GENERALIZATION OF PRE-IMAGE ITERATIONS FOR SPEECH ENHANCEMENT

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 extend the pre-image iteration method for speech de-noising by automatic determination of the kernel variance. The kernel variance needs to be adapted in different noise conditions. In previous work, the signal-to-noise ratio (SNR) was assumed to be known and the kernel variance was pre-defined using a development set. In the proposed method, a function is derived that maps a noise estimate to a potentially good value for the kernel variance. Hence, the SNR is not required to be known. Furthermore, the method is adapted for scenarios with colored noise, where – due to the properties of the noise – a different kernel variance for each frequency leads to better performance. We compare the proposed methods to the original pre-image iteration method and show an increase in performance in terms of the PEASS quality measures.

Index Terms— Speech enhancement, de-noising, pre-image iterations, kernel PCA

1. INTRODUCTION

Speech enhancement is important for many applications such as speech communications or speech recognition. Algorithms for speech enhancement can be divided into three main classes: Spectral-subtractive algorithms, statistical model-based algorithms and subspace algorithms [1, 2, 3, 4, 5, 6].

Subspace methods – which form the basis for this work – make use of principal component analysis (PCA) to enhance speech, i.e., PCA is applied on the magnitude spectrum and the phase of the noisy signal is used for the final transformation from frequency to time domain. The usage of the noisy phase generally works well but can affect the speech quality at low signal-to-noise ratios.

Recently, we proposed to use kernel PCA, a non-linear extension to PCA, for speech enhancement [7]. Kernel PCA has been successfully applied to image de-noising [8, 9]. In speech processing, kernel PCA has been used to extract robust features from reverberant speech [10]. In contrast to other speech enhancement methods we perform enhancement on complex spectral data. Thus we do not explicitly rely on the phase of the noisy signal. In [11], we derived a simplification of kernel PCA, called *pre-image iterations*, which improves the performance and is computationally less demanding.

The pre-image iteration algorithm performs enhancement in the spectral domain. It is based on the computation of a linear combination of noisy feature vectors extracted from the sequence of complex-valued short-term Fourier transforms (STFTs). The weights for the linear combination are computed by using a Gaussian kernel, which measures the similarity between two feature vectors. The de-noising ability strongly depends on the variance of the kernel. In previous work, the kernel variance is set for each signal-to-noise ratio (SNR) separately. The SNR is assumed to be known and the choice of the kernel variance is based on the performance on a development data set. Empirically, however, we found that a good value for the kernel variance is rather connected to the noise power and not to the SNR. The SNR rather influences the degradation of the speech signal, i.e., when speech components are masked by noise. Furthermore, for colored noise, it is problematic to use a single value for the kernel variance as the noise power varies over the frequency range.

In this paper, we automatically set the kernel variance by using a mapping function derived from the development data. This way, the kernel variance can be determined without knowing the SNR. The mapping function maps the noise power to a suitable value of the kernel variance. The values for the kernel variance used for the estimation of the mapping function are determined by a combination of objective quality measures. Using the noise power instead of the SNR favors a good noise reduction. In the case of colored noise we determine the kernel variance for each frequency separately to account for the non-uniform noise distribution over the frequency range. For each frequency, the equivalent noise power is estimated to determine the kernel variance from a mapping function learned from white noise.

The algorithm is tested on two databases at 0, 5, 10, and 15 dB SNR with two types of noise, i.e., on the *air-bone* database with additive white Gaussian noise (AWGN) and the *Noizeus* database with car noise. Compared to previous work where the variance was pre-determined for each SNR we achieve a better or similar performance. As benchmark we present performance results for spectral subtraction and the generalized subspace method. The performance achieved with pre-image iterations is superior, in addition the method does not create musical noise artifacts.

This paper is organized as follows: Section 2 describes pre-image iterations and the determination of a suitable kernel variance. Section 3 presents the experiments and the results. Section 4 concludes the paper.

2. PRE-IMAGE ITERATIONS FOR SPEECH ENHANCEMENT

In [11], we showed that *pre-image iterations* can be used for speech enhancement. Pre-image iterations are derived from kernel PCA, where data samples are transformed to a so-called feature space for processing. Depending on the kernel there may be no one-to-one mapping between feature space and input space and the sample in input space corresponding to a processed sample in feature space cannot be directly determined. Therefore, the sample has to be estimated and the estimate is called *pre-image*. Several methods have

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

been proposed to solve the pre-image problem (see [12]).

Pre-image iterations are based on the simplification of the iterative pre-image method of [8]. In [11], we neglected the kernel PCA coefficients and de-noising is performed by iteratively applying

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

where $\mathbf{z}_j^t$ is the enhanced sample in input space, t denotes the iteration step, $\mathbf{x}_i$ is the i^{th} original noisy sample, M is the number of noisy samples in one frequency band (see Section 2.2 for further details), and $k(\cdot,\cdot)$ defines the kernel function. The feature vectors $\mathbf{x}_i$ are extracted from the complex frequency domain representation (see Section 2.2). For enhancement of one specific sample $\mathbf{x}_j$, $\mathbf{z}_j^0$ is initialized by $\mathbf{x}_j$ which results in a robust convergence behavior. When the difference between $\mathbf{z}_j^{t+1}$ and $\mathbf{z}_j^t$ is below a given threshold, the iterations are terminated. Pre-image iterations are equivalent to forming convex combinations of noisy speech samples.

In [13], a regularization for pre-image estimation was proposed and the corresponding pre-image iteration equation is

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

where λ is the regularization parameter and $\mathbf{x}_j$ denotes the noisy sample which is enhanced. We use the Gaussian kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/c), \tag{3}$$

where parameter c denotes the variance of the kernel. This kernel determines the similarity between two data samples where the variance c is used to scale the degree to which the samples are treated as similar. The de-noising process is based on the fact that noise is random and that the feature vectors for noise are all relatively similar to each other. Consequently, the weights for the linear combination estimated by the kernel function are similar and the noise is averaged out (in the complex spectral domain). Speech components are rather dissimilar so they are maintained as long as the SNR is not too low. From the above equations it can be seen that the value of c considerably influences the de-noising performance.

2.1. Determination of the kernel variance

Two approaches are used for the automatic determination of the kernel variance, one for AWGN and one for colored noise. In the case of white noise, a function that maps the noise power to a suitable value of c is learned from the development set.

To find the mapping function, pre-image iterations are applied to each sentence in the development set with different values of c and the enhanced recordings are evaluated using the measures of the PEASS toolbox [14]. The PEASS toolbox returns four scores: one for the global quality (OPS - overall perceptual score), one for the preservation of the target signal (TPS - target perceptual score), one for the suppression of other signal (IPS - interference perceptual score), and one for the absence of additional artificial noise (APS - artifact perceptual score). The scores range from 0 to 100, larger values denote better performance. Ideally, all measures should be maximized, however, this is not possible, as, e.g, a good suppression of the interference (noise) conflicts with a good preservation of the target signal.

As optimization criterion S for finding the best setting of c, a linear combination of the four scores is used

$$S = 0.5 \cdot (\mathrm{OPS} + \frac{1}{3}(\mathrm{TPS} + \mathrm{IPS} + \mathrm{APS})). \tag{4}$$

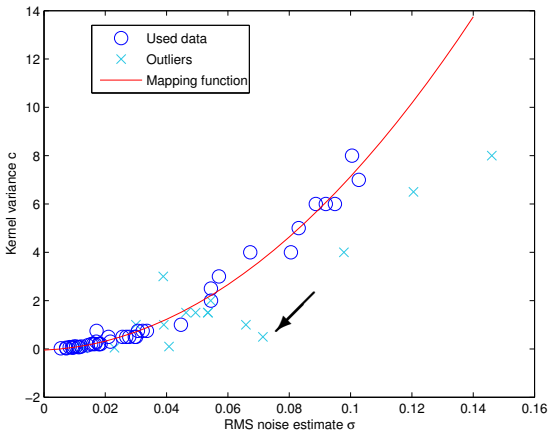

Fig. 1. Mapping function for the *air-bone* data derived by polynomial curve fitting with outliers removed from the data.

Additionally, the IPS score has to be greater than 10 to avoid the situation where S is large due to good TPS and APS scores but no de-noising is achieved.

For each speech utterance in the development set, the average noise power σ^2 is estimated from the beginning of the recording, assuming stationary noise and no speech within this region, i.e.,

$$\sigma^2 = \frac{1}{N}\sum_{n=0}^{N-1} y[n]^2, \tag{5}$$

where $y[n]$ is the noisy signal in time domain and N is the length of the frame used for computing the noise power. A frame length of $N = 2048$ samples was chosen experimentally. Figure 1 shows the values for c that are determined by the criterion S plotted as a function of the estimated noise root mean square (RMS) value σ. To fit the data a polynomial of degree two is used.

To improve the fit in Figure 1, outliers are removed. The data points marked by a cross are labeled as outliers since the values of c are not in the appropriate range for the noise estimate. For instance, for the data point marked with the arrow, the SNR is 0 dB and the predicted value of c is 0.5, which is not reliable as previous experiments have shown values around 4 to be a good setting for c at 0 dB and 0.5 rather suitable for 10 dB SNR.

For colored noise – car noise in our experiments – a single value for c for all frequencies is not suitable as the noise power is not equally distributed over the frequency range. To solve this problem, we first use a development set with utterances corrupted by white noise to derive a mapping function. For the utterances corrupted by colored noise we estimate the equivalent noise power at each frequency bin and use it to find the frequency-dependent kernel variance c_k.

The estimate of the noise power for each frequency bin is based on *Parseval's theorem* [15] that states that the mean of the squared magnitude values of the discrete Fourier transform of a signal $Y[k]$ is equal to the sum of the squared samples in time domain $y[n]$, i.e.,

$$\sum_{n=0}^{K-1} |y[n]|^2 = \frac{1}{K}\sum_{k=0}^{K-1} |Y[k]|^2 \tag{6}$$

where a frame of length K leads to a K-point Fourier transform. White noise is equally distributed over all frequencies and ideally

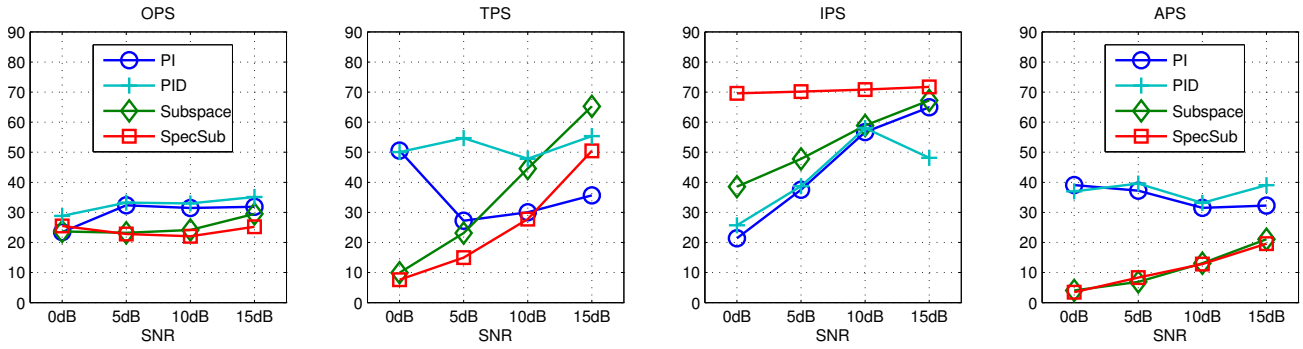

Fig. 2. Results of pre-image iterations (PI), pre-image iterations with automatic determination of the kernel variance (PID), the generalized subspace method (Subspace) and spectral subtraction (SpecSub) in terms of overall perceptual score (OPS), target perceptual score(TPS), interference perceptual score (IPS), and artifact perceptual score (APS) on the test set of the *airbone* database corrupted by additive white Gaussian noise (AWGN).

all Fourier coefficients are equally large. Therefore, in the ideal case, the energy of the time domain signal can be estimated from one Fourier coefficient, i.e.,

$$\sigma^2 = \frac{1}{N} \cdot \frac{1}{K} \sum_{k=0}^{K-1} |Y[k]|^2 = \frac{1}{N} |Y[k]|^2 \quad \forall k. \tag{7}$$

Relying on this relation, we take the Fourier coefficients from colored noise and derive the equivalent noise power σ_k^2 in time domain for each Fourier coefficient $Y[k]$.

In particular, 256-point STFTs are computed from 128-sample frames by application of zero-padding. The squared magnitude bins $|Y[k]|^2$ are averaged over the first 15 frames to get a more reliable estimate. Dividing the average by N gives the equivalent noise power σ_k^2 for the k^{th} frequency bin, that is subsequently used to derived a suitable value for c_k from the mapping function.

During processing, frequency bins are grouped to frequency bands as explained in Section 2.2. For the frequency bins within one band the values for c_k are averaged and this average is used for pre-image iterations within the band. This way the approach adapts to all kinds of stationary noise.

2.2. Feature extraction

The sample vectors $\mathbf{x}_i$ for pre-image iterations are extracted from the sequence of short-term Fourier transforms computed from the speech signal. First the 256-point STFT is computed from frames of 16 ms. The frames have an overlap of 50% and a Hamming window is applied. The resulting time-frequency representation is split on the time and on the frequency axis to reduce computational costs (see Figure 3, left side) which results in so-called *frequency bands*. Sample vectors are retrieved from these frequency bands by first extracting quadratic patches in an overlapping manner, where the size of each patch is 12×12 with an overlap of 11 (see Figure 3, right side). In previous experiments, windowing of the patches was beneficial, so a 2D Hamming window is applied. Then the patches are re-ordered in column-major order to form the sample vectors $\mathbf{x}_i$. The frequency bands cover a frequency range corresponding to 8 patches (i.e. 19 bins) and a time range corresponding to 20 patches (i.e. 31 bins). Bands are not overlapping along the frequency axis, along the time axis the overlap is 10 patches. This configuration was chosen due to good empirical results.

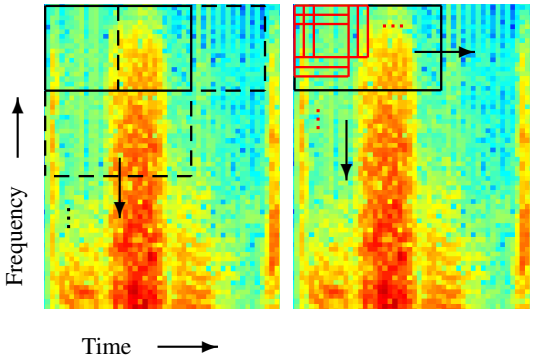

Fig. 3. Spectral detail of the clean utterance /t a sh e/. Left hand side: Extraction of frequency bands with time overlap of 10 patches. Right hand side: Extraction of 12×12 patches from one frequency band with an overlap of 10 in time and frequency.

After de-noising, the audio signal is resynthesized by reshaping the sample vectors $\mathbf{z}_j$ to patches. The patches of all frequency bands belonging to one time segment are rearranged using the overlap-add method with weighting as described in [16] generalized for the 2D domain. Then the STFT bins of overlapping time segments are averaged, the inverse Fourier transform is applied on the bins of each frame and the audio signal is synthesized with the weighted overlap-add method [16].

3. EXPERIMENTS

3.1. Data

The algorithm was tested on two databases: The *air-bone* database with utterances corrupted by AWGN and the *Noizeus* database with utterances corrupted by car noise. Both databases are tested at 0, 5, 10, and 15 dB SNR. The SNR was computed using the *active speech level* as described for the *Noizeus* database [17]. For estimation of the mapping function 20 dB were used in addition.

The *air-bone* database[1] consists of recordings of six German

[1]The recordings consist of two channels: one channel recorded by a standard microphone and a second recording by a bone-conductive microphone which is, however, not used in the described experiments.

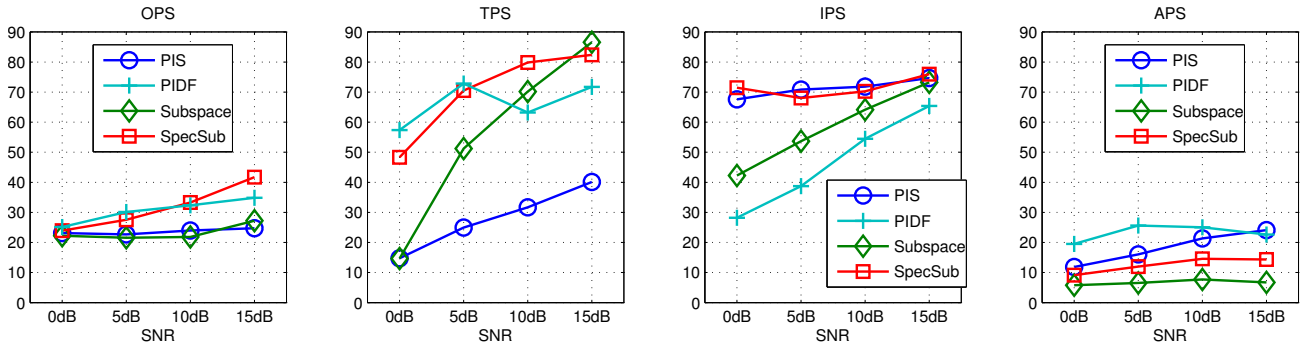

Fig. 4. Results of pre-image iterations with frequency-dependent scaling of the kernel variance (PIS), pre-image iterations with automatic frequency-dependent determination of the kernel variance (PIDF), the generalized subspace method (Subspace) and spectral subtraction (SpecSub) in terms of overall perceptual score (OPS), target perceptual score(TPS), interference perceptual score (IPS), and artifact perceptual score (APS) on the test set of the *Noizeus* database corrupted by car noise.

speaking individuals, three female and three male. Each read 20 sentences which results in a total of 120 sentences. The recording was done with a close talk microphone and 16 kHz sampling frequency. The development set contains two sentences per speaker, which results in twelve sentences per noise condition.

The *Noizeus* database [17] contains recordings of 30 IEEE sentences in English [18], spoken by three female and three male speakers (five sentences each). The utterances are sampled with 8 kHz and filtered by the modified intermediate reference system (MIRS) filters [19] to simulate telephone speech. For development, one sentence per speaker is used, resulting in six sentences per noise condition.

3.2. Results

The performance of the pre-image iteration approach with automatic determination of the kernel variance (PID) is evaluated using the PEASS toolbox and compared to the standard pre-image iteration approach (PI) as described in [11], to the generalized subspace method [6] and to spectral subtraction [2]. The regularization parameter λ in (2) is set to 0.25 for 0 dB and to 0.75 for the other tested SNRs in case of the *air-bone* database, in case of the *Noizeus* database no regularization is applied.

Figure 2 shows the results for the *air-bone* database. The overall performance improves slightly for all tested noise conditions. The preservation of the target speaker improves for all conditions except for 0 dB, the other scores are in a similar range.[2]

Figure 4 shows the performance of pre-image iterations with automatic frequency-dependent determination of the kernel variance (PIDF) on the *Noizeus* database with car noise. As reference, the PIS method is provided – an adaptation of PI where the variance is scaled logarithmically over the frequency range [20]. In addition, results for the generalized subspace method and for spectral subtraction are shown. The overall performance score for PIDF is larger than for PIS. The PIDF method results in better target preservation (higher TPS), however, the noise attenuation is weaker (lower IPS). In terms of APS the PIDF method is better indicating fewer artifacts.

In addition to objective evaluation, the enhanced utterances were evaluated by listening.[3] The variants of pre-image iterations have in common that they do not produce musical noise, but some residual noise is present around speech components. For the *Noizeus* database, noise is left in some frequency bands. This can be explained by a suboptimal estimation of the mapping curve. Listening to the utterances corrupted by AWGN used for development and inspection of the evaluation measures revealed that the values of c_k chosen by the score S do not lead to an optimal noise attenuation. Furthermore the Noizeus utterances corrupted by AWGN are filtered by the MIRS filter, hence the assumption that the noise is uniformly distributed over the frequencies is violated. Consequently, the noise power in the time domain suggests a lower noise level which leads to an underestimation of c_k for colored noise. These issues are subject to future work.

4. CONCLUSION

In this paper, we presented a generalization to the pre-image iteration method for speech enhancement. When applying pre-image iterations, the de-noising performance crucially depends on the setting of the kernel variance. In previous work, the kernel variance c was pre-determined depending on the SNR, which was assumed to be known. In this paper, the pre-image iteration method is extended by the automatic determination of c. In particular, we determine a mapping function from a development set. This enables to derive the kernel variance from a noise estimate of the utterance. Furthermore, we showed that we can use the mapping function to generalize to conditions with colored noise, where different kernel variances over the frequency range are necessary due to the non-uniform distribution of the noise power.

We tested the proposed method on the *air-bone* and the *Noizeus* database, which were corrupted by AWGN and car noise, respectively. The performance was evaluated by using the objective quality measures of the PEASS toolbox. The results of the developed methods are superior to the results achieved with the pre-image iteration method [11, 20]. Compared to the generalized subspace method and to spectral subtraction, the scores are in a similar range while the score measuring the artifacts is better for the developed methods, which do not suffer from musical noise.

In future, we aim to investigate further optimization criteria in combination with other evaluation measures for the derivation of the mapping function. Furthermore we plan experiments with other noise types, especially slowly changing noise.

[2]The outlier of the IPS for 15 dB can be neglected as the noise is low in this condition anyway.

[3]Audio examples are provided on `http://www2.spsc.tugraz.at/people/chrisl/audio/icassp2013`

5. REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*, CRC, 2007.

[2] M. Berouti, M. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 208–211, 1979.

[3] Robert J. McAulay and Marilyn L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 137–145, 1980.

[4] Yariv Ephraim and David 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, 1984.

[5] Yariv Ephraim and Harry L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251 –266, 1995.

[6] 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.

[7] 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.

[8] 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.

[9] Kwang In Kim, Matthias O. Franz, and Bernhard Schölkopf, "Iterative kernel principal component analysis for image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1351–1366, 2005.

[10] Tetsuya Takiguchi and Yasuo Ariki, "PCA-based speech enhancement for distorted speech recognition.," *Journal of Multimedia*, vol. 2, no. 5, pp. 13–18, 2007.

[11] Christina Leitner and Franz Pernkopf, "Speech enhancement using pre-image iterations," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4665–4668, 2012.

[12] Christina Leitner and Franz 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] 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.

[14] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker 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.

[15] Thomas F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall PTR, 2002.

[16] D. Griffin and Jae 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.

[17] Yi Hu and Philipos C. Loizou, "Subjective evaluation and comparison of speech enhancement algorithms," *Speech Communication*, vol. 49, pp. 588–601, 2007.

[18] IEEE Subcommitee, "IEEE recommended practice for speech quality measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225 – 246, 1969.

[19] ITU-T, "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 P.862*, 2000.

[20] Christina Leitner and Franz Pernkopf, "On kernel PCA, pre-image methods and pre-image iterations for speech enhancement," Tech. Rep., Graz University of Technology, 2012.