post-processing is not necessary.

sion. Figure 7 shows the results on the development set in four noise conditions when the threshold is varied from 3 to 5. The final values 4, 4, 4.25, and 4.5 were chosen for the noise conditions of 0, 5, 10, and 15 dB SNR, respectively. For these values, the APS is maximized and good artifact suppression, i.e. musical noise suppression, is achieved, while the overall quality is still in the upper range (or maximized as well).

Figure 8 shows the results for the test set with the optimal threshold settings. The recordings with suppressed musical noise achieve a better overall quality (OPS) than the original enhanced recordings and than the post-processing described in [7] for all SNR levels except of 0 dB. They also score better in terms of absence of artifacts (APS), which confirms that the musical noise is efficiently suppressed. Furthermore, our approach also achieves better interference suppression (IPS), however in terms of target preservation (TPS) it is slightly weaker than to the original subspace method. This can be explained by the fact, that speech components may be attenuated by the application of the mask. Listening to the processed utterances and inspection of the spectrograms (see Figure 6) confirm that there is less musical noise after post-processing while almost all speech components are preserved.[2] Only in the case of fricatives speech is attenuated due to the low energy - this is reflected by the score for target preservation.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method for musical noise suppression that relies on the knowledge gained from pre-image iterations executed on spectral data. Properties of the underlying signal can be inferred from the number of iterations until convergence and the spectral data can be segmented into speech and non-speech regions. We smooth the segmentation by applying image processing techniques and use the resulting binary mask to suppress musical noise in non-speech regions. This method can be applied as post-processing for any speech enhancement algorithm.

We applied the method on speech recordings corrupted by additive white Gaussian noise at different SNRs which have been enhanced by the generalized subspace method. For evaluation, we used the objective quality measures of the PEASS toolbox, which allow for an evaluation regarding four aspects: overall quality, preservation of the target signal, suppression of interfering signals, and absence of additional artificial noise. In terms of overall quality we achieve an improvement in almost all conditions. The score measuring the absence of additional noise increases. This confirms the reduction of musical noise. These results are consistent with the subjectively perceived quality that increases due to the attenuation of disturbing musical noise.

[2] Audio examples are provided on `http://www2.spsc.tugraz.at/people/chrisl/audio/eusipco2012`.

5. REFERENCES

[1] Y. Hu and P. 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] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113 – 120, 1979.

[3] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP*, 1979, pp. 208–211.

[4] Z. Goh, K.-C. 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, 1998.

[5] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *ICASSP*, 2002, vol. 1, pp. 537–540.

[6] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement systems," in *ICASSP 2009*, 2009, pp. 4409 –4412.

[7] C. Leitner and F. Pernkopf, "Speech enhancement using pre-image iterations," *ICASSP*, in Press, 2012.

[8] C. Leitner and F. Pernkopf, "Musical noise suppression for speech enhancement using pre-image iterations," *IWSSIP*, 2012.

[9] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046 –2057, 2011.

[10] S. Mika, B. Schölkopf, A. Smola, K.-R. 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.

[11] Tl. J. Abrahamsen and L. K. Hansen, "Input space regularization stabilizes pre-images for kernel PCA denoising," in *MLSP*, 2009.

[12] C. Leitner and F. Pernkopf, "The pre-image problem and kernel PCA for speech enhancement," in *Advances in Nonlinear Speech Processing*, vol. 7015 of *Lecture Notes in Computer Science*, pp. 199–206. 2011.

[13] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Pearson Prentice Hall, 2008.

[14] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236 – 243, 1984.