FULLY ADAPTIVE SVD-BASED NOISE REMOVAL FOR ROBUST SPEECH RECOGNITION

*Kris Hermus, Ioannis Dologlou, Patrick Wambacq and Dirk Van Compernolle**

Katholieke Universiteit Leuven – ESAT - PSI
Tel: +32 16 321829 - Fax: +32 16 321723
e-mail: Kris.Hermus@esat.kuleuven.ac.be
* Lernout & Hauspie Speech Products, Belgium

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

10.21437/Eurospeech.1999-429

ABSTRACT

This paper presents a new approach to improve the robustness of large vocabulary continuous speech recognition. The proposed technique – based on Singular Value Decomposition (SVD) – originates from classical signal enhancement, but it is adapted to the specific requirements imposed by the speech recognition process.

Additive noise reduction is obtained by altering the singular value spectrum of the signal observation matrix, thereby preserving speech signal components and suppressing noise-related components.

The basic algorithms are developed for white noise but they can easily be extended to the general coloured noise case. With the aid of a noise reference, non-stationary noise can be handled as well. All schemes are adaptive, and work in real-time.

Recognition experiments on a noise-corrupted database with large vocabulary, continuous speech (Resource Management) reveal that relative reductions of the WER[1] of more than 60 % are obtained.

1. INTRODUCTION

Current speech recognisers are able to transcribe fluently spoken language with a very high accuracy. WER's of less than 10 % are very common for speaker independent large vocabulary continuous speech recognisers.

However, substantial performance degradation is reported when the input speech is corrupted by disturbing sources. Without any compensation technique, recognisers become completely useless in real environments.

The basic reason for the robustness problem is the *mismatch* between the training and operating environments, leading to a discrepancy between the extracted parameters from the noisy signal and the clean signal acoustic models. Techniques to reduce this mismatch can mainly be classified as follows [1] :

- Adaptation of the speech signal to the training environment by *speech enhancement* techniques.
- Developing a noise-independent system by working with *noise resistant features*. Noise is not necessarily removed, but it is incorporated in both the parameters and the acoustic models.
- Adaptation of the acoustic models to the noisy input speech, which is called *model compensation*.

Despite the large research efforts, still a lot of restrictions on the allowed noise conditions are indispensable to obtain satisfactory robustness levels.

Indeed, removing noise in real environments – i.e. with unknown, rapidly varying noise sources – is very demanding. The need for an easy implementation (single microphone, ...) makes the problem even more challenging.

In this paper, we propose a new – SVD based – speech enhancement technique for robust speech recognition at moderate SNR ratios. Since the evaluation criterion is not an improvement in speech quality but an increase in recognition rate, adaptations to the classical schemes have to be made. Preserving the formant structure of the speech signal is one of the major issues.

We will explain that, starting from the classical signal enhancement algorithm, numerous schemes can be implemented with drastic reductions in error rates.

The next section describes the basic theory of SVD-filtering with its application to white noise removal. In section 3, the extension to the general coloured noise case is discussed. Experimental results can be found in section 4. Finally, a summary and conclusions are given in section 5.

2. THEORY of SVD-FILTERING

2.1. Introduction

Singular Value Decomposition (SVD) has been proven to be an efficient tool for signal processing techniques: image coding (eigenfaces), signal enhancement, image filtering, ...

Here we apply SVD to remove additive noise by altering the singular spectrum of the speech observation matrix.

[1] WER : Word Error Rate

2.2. Algorithm

The noisy signal is treated as a vector in a T-dimensional space, with noise and speech components lying in orthogonal subspaces. From the noisy signal, a Hankel matrix is constructed, whose singular spectrum is altered. The high energy components are supposed to contain only speech data, whereas the low energy components are supposed to contain only noise.

The noisy signal $y = y(0), y(1), \ldots$ consists of the desired signal x and the noise n; both components are considered being additive:

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

The basic noise reduction scheme consists of the following steps:

Framing All processing of the speech signal is frame based. Therefore, we create overlapping frames of T samples: $y(0), y(1), \ldots, y(T-1)$. The number of samples T is to be optimised.

Constructing the Hankel matrix From all frames, we construct a Hankel matrix Y:

$$Y = \begin{bmatrix} y(0) & y(1) & \ldots & y(M-1) \\ y(1) & y(2) & \ldots & y(M) \\ \vdots & \vdots & & \vdots \\ y(K-1) & y(K) & \ldots & y(T-1) \end{bmatrix} \tag{2}$$

with dimensions $K \times M$, with $K \geq M$ and $M+K = T+1$.

According to the assumption of additive noise, we can write Y as:

$$Y = X + N \tag{3}$$

SVD calculation We calculate the SVD of Y :

$$Y = U \Sigma V^T \tag{4}$$

Altering the singular value spectrum The largest singular components capture almost only signal information whereas the smallest ones contain almost only noise. By adapting the weights of the different singular components, noise reduction can be obtained:

$$\hat{X} = U\,(W\Sigma)\,V^T \tag{5}$$

with W a diagonal matrix containing the weights.

Restoring the Hankel structure Matrix $\hat{X}$ is not Hankel anymore; an easy extraction of the improved signal $\hat{x} = \hat{x}(0), \hat{x}(1), \ldots, \hat{x}(T-1)$ is impossible. However, the Hankel structure can be restored by constructing a new matrix $\overline{X}$ where every element from an anti-diagonal of $\hat{X}$ is replaced by the average value along that anti-diagonal.

2.3. FIR-filter representation

It was shown [2] that the overall procedure of altering the singular value spectrum is equivalent to a FIR-filtering operation on the noisy signal $y(k)$. Every singular component can be extracted – and hence given its own weight – by a FIR-filter, constructed from the corresponding right singular vector.

The interpretation in the frequency domain shows that the spectrum of $y(k)$ is decomposed into M components, and that each filtered component is given its proper weight.

2.4. Implementation

Depending on the applied weighting matrix W, different noise reduction algorithms can be developed. We discuss the basic Least Squares (LS) estimation, and the more powerful Minimum Variance (MV) estimation.

Least Squares estimation (rank reduction) We assume that x consists of p complex exponentials, such that X is of rank p. A *Least Squares / rank p* estimate of Y is obtained by setting the $M-p$ smallest eigenvalues to zero,

$$Y_{LS,p} = \begin{bmatrix} U_p & U_{M-p} \end{bmatrix} \begin{bmatrix} \Sigma_p & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_p^T \\ V_{M-p}^T \end{bmatrix} = U_p \Sigma_p V_p^T \tag{6}$$

with Σ_p containing the p largest singular values, and $Y_{LS,p}$ the best rank-p approximation of Y.

The Least Squares estimation has a major drawback. The method results in selecting some singular components, and suppressing the other components (binary approach). In the frequency domain, this is equivalent to filtering out some frequency bands, resulting in frequency dips and hence disappointing recognition results.

Minimum Variance estimation Much better results are obtained when every singular component adds up to the enhanced signal, giving the highest weights to the signal components, and the lowest weights to the components containing 'almost only' noise.

The weighting factor w_i for the i-th singular component with singular value σ_i equals

$$w_i = 1 - \frac{\sigma_{noise}^2}{\sigma_i^2} \tag{7}$$

where σ_{noise}^2 is estimated based on speechless frames.

The weighting matrix W becomes

$$W = diag\left(1 - \frac{\sigma^2_{noise}}{\sigma^2_1}, \ldots, 1 - \frac{\sigma^2_{noise}}{\sigma^2_M}\right) \qquad (8)$$

The mathematical formulation of this estimation, together with a geometrical interpretation, can be found in [3].

Since no singular components are removed from the signal, no frequency dips occur in the spectrum. The MV estimation does not destroy the formant structure and leads to much better recognition rates.

2.5. Applicability

The separation of signal and noise is based on the assumption that both components are orthogonal. Moreover, since the extraction of the signal is based on the correlation that is present in the signal, the noise should be totally uncorrelated to reach optimal performance. This means that the schemes as described above are *only* valid for the removal of *white noise*, i.e.

$$N^T N = \sigma^2_{noise} I \qquad (9)$$

3. COLOURED NOISE REMOVAL

The noise reduction schemes described so far are only valid for speech corrupted by white noise with a flat spectrum. However, the algorithms can be adapted easily for the removal of coloured noise if an estimate of the noise is available.

3.1. Noise Reference

An accurate estimate of the noise is indispensable to obtain significant noise reduction. Since we only allow a simple one-microphone recording, knowledge about the noise should be extracted from the noise during noise-only periods. This requires the presence of speechless frames in the input signal, such that a good estimation of the shape of the noise spectrum can be made.

For a fully automatic method we need:

speech/noise classifier A system that accurately detects those segments to be used for noise spectrum estimation. Misclassification leads to erroneous filtering of the signal, and probably makes the 'enhanced' signal unrecognisable until a new, correct noise shape is estimated.

stationarity Noise must be stationary during the speech segments, such that an estimate based on a current frame, is still valid for following frames with speech.

Once the noise spectrum is known, it can be included in the SVD method by applying a 'prewhitening' step.

3.2. Prewhitening

For the general coloured noise case, $N^T N \neq \sigma^2_{noise} I$. However, if N is available we can use the Cholesky factor from the QR-factorisation of N to prewhiten the noise.

To that end, we multiply the Hankel matrix Y with R^{-1} obtained from $N = QR$:

$$\hat{Y} = YR^{-1} = XR^{-1} + NR^{-1} \qquad (10)$$

such that the noise factor NR^{-1} of $\hat{Y}$ complies with (9). Indeed,

$$\left(NR^{-1}\right)^T \left(NR^{-1}\right) = Q^T Q = I \qquad (11)$$

The formulas derived for the white noise case can now be applied to $\hat{Y}$. Finally, a dewhitening step is performed by multiplying $\hat{Y}_{MV}$ or $\hat{Y}_{LS}$ with R.

3.3. Implementation by QSVD

Subsequent prewhitening and dewhitening can cause a loss of accuracy due to numerical instability. Working with the QSVD (Quotient SVD) of the pair (X, N) immediately leads to the SVD factorisation of YR^{-1}, without making quotients and products. More information can be found in [3].

4. RECOGNITION EXPERIMENTS

These noise reduction schemes were tested with the ESAT-speech recogniser on the Resource Management task (feb89 test set : continuous speech, 1 K words, Word Pair grammar, Context-Independent models).

For a description of the ESAT-speech recogniser, the reader is referred to [4].

4.1. Experimental scheme

Currently, the SVD-scheme precedes the normal MEL-cepstrum preprocessing. The frame-length and the order of the Hankel matrix are two important parameters that have to be optimised.

As the SVD noise removal can be seen as an adaptive filtering (with constant filter characteristics during one frame), the frame-length should be as small as possible to obtain a fast adaptation procedure. However, too short frame-lengths prohibit an accurate calculation of the SVD-filter bank. A frame-length of 30 msec (480 samples at 16 kHz sampling rate) is a good compromise.

The SVD exploits the speech signal correlation to separate it from the noise. This explains why the order M of the Hankel matrix must not be smaller than the order of the speech signal. A high frequency selectivity implies using a high M.
On the other hand, M should be kept as small as possible because the computational load of the SVD increases as $O(M^4)$.

During experiments, an order M of 8 was found to be optimal.

4.2. Experimental results

Four databases were constructed from the clean RM-data (with WER 4.88 %) by adding additive white and coloured noise (low-pass filtered white noise, cut-off frequency = 4 kHz), at 10 and 15 dB SNR.

Least Squares Estimation As already discussed, discarding some singular components[2], may suppress some frequency bands that were present in the original signal (decrease in intelligibility). At the same time, the formant structure can be lost, which compromises the discriminativity between the different phonemes. Mostly, this cannot be compensated by an overall increase in SNR of the speech signal (increase in speech quality).

Experiments revealed the shortcomings of the Least Squares estimation. Iteration of the basic SVD noise reduction step increases the signal enhancement, but the best obtained result was a relative reduction in WER of 15 %. Therefore, we concentrated on the more powerful Minimum Variance estimation.

Minimum Variance Estimation The WER's for the Minimum Variance estimation noise removal are summarised in table 1. Following specifications were used: frame length = 30 msec, order $Y = 8$.

These results are compared with those obtained by two other well-known techniques: Nonlinear Spectral Estimation [5] (NSE) and SNR-Normalisation [5] (SNR-N.). For SNR-N., which requires retraining with the noisy data, we normalised towards the optimal values of 12 and 18 dB for the 10 and 15 dB SNR database respectively.

Detailed analysis showed that (1) the formant structure was indeed preserved, (2) that the result is rather insensitive to the order of the Hankel matrix for M ranging from 8 to 20, and (3) that an ill-considered choice of the frame length can lead to significant deterioration of the results.

From the table we learn that our new approach is far superior to existing techniques as NSE and SNR-N. We get relative reductions of the WER of more than 65 % for white noise and of more than 55 % for coloured noise.

SNR ratio	Ref.	NSE	SNR-N.	MV
white noise				
10 dB	57.17	41.74	39.79	17.96
15 dB	25.77	21.36	21.71	8.79
coloured noise				
10 dB	56.19	34.21	39.24	23.55
15 dB	20.62	17.96	17.13	9.22

Table 1: *Word Error Rates for Minimum Variance SVD-noise removal. Ref: Reference (no noise reduction); NSE: Nonlinear Spectral Estimation; SNR-N.: SNR-Normalisation; MV: Minimum Variance.*

Preliminary tests on speech corrupted by non-stationary coloured noise give promising results, but at the same time indicate that removing rapidly varying noise requires an accurate speech/noise classifier.

[2] also called 'truncated SVD'

5. CONCLUSION

In this paper, we presented a new, SVD-based approach to tackle the severe problem of large vocabulary continuous speech recognition in noisy environments.

Starting from the basic SVD-noise reduction scheme, numerous powerful implementations are at hand, leading to significant noise reduction and remarkable WER reductions.

Due to the prewhitening of the noise, the technique remains valid for the general coloured noise case. Non-stationary noise can be removed with the aid of an accurate noise reference, estimated from noise-only frames. Moreover, the computational load is rather low, and the implementation in C works in real-time.

Recognition experiments showed that our new approach leads to drastic reductions in error rates, thereby outperforming well-known approaches as Nonlinear Spectral Estimation and SNR-Normalisation for robust speech recognition.

Further research focuses on improved (fast and accurate) adaptation to real environments, and on optimal integration of the SVD-scheme in the global preprocessing.

ACKNOWLEDGEMENT

This research is supported by the GOA Programme (K.U. Leuven Research Fund).

REFERENCES

1. Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16(3):261–291, April 1995.

2. P. C. Hansen and S. H. Jensen. FIR filter representations of reduced-rank noise reduction. In *IEEE Transactions on Signal Processing*, September 1996. Accepted for Publication.

3. S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sørensen. Reduction of broad-band noise in speech by truncated QSVD. *IEEE Transactions on Speech and Audio Processing*, 3:439–448, November 1995.

4. J. Duchateau, K. Demuynck, and D. Van Compernolle. Fast and accurate acoustic modelling with semi-continuous HMMs. *Speech Communication*, 24(1):5–17, April 1998.

5. T. Claes, F. Xie, and D. Van Compernolle. Spectral estimation and normalisation for robust speech recognition. In *Proc. International Conference on Spoken Language Processing*, volume IV, pages 1997–2000, Philadelphia, U.S.A., October 1996.