A Spatio-Temporal Speech Enhancement Scheme for Robust Speech Recognition

Erik Visser, Manabu Otsuka, Te-Won Lee
University of California, San Diego
Institute for Neural Computation
9500 Gilman Drive, La Jolla CA 92093-0523
Email: {visser,tewon,otsuka}@rhythm.ucsd.edu

Abstract

A new speech enhancement scheme is presented integrating spatial and temporal signal processing methods for robust speech recognition in noisy environments. The scheme first separates spatially localized point sources from noisy speech signals recorded by two microphones. Blind source separation algorithms assuming no a priori knowledge about the sources involved are applied in this spatial processing stage. Then denoising of distributed background noise is achieved in a combined spatial/temporal processing approach. The desired speaker signal is first processed along with an artificially constructed noise signal in a supplementary blind source separation step. It is further denoised by exploiting differences in temporal speech and noise statistics in a wavelet filterbank. The scheme’s performance is illustrated by speech recognition experiments on real recordings in a noisy car environment and compared to conventional techniques like beamforming and spectral subtraction.

1 INTRODUCTION

Despite the increasing importance of human computer interactions, information systems are still difficult to control manually in multi-task operating environments such as automobiles. Higher flexibility and safety standards can be achieved by using voice commands to interact with machines and computers. Although a number of commercial speech recognition systems are currently available, their performance usually degrades substantially under real-world conditions. In a car for example, noise sources are particularly numerous and uncertain like vibrations, fan and noise from open windows generating a spatially distributed background noise. Spatially isolated point sources like a passenger’s voice or music from a loud speaker further degrade voice commands which may also include reverberation. The driver’s utterance therefore needs to be enhanced before processing by a speech recognition system.

Single-microphone enhancement algorithms based on temporal information about the recorded signals are most frequently encountered. They often use a probabilistic framework with statistical models of a single speech signal corrupted by stationary Gaussian noise [1]. While reasonable performance is obtained when the noise is stationary, it deteriorates rapidly when noise power varies importantly or speech mixtures contain significant reverberation. Spatial information about signal mixtures can be exploited by using multiple microphones. In beamforming [2] for example, an array of microphones with a known geometry is used to suppress interfering signals. Here, source localization can be performed as well and reverberation be handled with adaptive algorithms [2]. However, these methods usually rely on a priori information about the acoustical environment and sources involved. Also, large microphone arrays are required for good performance whose implementation in cars is difficult and costly. The number of microphones can be drastically reduced by using recently developed source separation algorithms [3, 4]. These algorithms exploit spatial information about signal mixtures recorded at different microphone locations to explicitly separate interfering noise signals from the desired source signal without assuming any a priori source models.

In this paper, we propose a new robust speech enhancement scheme making synergistic use of spatial and temporal processing methods. Robustness is achieved primarily by assuming no a priori speech or noise models.

2 SPEECH ENHANCEMENT SCHEME

An analytical framework suitable for car recordings is considered with \( m \) different microphone mixture signals \( x(t) \) composed of \( m \) point source signals \( s(t) \) and additive background noise \( n(t) \)

\[
x(t) = \sum_{\tau=0}^{P} A(\tau) \ s(t - \tau) + n(t)
\]

where \( P \) is the convolution order, \( A(\tau) \) is a \( m \times m \) mixing matrix. A key distinction is made between spatially point sources \( s(t) \) and distributed background noise \( n(t) \).

---

\( ^{1} \) DENSO CORPORATION, Japan
Assuming little reverberation, signals originating from point sources can be viewed as identical when recorded at different microphone locations except for an amplitude factor and a delay. The unmixing strategy would consist in finding these latter parameters for each source and summing up the realigned and scaled mixture signals. However, background noise originates from a large number of spatially distributed sources resulting in no defined delay and amplitude differences between signals recorded at each microphone. Thus, a background noise unmixing strategy poses a singular problem. These different signal characteristics are addressed in subsequent stages of the speech enhancement scheme illustrated in Figure 1.

Figure 1: Proposed Speech Enhancement Scheme

Spatial information about interfering point sources is processed in the blind source separation units while the remaining stages remove distributed background noise by temporal information processing.

2.1 Blind Source Separation of Interfering Point Sources

In recent years, a number of signal processing algorithms have emerged implementing blind source separation (BSS) of mixture signals into its components by decorrelating their higher-order statistics [4]. However, the second order decorrelation approach presented in [3] yielded the most consistent separation performance in our experiments. The Multiple Adaptive Decorrelation (MAD) algorithm outlined in [3] is designed for separating $m$ records of $\mathbf{x}(t) = \sum_{\tau=0}^{\tau_m} \mathbf{A}(\tau) \mathbf{s}(t-\tau) = [x_1(t)x_2(t)\cdots x_m(t)]$ into $m$ original sources $\mathbf{s}(t) = [s_1(t)s_2(t)\cdots s_m(t)]$ by finding a sequence of $m \times m$ unmixing filter matrices $\mathbf{W}(\tau)$ such that $\hat{\mathbf{s}}(t) = \sum_{\tau=0}^{\tau_m} \mathbf{W}(\tau) \mathbf{x}(t-\tau)$, $Q$ being the filter length. The unmixing filter computation is executed in the frequency domain where $\mathbf{X}(\omega,t) \sim \mathbf{A}(\omega) \mathbf{S}(\omega,t)$, $\mathbf{X}(\omega,t)$ being the spectrogram obtained by consecutively computing the Short Time Fourier Transform of length $T$ (where $T > > P$, the convolution order), of $x(t)$ at each time instant $t$ in an overlap-shift fashion [3]. If the cross correlation of the measurements is denoted by $\hat{R}_x(\omega,t) = \mathbb{E} [\mathbf{X}(\omega,t) \mathbf{X}^H(\omega,t)]$ and that of the sources by $\hat{A}_s(\omega,t) = \mathbb{E} [\mathbf{S}(\omega,t) \mathbf{S}^H(\omega,t)]$, $\mathbf{W}(\omega)$ is found by minimizing

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

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

$$\hat{\mathbf{W}}_n(\omega) = 1$$

Constraint 2 imposes that the filter length $Q$ be much smaller than $T$ to solve the frequency permutation problem [3]. Also scaling issues are solved by fixing the diagonal elements of the filter matrices to unity (Constraint 3). The final learning rule is $\Delta \mathbf{W}(\omega) \sim \hat{E}(\omega,t) \mathbf{W}(\omega) \hat{R}_x(\omega,t)$ where $\hat{E}(\omega,t) = \mathbf{W} \hat{R}_x(\omega,t) \mathbf{W}^H - \Lambda_s(\omega,t).$ The MAD algorithm has been implemented in C code and is executable in near real-time on a 550 MHz PC. The approach has shown robust performance in a number of applications. Moreover, constraint (3) in problem (1) ensures that the dominant speaker voice will be separated at the microphone position at which its amplitude is highest. This makes an additional algorithmic step to determine which of the separated sources is the speaker’s voice unnecessary. In [3], a learning rule for removing the distributed background noise $n(t)$ at the same time is presented. This would however require a number of microphones much larger than the number of point sources. Therefore, an alternative procedure for reducing noise using the BSS algorithm in an additional step is presented next.

2.2 Background Denoising using Source Separation

To achieve this, the separated speaker’s signal from the previous BSS stage and an artificially generated, pure noise signal are used as new inputs to the BSS algorithm. The basic idea is that distributed background noise contained in a single channel is transformed into a pseudo-point source and can thus be separated using spatial source separation. The new artificial noise signal should approximate background noise contained in the previously separated speaker’s signal up to a delay and amplitude scaling factor. It can be constructed in the frequency domain by estimating the noise power from pure noise intervals in the speaker’s signal and taking the complete phase information of the speaker’s signal. The signal segmentation into noise and speech intervals can be done using standard voice activity detection algorithms [6] to obtain an estimate of the instationary noise power. Using the estimated noise power spectrum in these intervals and phase information from the speaker’s signal, a noise signal is generated which strongly correlates in time and frequency with the noise background contained in the separated speaker’s signal.

Consider the separated noisy speaker signal $\hat{s}_1(t), t \in [t_0, t_1]$, and noise-only interval $s_n(t) = \hat{s}_1(t), t \in [t_0, t_1]$. The Fourier transform of $s_n(t)$ yields $S_n(\omega) = \mathbb{E}_s e^{i \phi_s(\omega)}$. By consecutively computing the
Fourier transform of length $T_r = t_1 - t_0$ of $\hat{s}_1$
\[ S_{T_r}(\omega) = |S_r(\omega)| e^{i \Phi_{S_{T_r}}(\omega)} \quad (4) \]
and replacing its magnitude spectrum by the noise magnitude yielding
\[ S_{T_r}(\omega) = |S_n(\omega)| e^{i \Phi_{s_{T_r}}(\omega)}, \quad (5) \]
a pure noise signal $S_n(\omega)$ is constructed having the noise power characteristics but the speaker signal’s phase. The time-domain version $s^n(t)$ of $S_n(\omega)$ is constructed by taking the inverse Fourier transform of length $T_r$ and repeating the procedure on $\hat{s}_1(t)$ in an overlap-add fashion [3], i.e. taking Fourier/inverse Fourier transforms of window length $T_r$ and shifting the window by $\frac{T_r}{2}$. Using $s^n(t)$ and $\hat{s}_1(t)$ as inputs to the MAD algorithm denoises $\hat{s}_1(t)$ yielding the BSS denoised speaker signal $y(t)$. As opposed to spectral subtraction, musical noise artifacts resulting from power over-subtraction [1] are minimized since the generated artifactual spectrum would increase the correlation between signals again.

2.3 Background Denoising using a Wavelet Filterbank
The previous denoising step may not remove all noise since its power may be underestimated on short, pure noise sample intervals especially in the low frequency subbands. Hence a complementary denoising approach exploiting the statistical difference of speech and noise is added in a final step.

The statistical distribution of background noise in cars can be approximated by a Gaussian distribution by invoking the Central Limit Theorem [6]. Speech signals on the other hand have a much sparser distribution. 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 have a high probability of being zero and can thus be eliminated by applying a coring or shrinkage function [5]. The continuous wavelet transform provides such a signal mapping into sparse subspaces using linear filters. Studies have shown that wavelet coefficients are naturally sparse [5, 7]. The transform is efficiently implemented by using an oversampled, shift-invariant multiresolution filter bank [7]. Moreover it provides an orthogonal frequency decomposition with a constant subband frequency width - center frequency ratio property in the mel scale used in speech recognition [6]. Therefore a sparse representation and physiologically intuitive frequency subdivision are obtained at the same time. Figure 2 illustrates the denoising wavelet filterbank. After computing the noisy wavelet coefficients $y_i$ for each subband $i$ of the BSS denoised speaker signal $y(t)$, the objective is to find denoised coefficients $y^n_i$ for each subband $i$ such that the joint probability $P(y_i - y^n_i | \sigma_i) P(y^n_i) \phi$ of the independent noise $(y_i - y^n_i)$ and speech $(y^n_i)$ distributions is maximized, $\sigma_i$ being the noise level. The priors are given by the Gaussian $P(y_i - y^n_i | \sigma_i) \approx e^{-\frac{(y_i - y^n_i)^2}{2 \sigma_i^2}}$ and the Laplacian distribution $P(y^n_i) \approx e^{-|y^n_i|}$. When the log-likelihood is considered, one obtains:
\[ y^n_i = \arg \min_{y_i} \left( \sum_j \frac{(y_i(j) - y^n_i(j))^2}{2 \sigma_i^2} + \sum_j |y^n_i(j)| \right) \]
whose approximate analytical solution is given by the shrinkage function $k$ [5]:
\[ y^n_i = k(y_i) = \text{sign}(y_i) \max \left( 0, \left| y_i \right| - \sqrt{2 \sigma_i^2} \right) \]
The noise thresholds $\sigma_i$ are estimated from the values of the wavelet coefficients in pure noise time intervals [5].

3 EXPERIMENTS
Speech recognition experiments were conducted with speech data recorded in a real noisy car environment. Worst case noise scenarios were considered so that quantitative speech enhancement performance of the spatial and temporal processing stages of the proposed scheme could be adequately measured and compared to conventional techniques given by beamforming followed by spectral subtraction. The driver was uttering digit sequences and the passenger was talking simultaneously on his cell phone while driving at 40 MPH with open windows, radio and fan turned on. Two microphones were attached on each side of the rear window mirror to record the speech mixtures. The distance between microphones was 15 cm and the recorded speech data was sampled at 8 kHz. Before each test utterance, both driver and passenger were silent for 1 sec to estimate the noise background constituted by fan, wind and road noise as well as music from the radio. The noise power for the new artificial noise signal $s^n(t)$ was estimated from pure noise, 1 sec pre-speech intervals as in spectral subtraction. Separated driver’s signal $\hat{s}_1(t)$ and the latter artificial noise signal were decorrelated by the MAD algorithm [3]. Wavelet coefficient shrinkage was only done in the three
lowest of 8 frequency subbands to avoid phase distortions caused by the shrinkage function in high frequency bands. Figure 3 illustrates the obtained speech enhancement at each stage.

![Figure 3: Illustration of Speech Enhancement](image)

3.1 Speech Recognition Results

The speech recognizer as well as a multiple noise condition database for training the HMM models was provided by the AURORA 2 benchmark dataset [8]. The speech feature extraction front-end FE.y2.0 [8] was used for computing the 39 Mel-Frequency Cepstral Coefficients (MFCC) (including energy, velocity and acceleration coefficients). The test dataset for the speech recognizer was given by 40 continuous digit sequences with a total number of 147 digits.

The digit recognition accuracy dropped to 46.9 % when speech recognition was applied to the raw mixture signal $s_1$ recorded by the microphone on the driver’s side. Interfering passenger’s voice as well as strong background noise in the mixture contributed to this unacceptable performance. The conventional strategy of beamforming i.e. finding the delay between microphones and in phase summation of the recorded mixtures, followed by single channel spectral subtraction yielded a final accuracy of only 56.9 %. Hence passenger’s voice, car vibrations and fan noise were insufficiently removed with these traditional techniques. By comparison, using the separated noisy driver signal $s_1(t)$ from the first BSS stage achieved an accuracy of 72.1 %. While applying spectral subtraction after this stage improved accuracy to 72.8 %, denoising using BSS instead achieved an accuracy of 74.8 % (recognition on signal $y(t)$ ). Hence, although noise power was estimated in both cases using the same pure noise, 1 sec interval, the BSS based denoising method causes less artifacts and/or better noise power removal than spectral subtraction. A final recognition accuracy of 79.6 % was obtained using the driver’s signal $y^a(t)$ output from the wavelet filterbank. This latter accuracy indicates that significant noise power was contained in the low frequency bands (car vibrations) and underestimated by the denoising procedure using BSS or spectral subtraction. A maximum achievable baseline accuracy of 92.5 % was determined with a recorded clean testset of similar size, quantifying the effects of microphones, speakers and noise scenarios different from those in the AURORA training database.

4 Conclusions

A new speech enhancement scheme for robust speech recognition has been presented. The approach uses a combined spatial and temporal processing strategy to handle reverberation, highly interfering sources and background noise without the need of microphone arrays nor a priori speech or noise models. In experiments with real noisy car recordings, more than 30 percent recognition accuracy improvement was obtained with the proposed scheme compared to 10 percent with conventional techniques. Beamforming performed significantly worse than BSS with only two microphones being considered while spectral subtraction was inferior to combined BSS denoising and wavelet coefficient shrinkage. Future improvements to the scheme include source localization at the front-end to optimize separation of interfering sources, and multi-model based, Wiener filtering at the back-end for additional denoising fine-tuning. The scheme is expected to yield robust recognition accuracies over 90 % in a large number of real-life situations.

References