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

SINGLE CHANNEL SPEECH ENHANCEMENT USING PRINCIPAL COMPONENT ANALYSIS AND MDL SUBSPACE SELECTION

Rolf Vetter, Nathalie Virag, Philippe Renevey and Jean-Marc Vesin
Signal Processing Laboratory
Swiss Federal Institute of Technology, Lausanne
CH–1015 Lausanne, Switzerland
Rolf.Vetter@epfl.ch

10.21437/Eurospeech.1999-529

ABSTRACT

We present in this paper a novel subspace approach for single channel speech enhancement and speech recognition in highly noisy environments. Our algorithm is based on principal component analysis and the optimal subspace selection is provided by a minimum description length criterion. This choice overcomes the limitations encountered with other selection criteria, like the overestimation of the signal–plus–noise subspace or the need for empirical parameters. We have also extended our subspace algorithm to take into account the case of colored noise. The performance evaluation shows that our method provides a higher noise reduction and a lower signal distortion than existing enhancement methods and that speech recognition in noise is improved. Our algorithm succeeds in extracting the relevant features of speech even in highly noisy conditions without introducing artefacts such as "musical noise".

1. INTRODUCTION

The performance of automatic speech processing systems degrades drastically in noisy environments. Therefore, several single channel enhancement algorithms using the Discrete Fourier Transform (DFT), such as subtractive–type approaches [1, 2] or Wiener filtering, have been developed. The major problem with most of these methods is that they suffer from a distortion called "musical noise". To reduce this distortion, the DFT can be replaced by the Discrete Cosine Transform (DCT) [3] or the Karhunen–Loève Transform (KLT) [4]. The basic idea in these subspace approaches is to observe the data in a large m–dimensional space of delayed coordinates. Since noise is assumed to be random, it extends approximately in uniform manner in all the directions of this space, while in contrast, the dynamics of the deterministic system underlying the speech signal confine the trajectories of the useful signal to a lower–dimensional subspace of dimension $p < m$. Consequently, the eigenspace of the noisy signal is partitioned into a noise and a signal–plus–noise subspace. Enhancement is obtained by nulling the noise subspace and optimally weighting the signal–plus–noise subspace.

In this paper we propose a novel subspace approach for single channel speech enhancement and recognition in highly noisy environments based on the KLT, and implemented via Principal Component Analysis (PCA) [5]. This choice is motivated by the fact that the KLT provides an optimum compression of information, while the DFT and the DCT are suboptimal. The main problem in subspace approaches is the optimal choice of the different parameters. We propose therefore a novel approach for the optimal subspace partition using the Minimum Description Length (MDL) criterion [6]. This criterion provides consistent parameter estimates and allows us to implement an automatic noise reduction algorithm that can be applied almost blindly to the observed data.

2. PROPOSED SUBSPACE APPROACH

2.1. Principal Component Analysis

Consider a speech signal $s(t)$ corrupted by an additive stationary background noise $n(t)$. The observed noisy signal can be expressed as follows:

$$x(t) = s(t) + n(t) \qquad (1)$$

Our noise reduction algorithm operates on a frame–by–frame basis and the general enhancement scheme is represented in Figure 1.

A very efficient and robust implementation of the subspace approach is provided by the PCA of the following m–dimensional vector, obtained by an embedding in the space of the delayed coordinates [5]:

$$\begin{gathered}\mathbf{x}(t) = [x(t), x(t-1), \ldots, x(t-lag)]^T \\ t = lag, \ldots, N-1\end{gathered} \qquad (2)$$

where N is the frame size, $m = (lag + 1)$ is the embedding dimension and lag has to be chosen to get an optimal value for m. We assume that the data set has zero mean and consider the following orthogonal transformation:

$$\mathbf{x}(t) = \sum_{j=1}^{m} a_j(t)\mathbf{\Phi}_j \qquad (3)$$

The goal of PCA is to find a set of vectors $\mathbf{\Phi}_j$ for $j = 1, \ldots, m$, to obtain a maximum decrease rate of the variance of the projections $\langle \mathbf{x}(t)^T \mathbf{\Phi}_j \rangle$. This results in an eigenvalue problem of the estimated covariance matrix $\mathbf{C} = \langle \mathbf{x}(t)\mathbf{x}(t)^T \rangle$. Since $\mathbf{C}$ is a $m \times m$ symmetric non-negative matrix it determines a complete set of orthogonal eigenvectors, associated with real, non-negative, eigenvalues. These eigenvalues can be ordered $\lambda_1 \geq \lambda_2 \geq \lambda_3, \ldots, \geq \lambda_m$ and the statistical variance of the data set in the direction of the the j^{th} eigenvector $\mathbf{\Phi}_j$ is proportional to the eigenvalue λ_j. The $a_j(t)$ coefficients, called the *principal components*, can be found by projecting the data vectors onto each eigenvector in turn:

$$a_j(t) = \mathbf{x}^T(t)\mathbf{\Phi}_j \quad j = 1, \ldots, m \qquad (4)$$

Noise reduction can be achieved by reconstructing the initial data using only the p weighted eigenvectors of the signal–plus–noise subspace, as proposed by Ephraim et al. in [4]:

$$\hat{\mathbf{s}}(t)_{Eph95} = \sum_{j=1}^{p} g_j a_j(t)\mathbf{\Phi}_j \quad p < m \qquad (5)$$

where g_j is a weighting function given by:

$$g_j = \exp\{-\nu\sigma_n^2/\lambda_j\} \quad j = 1, \ldots, p \qquad (6)$$

with $\nu = 5$. The parameters m and p are generally chosen in such a way that the noise is essentially relegated to the residuals of the signal approximation given by Equation (5).

2.2. Subspace Partitioning

The optimal design of a PCA–based noise reduction algorithm for speech enhancement is a difficult task. The parameters m and p should be chosen in optimal manner through appropriate selection rules. Furthermore, the use of a weighting function g_j in Equation (5) introduces a considerable amount of speech distortion. Therefore, in order to simultaneously maximize noise reduction and minimize signal distortion, we present in this paper a more promising approach consisting in a partition of the eigenspace of the noisy data into 3 different subspaces (see Figure 1):

1. A noise subspace which contains mainly noise contributions. These components are nulled during reconstruction.

2. A signal subspace which contains principal components $a_j(t)$ with a high signal–to–noise ratio $\mathrm{SNR}_j \gg 1$. Components of this subspace are not weighted since they contain mainly components from the original signal. This allows a minimization of the signal distortion.

3. A signal–plus–noise subspace which includes the components $a_j(t)$ with $\mathrm{SNR}_j \approx 1$. The estimation of its dimension can only be done with a high error probability. Consequently, principal components with $\mathrm{SNR}_j < 1$ may belong to it and a weighting is applied during reconstruction.

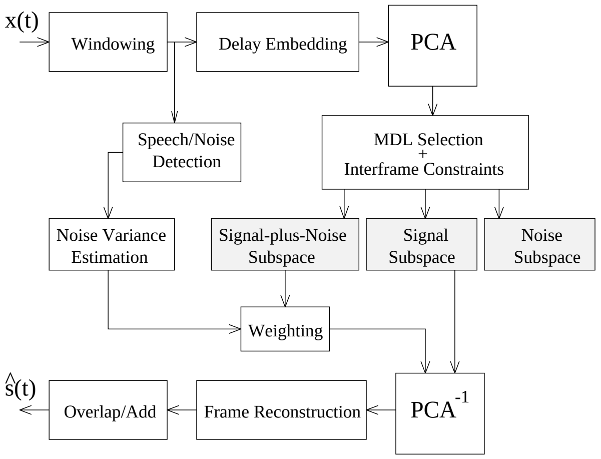


Figure 1: *The proposed enhancement algorithm.*

Using this new partition, the reconstructed signal is given by:

$$\hat{\mathbf{s}}(t) = \sum_{j=1}^{p_1} a_j(t)\,\mathbf{\Phi}_j + \sum_{j=p_1+1}^{p_2} g_j a_j(t)\,\mathbf{\Phi}_j \quad p < m \qquad (7)$$

where g_j is the weighting function given by Equation (6). We note that the proposed approach requires the determination of the parameters p_1 and p_2. The parameter p_1 should provide a very parsimonious representation of the signal whereas p_2 should also select components with $\mathrm{SNR}_j \approx 1$. We will see in the next section that the framework of MDL allows us to deduce an appropriate estimator.

A crucial point is the adequate choice of the embedding dimension m of the PCA. In this paper we use a rule for the determination of m that has been proposed in the context of singular spectrum analysis [7]. It is applicable if the useful signal is constituted of quasi–periodic contributions of a bandwidth Δf_x and is given by:

$$m < min\{1/\Delta f_x \quad (N/3 + 1)\} \qquad (8)$$

For speech signals, we found that an optimal value for m is in the range from 40 to 80.

2.3. Subspace Selection based on MDL

The determination of the number of relevant principal components p_1 and p_2 in Equation (7) requires the use of a truncation criterion applicable for short time series. Among the possible selection criteria, the MDL criterion has been shown in multiples domains to be a consistent model order estimator especially for short time series [5, 8]. MDL selects the model that produces the minimum code length for the given data. If we apply the MDL criterion proposed in [6] to the PCA selection problem, we obtain in the case of additive white Gaussian noise:

$$\begin{aligned} MDL(p_i) &= -\ln \left\{ \frac{\prod_{j=p_i+1}^{m} \hat{\lambda}_j^{\frac{1}{m-p_i}}}{\frac{1}{m-p_i} \sum_{j=p_i+1}^{m} \hat{\lambda}_j} \right\}^{(m-p_i)N} \\ &\quad + M \cdot \left(\tfrac{1}{2} + \ln[\gamma]\right) \\ &\quad - \tfrac{M}{p_i} \sum_{j=1}^{p_i} \ln \left[\hat{\lambda}_j \sqrt{2/N}\right] \end{aligned} \tag{9}$$

where $i = 1, 2$ and $M = p_i m - p_i^2/2 + p_i/2 + 1$ is the number of free parameters. The parameter γ determines the selectivity of MDL. Accordingly, p_1 and p_2 are given by the minimum of $MDL(p_i)$ with $\gamma = 64$ and $\gamma = 1$ respectively. Furthermore, interframe constraints have been introduced to increase the consistency of the MDL estimator.

2.4. Signal Reconstruction

The reconstructed speech signal computed with Equation (7) for $t = lag, \ldots, N-1$ leads to the following $m \times (N-m-1)$ matrix:

$$\begin{bmatrix} \hat{s}(lag) & \hat{s}(lag-1) & \ldots & \hat{s}(0) \\ \hat{s}(lag+1) & \hat{s}(lag) & \ldots & \hat{s}(1) \\ \ldots & \ldots & \ldots & \hat{s}(2) \\ \ldots & \ldots & \ldots & \ldots \\ \hat{s}(2*lag) & \ldots & \ldots & \hat{s}(lag) \\ \ldots & \ldots & \ldots & \ldots \\ \hat{s}(N-1) & \ldots & \ldots & \hat{s}(N-1-lag) \end{bmatrix}$$

We observe that there exists m different candidates for each time sample of the reconstructed signal. Therefore, we performed a statistical averaging over all m candidates that further improved the accuracy of the reconstructed sample. Experimental results showed that improvements up to 3 dB can be obtained.

3. PERFORMANCE EVALUATION

3.1. Compared Algorithms and Databases

For the performance evaluation, we have compared the following single channel enhancement algorithms:

1. *NSS*: nonlinear spectral subtraction using the DFT [2].
2. *Eph95*: subspace approach by Ephraim et al. using the KLT [4]. This approach has been developed for white Gaussian noise only.
3. *PCA–MDL*: proposed subspace approach.

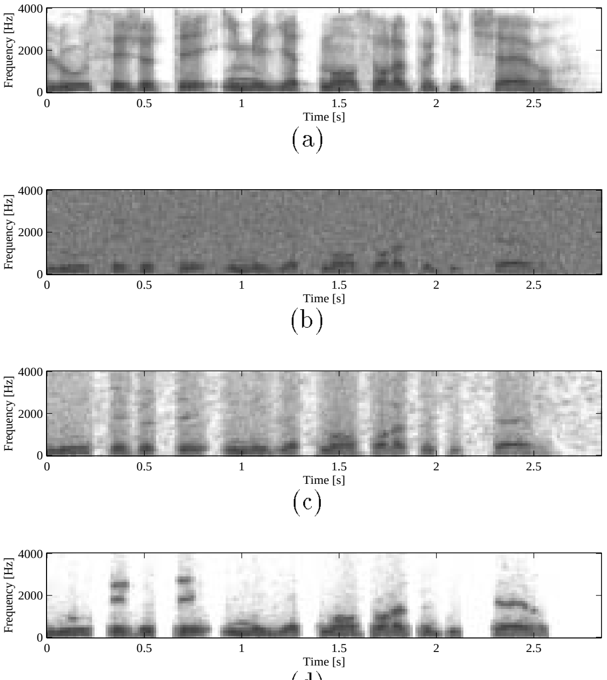
The testing database has been created by adding different types of background noises from the Noisex database to the clean speech signals, at SNRs ranging from -6 dB to ∞ dB. The sampling frequency is 8 kHz. The frame size is N=400 and we apply Hanning windowing with 50 % overlap.

3.2. Enhancement in White Gaussian Noise

We have based our performance evaluation on the segmental SNR, the Itakura–Saito distortion measure (IS), the observation of the spectrograms as well as informal listening tests. We have observed that generally subspace approaches based on the PCA (*Eph95* and *PCA–MDL*) outperform linear and nonlinear subtractive–type methods using DFT. In particular, the use of a subspace approach significantly reduces the "musical noise".

Noisy		Eph95		PCA–MDL	
SNR	IS	SNR	IS	SNR	IS
0 dB	6.2	6.5 dB	4.1	8.8 dB	3.2
6 dB	5.1	10.5 dB	3.2	12.6 dB	3.1
18 dB	2.2	21.9 dB	1.1	22 dB	0.9

Table 1: *Segmental SNR and Itakura–Saito measure in the case of white Gaussian noise.*



(a)

(b)

(c)

(d)

Figure 2: *Speech spectrograms: (a) original French speech signal: Un loup c'est jeté immédiatement sur la petite chèvre , (b) noisy signal (additive white Gaussian noise at an input SNR = 3 dB), enhanced signals using (c) Eph95, (d) PCA–MDL.*

If we compare the subspace approaches, we can see in Table 1 that our method provides similar performance with respect to *Eph95* for high input SNRs. However, it leads to a higher noise reduction and a lower signal distortion (smaller value of IS) for low SNRs. This effect is confirmed in the spectrogram of Figure 2. Indeed, our method better extracts the relevant features of the speech signal. These results highlight the efficiency and consistency of the MDL based subspace algorithm. Furthermore, this approach does not require parameter tuning based on empirical considerations. One important additional feature of our method is that it is highly efficient in detecting speech pauses, even in very noisy conditions.

3.3. Enhancement in Colored Noise

In order to be able to apply the MDL selection approach to colored noises, we have to modify the covariance matrix $\mathbf{C}$ of the noisy data by taking into account the covariance matrix of noise computed during speech pauses. This leads to the results presented in Table 2 for helicopter cockpit noise. We can see, that even in this case, our method provides good performance over subtractive–type algorithms.

Noisy		NSS		PCA–MDL	
SNR	IS	SNR	IS	SNR	IS
0 dB	3.1	5.2 dB	3	6.7 dB	2.4
6 dB	2.1	10.1 dB	1.9	10.9 dB	1.1
18 dB	0.5	20.2 dB	0.4	20.5 dB	0.3

Table 2: *Segmental SNR and Itakura–Saito measure in the case of helicopter cockpit noise.*

3.4. Speech Recognition in Noise

We have applied our enhancement algorithm as a preprocessing stage to speech recognition in noise. We have used a speech recognizer which has been designed and trained on clean speech for the isolated digit recognition. The recognizer has been built up by the HTK HMM toolkit version 2.1. The features for speech recognition are the 12 MFCC and the energy, together with the first and second order derivatives of these 13 parameters. The training database is constituted of 400 recordings of 7 digits. The general model for the isolated digit recognition consists of a model for silence between the digits (3 emitting states). The testing database contains 50 sequences of 7 digits with additive white Gaussian noise.

Input SNR	Noisy	NSS	Eph95	PCA–MDL
−6 dB	16 %	20 %	27 %	37 %
0 dB	20 %	31 %	39 %	44 %
6 dB	35 %	50 %	60 %	68 %

Table 3: *Correctness of recognition in the case white Gaussian noise.*

Table 3 gives the recognition results in terms of correctness for the compared algorithms. These results underline that our method allows an extraction of the relevant features of speech even in highly noisy conditions.

4. CONCLUSION

We have presented in this paper a novel subspace approach for single channel speech enhancement and speech recognition in highly noisy environments. This approach is based on PCA and an associated MDL subspace selection. The performance evaluation based on segmental SNR, Itakura–Saito distortion measure, observation of the spectrograms, as well as informal listening tests, show clearly that our algorithm provides a lower signal distortion and a higher noise reduction that existing enhancement methods based on traditional subspace approaches. Finally, our enhancement method has been tested as a preprocessing stage to speech recognition for several noises and SNRs, showing a significant improvement for recognition over the methods under comparison.

REFERENCES

[1] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system", *IEEE Trans. on Speech and Audio Proc.*, Vol. 7, No. 2, pp. 126–137, March 1999.

[2] P. Lockwood and J. Boudy, "Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and projection, for robust recognition in cars", *Speech Communications*, Vol. 11, No. 2-3, pp. 215–228, June 1992.

[3] I.Y. Soon, S.N. Koh, and C.K. Yeo, "Noisy speech enhancement using Discrete Cosine Transform", *Speech Communication*, Vol. 24, No. 3, pp. 249–257, June 1998.

[4] Y. Ephraim and H.L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Trans. on Speech and Audio Proc.*, Vol. 3, No. 4, pp. 251–266, July 1995.

[5] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1991.

[6] K. Judd and A. Mees, "On selecting models for nonlinear time series", *Physica D*, Vol. 82, pp. 426–444, 1995.

[7] R. Vautard, P.Yiou and M. Ghil, "Singular spectrum analysis: A toolkit for short, noisy chaotic signals", *Physica D*, Vol. 58, pp. 395–424, 1992.

[8] R. Vetter, P. Celka, J.-M. Vesin, G. Thonet, E. Pruvot, M. Fromer, U. Scherrer and L. Bernardi, "Subband modeling of the human neurocardiovascular system: new insights into cardiovascular regulation", *Ann. Biomed. Eng.*, Vol. 26, pp. 293–307, 1998.