SPEECH ENHANCEMENT IN A NOISY CAR ENVIRONMENT

Erik Visser[*], *Te-Won Lee, Manabu Otsuka*[†]

University of California, San Diego
Institute for Neural Computation
La Jolla, CA 92093-0523
{visser,tewon,motsuka}@rhythm.ucsd.edu

ABSTRACT

We present a new speech enhancement method for robust speech recognition in a noisy car environment. The method is based on the combination of two important building blocks, namely blind source separation given two microphone signals and speech denoising using a hybrid Wavelet - Independent Component Analysis (ICA) filterbank. The first block separates point sources such as the passenger's voice signal whereas the second block eliminates distributed noise signals such as road and wind noise. We performed experiments with real recordings taken while driving in a noisy automobile environment. The method works nearly real-time and achieves good separation results.

1. INTRODUCTION

Human computer interactions are becoming increasingly important in todays technological society and people are getting used to interacting with computers on a daily basis. In car environments, input modes for steering control or information retrieval are traditionally limited to hand activated devices such as the steering wheel and buttons on the board panel. However, the natural way of using human voice commands has been given considerable attention by car manufacturers in recent years. Although some commercial products are currently available, the performance of those systems usually degrades substantially under real-world conditions. For example, a speech recognition system in an automobile may process voice commands when spoken in a quiet situation, but the system recognition performance may be unacceptable in the presence of interferring sounds such as the car engine noise, music, and other voices such as the passenger's voice.

In this paper, we consider a real-time speech enhancement system for robust speech recognition that makes use of multiple microphones for speech source separation as well as the intrinsic structures of speech signals for speech denoising. As illustrated in Figure 1, the method involves two main stages: a) Blind Source Separation (BSS), which exploits the time-correlation of speech signals captured by two microphones and b) denoising, which uses the different statistics in speech and noise signals to separate them.

Our motivation for combining the BSS and denoising method is to exploit the strength of each individual method. The BSS algorithm achieves good separation results when two distinct point sources are recorded at the same time [4]. However, in real car environments we have to deal with distributed noise sources such as wind and road noise where this assumption is not valid. Denoising algorithms usually work with the assumption of additive Gaussian noise which is an appropriate model for the noise considered in this framework. However denoising cannot deal with point sources when the sources have similar statistical characteristics such as two mixed speech signals. In combining these two methods, we first separate a strongly interferring source signal (such as the passenger's voice) and then separate out the remaining distributed noise signals by an adaptive denoising scheme.

Methods for noise level estimation, driver's speech detection and automatic speech recognition will not be discussed in this paper since conventional methods are available [1, 2, 3, 12]. The focus of this paper is to develop and implement the two previously mentioned subschemes where we concentrate on the new adaptive denoising part applied to real recordings. The challenge of our design is to perform in real-time and to enhance the speech signal under difficult noise conditions.

2. BLIND SOURCE SEPARATION

The BSS approach adopted in this paper is the Multiple Adaptive Decorrelation (MAD) algorithm [4] designed for separation of non-stationary convolved signal mixtures. Although several BSS algorithms exist, we choose this one due to its real time performance and good separation results.

In the MAD framework, it is assumed that the original

[*] corresponding author. Category:Applications
[†] on visit from DENSO Corporation, Japan

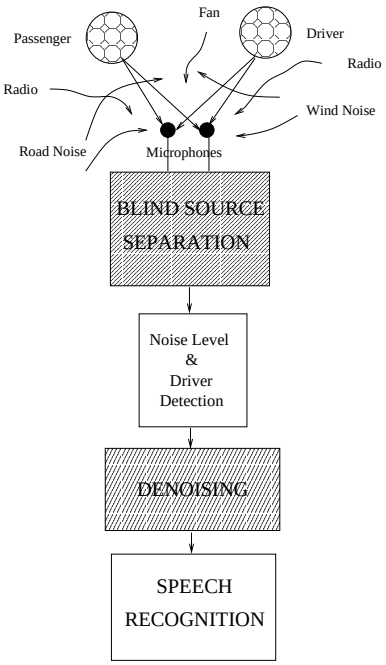

Fig. 1. Speech enhancement schematic representation

sources $\mathbf{s}(t)$ can be recovered from the measurements

$$\mathbf{x}(t) \quad = \quad \sum_{\tau=0}^{P} \mathbf{A}(\tau)\ \mathbf{s}(t-\tau)$$

by finding a sequence of unmixing filter matrices $\mathbf{W}(\tau)$ such that

$$\hat{\mathbf{s}}(t) \ = \ \sum_{\tau=0}^{Q} \mathbf{W}(\tau)\ \mathbf{x}(t-\tau).$$

The search is executed in the frequency domain where

$$\mathbf{x}(\omega,t) \quad \simeq \quad \mathbf{A}(\omega)\ \mathbf{s}(\omega,t), \quad T >> P,$$

T being the length of the Fourier transform window which is applied in an overlap-add fashion [4]. If the cross correlation of the measurements is denoted by

$$\hat{R}_x(\omega,t) \quad = \quad \mathsf{E}\left[\mathbf{x}(\omega,t)\ \mathbf{x}^H(\omega,t)\right]$$

and that of the sources by

$$\hat{\Lambda}(\omega,t) \quad = \quad \mathsf{E}\left[\mathbf{s}(\omega,t)\ \mathbf{s}^H(\omega,t)\right],$$

$\mathbf{W}(\omega)$ is found by minimizing

$$\hat{\mathbf{W}},\hat{\mathbf{\Lambda}}_s \quad = \quad \arg\min_{\hat{\mathbf{W}},\hat{\mathbf{\Lambda}}_s} \sum_{\omega=1}^{T} ||\mathbf{W}\ \hat{R}_x(\omega,t)\ \mathbf{W}^H \ - \ \Lambda_s(\omega,t)||^2$$

$$s.t. \quad \mathbf{W}(\tau)=0, \forall\tau \ > \ Q,\ Q << \ T, \qquad (1)$$

$$\mathbf{W}_{ii}(\omega)=1 \qquad (2)$$

The constraints (1) impose that the filter length Q be much smaller than T to solve the frequency permutation problem [4]. Also scaling issues are solved by fixing the diagonal elements of the filter matrices to unity (constraint (2)). Since the source correlation is updated as

$$\hat{\Lambda}_s(\omega,t) \quad = \quad diag\left[\mathbf{W}(\omega)\ \hat{R}_x(\omega,t)\ \mathbf{W}^H\right],$$

the cost basically minimizes the off-diagonal elements of the cross correlation matrix $\hat{R}_x(\omega,t)$. One finally obtains the learning rule [4]

$$\Delta\mathbf{W}^*(\omega) \quad = \quad 2\ \mu\ E(\omega,t)\ \mathbf{W}(\omega)\ \hat{R}_x(\omega,t)$$

where μ is the learning rate and

$$E(\omega,t) \quad = \quad \mathbf{W}\ \hat{R}_x(\omega,t)\ \mathbf{W}^H \ - \ \hat{\Lambda}_s(\omega,t).$$

The frequency domain version of the MAD algorithm has been implemented in C code and is executable in real-time on a 550 MHz PC. We performed several experiments with real recordings. An example of a driver's voice corrupted by a passenger's voice in a noisy car environment is available from http://rhythm.ucsd.edu/∼visser/ICA2001. Two microphones attached on each side of the center rear mirror were used to record the files sampled at 16 kHz. The driver and the passenger were speaking at the same time while driving at 40 MPH with open windows, high fan noise and music in the background. The algorithm successfully separated the passenger's voice from the driver's voice. However, as

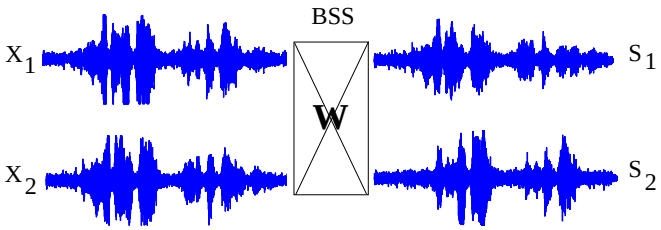

Fig. 2. Recorded mixtures and separated sources

illustrated in Figure 2, both separated source files still contain the original noise originating from the open window, vibrations of the car, fan or music from the radio. BSS cannot separate these background signals from the voices since the latter are spread out in space and therefore are undistinguishable at either microphone. This noise needs to be removed by a denoising approach based on different statistical properties of noise and speech.

3. ADAPTIVE DENOISING

Music has a less sparse probability density distribution than speech and the other noise sources encountered such as the

wind, fan or vibration noise can generally be approximated by a Gaussian distribution due to the central limit theorem [3]. This is a reasonable assumption in particular when noise signals appear at the same time. Thus by transforming the original signal into a space where super-Gaussian distributions or sparseness are emphasized, speech components will have large values while noise coefficients will be small and can thus be eliminated by applying a coring or shrinkage function [5].

The main questions are how to find the optimal basis transformation and how the transformation should be implemented to allow real-time performance. A natural way to look for optimal basis functions is to learn them from samples of speech data using Independent Component Analysis (ICA) [6]. Whereas these ICA learning approaches have produced bandpass filter-like basis functions, the implementation of matrix operations to perform the basis transform

$$\mathbf{a} \quad = \quad \mathbf{W}\,\mathbf{x}$$

is computationally costly $O(N^2)$. Non-obvious technical questions are how to choose the length and overlap between neighboring speech segments since the matrix based approach in [6] is not shift-invariant and suffers from blocking effects leading to poor reconstructed speech quality. This shift dependence is also reflected in the redundancy of basis functions which often are time-shifted versions of the same bandpass filter. Moreover the learned basis functions in [6] cover the frequency range in a continuous and overlapping manner and thus also exhibit a redundancy in frequency content.

Therefore the computation of sparse coefficients should rather be implemented by using *linear filters*. The wavelet transform provides such a signal mapping into sparse subspaces [11]. It can be efficiently implemented by using a critically sampled multiresolution filter bank [7] or its oversampled, shift-invariant equivalent [8]. Studies have further shown that wavelet coefficients are naturally sparse [9]. Coding efficiency derived from ICA basis functions is not significantly different from the one obtained with wavelets [10]. Moreover the wavelet transform implements an orthogonal frequency decomposition with a constant subband frequency width - center frequency ratio property. The subdivision yields large subband-width at high frequencies and low bandwidth at low frequencies in analogy to the mel scale used in speech recognition. Therefore a sparse representation and physiologically intuitive frequency subdivision are obtained at the same time. Finally the orthogonality properties of wavelet decompositions are useful for eliminating unwanted frequency bands as is frequently done with speech for recognition purposes by highpass filtering the data.

The approach adopted in this paper is to efficiently compute initial shift-invariant wavelet coefficients and sparsify them further if necessary by directly minimizing a sparseness measure like in ICA. An bandpass *filter* will be used to filter the wavelet coefficients into sparser subbands. Figure 3 illustrates the denoising scheme. For each subband

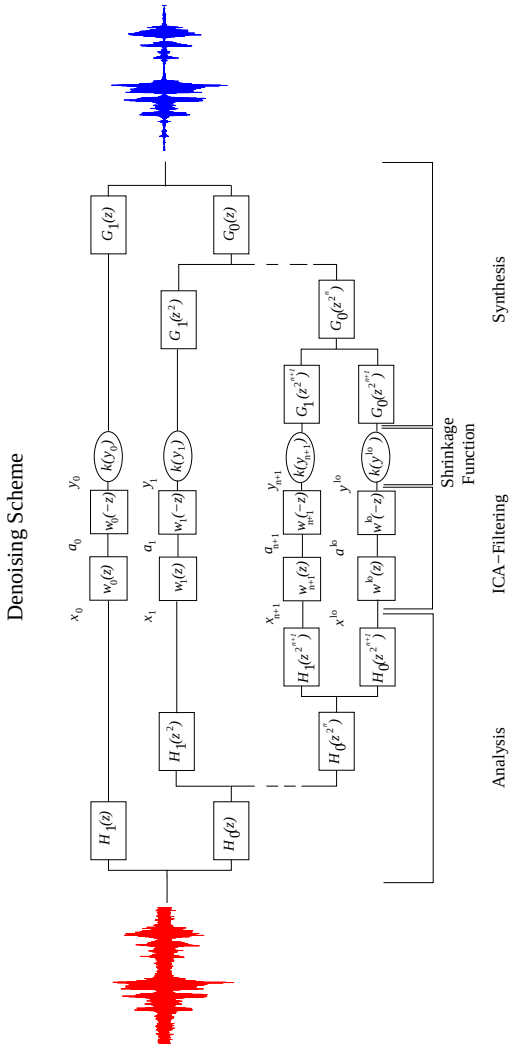

Fig. 3. Filterbank: signal analysis is implemented using upsampled versions of orthogonal low- and highpass filters H_0 and H_1, respectively. For each subband i, wavelet coefficients x_i are then filtered by ICA filters w_i. A shrinkage function k is applied. Signal reconstruction is done using upsampled versions of synthesis filters G_0 and G_1 which are obtained from the time reversed analysis filters [11]

$i = 0, 1, \cdots, n+1$ and the low frequency band lo, we have ICA subspace projected wavelet coefficients

$$a_i \quad = \quad w_i * x_i$$

and "reconstructed" wavelet coefficients

$$y_i \quad = \quad w_i^* * a_i$$

x_i is the wavelet coefficient vector and w_i the ICA filter (w_i^* is the reversed time version of w_i). In the Fourier domain, this yields

$$\begin{aligned} Y_i(j\omega) &= W_i(j\omega)\, W_i^*(j\omega)\, X_i(j\omega) \\ &= |W_i(j\omega)|\, X_i(j\omega) \end{aligned}$$

where $W_i(j\omega)$ is the complex Fourier transform of the filter w_i, $W_i^*(j\omega)$ the complex conjugate of W_i, Y_i and X_i the Fourier transforms of y_i and x_i respectively.

The design criterion for the filters w_i is to maximize the a posteriori likelihood of y_i

$$\hat{w_i} = \max P(y_i|a_i, w_i)\, P(a_i) \quad (3)$$

where the priors are given by

$$\begin{aligned} P(y_i|a_i, w_i) &\simeq e^{-\frac{(y_i - w_i^* * a_i)^2}{2\,\sigma_i^2}}, \\ P(a_i) &\simeq e^{-S(a_i)} \end{aligned}$$

Thus, for each subband i, we assume a Gaussian distribution with standard deviation σ_i for the noise signal and a super-Gaussian distribution modeled by the function S for the coefficients a_i. Problem (3) can be reduced to (4) by considering the log likelihood:

$$\hat{w_i} = \min \frac{(y_i - x_i)^2}{2\,\sigma_i^2} + \sum_j S\big(a_i(j)\big) \quad (4)$$

If the noise level σ_i is low, the first term in eq. (4) prevails and thus little can be done to improve the sparseness of a_i. In this case, a shrinkage function which constitutes an approximate analytical solution to problem (4) can be applied. For example, if a Laplacian distribution

$$S(a) = |a|$$

is assumed, the shrinkage function k is given by

$$y_i = k(y_i) = \text{sign}(y_i)\, \max(|y_i| - \sqrt{2}\,\sigma_i) \quad (5)$$

If σ_i is large, the second term in (4) is emphasized and wavelet coefficients can be sparsified. It is important that filters w_i are reversed since it will finally only introduce an amplitude modulation of x_i. A simple filter would introduce phase distortion in the synthesis coefficients leading to violation of the orthogonal reconstruction principle. Also direct sparsification of y instead of a will lead to whitening of the signal since in that case, only the amplitude of the signal can be changed whereas in problem (4), the filter phase acts as an additional degree of freedom. Once sparse a are found, an additional shrinkage function can be used to remove small coefficients. Note that the first term also prevents the filter from simply whitening the original signal x or scaling its mean amplitude down to levels where sparseness is minimized. It preserves relevant sparse features present in the original signal. The quality of y_i depends on an accurate estimation of the noise level σ_i which can be supplied by conventional techniques [1, 12].

4. REAL RECORDING EXAMPLE FOR DENOISING

The recorded BSS car files in section 2 were used to illustrate the filterbank performance. The resulting frequency subdivision as well as the subband coefficients before and after denoising are shown in Figure 4. 12 tap Daubechies

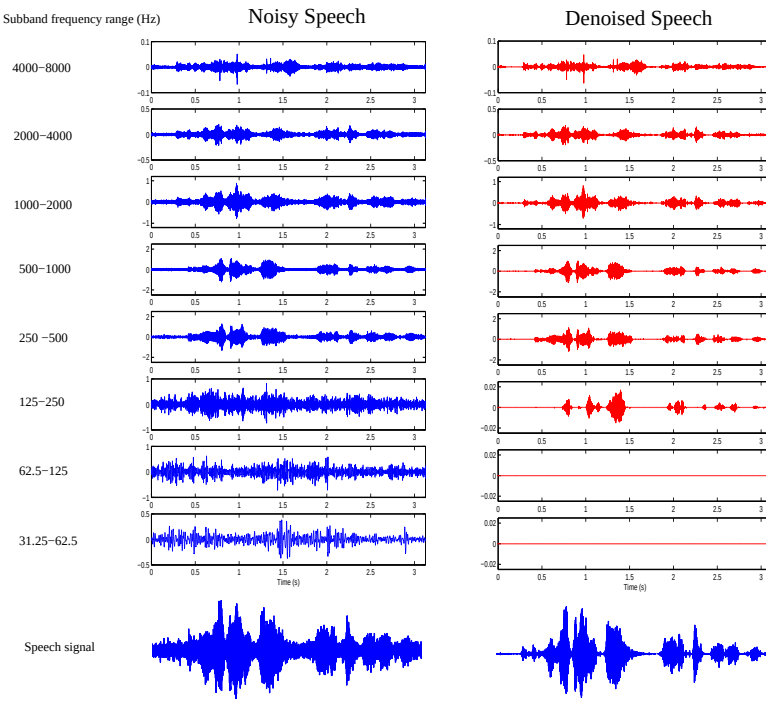

Fig. 4. Denoising: subbands # $0, \cdots, 6, lo$ (From top to bottom)

wavelet filters were chosen for H_0,H_1, G_0 and G_1 [11]. It can be seen that the lowest frequency subbands should not contain significant amounts of speech information as speech features are usually concentrated in the 125-4000 Hz range. Figure 4 shows that coefficients in subbands 0-4 are sparse enough. These coefficients were shrinked by applying (5) i.e. a Laplacian probability distribution for speech was assumed. When considering subbands 5, 6 and lo, it was initially attempted to completely discard them during reconstruction. No speech quality loss was noticed for subbands 6 and lo while canceling subband 5 resulted in a loss of low frequency components in the driver's voice. The strong road noise level therefore masks sparse speech coefficients and ICA filters have to be designed. The cost (4) was minimized by using a Sequential Quadratic Programming (SQP) optimization routine in Matlab. The filter length was 688 taps and was linearly interpolated by using 120 equally time spaced optimization parameters. The function $S(x) = \log(1 + (\frac{x}{\sigma})^2)$ was used as the sparseness cost term. It can be made as sparse as the Laplacian distribution

but has the advantage of being continuously differentiable.

One can see the gradual sparseness improvement after filtering with ICA filter w_5 shown in Figure 5. As the per-

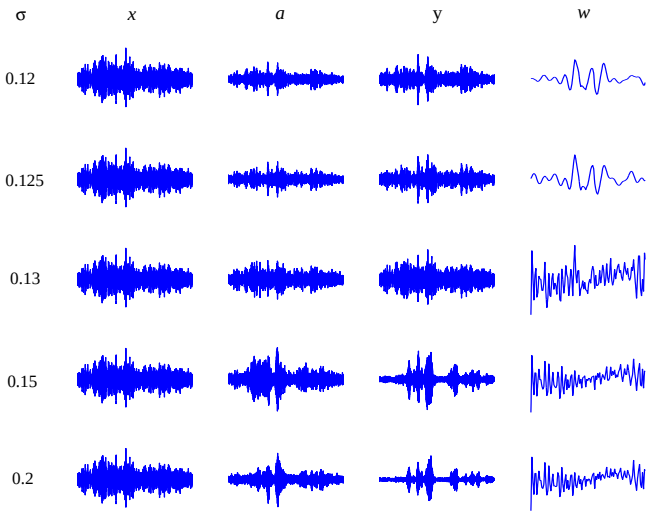

Fig. 5. Learning ICA filters for various estimated noise levels σ (subband # 5)

fect reconstruction of y from x is relaxed to accommodate a stronger noise level, the filter learns the noise features to deconvolve them from the underlying sparse speech coefficients. The initial filter looks more like a Gabor filter which represented the original wavelet bandpass filter. The final learned filter looks more like the road noise features and thus allows to uncover the underlying speech signal. The remaining sparse coefficents are clearly identified as speech signals when observing their time alignment with the high frequency subbands.

The implementation of the filterbank was done in Matlab. Estimation of the noise level was done manually at this point so as to guarantee a good sound quality. For computationally effective implementations, we consider developing a lookup table with filters computed off-line for different car speeds, car environments and noise levels, stored in a database and implemented in C code. The actual filter will then be picked on-line from the table according to the estimated noise levels, car speed and functional state of other devices in the car environment (radio or fan switched on, open window). If filters become too long and thus computation of time convolution lengthy, implementation of the filterbank can be done in the Fourier domain, thus reducing the number of multiplications involved [11].

The SNR was improved significantly as it can be seen in Figure 4 (denoised audio files are downloadable from http://rhythm.ucsd.edu/∼visser/ICA2001). As opposed to conventional denoising methods like spectral subtraction [12], no artifacts like musical noise are generated and non stationary noise sources can be dealt with. Indeed, if the noise environment and level are estimated online, the shrinkage function thresholds and ICA filters can be adjusted accordingly. Whereas the phase information of the original files was slightly affected by the nonlinear shrinkage function, we are confident that the speech recognition word error rate will be significantly decreased. We also note that the cepstral coefficients are rather insensitive to mild phase distortion [3].

5. CONCLUSION

A real-time two stage speech enhancement scheme was proposed. Its performance was illustrated on speech data recorded in a noisy car environment. An automatic procedure to estimate the noise level and discriminate between the driver's and passenger's voice needs to be added to complete the application. The first procedure can be arranged by recording during speaker silence and estimating the signal amplitude [1, 12]. The latter procedure can be solved by applying known solutions to direction of arrival estimation since the location of the microphones are known [3], or more advanced techniques using microphones arrays [2]. Emphasis was put on the separation of both the distinct point sources as well as distributed noise signals while maintaining near real-time performance. Taking advantage of BSS to efficiently remove a point source and of the new denoising algorithm to adaptively denoise the remaining signals in different subbands with a sparseness objective, thereby exploiting the intrinsic statistical structure of the speech signal, is key to this method. Optimizations can be done using filters from lookup tables and smoother shrinkage functions for better denoising quality. Our current research focuses on quantitative studies of word error rate reduction of speech recognition systems on the Aurora benchmark dataset and compares the proposed speech enhancement scheme to conventional techniques.

6. REFERENCES

[1] Mokbel, C.E., Chollet, G.F.A., Automatic Word Recognition in Cars. In *IEEE Trans. on Speech & Audio Processing*, Vol. 3, 5, pp. 346-356, September 1995

[2] Nagai, T., Kondo, K., Kaneko, M., Kurematsu, A., Estimation of Source Location based on 2-D MUSIC and its Application to Speech Recognition in Cars. In *Proceedings of the Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 5, Salt Lake City, May 2001

[3] Rabiner, L., Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, 1993