The Pre-Image Problem and Kernel PCA for Speech Enhancement

Christina Leitner[⋆] and Franz Pernkopf

Signal Processing and Speech Communication Laboratory,
Graz University of Technology,
Inffeldgasse 16c, 8010 Graz, Austria
{christina.leitner,pernkopf}@tugraz.at
http://www.spsc.tugraz.at

Abstract. In this paper, we use kernel principal component analysis (kPCA) for speech enhancement. To synthesize the de-noised audio signal we rely on an iterative pre-image method. In order to gain better understanding about the pre-image step we performed experiments with different pre-image methods, first on synthetic data and then on audio data. The results of these experiments led to a reduction of artifacts in the original speech enhancement method, tested on speech corrupted by additive white Gaussian noise at several SNR levels. The evaluation with perceptually motivated quality measures confirms the improvement.

Keywords: Kernel PCA, speech enhancement, pre-image problem

1 Introduction

Subspace methods for speech enhancement are based on the assumption that a speech signal only lives in a subspace of a given signal space. Noise on the other hand is present in the entire space. Noise reduction is done by retrieving only the signal components that belong to the speech subspace and setting the other components to zero. Usually this is realized by applying principal component analysis (PCA) to an estimated covariance matrix of the speech signal [9].

PCA can be easily extended to a non-linear algorithm by using kernels. Kernel methods work in a high-dimensional feature space and can solve problems that are not solvable in the original (input) space. After processing data in the feature space we are, however, often not interested in the sample in feature space but in the corresponding sample in input space, the so-called pre-image. Due to the non-linear mapping the pre-image does not always exist, and if it exists it is not necessarily unique. Several methods were proposed to solve this "pre-image problem" [10, 7, 1, 4].

In this paper, we show how the non-linear extension of PCA, kernel PCA (kPCA), can be used for speech de-noising. Our algorithm works directly on the

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

complex valued coefficients of the short-term Fourier transform (STFT) instead of the magnitude spectrum as done in many speech enhancement methods. Our algorithm is robust to musical noise, however, we observe a buzz-like artifact.

The investigation of our method revealed that the used iterative pre-image method often converges to the same sample within one subregion of the time-frequency representation. This results in a regular pattern that can be perceived as a buzz-like artifact. A similar phenomenon was reported for PSOLA methods where the repetition of the same segment is described to create a "buzziness" of the signal [12]. In order to gain better understanding and to eliminate this artifact we tested different pre-image methods, first on synthetic and then on audio data.

The paper is organized as follows: Section 2 introduces kernel PCA for de-noising. Section 3 describes our experiments on synthetic and audio data and presents the results. Section 4 concludes the paper.

2 Kernel PCA

PCA is used in data compression and de-noising to extract relevant information from data. For de-noising, eigenvalue decomposition is applied to the covariance matrix of the data and the data samples are projected on the eigenvectors corresponding to the largest eigenvalues.

PCA can be easily extended to kernel PCA by substituting inner products by kernels [11]. By doing so, the data $\mathbf{x}$ is implicitly mapped to a high-dimensional feature space and all operations are done in this feature space. Due to the formulation in terms of kernels the mapping $\mathbf{\Phi}(\mathbf{x})$ never has to be computed explicitly. The projection of a sample in feature space can be formulated as follows:

$$\mathbf{P}_n\mathbf{\Phi}(\mathbf{x}) = \sum_{k=1}^{n} \beta_k \mathbf{V}^k, \tag{1}$$

where $\mathbf{P}_n$ denotes the projection operator and β_k is the coefficient of the projection onto the eigenvector $\mathbf{V}^k$, that can be reformulated as

$$\beta_k = (\mathbf{V}^k)^{\mathrm{T}}\mathbf{\Phi}(\mathbf{x}) = \sum_{i=1}^{M} \alpha_i^k k(\mathbf{x}, \mathbf{x}_i), \tag{2}$$

where α_i^k is the i^{th} entry of the eigenvector solving the eigenproblem $M\lambda\alpha = \mathbf{K}\alpha$ with $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. This eigenproblem is equivalent to the problem $\lambda\mathbf{V} = \bar{\mathbf{C}}\mathbf{V}$ with the covariance matrix in feature space $\bar{\mathbf{C}} = \frac{1}{M}\sum_{j=1}^{M} \mathbf{\Phi}(\mathbf{x}_j)\mathbf{\Phi}(\mathbf{x}_j)^{\mathrm{T}}$ [11].

We are, however, not interested in the de-noised sample in feature space but in the de-noised sample in input space. Due to the non-linear mapping, the existence of such a sample is not guaranteed, and if the sample exists it is not necessarily unique [10]. Hence, our goal is to find a sample $\mathbf{z}$ that satisfies $\mathbf{\Phi}(\mathbf{z}) = \mathbf{P}_n\mathbf{\Phi}(\mathbf{x})$. For non-invertible kernel functions like the Gaussian kernel

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

where c denotes the kernel's variance, $\mathbf{z}$ can be found by minimizing the distance between the feature map of $\mathbf{z}$ and the projection in feature space $\rho(\mathbf{z}) = \|\mathbf{\Phi}(\mathbf{z}) - \mathbf{P}_n\mathbf{\Phi}(\mathbf{x})\|^2$. Mika et al. [10] showed that this can be solved with the iterative update rule

$$\mathbf{z}_{t+1} = \frac{\sum_{i=1}^{M} \gamma_i k(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{M} \gamma_i k(\mathbf{z}_t, \mathbf{x}_i)}, \tag{4}$$

where $\gamma_i = \sum_{k=1}^{n} \beta_k \alpha_i^k$. Note that the resulting pre-image $\mathbf{z}$ is always a linear combination of the input data $\mathbf{x}_i$. The algorithm is sensitive to initialization, however this can be tackled by reinitializing with different values.

For the derivation of kPCA the data is assumed to be centered in feature space, i.e., $\sum_{i=1}^{M} \mathbf{\Phi}(\mathbf{x}_i) = 0$. With real data this assumption does not hold and centering has to be done explicitly. Centering can easily be done by modifying the kernel matrix $\mathbf{K}$ to get the centered kernel matrix

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_M\mathbf{K} - \mathbf{K}\mathbf{1}_M + \mathbf{1}_M\mathbf{K}\mathbf{1}_M, \tag{5}$$

where $\mathbf{1}_M$ is a matrix with all entries equal to $1/M$ (for more details, see [11]). The remaining PCA steps can then be conducted on the centered kernel matrix.

3 Experiments

In order to investigate the effects of centering and pre-image reconstruction we tested several implementations with centered and uncentered data and different pre-image methods. The compared pre-image methods were the following:

1. Iterative method by Mika et al. as denoted in Eq. (4).
2. Iterative method with additional normalization of the weighting coefficient $\tilde{\gamma}_i = \gamma_i + 1/N(1 - \sum_{j=1}^{M} \gamma_k)$ proposed by Kwok and Tsang [7].
3. Iterative method with regularization by Abrahamsen and Hansen [1]

$$\mathbf{z}_{t+1} = \frac{\frac{2}{c}\sum_{i=1}^{M} \tilde{\gamma}_i k(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i + \lambda\mathbf{x}_0}{\frac{2}{c}\sum_{i=1}^{M} \tilde{\gamma}_i k(\mathbf{z}_t, \mathbf{x}_i) + \lambda} \tag{6}$$

   where λ is a non-negative regularization parameter and $\mathbf{x}_0$ the noisy sample.
4. Non-iterative method described by Honeine and Richard [4]

$$\mathbf{z} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{X} - \lambda\mathbf{K}^{-1})\gamma \tag{7}$$

   where $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_n]$ and $\gamma = [\gamma_1\ \gamma_2\ \ldots\ \gamma_n]^{\mathrm{T}}$ with γ_i from above.

Experiments were first done on synthetic toy data and then on the audio data used for speech enhancement.

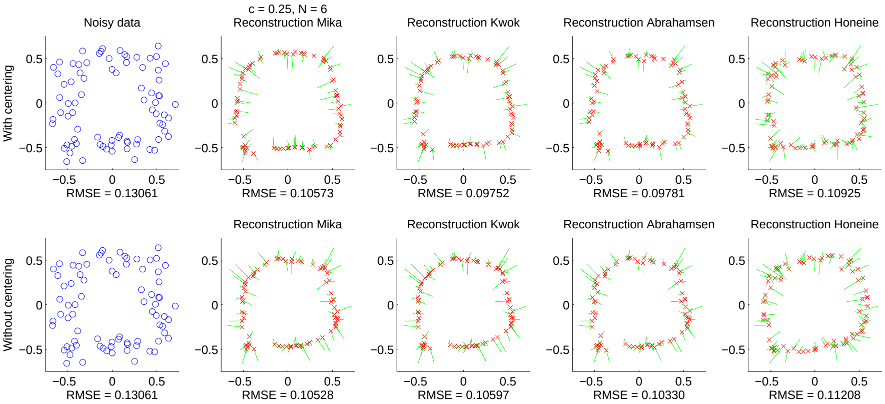

Fig. 1. De-noising with different pre-image methods on centered and uncentered data. The green lines illustrate the distance between de-noised (red) and noisy samples.

3.1 Comparison of Pre-Image Methods on Synthetic Data

Four synthetic datasets were generated: the “square” dataset, the “sine” dataset, the “spiral” dataset, and another dataset with complex valued data. The square dataset consists of samples along a 1×1 square, where the samples for each edge are drawn from a uniform distribution and corrupted by additive white Gaussian noise of variance 0.0025. The sine dataset is defined by samples with the coordinates $(x, \sin(2\pi x))$ with x uniformly distributed on the interval $[0, 4\pi]$ plus additive white Gaussian noise of variance 0.0025. The spiral dataset is given by samples with the coordinates $(At\cos(t), At\sin(t))$ where $A = 0.1$ and t is uniformly distributed on the interval $[0, 4\pi]$. White Gaussian noise of variance 0.005 is added. As for speech enhancement, the same data was used for training and testing, i.e., for eigenvalue decomposition and projection.

Table 1. RMSE for different pre-image methods on synthetic data. N denotes the number of components used for projection and c is the variance of the Gaussian kernel.

				Centered				Uncentered			
Dataset	N	c	Noisy	Mika	Kwok	Abr.	Hon.	Mika	Kwok	Abr.	Hon.
Square	7	0.25	0.142	0.123	0.115	0.117	0.127	0.116	0.116	0.116	0.126
Sine	15	0.1	0.071	0.063	0.060	0.060	0.084	0.060	0.060	0.060	0.060
Spiral	20	0.1	0.100	0.091	0.089	0.087	0.111	0.091	0.089	0.087	0.111
Complex	7	0.25	0.142	0.123	0.115	0.117	0.126	0.115	0.115	0.115	0.126

For comparison, the root mean squared error (RMSE) between reconstructed samples and noise-free reference samples was computed. For each dataset the RMSE was averaged over 100 realizations. Tab. 1 shows selected results for the

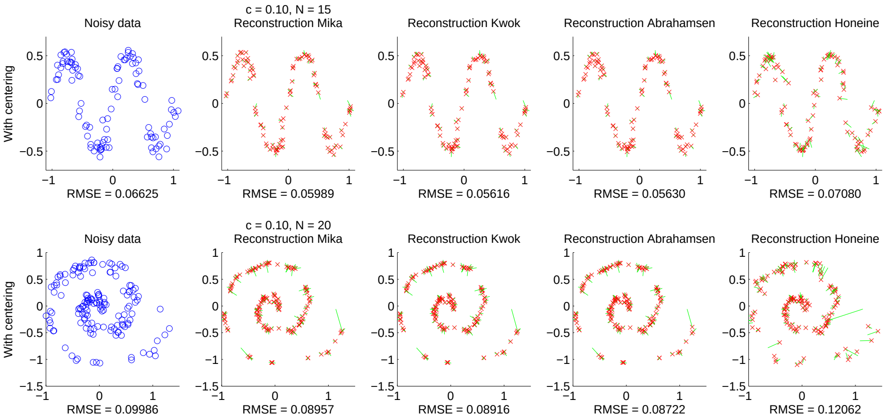

Fig. 2. De-noising of the sine and the spiral dataset.

four datasets with and without centering. Fig. 1 illustrates the de-noising and projection onto 6 principal components for one realization of the square dataset. Fig. 2 shows de-noising for the sine and the spiral dataset with projection on 15 and 20 components, respectively (plots for uncentered data are omitted due to their similarity).

As our speech enhancement algorithm works on complex valued data we did one further experiment on complex valued data. We generated a dataset similar to the square dataset where the real part of a complex number corresponds to the first coordinate and the imaginary part to the second. The results are shown in Tab. 1 and illustrated in Fig. 3. The example demonstrates that the pre-image methods can be applied to complex valued data as well.

From these experiments, it can be concluded that the methods of Kwok and Tsang [7] and Abrahamsen and Hansen [1] yield the best results. In contrast to Mika et al. [10] they perform normalization which seems to be necessary to achieve good reconstruction of the pre-image. The method described by Honeine and Richard does not perform as good. It has to be noted that our experiment is different from [4], because we use the same data for training and testing while they use different datasets. We encountered no problems of stability of the iterative algorithms, as we always use the noisy sample for initialization which is very robust.

Further, the question if centering should be applied or not cannot be clearly answered from the toy examples. Depending on the number of components either the implementation with or without centering performs better.

3.2 Speech Enhancement

As described in [8], our speech enhancement algorithm works on the complex STFT coefficients. To extract feature vectors for the kernel PCA we apply the following procedure: First, the STFT is calculated from frames of 256 samples with

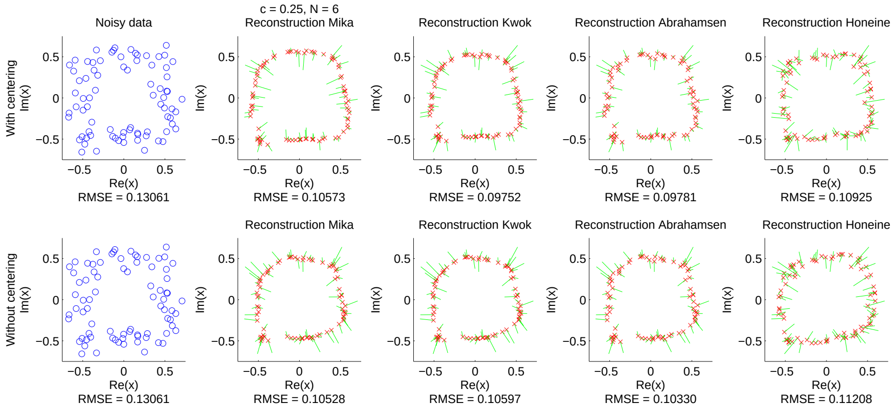

Fig. 3. De-noising of the complex square dataset.

50% overlap. The resulting time-frequency representation is cut into segments of 0.25 seconds to avoid large kernel matrices, as they increase computation times. To compensate for different energy levels in the frequency range, each time segment is processed in overlapping frequency bands. On each frequency band kPCA is applied independently. Sample vectors are retrieved by splitting the frequency bands into overlapping patches of size 12 × 12 with overlap 1. In previous experiments we achieved better de-noising when the patches are windowed. Hence, a 2D Hamming window is applied and the patches are rearranged as vectors to obtain samples for kPCA, i.e., each patch is one sample. The noisy data is projected on the eigenvector corresponding to the largest eigenvalue and the de-noised sample in input space is computed.

For resynthesis, patches at the same time-frequency position but of different frequency bands are averaged. Then the patches are summed up in an overlapping manner. To compensate for windowing they are weighted with the standard method from [3] modified for the 2D domain. The time segments are merged and the inverse Fourier transform is applied at each time instant. Finally, the signal is synthesized using the weighted overlap-add method of [3].

The pre-image methods were tested and evaluated by listening and by visual inspection of the spectrogram. Following observations were made:

1. The pre-image method of Mika et al. often fails to converge and the audio signal is mostly zero.
2. The pre-image method of Kwok and Tsang is stable. The audio signal is de-noised, however a buzz-like artifact occurs.
3. With the pre-image method of Abrahamsen and Hansen weak or good de-noising is achieved, depending on the value of λ. The same artifact as for the pre-image method of Kwok and Tsang appears.
4. The pre-image method of Honeine and Richard returns no meaningful audio signal. If the regularization parameter λ is set to zero the audio signal con-

tains similar artifacts as the pre-image method of Kwok and Tsang, but the speech signal is heavily attenuated.

The outcome of these experiments led to the idea to combine the methods from Kwok and Tsang and Honeine and Richard to reduce the buzz-like artifact. Indeed, a subtraction of the signal of Honeine and Richard's method from the signal of Kwok and Tsang's method in time domain results in a signal of better quality as the buzz-like artifact is significantly reduced. Furthermore, we tested this approach and the approach using Kwok and Tsang's method on a database.

The database contains recordings of six speakers (three female, three male), each speaker uttered 20 sentences which makes 120 sentences in total. Recording was done with a close-talk microphone and 16 kHz sampling frequency. White Gaussian noise was added at 5, 10 and 15 dB SNR.

For evaluation we used speech quality measures that showed good correlation with the results of subjective listening tests [5]. These measures are: the perceptual evaluation of speech quality measure (PESQ), the log-likelihood ratio (LLR) and the frequency-weighted segmental SNR (fwsegSNR).

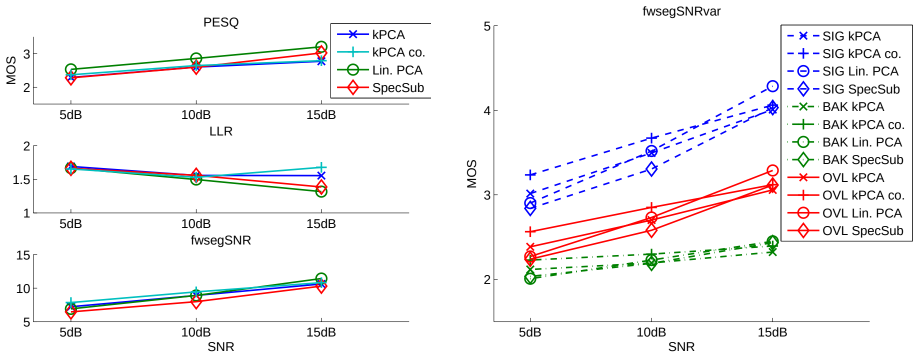

Fig. 4. Comparison of kPCA with linear PCA (lin. PCA) and spectral subtraction (SpecSub). For both kPCA implementations the values for c are 2, 0.5, and 0.25 for 5, 10, and 15 dB, respectively. The kPCA method with combined pre-imaging (kPCA co.) outperforms the original kPCA method in almost all conditions.

We compared our algorithms with the linear PCA method (lin. PCA) of Hu and Loizou [6] and with spectral subtraction (SpecSub) [2] as implemented in [9]. The results are shown in Fig. 4. In addition to the mentioned measures, a variant of the frequency-weighted segmental SNR (fwsegSNRvar) is given that returns three values: one for the signal quality only (SIG), one for the background intrusion (BAK), and one for the overall quality (OVL). It can be seen that our algorithms achieve a similar performance as linear PCA and spectral subtraction. Furthermore the combined approach that makes use of both pre-image methods, namely Kwok and Tsang's [7] and Honeine and Richard's [4], almost always scores better than the approach using Kwok and Tsang's method only.

4 Conclusion

In this paper, we compared different approaches to solve the pre-image problem in kPCA, using synthetic data and audio data for speech enhancement. For synthetic data the iterative methods behave similar, whereas the non-iterative method performs worse. When applied to audio data, the results are different: Only the iterative methods of Kwok and Tsang and Abrahamsen and Hansen result in a meaningful audio signal. The method of Honeine and Richard mainly models the artifact, however can be used in combination with the method of Kwok and Tsang to improve the audio quality.

We tested the approaches with Kwok and Tsang's method and the combined method on audio data corrupted by additive white Gaussian noise. The evaluation with objective quality measures shows that our algorithms achieve similar performance as linear PCA and spectral subtraction. While these methods are affected by musical noise, our first approach results in a buzz-like artifact, that is significantly reduced by combining the pre-image methods.

References

1. Abrahamsen, T.J., Hansen, L.K.: Input Space Regularization Stabilizes Pre-Images for Kernel PCA De-Noising. IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2009)
2. Berouti, M., Schwartz, M., Makhoul, J.: Enhancement of Speech Corrupted by Acoustic Noise. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1979 pp. 208–211 (1979)
3. Griffin, D., Lim, J.: Signal Estimation from Modified Short-Time Fourier Transform. IEEE Transactions on Acoustics, Speech and Signal Processing 32(2), 236 – 243 (1984)
4. Honeine, P., Richard, C.: Solving the Pre-Image Problem in Kernel Machines: A Direct Method. IEEE International Workshop on Machine Learning for Signal Processing (MLSP) (2009)
5. Hu, Y., Loizou, P.: Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech, and Language Processing 16(1), 229 –238 (2008)
6. Hu, Y., Loizou, P.C.: A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise. IEEE Transactions on Speech and Audio Processing 11, 334–341 (2003)
7. Kwok, J.T., Tsang, I.W.: The Pre-Image Problem in Kernel Methods. IEEE Transactions on Neural Networks 15, 408–415 (2004)
8. Leitner, C., Pernkopf, F., Kubin, G.: Kernel PCA for Speech Enhancement. Interspeech 2011 (accepted) (2011)
9. Loizou, P.C.: Speech Enhancement: Theory and Practice. CRC (2007)
10. Mika, S., Schölkopf, B., Smola, A., Müller, K.R., Scholz, M., Rätsch, G.: Kernel PCA and De-Noising in Feature Spaces. Advances in Neural Information Processing Systems 11 pp. 536–542 (1999)
11. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Tech. rep., Max Planck Institute for Biological Cybernetics (1996)
12. Zölzer, U. (ed.): DAFX - Digital Audio Effects. John Wiley & Sons (2002)