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 present a new method to de-noise speech in the complex spectral domain. The method is derived from kernel principal component analysis (kPCA). Instead of applying PCA in a high-dimensional feature space and then going back to the original input space by using a solution to the pre-image problem, only the pre-image step is applied for de-noising. We show that the de-noised audio sample is a convex combination of the noisy input data and that the resulting algorithm is closely related to the soft k-means algorithm. Compared to kPCA, this method reduces the computational costs while the audio quality is similar and speech quality measures do not degrade.

Index Terms— Speech enhancement, kernel PCA, pre-image problem

1. INTRODUCTION

Subspace methods, one class of speech enhancement algorithms, are based on the assumption that the noisy speech signal lives in a signal space that can be separated in a speech plus noise and in a noise subspace. For de-noising, subspace methods try to retrieve the components only living in the speech (plus noise) subspace, with additional filtering to attenuate the noise components. This is usually done by the application of the principal component analysis (PCA) or equivalently the Karhunen-Loève-transform (KLT). Subspace methods have been developed for white [1] and colored noise [2].

Recently, we used kernel PCA, which is a non-linear extension to PCA, for speech de-noising [3]. Our approach, however, differs from the standard PCA approaches in several points. We apply kernel PCA on feature vectors extracted from the complex coefficients of the short-term Fourier transform (STFT) while standard PCA is applied in the time domain. Furthermore, standard PCA uses the covariance matrix and kernel PCA is applied on the kernel matrix. Complex-valued data can be easily handled by using a Gaussian kernel.

Kernel PCA implicitly performs a non-linear transformation of the data to a higher-dimensional feature space, where PCA is executed. For speech de-noising the data has to be transformed back to input space to get the de-noised audio samples. This inverse transformation is not straight-forward. Several solutions have been proposed to solve this so-called pre-image problem. For Gaussian kernels, the solution is commonly computed iteratively. The de-noised sample is determined as weighted sum of noisy samples where the weights depend on the kPCA and the kernel [4]. This pre-image method can be refined by additional normalization [5] or by regularization [6].

Previous experiments on synthetic data and real audio data have shown that the applied pre-image method can crucially influence the outcome of the de-noising process [7]. In this paper, we go one step beyond: We drop the kPCA altogether and only rely on the pre-image method to de-noise speech.

This paper is organized as follows: Section 2 describes the motivation and the implementation of our approach. Section 3 presents the experiments, the evaluation and the results. Section 4 concludes the paper.

2. FROM KERNEL PCA TO PRE-IMAGE DE-NOISING

Principal component analysis decomposes data into components assigned to different directions in the transformation space according to their variance. Decomposition is done by eigenvalue decomposition (EVD) of the covariance matrix, where the eigenvectors span the transformation space and the eigenvalues indicate the amount of variance in each direction. For de-noising, components into directions of small variance are assumed to contain information about noise only. Therefore, they are neglected and the data is projected onto the eigenvectors corresponding to the largest eigenvalues. The number of projection components determines the degree of de-noising.

We empirically observed that the number of used components had only a minor, almost no, effect on the outcome of the de-noising process when using kPCA on spectral data.

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

The de-noising quality was rather the same whether projection was performed on one or more components. However, besides the chosen variance of the kernel, the applied pre-image method heavily influenced the results [7]. Therefore, we further investigated the contribution of the projection and the pre-image reconstruction on the de-noising process.

When a Gaussian kernel is used, the pre-image $\mathbf{z}$ of a sample in feature space $\mathbf{\Phi}(\mathbf{z})$ can be computed iteratively by

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

where γ_i are the weighting coefficients derived from the projection of the kPCA,

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

is the Gaussian kernel with its variance c, t denotes the iteration index, and M is the number of samples. (For more details see [4, 5, 6].) Note that the pre-image is always a linear combination of the (noisy) input samples $\mathbf{x}_i$. The Gaussian kernel serves as similarity measure between two samples. If the samples are equal, it is one, if they are very distinct, it is close to zero. The variance c is used as parameter to define the extent to which samples are judged to be similar. In [3], we initialize $\mathbf{z}_0$ to the noisy sample $\mathbf{x}_0$ and iterate (1) until convergence.

Since the influence of the number of components is only minor, we neglect the weighting coefficients γ_i, i.e., set them to one, and compute the preimage $\mathbf{z}$ – the de-noised sample – by a linear combination of noisy samples that only depends on the kernel function. As no EVD for the kPCA has to be computed the computational costs are reduced. The applied equation can be reformulated as

$$\mathbf{z}_{t+1} = \sum_{i=1}^{M} \tilde{k}(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i, \tag{3}$$

where

$$\tilde{k}(\mathbf{z}_t, \mathbf{x}_i) = \frac{k(\mathbf{z}_t, \mathbf{x}_i)}{\sum_{j=1}^{M} k(\mathbf{z}_t, \mathbf{x}_j)}. \tag{4}$$

As the kernel function can only take values between zero and one, $\tilde{k}(.,.)$ is also constrained to values within the interval $[0, 1]$. Furthermore it is normalized such that $\sum_{i=1}^{M} \tilde{k}(\mathbf{z}_t, \mathbf{x}_i) = 1$. Due to these constraints, the pre-image $\mathbf{z}$ can be seen as a convex combination of the training samples $\mathbf{x}_i$ [8]. In other words the de-noised sample lies in a convex hull spanned by the noisy samples.

Another interesting aspect of this approach is its close relation to the soft k-means algorithm [9]. The soft k-means algorithm is used for clustering. The mean of one cluster is defined as

$$\mathbf{m}_k = \frac{\sum_{i=1}^{M} \frac{\exp(-\beta d(\mathbf{m}_k, \mathbf{x}_i))}{N_i} \mathbf{x}_n}{\sum_{i=1}^{M} \frac{\exp(-\beta d(\mathbf{m}_k, \mathbf{x}_i),)}{N_i}} \tag{5}$$

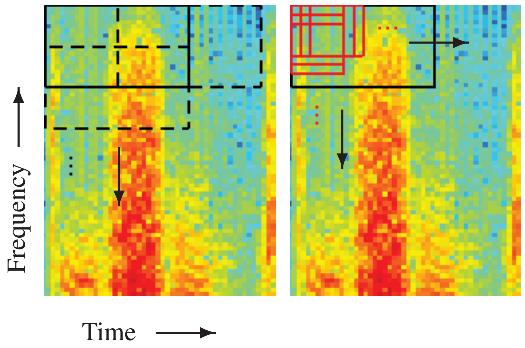

Fig. 1. Spectral detail of the clean utterance /t a sh e/. Left hand side: Extraction of frequency bands with hopsize 4. Right hand side: Extraction of 12 × 12 patches with hopsize 2 within a frequency band.

where

$$N_i = \sum_{l=1}^{K} \exp(-\beta d(\mathbf{m}_l, \mathbf{x}_i)). \tag{6}$$

$d(\mathbf{m}_l, \mathbf{x}_i)$ is the distance between the two points $\mathbf{m}_l$ and $\mathbf{x}_i$, β is the so-called stiffness parameter and K is the number of clusters. When the squared Euclidean distance is used, the exponential term is equivalent to the Gaussian kernel, where $c = 1/\beta$. So the soft k-means update of the cluster mean is the same as the update of the pre-image $\mathbf{z}$ apart from the normalization factor N_i, (which is different for every x_i).

Besides the proposed pre-image iteration we extended (1) with an additional regularization [6], that simplifies to

$$\mathbf{z}_{t+1} = \frac{\frac{2}{c}\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i)\mathbf{x}_i + \lambda \mathbf{x}_0}{\frac{2}{c}\sum_{i=1}^{M} k(\mathbf{z}_t, \mathbf{x}_i) + \lambda} \tag{7}$$

when the weighting coefficients γ_i are neglected. Here, λ is the regularization parameter which determines the trade-off between the noisy sample $\mathbf{x}_0$ (of which the pre-image should be found) and the convex combination.

3. EXPERIMENTS

In the previous application of kernel PCA for speech enhancement we extracted the feature vectors from the sequence of STFT coefficients. For the experiments using only the pre-image iteration we used the same features. First the STFT 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 on the time axis to reduce computational cost, and on the frequency axis to compensate for different energy levels (see Figure 1, left side). The retrieved frequency bands are split into overlapping patches of size 12 × 12 with overlap 11 (see Figure 1, right side). The height of the frequency bands is chosen to equal 8 patches with an overlap of 4 patches between adjacent bands.

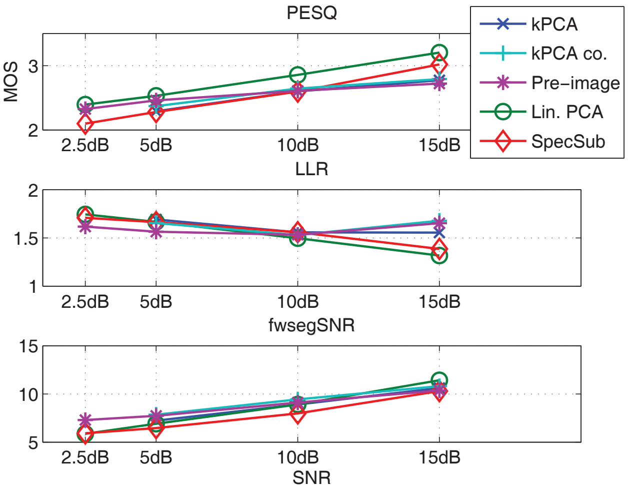

Fig. 2. Comparison of the pre-image iteration method (Pre-image) to kernel PCA (kPCA), kernel PCA with combined pre-imaging (kPCA co.), linear PCA (lin. PCA), and spectral subtraction (SpecSub) using the perceptual evaluation of speech quality (PESQ) measure, the log-likelihood ration (LLR), and the frequency-weighted segmental SNR (fwsegSNR). For the kPCA and the pre-image implementation the values for c are 3, 2, 0.5, and 0.25 for 2.5, 5, 10, and 15 dB SNR, respectively. The regularization parameter λ is set to 0.5 for the pre-image method.

This configuration of patches and bands led to good results in previous work [3]. In previous experiments, windowing of the patches was beneficial, so a 2D Hamming window is applied and then the patches are rearranged to vectors. These vectors are used for the pre-image method, that is applied on each frequency band independently, i.e., all pre-image estimates are only based on linear combinations of samples retrieved from the same frequency band.

For resynthesis, patches at the same time frequency position but from overlapping frequency bands are averaged. Patches are rearranged using the overlap-add method with weighting as described in [10] generalized for the 2D domain. The overlapping time segments are averaged, the inverse Fourier transform is applied and the audio signal is synthesized with the weighted overlap-add method [10].

The method relying on the pre-image iteration only was compared to previous results in [3] using the kernel PCA and the pre-image method of [5] (called kPCA) and another implementation using a combination of pre-image methods [7] (labeled as kPCA co.). The first approach suffers from a buzz-like artifact that could be significantly reduced by the second approach. The major advantage of both approaches is that they are free from musical noise. The enhanced signal of the approach presented in this paper sounds very similar to kPCA co. but is computationally more efficient.[1] It is faster by a factor of 1.5. For additive white Gaussian noise of 10 dB SNR almost no difference can be heard. Visual inspection of the spectrogram revealed that the new approach has a slightly higher low pass behavior and that a little more residual noise is left. With additional regularization in (7) (see [6]), the audio signal sounds similar as without regularization but with slightly more background noise that changes with the value of λ.

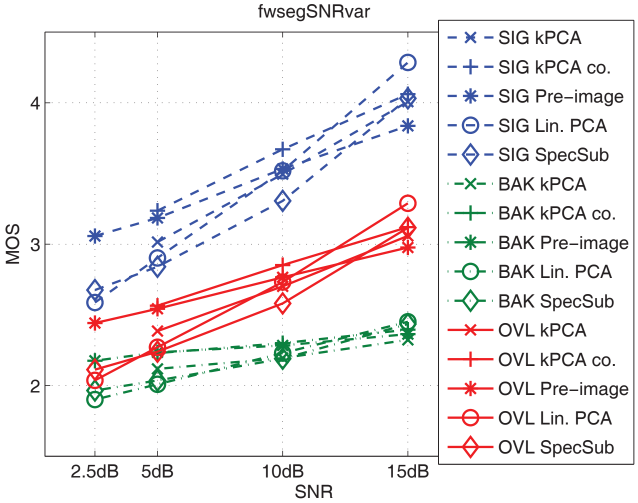

Fig. 3. Comparison of the pre-image iteration method (Pre-image) to kernel PCA (kPCA), kernel PCA with combined pre-imaging (kPCA co.), linear PCA (lin. PCA), and spectral subtraction (SpecSub) using a variant of the frequency-weighted segmental SNR that separately evaluates the speech quality (SIG), the background intrusion (BAK), and the overall quality (OVL).

In the following, we evaluated the new approach on a database consisting of recordings from six speakers (3 male, 3 female). Each uttered 20 sentences which leads to 120 sentences in total. Recording was performed with a close-talk microphone and 16kHz sampling frequency. White Gaussian noise was added at 2.5, 5, 10, and 15 dB SNR. For evaluation, we used speech quality measures that were reported to have high correlation with results of subjective listening tests [11]. The used measures are: the perceptual evaluation of speech quality measures (PESQ), the log-likelihood ratio (LLR), the frequency-weighted segmental SNR (fwsegSNR), and a variant of the frequency-weighted segmental SNR (fwsegSNR-var), that separately evaluates the signal quality (SIG), the background intrusion (BAK), and the overall quality (OVL).

In addition, we compared our algorithm to the linear PCA method (lin. PCA) of Hu and Loizou [2] and with spectral subtraction (SpecSub) as described in [12] and provided in

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

[13]. We found that the method with regularization in (7) scores significantly better than (3), so experiments are limited to this approach. Fig. 2 and 3 show the performance results in terms of PESQ, LLR, fwsegSNR, and fwsegSNRvar. It can be seen that the results of the kernel PCA related algorithms lie in the range of the results achieved by linear PCA and spectral subtraction. The pre-image iteration with regularization achieves similar results as the kPCA approaches, although the used resources are significantly reduced. This outcome supports the conjecture that when using complex spectral data the projection in kPCA is of minor importance and that the main contribution for de-noising stems from the pre-image iteration, i.e. the convex combination of noisy samples, as well as averaging effects due to the feature extraction approach.

4. CONCLUSION AND FUTURE WORK

In this paper, we derived a new method for speech de-noising from kernel PCA that is applied in the spectral domain. Kernel PCA is equivalent to PCA applied in a high-dimensional feature space using a non-linear mapping. The problem of inverse transformation that is necessary for de-noising is known as the pre-image problem. We showed that solutions to this problem can be also used for speech de-noising even when the information from kPCA, namely the projection coefficients, is neglected.

The de-noised samples of the new algorithm are convex combinations of the noisy samples, in other words they lie in the convex hull spanned by the noisy samples. This is intuitively meaningful as vectors that live in a region spanned by speech vectors are likely to represent speech vectors as well. Furthermore, this algorithm is closely related to the soft k-means algorithm. This leads to the interpretation that the noise is averaged out by forming a linear combination of noisy samples.

The method was tested on audio data corrupted by additive white Gaussian noise at different SNRs. The audio quality is similar to results from a previous kernel PCA implementation, however, the computation is faster by a factor of 1.5. For evaluation, speech quality measures were computed and compared to other kPCA implementations, standard PCA and spectral subtraction. With the proposed method, similar results are achieved, however, it is not affected by musical noise, as is the case for linear PCA and spectral subtraction.

In future, we plan to evaluate several modifications related to the pre-image iteration approach. First, we want to test whether an additional processing of the samples used for the convex combination, such as smoothing, improves the result. Second, we aim to evaluate a supervised method where the de-noised samples are constructed from clean speech samples from a database. For this method, the speaker identity has to be known and clean speech has to be available.

As objective quality measures cannot fully replace a subjective evaluation, we plan to do a subjective listening test. Furthermore, we want to extend the experiments to scenarios with different noise types such as babble noise and perform experiments on a publicly available database like the NOIZEUS database.

5. REFERENCES

[1] 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.

[2] 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.

[3] Christina Leitner, Franz Pernkopf, and Gernot Kubin, “Kernel PCA for speech enhancement,” *Interspeech 2011* , pp. 1221–1224, 2011.

[4] 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.

[5] 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.

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

[7] 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.

[8] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[9] David MacKay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[10] 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.

[11] 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.

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

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