Dual Channel Speech Enhancement using Coherence Function and MDL-based Subspace Approach in Bark Domain

Rolf Vetter, Philippe Renevey, Jens Krauss

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

Abstract
A novel algorithm for dual channel speech enhancement is presented. It combines the coherence function and a subspace approach in the Bark domain together with an optimal subspace selection through the minimum description length (MDL) criterion. The coherence function allows one to exploit the spatial diversity of the sound field. The processing in the Bark domain permits to take into account of masking properties of the human auditory system while the MDL-based subspace approach ensures statistical robustness. Performance evaluation in real sound fields has highlighted the ability of the algorithm to enhance noisy signals and improve intelligibility for various experimental conditions.

1. INTRODUCTION
Verbal communications take often place in adverse noisy environments, such as for example background noise generated by competing speakers, low signal-to-noise ratios and reverberation related degradation of speech. Due to the great significance of this problem, many noise reduction schemes have been proposed over the last decades (for a review see for example [1, 2, 3, 4]). Basically they can be divided into single and multiple-channel approaches. Recent studies have shown that single-channel speech enhancement algorithms are still unable to improve intelligibility, even if they can now at least enhance signal quality without reducing intelligibility [5]. In contrast, multiple-microphone noise reduction schemes have been shown repeatedly to increase speech intelligibility even if there remain some theoretical and practical issues to be solved [1, 5]. The performance of multiple channel speech enhancement algorithms improve with an increasing number of microphones. However, a larger number of microphones implies higher costs and an increasing demands in computational load. Moreover, due to the large dimensions of such systems, they are often not cosmetically acceptable. A promising trade-off seems to be dual channel approaches. They are able to retain the spatial diversity of the sound field while requiring low computational means. Therefore we focus in this paper on dual channel speech enhancement. More particularly, we propose a merging of single channel subspace approaches [6, 7, 8] and dual channel speech enhancement based on spatial coherence of noisy sound field [9]. In order to take into account in optimal manner of noise masking properties of the human auditory system, the coherence is evaluated in the Bark domain and it is combined with an MDL-based subspace approach. Consequently, the proposed algorithm \(^1\) yields the robustness subspace approaches [6, 7, 8] together with the low computational requirements of the DCT [10] and the high perceptual performance due to the inclusion of noise masking [3] and spatial diversity of real sound fields [9].

2. PROPOSED SUBSPACE APPROACH

2.1. Acoustical Basis
A speech signal \(s(t)\) uttered by a speaker is submitted to modifications due to its propagation. Additionally, some noise is added so that the two resulting signals, which are available on the microphones can be written as:

\[
x_1(t) = s_1(t) + n_1(t) \\
x_2(t) = s_2(t) + n_2(t)
\]

where \(N_t\) is the number of observed samples. The present work is based on the following fundamental assumptions:

(A1) The microphones are in the direct sound field of the signal of interest, (A2) whereas they are in the diffuse sound field of the noise sources. Assumption (A1) requires that the distance between speaker of interest and microphones is smaller than the critical distance whereas (A2) requires that the distance between noise sources and microphones is larger than the critical distance [4]. This is a plausible assumption for a large number of applications. As an example, consider a moderately reverberating room with a volume of 125 m\(^3\) and a reverberation time of 0.2 seconde, which yields a critical distance of \(r_c = 1.4\) m. Consequently, assumption (A1) is verified if the the speaker is nearer than \(r_c\) while (A2) requires that the noise sources are at a distance larger than \(r_c\). The consequence of (A1) is that the contributions of the signal of

\(^1\) Subject of a patent application
interest \( s_1(t) \) and \( s_2(t) \) in the recorded signal are highly correlated. In contrast, (A2) together with a sufficient distance between microphones implies that the contributions of noise \( n_1(t) \) and \( n_2(t) \) in the recorded signal are weakly correlated. Since signal and noise have generally non-uniform distribution in the time-frequency domain, it is advantageous to perform a correlation measure with respect to frequency and time. This leads to the concept of time adaptive coherence function [9], which is combined in this work with an MDL-based subspace approach [8].

### 2.2. Global Subspace Approach

The principle of the applied subspace approach can be resumed as follows [8]. The algorithm operates on a frame-to-frame basis to satisfy quasi-stationary characteristics of speech. For each frame, the noisy speech signal is observed in some multidimensional dual domain, which is partitioned into a signal, a signal-plus-noise and a noise subspace (see Figure 1). Classically, components of the dual domain are obtained by applying the eigenvectors or eigenfilters computed by the Karhunen–Loève Transform (KLT) on the delay-embedded noisy data [6, 7]. In order to avoid the large computational means required for these operations, masking properties of the human auditory system have been used in [8] to substitute the eigenfilters by the so-called Bark filters [2]. Thus, the subspace partitioning is achieved in the Bark domain and signal reconstruction is performed using unaltered components of the noise subspace, weighted components of the signal-plus-noise subspace and unaltered components of the signal subspace. Herein, we propose an extension of this subspace approach to dual channel speech enhancement. The main improvement to be obtained resides in the exploitation of the spatial diversity of speech and noise sound fields, namely of the coherence function between the two recorded signals. This coherence function is used to improve the weighting of the signal-plus-noise subspace during reconstruction (see Figure 1). The components of the dual domain, namely the Bark components are obtained in the proposed approach as follows [8]:

\[
X_i(k)_{bar} = \sum_{j=-b+1}^{b/2} G(j, k) \{X_i(k-j)\}^2
\]

where \( l = 1, 2, k = 0, \ldots, N - 1 \) and \( b + 1 \) is the processing-width of the filter, \( G(j, k) \) is the Barkfilter whose bandwidth depends on \( k \) and \( X_i(k) \) are DCT components defined as:

\[
X_i(k) = \alpha(k) \sum_{t=0}^{N-1} x(t) \cos \left( \frac{\pi(2t+1)k}{2N} \right)
\]

where \( \alpha(0) = \sqrt{1/N} \) and \( \alpha(k) = \sqrt{2/N} \) for \( k \neq 0 \). The DCT is applied instead of the DFT because of its higher energy compacting for speech [10].

### 2.3. Subspace Selection

A crucial point in a subspace approach is the adequate choice of the dimensions of the signal–plus–noise \( (p_2) \) and the signal subspace \( (p_1) \). Among the possible selection criteria, the MDL criterion has been shown to provide high performance in speech enhancement applications [7, 8]. The following criterion has been applied in [8]:

\[
MDL(p_i) = -\ln \left\{ \prod_{j=p_i+1}^{N} \lambda_j^{\frac{N-p_i}{N-p_i}} \right\}^{(N-p_i)/N} + M \cdot \left( \frac{1}{p_i} + \ln \left[ \gamma \right] \right) - \frac{M}{p_i} \sum_{j=p_i}^{N} \ln \left[ \lambda_j \sqrt{2/N} \right]
\]

where \( i = 1, 2 \), the number of free parameters \( M = p_i N - p_i^2/2 + p_i/2 + 1 \) and \( \lambda_j \) is the average of the Bark components of the two channels given by Equation (1) rearranged in decreasing order. The parameter \( \gamma \) 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 \( \gamma \) involves that the parameter \( p_1 \) provides a very parsimonious representation of the signal whereas \( p_2 \) selects also components with \( SNR_j \approx 1 \). Notably, it has to be pointed out that the MDL criterion allows one to obtain an efficient and reliable speech/noise detector, which is used in our algorithm to update noise related parameters.

### 2.4. Reconstruction of Enhanced Signal Using Time Domain Masking and Coherence Function

The enhanced signal is obtained by applying the inverse DCT to components of the signal and weighted compo-

---

Figure 1: The proposed enhancement algorithm.
ponents given by Equation (1) in decreasing order and $g_j$ is an appropriated weighting function. It is obtained by auto-regressive moving average time domain masking of the form

$$g_j(k) = \kappa a g_j(k - 1) \sum_{i=0}^{n} \kappa_{b} g_j(k - i)$$

where the non-filtered weighting function uses the coherence function as well as the local SNR $j$ of each Bark component as follows:

$$\hat{g}_j = \exp \left\{ -C_j^{-1} \text{SNR}_j^{-1} \nu_j \right\} \quad j = p_1 + 1, \ldots, p_2$$

The coherence function is evaluated in the Bark domain by

$$C_j = \frac{P_{a, a}(j)}{P_{a, x_1}(j) + P_{x_2, x_2}(j)}$$

It is computed successively from frame to frame and contributes in this way to a smoothing of the enhanced signal.

$$P_{a, x_i}(j) = (1 - \lambda_i) P_{a, x_{i-1}}(j) + \lambda_i X_{a}(j) X_{a}^{Bark} X_{a}(j) X_{a}^{Bark}$$

with $p, q = 1, 2$. The parameter $\nu$ is adjusted through a nonlinear probabilistic operator in function of the global SNR. Eventually, the final step consists in an optimal merging of the two enhanced signals.

$$\tilde{s}(t) = w_1 \tilde{s}_1(t) + w_2 \tilde{s}_2(t)$$

where $w_1$ and $w_2$ are chosen to optimize the posterior SNR.

3. PERFORMANCE EVALUATION

3.1. Database and Compared Algorithms

In order to assess the ability of the proposed algorithm to enhance noisy signals a large number of simulations have been performed. Herein, results for real noisy sound fields according to the experimental setup of Figure 2 are presented. The noise source is situated at 1.5m whereas the signal source is located at 1m of the two microphones. Different types of background noises from the Noisex database and phonoetically equilibrated french sentences have been used. The variance of noise has been adjusted to obtain SNRs in the recoded signals, ranging from 0 dB to 20. The sampling frequency has been chosen at 8 kHz. To illustrate the performance improvement provided by the proposed dual channel approach with respect to a single channel approaches the following algorithms have been compared: (i) single channel approach based on non-causal Wiener Filtering (Wien–Nc) [2], (ii) proposed dual channel approach (Coh–MDL). The perfor-

![Figure 2: Experimental setup for performance assessment in real environment.](image)

![Figure 3: Speech spectrograms: (a) original French speech signal: Un loup s’est jeté immédiatement sur la petite chèvre, (b) one of the recorded noisy signals (non-stationary factory noise at an segmental input SNR = 10 dB), enhanced signals using (c) WienNc, (d) COH–MDL.](image)
fined by the American National Standard ANSI S3.5-1997, the observation of the spectrograms as well as informal listening tests. In order to obtain a relevant performance assessment, the mean value of $SNR, IS$ and $SII$ have been computed after discarding frames without any speech activity. Globally, we have observed that the proposed dual channel subspace approach outperforms classical single channel algorithms such as for example Wien–$NC$. This observation is confirmed quantitatively in Table 1 and 2. Indeed, one can note that $COH$–MDL provides smaller $IS$-values and larger $SII$-values than Wien–$NC$ for similar experimental conditions. This observation has been confirmed qualitatively by informal listening tests. They have pointed out that the inclusion of the coherence function improves the perceptual performance of the single channel MDL–Bark subspace approach presented in [8]. The analysis of Figure 3c highlights furthermore that Wien–$NC$ provides a considerable amount of residual ”musical noise”. In contrast, Figure 3d underlines the high performance of the proposed approach.

## 4. CONCLUSION

We have presented in this paper a dual channel extension of a computational efficient single channel speech enhancement algorithm. The proposed method is based on a subspace approach in the Bark domain with a subspace selection provided by the MDL criterion. The main contribution of this work resides in the inclusion of spatial diversity, namely the coherence between the two microphone signals to improve speech enhancement and intelligibility in subspace approaches. High perceptual performance haven been obtained in real sound fields. This feature together with the low computational requirements of the algorithm promotes it as a promising solution for real time speech enhancement in natural environments.

## 5. References


