EVALUATION OF SPEECH ENHANCEMENT BASED ON PRE-IMAGE ITERATIONS USING AUTOMATIC SPEECH RECOGNITION

Christina Leitner

DIGITAL
JOANNEUM RESEARCH
Forschungsgesellschaft mbH
Steyrergasse 17, 8010 Graz, Austria

Juan A. Morales-Cordovilla, Franz Pernkopf

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

ABSTRACT

Recently, we developed pre-image iteration methods for single-channel speech enhancement. We used objective quality measures for evaluation. In this paper, we evaluate the de-noising capabilities of pre-image iterations using an automatic speech recognizer trained on clean speech data. In particular, we provide the word recognition accuracy of the de-noised utterances using white and car noise at 0, 5, 10, and 15 dB signal-to-noise ratio (SNR). Empirical results show that the utterances processed by pre-image iterations achieve a consistently better word recognition accuracy for both noise types and all SNR levels compared to the noisy data and the utterances processed by the generalized subspace speech enhancement method.

***Index Terms*—** Speech enhancement, speech de-noising, pre-image iterations, automatic speech recognition

1. INTRODUCTION

Speech enhancement is important in the field of speech communications and speech recognition. Many methods have been developed, including spectral subtraction [1], statistical model-based methods such as estimators of the short-time spectral amplitude [2], and subspace methods based on principal component analysis (PCA) [3, 4]. Recently, we proposed pre-image iterations (PI) for speech enhancement which are derived from *kernel PCA*, the non-linear extension of PCA [5]. Inspired by subspace methods, a significant difference of PI is the use of complex-valued spectral data as feature vectors. Furthermore, PI exhibit a similarity to non-local filtering, a technique applied for image de-noising. While many de-noising algorithms often compute the value of the de-noised pixel solely based on the value of its surrounding pixels, non-local filters average over pixels that are located all over the image but have a similar neighborhood. This approach is favorable if images contain repetitive patterns such as textures. Although popular for image de-noising, non-local filtering has only recently gained attention in speech enhancement [6].

In this paper, we use automatic speech recognition (ASR) to evaluate the performance of pre-image iterations. So far, we have evaluated PI using objective quality measures like the measures of the PEASS toolbox [7]. While a speech enhancement method might show good perceptual speech quality this does not necessarily mean that intelligibility or word accuracy of a recognizer are improved [8]. We investigate how an off-the-shelf recognizer performs on noisy speech before and after processing by different enhancement algorithms. We realize this by feeding the enhanced utterances into a pre-trained speech recognizer and by comparing the word accuracy of unprocessed and enhanced noisy data. The focus of this paper is strictly on the evaluation of different enhancement methods and not on optimization of the recognition results. Therefore, the speech recognizer is assumed to be not tunable and it is not adapted to the enhanced data.

We evaluate two approaches of PI: In the first method, the tuning parameter c – the variance of the kernel – is set depending on the SNR [5], which is assumed to be known. In the second method, c is determined from a mapping function using a noise estimate. The mapping function is derived from a development set and maps noise estimates to values of c [9]. This method is favorable when the recordings have different noise levels. Experiments are performed on speech data corrupted by white and car noise at 0, 5, 10, and 15 dB SNR. As a benchmark, results for spectral subtraction [1], the generalized subspace method [4], and the minimum mean-square error (MMSE) log-spectral amplitude estimator [2] are provided. The word accuracy achieved on the data enhanced by PI is superior to the word accuracy achieved by the generalized subspace method, similar to the word accuracy by spectral subtraction and mostly lower than the accuracy for the MMSE log-spectral amplitude estimator. The pre-image iteration methods as well as the MMSE log-spectral amplitude estimator produce fewer artifacts which results in higher word accuracies especially in low SNR conditions.

We gratefully acknowledge funding by the European Commission in the context of the European project DIRHA under the project number FP7-ICT-2011-7-288121.

The paper is organized as follows: Section 2 introduces the pre-image iteration methods. In Section 3, the databases and the recognition system are described. In Section 4, the results are discussed. Section 5 concludes the paper.

2. PRE-IMAGE ITERATIONS

In [5], 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 original 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 been proposed to solve the pre-image problem (see [10]).

Pre-image iterations are based on the simplification of the iterative pre-image method of [11]. In [5], 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 2.1 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. 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.

We employ a regularization for pre-image estimation as proposed in [12]. 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 scales the similarity and influences the de-noising performance. 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.

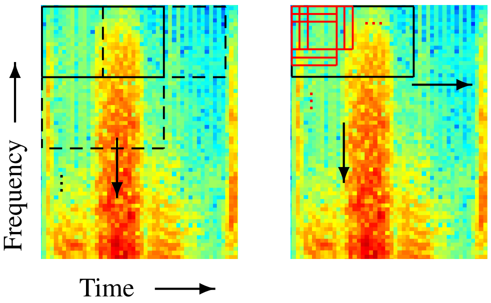

Fig. 1. Spectral detail of the clean utterance /t a sh e/. Left hand side: Extraction of frequency bands with time overlap of 10 patches. Right hand side: Extraction of 12 × 12 patches from one frequency band with an overlap of 10 in time and frequency.

2.1. Feature Extraction and Synthesis

The sample vectors $\mathbf{x}_i$ for pre-image iterations are extracted from the sequence of short-term Fourier transforms (STFTs) 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 1, 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 1, 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.

After de-noising, the audio signal is re-synthesized by reshaping the sample vectors to patches. The patches of all frequency bands belonging to one time segment are rearranged using the overlap-add method with weighting as in [13] 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 [13].

2.2. Automatic Determination of the Kernel Variance

As the performance of the pre-image iterations strongly depends on the kernel variance c, we adapt c for varying noise conditions and levels. Schemes for determining c directly from the processed utterance have been proposed in [9]. In particular, two approaches for additive white Gaussian noise (AWGN) and colored noise have been developed which are shortly summarized in the following:

(i) For white noise, a function for mapping the noise power estimate to a suitable value of c is learned using development data. A weighted composition of the four scores of the PEASS toolbox [7] – overall perceptual score, target perceptual score, interference perceptual score, and artifact perceptual score – is used to determine well-performing values of c for the individual noise levels. The c and noise level values from development data are fitted by a polynomial of second order. This function is used for the mapping. The noise power is estimated at the beginning of the utterance assuming absence of speech and stationary noise.

(ii) For colored noise, a single value of c for all frequency bands is insufficient for substantial de-noising. For this reason we derive the averaged noise power estimate for each frequency band individually. These estimates are used in the mapping function derived for white noise to obtain values of c for each frequency band.

3. EXPERIMENTAL FRAMEWORK

3.1. Database Description

The enhancement algorithms are evaluated in terms of ASR performance using two different databases, the Bavarian Archive for Speech Signals (BAS) *PHONDAT-1* database [14] for training and the *airbone* database [10] for testing.

The training corpus consists of 4999 clean utterances of the BAS *PHONDAT-1* database sampled at 16 kHz. These utterances correspond to 50 speakers resulting in around 100 utterances per speaker and 1504 different words in total.

The test corpus consists of 120 utterances of the *airbone* database which are read by six speakers – three male and three female – of the Austrian variety of German. The number of tested words is 50 and more than the half of them do not occur in the training set. The recordings are sampled at 16 kHz. These utterances are contaminated with two different types of noise, namely white and car noise, at different SNRs, i.e., 0, 5, 10, and 15 dB. The car noise is taken from the *NOISEX-92* database [15]. The SNR computation is based on the *active speech level* such that only samples where speech is present are used to estimate the energy of the signal [16]. A small development set – consisting of two sentences per speaker – is used to determine the SNR-dependent value of the kernel variance c. The mapping function for estimating the kernel variance as described in Section 2.2 is derived from the same development set.

Condition	0 dB	5 dB	10 dB	15 dB	Average
Noisy	0.00	15.56	38.89	65.56	30.00
PI	27.22	53.89	68.33	72.59	57.15
PID	35.93	58.70	72.22	77.59	61.11
Subspace	2.59	4.63	16.30	42.96	16.62
Subspace$_{\text{MNS}}$	22.96	36.48	46.85	68.89	43.80
SpecSub	25.74	53.15	73.89	85.56	59.59
LogMMSE	37.78	58.15	74.63	89.07	64.91
Clean	97.78				

Table 1. WAcc achieved on the noisy data, after enhancement (i) by pre-image iterations with SNR-dependent setting of the kernel variance (PI), (ii) by pre-image iterations with automatic determination of the kernel variance (PID), (iii) by the generalized subspace method (Subspace), (iv) by the generalized subspace method with post-processing using musical noise suppression (Subspace$_{\text{MNS}}$), (v) by spectral subtraction (SpecSub), and (vi) by the MMSE log-spectral amplitude estimator (LogMMSE), evaluated on data corrupted by AWGN at 0, 5, 10, and 15 dB SNR.

3.2. Recognition system

The automatic speech recognizer is based on the *Hidden Markov Toolkit* (HTK). The front-end (FE) and the back-end (BE) are both derived from the standard recognizer of the Aurora-4 database [17]. The FE computes 13 Mel frequency cepstral coefficients (MFCCs) by using a sampling frequency of 16kHz, a frame shift of 10 ms, and a window length of 32 ms. Cepstral mean normalization is employed. Furthermore, delta and delta-delta features are computed leading to a feature vector of 39 components. For training, the BE uses a dictionary based on 34 SAMPA-monophones. The transcriptions in this dictionary are derived from more detailed transcriptions based on 44 SAMPA-monophones by clustering less common monophones in the corpus. For each triphone, a hidden Markov model (HMM) with 6 states and Gaussian mixture models of 8 components per state is trained. To reduce the complexity and to overcome the lack of training data for some triphones, tree-based clustering based on monophone-classification is applied. With tree-based clustering also triphone models that have not been observed in the training data can be created. The grammar used for training is probabilistically modeled. In contrast to that, a rule-based grammar is applied for testing as the utterances of the airbone database obey very strict grammar rules.

4. RESULTS AND DISCUSSION

Table 1 and 3 show the word accuracy (WAcc) in percent achieved on the noisy, enhanced, and clean data. The word accuracy is defined as WAcc $= \frac{N-S-D-I}{N} \times 100\%$, where N is the number of words, S is the number of substitutions, D is the number of deletions and I is the number of insertions.

PID	0 dB	5 dB	10 dB	15 dB
Noisy	*	*	*	*
Subspace	*	*	*	*
SpecSub	*			-
LogMMSE	-		-	-

Table 2. Results of the statistical significance test between PID and the reference methods for the WAcc in Table 1. The asterisk indicates a significantly better performance of PID with a significance level of 0.01, while the minus sign indicates a lower performance.

Table 1 shows the WAcc for pre-image iterations with SNR-dependent setting of the kernel variance (PI)[1] and for pre-image iterations with automatic determination of the kernel variance (PID) for AWGN as described in Section 2.2. Table 3 shows the results of pre-image iterations with frequency-dependent determination of the kernel variance (PIDF) developed for colored noise. In the presented experiments car noise was used. As a benchmark word accuracies achieved by the generalized subspace method [4], by spectral subtraction [1], and by the MMSE log-spectral amplitude estimator [2] are provided.

In addition to the WAcc, we evaluated if the performance difference between the pre-image iteration methods and the reference methods is statistically significant. We use a *matched pairs test* as recommended in [18]. For all evaluations, we employ a significance level of 0.01. Table 2 and 4 show the results of the significance test between the pre-image iteration methods and the reference methods.

Condition	0 dB	5 dB	10 dB	15 dB	Average
Noisy	1.30	25.93	62.78	85.19	43.80
PIDF	34.95	62.04	81.48	89.26	66.93
Subspace	8.52	27.04	66.85	81.48	45.97
SpecSub	29.26	61.11	79.26	90.74	65.23
LogMMSE	52.78	75.74	86.11	94.07	77.17
Clean	97.78				

Table 3. WAcc achieved on the noisy data, after enhancement (i) by pre-image iterations with frequency-dependent determination of the kernel variance (PIDF), (ii) by the generalized subspace method (Subspace), (iii) by spectral subtraction (SpecSub), and (iv) by the MMSE log-spectral amplitude estimator (LogMMSE), evaluated on data corrupted by car noise at 0, 5, 10, and 15 dB SNR.

PIDF	0 dB	5 dB	10 dB	15 dB
Noisy	*	*	*	
Subspace	*	*	*	*
SpecSub	*			
LogMMSE	-	-	-	-

Table 4. Results of the statistical significance test between PIDF and the reference methods for the WAcc in Table 3. The asterisk indicates a significantly better performance of PIDF with a significance level of 0.01, while the minus sign indicates a lower performance.

The WAcc for the noisy data clearly states that the recognizer performance suffers from the noise contamination. The enhancement based on pre-image iterations successfully increases the WAcc in comparison to the noisy data. The WAcc of the pre-image iteration methods is always superior to the WAcc of the generalized subspace method, similar to the WAcc of spectral subtraction and mostly lower than the WAcc of the MMSE log-spectral amplitude estimator. The superior performance is significant for the generalized subspace method, for spectral subtraction at 0 dB SNR and the noisy data except for 15 dB SNR. The good WAcc of the pre-image iteration methods constitutes a difference to the results achieved with objective quality measures such as PESQ, where the scores of the reference methods are rather slightly higher than the scores of the pre-image iteration methods (cf. [10]). The comparison of PI to PID reveals that the PID method always achieves higher word accuracies. This confirms that the automatic determination of the kernel variance is preferable over using a fixed value for one noise condition. The results for the experiments with car noise show that this type of noise is less harmful to the performance of the recognizer. This can be explained by the fact that the noise energy is concentrated below 1kHz, where the speech components are relatively strong and the distortion by the noise therefore is limited.

Listening to the utterances processed by the generalized subspace method and by spectral subtraction reveals that musical noise is very prominent. The utterances enhanced by the pre-image iteration methods and the MMSE log-spectral amplitude estimator are less affected by such artifacts. This explains the better performance of pre-image iteration methods and the MMSE log-spectral amplitude estimator especially in low SNR conditions. To test the hypothesis that musical noise is problematic for the speech recognizer we further evaluated the WAcc on data corrupted by AWGN, enhanced by the generalized subspace method and subsequently post-processed by the musical noise suppression (MNS) method proposed in [10]. The results are included in Table 1 and denoted as $\text{Subspace}_{\text{MNS}}$. The WAcc is much better after the MNS and the performance difference is significant. Hence, the musical noise is indeed a problem for the recognizer and speech enhancement methods introducing too many artifacts may be counterproductive, as shown for the generalized subspace method, where the WAcc is even lower than the WAcc for the noisy data.

Finally the high WAcc on clean data suggests that the recognizer trained on the BAS database generalizes well to the test data of the *airbone* database, although the speakers have different accents (German and Austrian) and the vocabulary is not entirely the same.

[1] One value for c is derived from the development set and applied for all sentences of one SNR condition.

5. CONCLUSION

Pre-image iterations have been shown to be useful for speech de-noising. So far, they have only been evaluated by objective quality measures and by informal listening tests [5, 9]. In this paper, we evaluated the performance achieved by an automatic speech recognizer tested on speech utterances corrupted by noise and subsequently enhanced by the pre-image iteration method. Furthermore, results after enhancement by the generalized subspace method, by spectral subtraction, and by the MMSE log-spectral amplitude estimator are presented. The speech recognizer is trained on clean data and is not adapted to the data used for the enhancement experiments. This way, the effects of testing noise contaminated and subsequently enhanced data can optimally be analyzed.

Experiments were performed on data corrupted by additive white Gaussian noise and by car noise at 0, 5, 10, and 15 dB SNR. The data enhanced by pre-image iterations results in higher word accuracies compared to noisy data. The WAcc of the PI methods is superior to the WAcc of the generalized subspace method, similar to the WAcc of spectral subtraction and mostly lower than the WAcc of the MMSE log-spectral amplitude estimator. The enhancement methods produce different types of artifacts that affect the speech recognizer differently. The generalized subspace method and spectral subtraction produce musical noise which decreases the WAcc especially in low SNRs. The conjecture that musical noise is a major impairment for the speech recognizer is confirmed in a further experiment where the WAcc of enhanced data is compared to enhanced data subsequently processed by a musical noise suppression algorithm. The WAcc after musical noise suppression is higher than the WAcc of the data only processed by the speech enhancement algorithm. This means that the attenuation of musical noise improves the recognition performance.

In future, we would like to extend the pre-image iteration method by a noise tracker to generalize the method from stationary noise to other noise types such as babble noise.

REFERENCES

[1] 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.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443 – 445, 1985.

[3] Y. Ephraim and H. 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.

[4] 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.

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

[6] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1584–1599, 2011.

[7] 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.

[8] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC, 2007.

[9] C. Leitner and F. Pernkopf, "Generalization of pre-image iterations for speech enhancement," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7010–7014, 2013.

[10] C. Leitner, *Kernel PCA and Pre-Image Iterations for Speech Enhancement*, Ph.D. thesis, Graz University of Technology, 2013.

[11] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and denoising in feature spaces," *Advances in Neural Information Processing Systems 11*, pp. 536–542, 1999.

[12] T. J. Abrahamsen and L. K. Hansen, "Input space regularization stabilizes pre-images for kernel PCA denoising," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2009.

[13] D. W. Griffin and J. S. 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.

[14] F. Schiel and A. Baumann, "Phondat 1, corpus version 3.4," 2006.

[15] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

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

[17] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task," Tech. Rep., STQ AURORA DSR, Working Group, 2002.

[18] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 532–535, 1989.