SINGLE CHANNEL SPEECH ENHANCEMENT USING MDL-BASED SUBSPACE APPROACH IN BARK DOMAIN

Rolf Vetter

Centre Suisse d'Electronique et de Microtechnique, CSEM
CH–2007 Neuchâtel, Switzerland
e-mail: Rolf.Vetter@csem.ch

ABSTRACT

We present in this paper a novel algorithm for single channel speech enhancement. It is based on a subspace approach in the Bark domain and an optimal subspace selection by the minimum description length (MDL) criterion. The processing in the Bark domain allows us to take into account in an optimal manner the masking properties of the human auditory system. The subspace selection provided by the MDL criterion overcomes the limitations encountered with other selection criteria, like the overestimation of the signal–plus–noise subspace or the need for empirical parameters. Together, the resulting MDL-subspace approach in the Bark domain provides maximum noise reduction while minimizing signal distortions. The performance of our algorithm is assessed in white and colored noise. It shows that our algorithm provides high performance for a large scale of input signal-to-noise ratio.

1. INTRODUCTION

Speech enhancement is often necessary to reduce listener's fatigue or to improve the performance of automatic speech processing systems. 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" [7]. To reduce this distortion one can replace the DFT by the Discrete Cosine Transform (DCT) [3]. Further improvements have been achieved by using perceptual properties of the human auditory system [1] or eigenspace approaches based on Karhunen–Loève Transform (KLT) [4, 5]. Notably, it has been shown that highest performance is obtained when using KLT with an associated subspace selection using the Minimum Description Length (MDL) criterion [6, 5]. Nevertheless, such an approach is not appropriate for real time implementation since the eigenvectors or eigenfilters have to be computed during each frame, which implies high computational requirements. To circumvent this drawback we use prior knowledge about perceptual properties of the human auditory system, that is, we substitute eigenfilters in the KLT approach by the so-called Bark filters [7]. Since it has been shown that the discrete cosine transform (DCT) outperforms the discrete Fourier Transform (DFT) in terms of speech energy compaction [3], we perform the Bark filtering in the DCT domain. Thus, our approach consists of a MDL based subspace approach in the DCT-Bark domain. As a such, it represents a merging of three well known single channel speech enhancement algorithms, namely KLT based subspace approaches [4, 5], speech enhancement based on masking properties of the human auditory system [1] and speech enhancement using DCT [3]. Our algorithm yields the robustness of the KLT based subspace approach together with the low computational requirements of the DCT and the high perceptual performance due to the inclusion of noise masking. The statistical robustness of the algorithm is ensured by the MDL criterion which provides a consistent parameter estimator 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. Global Framework for a Subspace Approach

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

$$x(t) = s(t) + n(t) \qquad t = 0, \ldots, N_t - 1 \qquad (1)$$

where N_t is the number of observed samples. We propose in this paper a subspace algorithm operating on a frame-by-frame basis with a frame length of N samples (see Figure 1). In a general way we can formulate the basic idea in subspace approaches as follows: the noisy data is observed in a large m–dimensional space of a given dual domain (for example eigenspace computed by KLT [4]). If the noise is random and white, it extends approximately in a uniform manner in all the directions of this dual domain, 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 $p < m$. As a consequence, 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 [4].

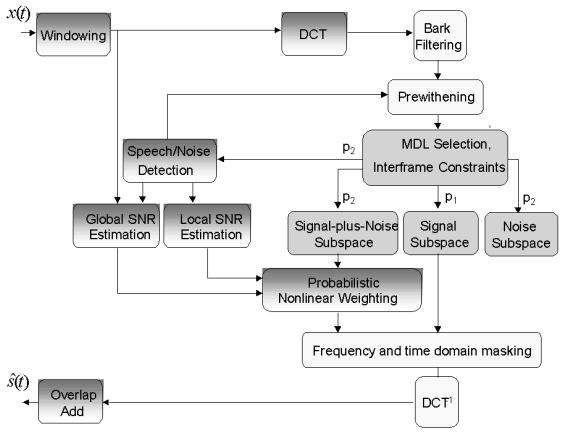

Figure 1: *The proposed enhancement algorithm.*

The optimal design of such a subspace algorithm is a difficult task. The subspace dimension p should be chosen during each frame in an optimal manner through an appropriate selection rule. Furthermore, the weighting of the signal–plus–noise subspace introduces a considerable amount of speech distortion. In order to simultaneously maximize noise reduction and minimize signal distortion, we have presented recently a promising approach consisting in a partition of the eigenspace of the noisy data into 3 different subspaces [5](see Figure 1):

1. A noise subspace of dimension $m - p_2$, which contains mainly noise contributions. These components are nulled during reconstruction.

2. A signal subspace of dimension p_1 containing components with high signal–to–noise ratios $\text{SNR}_j \gg 1$. 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 of dimension $p_2 - p_1$, which includes the components with $\text{SNR}_j \approx 1$. The estimation of its dimension can only be done with a high error probability. Consequently, components with $\text{SNR}_j < 1$ may belong to it and a weighting is applied during reconstruction.

In classical subspace approaches components of the dual domain are obtained by applying the eigenvectors or eigenfilters computed by KLT on the delay-embedded noisy data [4, 5]. To avoid the large computational means required for these operations, we use in this paper masking properties of the human auditory system in order to substitute the eigenfilters by the so-called Bark filters [5].

2.2. Bark Filtering using Masking Properties of the Human Auditory System

Noise masking is a well known feature of the human auditory system. It denotes the fact that the auditory system is incapable to distinguish two signals close in the time or frequency domains. This is manifested by an elevation of the minimum threshold of audibility due to a masker signal, which has motivated its use in the enhancement process to mask the residual noise and/or signal distortion [1]. The most applied property of the human ear is *simultaneous masking*. It denotes the fact that the perception of a signal at a particular frequency by the auditory system, is influenced by the energy of a perturbating signal in a critical band around this frequency. Furthermore, the bandwidth of a critical band varies with frequency, beginning at about 100 Hz for frequencies below 1 kHz, and then increasing up to 1 kHz for frequencies above 4 kHz. From the signal processing point of view the simultaneous masking is implemented by a critical filterbank, the so-called *Bark filterbank*, which gives equal weight to portions of speech with the same perceptual importance [7]. This prior knowledge about the human auditory system can be used to replace the eigenfilters in the KLT approach by Bark filtering. In order to have a maximum energy compacting the filtering is processed in the DCT domain [3]. Since Bark filtering is based on energy considerations we use the square of the DCT components. Finally we obtain the Bark components by

$$X(k)_{Bark} = \sum_{j=-b/2}^{b/2} G(j,k) \left\{X(k)\right\}^2 \quad k = 0, \ldots, N-1 \tag{2}$$

where $b+1$ is the processing-width of the filter, $G(j,k)$ is the Barkfilter whose bandwidth depends on k and $X(k)$ are DCT components defined as:

$$X(k) = \alpha(k) \sum_{t=0}^{N-1} x(t) cos \left\{ \frac{\pi(2t+1)k}{2N} \right\} \tag{3}$$

where $\alpha(0) = \sqrt{1/N}$ and $\alpha(k) = \sqrt{2/N}$ for $k \neq 0$. More details about DCT and its application in speech enhancement can be found in [3]. At this point it is important to note that by computing dual domain components as given by Equation (2), we obtain a dual domain of dimension $m = N$.

2.3. Subspace Selection based on MDL

A crucial point in the proposed algorithm is the adequate choice of the dimensions of the signal–plus–noise (p_2) and the signal subspace (p_1). It requires the use of a truncation criterion applicable for short time series. Among the possible selection criteria, the MDL criterion has been shown in multiple domains to be a consistent model order estimator, especially for short time series [8, 9, 5]. This high reliability and robustness of the MDL criterion constitutes the primer motivation for its use in our approach. To achieve this task, we assume that the Bark components given by Equation (2) rearranged in decreasing order constitute a liable approximation of the principal components

of speech. Under this assumption the following expression is obtained for the MDL in the case of additive white Gaussian noise [5]:

$$\begin{aligned} MDL(p_i) &= -\ln\left\{\frac{\prod_{j=p_i+1}^{N}\lambda_j^{\frac{1}{N-p_i}}}{\frac{1}{N-p_i}\sum_{j=p_i+1}^{N}\lambda_j}\right\}^{(N-p_i)N} \\ &\quad +M\cdot\left(\tfrac{1}{2}+\ln[\gamma]\right) \\ &\quad -\tfrac{M}{p_i}\sum_{j=1}^{p_i}\ln\left[\lambda_j\sqrt{2/N}\right] \end{aligned} \tag{4}$$

where $i = 1, 2$, $M = p_i N - p_i^2/2 + p_i/2 + 1$ is the number of free parameters and λ_j for $j = 0, \ldots, N-1$ are the Bark components given by Equation (2) rearranged in decreasing order. The parameter γ determines the selectivity of MDL. Accordingly, the dimension of the signal p_1 and the signal–plus–noise subspace p_2 are given by the minimum of $MDL(p_i)$ with $\gamma = 64$ and $\gamma = 1$ respectively. This choice of γ involves that the parameter p_1 provides a very parsimonious representation of the signal whereas p_2 selects also components with $\text{SNR}_j \approx 1$. In order to illustrate the efficiency of the MDL based subspace selection we show an example in Figure 2e. Its analysis highlights an important feature of our method, namely a null signal subspace for frames without any speech activity, which yields very reliable speech/noise detector. This information is then used in our algorithm to update the Bark spectrum and the variance of noise during frames without any speech activity, which ensures eventually an optimal signal prewithening and weighting. Notably, it has to be pointed out that the prewhitening of the signal is important since MDL assumes white Gaussian noise [5].

2.4. Reconstruction of Enhanced Signal

The enhanced signal is obtained by applying the inverse DCT to components of the signal and weighted components of the signal–plus–noise subspace. Using the definition of the inverse DCT [3] it can be written as:

$$\hat{\mathbf{s}}(t) = \sum_{j=1}^{p_1} a_{I_j}(t)\,\Phi_j + \sum_{j=p_1+1}^{p_2} g_j\, a_{I_j}(t)\,\Phi_j \tag{5}$$

with

$$a_k(t) = \alpha(k)cos\left\{\frac{\pi(2t+1)k}{2N}\right\}\Phi_j = \lambda_j^{1/2}\exp^{arg\{X(I_j)\}} \tag{6}$$

where λ_j for $j = 1, \ldots, N$ are the Bark components given by Equation (2) rearranged in decreasing order, I_j is the index of rearrangement and g_j is an appropriated weighting function given by:

$$g_j = \exp\{-\nu/SNR_j\} \quad j = p_1+1, \ldots, p_2 \tag{7}$$

where the parameter ν is adjusted through a nonlinear probabilistic operator in function of the global SNR and SNR_j for $j = 0, \ldots, N-1$ is the estimated signal-to-noise ratio of each Bark component.

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 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. *BARK–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 0 dB to 20 dB. The sampling frequency is 8 kHz, the frame size N=256 samples and we apply Hanning windowing with 50 % overlap. The performance evaluation is based on the segmental signal-to-noise ratio (SNR) and the Itakura–Saito distortion measure (IS), the observation of the spectrograms as well as informal listening tests. To obtain a relevant performance assessment we have computed the mean value of SNR and IS after discarding frames without any speech activity.

Generally, we have observed that subspace approaches outperform linear and nonlinear subtractive–type methods using DFT. In particular, subspace approaches yield a considerable reduction of the so-called "musical noise". This observation is confirmed quantitatively in Table 1 in the case of colored noise, since smaller IS values have been obtained for *BARK–MDL* than for *NSS*. In a qualitative way, this observation has been confirmed by informal listening tests but also through inspection of the spectrograms in Figure 2. Indeed, the analysis of Figure 2c highlights that *NSS* provides a considerable amount of residual "musical noise". In contrast, Figure 2d underlines the high performance of the proposed approach since it extracts the relevant features of the speech signal and reduces the noise to a tolerable level. This high performance confirm previous results on the efficiency and consistency of the MDL based subspace algorithms[5].

Noisy		Bark–MDL		Eph95	
SNR	IS	SNR	IS	SNR	IS
5 dB	4.16	11.4 dB	1.68	10.2 dB	1.81
10 dB	3.12	13.5 dB	0.95	13.9 dB	1.31
15 dB	1.81	17.3 dB	0.68	17.5 dB	0.74

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

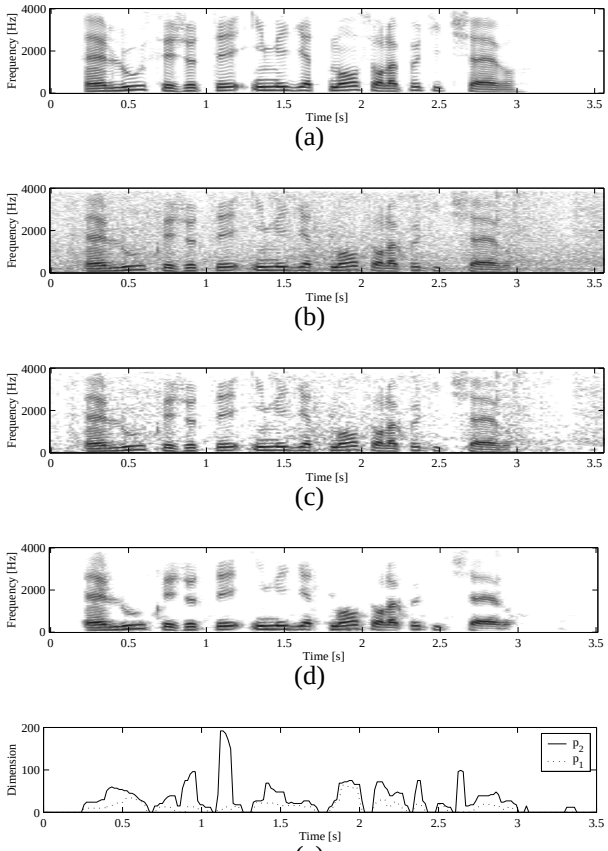

Figure 2: *Speech spectrograms: (a) original French speech signal: Un loup s'est jeté immédiatement sur la petite chèvre , (b) noisy signal (non-stationary factory noise at an segmental input $SNR = 10$ dB), enhanced signals using (c) NSS, (d) BARK–MDL, (e) signal and signal–plus–noise subspace dimension estimated by MDL.*

If we compare the subspace approaches, we can see in Table 1 that our method provides similar performance with respect to *Eph95*. However, it has to be pointed out that the computational load of *BARK-MDL* is reduced by an order of magnitude with respect to *Eph95*. Furthermore, an important additional feature of our method is that it is highly efficient and robust in detecting speech pauses, even in very noisy conditions. This can be observed in Figure 2e, for the signal subspace dimension is zero during frames without any speech activity.

4. CONCLUSION

We have presented in this paper a novel subspace approach for single channel speech enhancement in adverse noisy environments. Our method is based on a subspace approach in the Bark domain with a subspace selection provided by the MDL criterion. The performance evaluation based on segmental SNR, Itakura–Saito distortion measure, observation of the spectrograms, as well as informal listening tests, shows that our algorithm provides similar performance as eigenspace approaches based on KLT.

Noisy		Bark–MDL		NSS	
SNR	IS	SNR	IS	SNR	IS
5 dB	1.78	8.9 dB	0.83	7.4 dB	1.14
10 dB	0.85	12.9 dB	0.32	11.9 dB	0.41
15 dB	0.26	16.8 dB	0.08	16.5 dB	0.12

Table 2: *Segmental SNR and Itakura–Saito measure in the case of non–stationary factory noise.*

However, since our algorithm operates in the DCT domain its computational requirements are very low. This feature together with the high robustness and perceptual performance promote our algorithm as a promising solution of for real time speech enhancement in real world conditions.

5. 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. Ephrahim and H.L. Van Trees, "A signal subspace approach for speech enhancement", *IEEE Trans. on Speech and Audio Proc.*, Vol. 3, pp. 251–266, 1995.

[5] R. Vetter, N. Virag, P. Renevey, and J.-M. Vesin, "Single channel speech enhancement using principal component analysis and MDL subspace selection", *in Eurospeech'99, Budapest, Hungary*, 1999.

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

[7] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, New York, 1993.

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

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