EXPERIMENTS WITH AN EXTENDED ADAPTIVE SVD ENHANCEMENT SCHEME FOR SPEECH RECOGNITION IN NOISE

Christian Uhl and Markus Lieb

Philips Research Laboratories, Weisshausstrasse 2, D-52066 Aachen, Germany
e-mail: {christian.uhl, markus.lieb}@philips.com

ABSTRACT

An extension to adaptive signal subspace methods is presented, based on singular value decomposition (SVD) with an online estimation of the noise variance. With this approach aiming at automatic speech recognition (ASR) in adverse environmental conditions no speech detection has to be performed. A comparison of different SVD approaches and nonlinear spectral subtraction within ASR experiments of different applications is conducted for weakly correlated noise scenarios. Better performance in the case of signal subspace speech enhancement with respect to both accuracy as well as robustness of parameter tuning are reported.

1. INTRODUCTION

The performance of commonly used speech recognizers is affected by adverse environmental conditions. Suppression of additive noise already during the feature extraction stage is one important component in the development of *robust* speech recognition systems. Conventional methods, working in the spectral domain, like Wiener filter and spectral subtraction, encounter wide usage in speech processing and recognition systems [1]. However, drawbacks are associated with such spectral subtraction type speech enhancement approaches: (1) Spectral noise shape estimates have to be obtained during non-speech, noise-only, periods. Therefore, reliable speech/non-speech detection is a crucial issue for spectral subtraction, whilst mis-classification leads to significant loss in performance. (2) “Musical tones” due to the subtraction process degrade the intelligibility as well as the performance of automatic speech recognition (ASR). (3) Parameters, like time-constants, over-subtraction-values or target noise-floor usually have to be tuned individually for different environmental conditions. Thus, unique default settings for different applications always comprise a trade-off.

Our study deals with an alternative approach located in the time domain, subspace based speech enhancement, which has been investigated and applied successfully recently [2, 3]. The signal is embedded via delay coordinates into an high-dimensional space and the goal is to find a projection of the signal onto the clean signal subspace suppressing additive noise components.

The paper is organized as follows: After a short description of the general SVD approach, we present an algorithmic extension, allowing for an adaptive un-supervised estimation of SVD related parameters. Experimental results under various additive noise conditions are presented for the TIMIT mono-phone decoding task, as well as for a large vocabulary continuous dictation task.

2. SVD BASED SPEECH ENHANCEMENT

The observed signal $y(t)$ is assumed to consist of a clean speech signal $x(t)$ and an additive noise part $n(t)$:

$$y(t) = x(t) + n(t). \tag{1}$$

Toeplitz structured matrices $A(t)$ of dimension $(N+1) \times (M+1)$ are built from the signal amplitude $a(t)$ by introducing M delay coordinates in an interval of length N:

$$A(t) = \begin{pmatrix} a(t) & \dots & a(t-M) \\ \vdots & & \vdots \\ a(t+N) & \dots & a(t+N-M) \end{pmatrix} \tag{2}$$

Therefore we can rewrite Eq. 1 in terms of Toeplitz matrices:

$$Y(t) = X(t) + N(t) \tag{3}$$

In the case of Gaussian white noise, i.e. a diagonal correlation matrix for $N(t)$

$$E\{N^T(t)N(t)\} = \sigma_N^2 I, \tag{4}$$

SVD based methods can be applied for signal enhancement.

With the SVD of $Y(t)$: $Y = U\Sigma V^T, \quad \Sigma = \text{diag}\,(\sigma_i)$, different signal estimators $\hat{X}(t)$ depending on the cost function to be optimized have been proposed [3]:

$$\hat{X} = UG\Sigma V^T, \quad G = \text{diag}\,(g_i) \tag{5}$$

1. *Minimum Variance (MV) Estimator*

$$g_i = 1 - \frac{\sigma_N^2}{\sigma_i^2} \tag{6}$$

2. *Time Domain Constrained (TDC) Estimator*

$$g_i = \frac{1 - \sigma_N^2/\sigma_i^2}{1 - (1-\gamma)\sigma_N^2/\sigma_i^2} \tag{7}$$

3. *Spectral Domain Constrained (SDC) Estimator*

$$g_i = (1 - \frac{\sigma_N^2}{\sigma_i^2})^{\beta_1} \quad \text{or} \quad g_i = \exp[\frac{-\beta_2 \sigma_N^2}{\sigma_i^2 - \sigma_N^2}] \tag{8}$$

Thereby, the parameters $\gamma \geq 0$ and $\beta_i \geq 1/2$ are associated with constraints of the residual noise. The parameter σ_N^2 represents the noise variance.

From the estimate $\hat{X}$ one obtains an estimate $\hat{x}(t)$ of the signal by averaging over anti-diagonal elements of $\hat{X}$ [3].

3. ADAPTIVE ONLINE ESTIMATION OF THE NOISE VARIANCE

The noise variance usually is estimated during non-speech periods, thus, reliable voice activity detection is required. To avoid this, we propose the following adaptive procedure to obtain an estimate for σ_N^2 continuously:

The squared values of the singular values of the observation matrix Y are estimates of the eigenvalues of the correlation matrix C_{YY} of the observation, $\sigma_i^2 = \lambda_i$ with $\lambda_i =$ EV of C_{YY} Since the correlation matrix is given as

$$C_{YY} = Y^T Y = X^T X + N^T N = X^T X + \sigma_N^2 I, \tag{9}$$

the corresponding eigenvalues are given by $\lambda_i^Y = \lambda_i^X + \lambda_i^N$. If we assume that the clean speech eigenvalues λ_i^X vanish for indices large enough ($i > C$), the remaining part of the eigenvalue spectrum can serve as the noise variance for white noise (see Fig. 1 for a schematic sketch):

$$\sigma_N^2 = \lambda_i^N = \lambda_i^Y \text{ for } i \geq C \tag{10}$$

Therefore the saturation value of the singular value spectrum represents an estimate of the noise variance σ_N^2. We estimate the noise variance for each frame, by the following iteration (starting with $\sigma_N = \sigma_M$)

$$\sigma_N = \frac{1}{M - C + 1} \sum_{i=C}^{M} \sigma_i, \quad \text{with} \quad \sigma_{C-1} > \theta \cdot \sigma_N \tag{11}$$

smoothing by a recursive filter,

$$\bar{\sigma_N} = (T \cdot \bar{\sigma_N} + \sigma_N)/(T+1) \tag{12}$$

with a threshold θ and a time constant T.

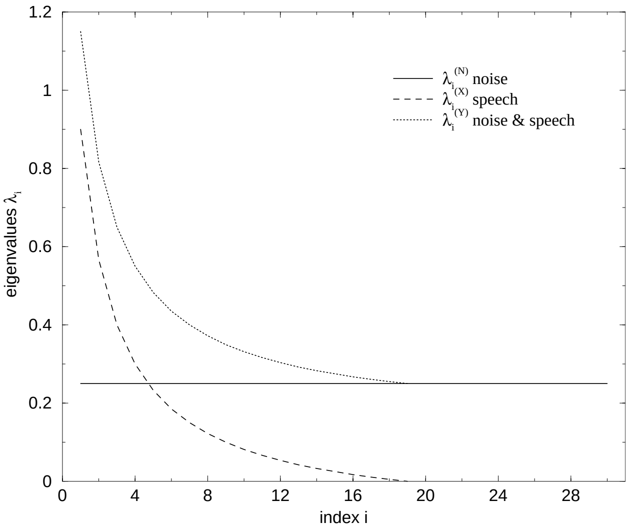

Fig. 1. Eigenvalue spectra λ_i of correlation matrices for clean speech, white noise and noisy speech.

4. SPEECH RECOGNITION EXPERIMENTS

4.1. TIMIT

To evaluate the potential of the MV-SVD method experiments have been performed on the context independent phoneme classification task TIMIT and comparisons to standard non-linear spectral subtraction (NSS) are drawn. The baseline system contains a standard MFCC frontend working on 8kHz down-sampled data, delivering 24 component feature vectors, including 12 delta features, on a 16msec frame spacing. Recursive cepstral mean subtraction is enabled. Context independent phonemes are modeled by 5-state left-to-right HMMs [4] with Laplacian mixture distributions, whilst the phone insertion penalty has been adjusted per experiment to obtain insertion rates around 10% for sake of comparability with results published before. The baseline performance of 63.7% accuracy for the clean 8kHz case compares well to state-of-the-art publications dealing with this task [5]. For the standard MV-SVD approach as well as for the NSS tests an optimization of the enhancement parameters individually for each noise conditions turned out to be crucial in order to achieve best performance, while for the adaptive MV-SVD algorithm a common parameter set could be used.

In Table 1 results are summarized showing phone-accuracies achieved for various Gaussian white noise conditions. Enhancement methods are applied during recognition only, i.e. models trained under clean conditions are used throughout the tests.

The results show the competitive behavior of the subspace methods compared to NSS. While the plain MV-SVD

Condition	Phone Accuracy (%)			
	no Enh.	NSS	SVD	SVD adapt.
SNR 6dB	21.70	24.89	23.67	35.45
SNR 12dB	31.69	34.93	33.90	42.67
SNR 18dB	42.90	46.86	44.40	45.09

Table 1. Experimental results on TIMIT databases for artificially added Gaussian noise

cannot reach the NSS performance, the adaptive MV-SVD outperforms the spectral subtraction at low SNR conditions significantly.

Still there remains a big performance gap to the clean scenario and also to matched training/test scenarios, where the three enhancement methods performed with no significant performance difference.

4.2. Large Vocabulary

To further validate these findings comparisons of NSS, MV-SVD, adaptive MV-SVD and the variants TDC/SDC of SVD are investigated in a completely different ASR application.

4.2.1. Task

The task consists of a large vocabulary continuous speech recognition task for English dictation with context dependent phonemes, generalized by CART. Our test-set consists of each five randomly chosen dictations from eight speakers (6597 words) with resulting word error rates (WERs) varying (for 95% confidence) with ± 1.2%. The sound files were degraded by adding artificially white noise with a resulting signal to noise ratio of approx. 8dB. The recognition experiments are conducted with speaker-independent references trained on clean data.

4.2.2. Comparison NSS and MV-SVD

With this setup the clean speech WER is given by 16.5% and drops for the degraded signal (8dB SNR) to 73.5%. Applying NSS with parameter optimization for each speaker 49.7% WER were obtained.

Fig. 2 shows the WER in dependence of the assumed noise variance σ_N obtained by applying the MV variant of the SVD subspace method. The number of delay coordinates was chosen as $M = 28$, the frame width as $N = 200$. A distinct minimum can be observed with a minimal WER of 39.9%, well below the WER obtained by NSS.

For a fixed choice of the noise variance σ_N and varying the number of delay coordinates M, the corresponding WERs are plotted in Fig. 3. The WER decreases with increasing number of delay coordinates. For dimensions large enough, the WER saturates, i.e. further enlargement of the subspace does not further improve the recognition performance. Enlargement of the width of the frames did not considerably alter the performance of the enhancement, i.e. with $N = 200$ the saturation is already reached.

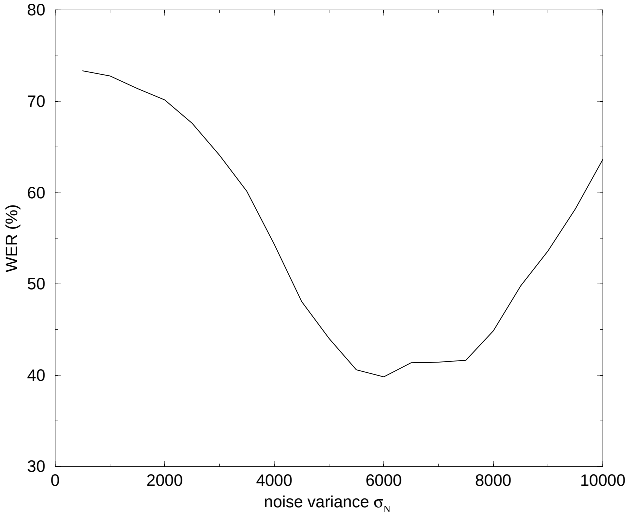

Fig. 2. WER as a function of adjusted noise variance σ_N.

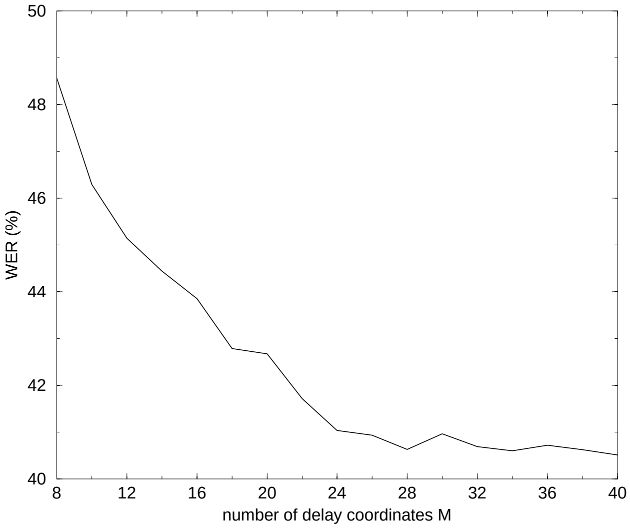

Fig. 3. WER varying the number of delay coordinates M.

4.2.3. Adaptive SVD variants

The only free parameter remaining in the MV-SVD algorithm, the noise variance σ_N, can be estimated online, as described in Sect. 3. Table 2 shows the resulting WERs in dependence of the time constant of the recursive filter for smoothing the noise variance estimation in the case of MV-

SVD speech enhancement.

T (frames)	0	10	20	50	100	200
WER (%)	45.4	42.7	42.3	42.3	41.9	41.8

Table 2. Online estimation of σ_N, variation of the time constant of the recursive filter, MV-SVD approach

For a constant T chosen large enough the WER is below 42%, still better than the spectral subtraction approach, and only slightly worse than the σ_N-optimized approach.

Finally, Figure 4 shows WERs for the more general variants of the adaptive SVD subspace approach depending on one parameter each: the TDC approach with variations of γ and the SDC approaches with variations of β_1 and β_2. The noise variance was estimated online with a time constant of the recursive filter $T = 100$. Improvements to the MV variant can be observed, with the best performance (39.3% WER) by the SDC1 variant and $\beta_1 = 3$.

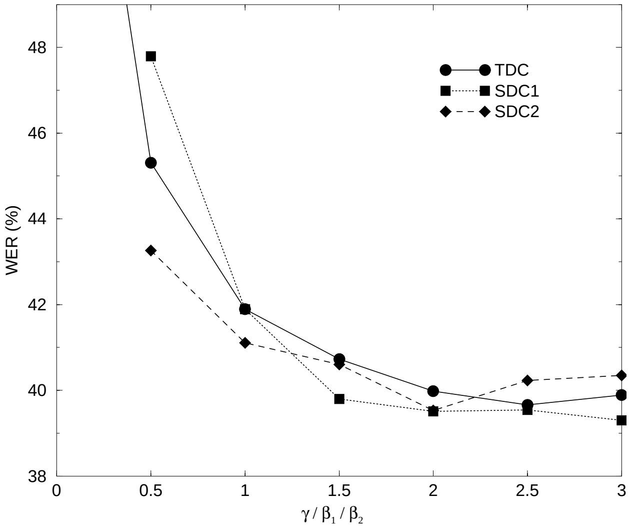

Fig. 4. Variation of γ for the TDC approach and variations of β for the SDC approaches, both with adaptive σ_N-estimation

4.3. Dictation via Telephone

The MV-SVD subspace method was also applied to real world data: dictations via telephone with a recognition based on clean speech trained references with the same setup as in Sect. 4.2 for 14 test speakers. The enhancement was performed by the adaptive approach with time constant $T = 100$. Tab. 3 shows some results with and without MV-SVD based speech enhancement.

An improvement up to 17.2% relative is observed, in cases of no improvement the enhancement does not (or only insignificantly) degrade the performance. This is remarkable, since the SVD approach assumes white noise degradation, and still – also for this real world scenario – an improvement is present.

WER (%)						
Speaker	A	B	C	D	. . .	avr.
no enhanc.	8.5	9.7	21.9	28.2	. . .	19.0
MV SVD	8.6	9.2	19.6	23.4	. . .	18.2

Table 3. WERs for dictations via telephone with and without adaptive MV SVD speech enhancement

5. CONCLUSIONS

Our findings confirm results [6] of a better noise robust ASR performance with the MV-SVD approach compared to NSS for weakly correlated noise scenarios. The main advantage relies on its robust performance with respect to parameter tuning. Especially the proposed adaptive approach needs no speech detection, essentially no parameter tuning and outperforms NSS in our applications.

Acknowledgment

We would like to thank D. Klakow and A. Fischer for helpful discussions and suggestions.

6. REFERENCES

[1] A. Fischer and V. Stahl, "On Improvement Measures for Spectral Subtraction applied to Robust Automatic Speech Recognition in Car Environments," in *Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.

[2] 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, July 1995.

[3] P. S. K. Hansen, *Signal Subspace Methods for Speech Enhancement*, Ph.D. thesis, Technical University of Denmark, Dep. of Math. Modelling, 1997.

[4] M. Lieb and R. Haeb-Umbach, "LDA derived cepstral trajectory filters in adverse environmental conditions," in *Proc. ICASSP*, 2000.

[5] J.-C. Junqua, "Impact of the unknown communication channel on automatic speech recognition: A review," in *Proc. EUROSPEECH*, 1997.

[6] K. Hermus, I. Dologlou, P. Warmbacq, and D. Van Compernolle, "Fully Adaptive SVD-Based Noise Removal for Robust Speech Recognition," in *Proc. EUROSPEECH*, Budapest, Hungary, Sept. 1999, pp. 1951–1954.