Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE @ DIRECT®

Speech Communication 41 (2003) 393–407

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments

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

[a] Institute for Neural Computation, University of California, San Diego, 9500 Gilman Drive, Dept 0523, La Jolla, CA 92093-0523, USA

[b] DENSO Corporation, Research Laboratories, 500-1 Minamiyama Komenoki, Nisshin Aichi 470-0111, Japan

Received 1 April 2002; received in revised form 1 December 2002; accepted 5 December 2002

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. In comparison to a common multi-microphone technique like beamforming with spectral subtraction, the scheme is shown to enable more accurate speech recognition in the presence of a highly interfering point source and strong background noise.

Keywords: Speech enhancement; Robust speech recognition; Blind source separation; Noisy environments

1. Introduction

Human computer interactions are becoming increasingly important in modern society and people are getting used to interacting with computers on a daily basis. In automobile environments for example, higher flexibility and safety standards can be achieved by using human voice commands to retrieve information from navigation systems or execute simple control tasks. A number of commercial speech recognition systems with various vocabulary size are currently available. However the performance of those systems usually degrades substantially under real-world conditions. In a car, the number of noise sources is very large, quasi-infinite, since vibrations, road, fan and wind noise from open windows generate a continuous, distributed noise background. In addition, several highly interfering point sources like the passenger's voice or music from a loud speaker disturb the driver's voice commands to a speech recognition system. Also the driver's voice may be

[*] Corresponding author.
E-mail addresses: visser@salk.edu (E. Visser), mootsuka@rlab.denso.co.jp (M. Otsuka), tewon@salk.edu (T.-W. Lee).

0167-6393/03/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0167-6393(03)00010-4

distorted by reverberation. The signal uttered by the driver will thus be recorded as a noisy convolved mixture by a microphone on the front panel.

There are two basic ways to improve speech recognition performance in the presence of disturbances: (a) adding a front-end speech enhancement unit improving the spectral quality of the recorded signal and (b) training the speech models of the recognizer engine (in most instances Hidden Markov models (HMM) (Rabiner and Juang, 1993)) on noisy speech databases. The latter option may yield robust recognition accuracy if sufficient noise scenarios are included in the training phase but performance will still deteriorate substantially if noise sound pressure levels are too high or highly interfering speech signals are present. However efficient preprocessing of these perturbations provides a promising solution to this task and is the focus of this paper.

Speech enhancement schemes currently available in the literature can be subdivided into single-microphone and multiple microphone methods. Single-microphone algorithms are most commonly encountered and are solely based on temporal/spectral information about the recorded signals. A variety of schemes combining different time/spectral and cepstral domain based speech processing methods have been recently proposed for robust speech recognition purposes (Droppo et al., 2001; Lieb and Fischer, 2001; Zhu et al., 2001; Macho et al., 2002; Adami et al., 2002). The traditional framework used in single-microphone enhancement techniques is a probabilistic one with statistical models of a speech signal corrupted by additive Gaussian noise (Ephraim and Malah, 1984; Dembo and Zeitouni, 1988). The noise signal estimate is commonly adapted from the most recent recording, i.e. a few seconds before the command is spoken, or voice activity detection algorithms are used to estimate noise power from noisy speech silence intervals. This approach works well when the noise signal is reasonably stationary. Perceptually inspired processing techniques (Hermansky and Morgan, 1994) and variations of cepstral mean subtraction approaches for speech recognition (Atal, 1974) have been successfully applied to handle convolutional noise and speech reverberation as well. However performance is unsatisfactory when strongly reverberated speech signals recorded in non-stationary noise environments are considered or the desired speaker signal is corrupted by highly interfering speech sources.

To overcome the limitations of single-microphone temporal processing methods, spatial information can be exploited by using multiple microphones. In beamforming (Brandstein and Silverman, 1997; Johnson and Dudgeon, 1993) for example, an array of microphones with a known geometry enabling both spatial and temporal measurements of sounds is used to *suppress* interfering signals. Acoustic room modeling and source localization can be performed as well as reverberation be handled to some extent with adaptive algorithms (Brandstein and Silverman, 1997; Johnson and Dudgeon, 1993). Multiple-microphone configurations for speech processing are going to play an ever increasing role in multimedia systems, video-conferencing facilities, computer interfaces etc. (Brandstein and Silverman, 1997; Dahl and Claesson, 1999; Ward et al., 1998; Silverman et al., 1997; Fischer and Simmer, 1996; Mahieux et al., 1996). However, for the successful deployment of microphone setups, it is important that they be able to perform their functions in a robust manner in challenging and uncertain acoustic environments. Also, large microphone arrays are generally required for good performance and implementation of such infrastructure in cars is difficult and costly.

The number of microphones can be drastically reduced by using recently developed second or higher-order decorrelation based, blind source separation algorithms (Parra and Spence, 2000; Bell and Sejnowski, 1995; Lee et al., 1997). These signal processing algorithms exploit spatial information about signal mixtures recorded at a limited number of microphone locations to explicitly *separate* interfering noise signals from the desired source signal. Since they assume no a priori information about the interfering sources, they are particularly suited for environments where the number of disturbance scenarios is virtually unlimited like inside a car where vibrations and other noise originate from many different sources such as an open window. Probabilistic denoising approaches using multiple models for both speech and noise sources (Attias et al., 2001) would require a very large model database to work reliably in such unknown situations.

In this paper, we propose a new robust speech enhancement system for noisy car environments. Our approach consists of making synergistic use of spatial and temporal processing for preprocessing of the noisy speech signal using standard hardware and software requirements and a small microphone setup with two microphones. Moreover these processing methods are largely "blind", i.e. they assume no a priori information about the speech and noise sources involved. We believe that effective use of spatial and temporal processing, as outlined in this paper, provides a promising solution to the challenging problem of robust speech recognition in noisy car environments.

The paper is organized as follows. Section 2 defines the speech enhancement problem analytically. In Section 3 a spatio-temporal speech enhancement scheme is proposed and its subsequent components explained in detail. Section 4 addresses the application of the scheme to real car recordings and speech recognition experiments. The paper concludes with Section 5.

2. Problem formulation

We consider the case where m mixture signals $x_1(t), x_2(t), \ldots, x_m(t)$ composed of m point source signals $s_1(t), s_2(t), \ldots, s_m(t)$ and additive background noise $n(t)$ are recorded at m different microphone locations

$$\boldsymbol{x}(t) = \boldsymbol{A}(t)\boldsymbol{s}(t) + \boldsymbol{n}(t). \tag{1}$$

The $m \times m$ matrix $\boldsymbol{A}(t)$ is called the mixing matrix. One of the sources $s_i(t)$, $i \in [1, m]$ represents the desired speaker signal. The formulation in (1) considers no reverberation and will be referred to as the instantaneous mixture problem. In most practical situations the recorded microphone signals however contain a significant amount of reverberation. Eq. (1) then becomes

$$\boldsymbol{x}(t) = \sum_{\tau=0}^{P} \boldsymbol{A}(\tau)\boldsymbol{s}(t-\tau) + \boldsymbol{n}(t), \tag{2}$$

where P is the convolution order and depends on the environment acoustics. The framework defined in (2) will enable us to address the real car recordings. An important distinction is made between spatially point sources and distributed background noise. 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. This method does not work for background noise however since it originates from a complex combination of 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 characteristics of point source and distributed signals call for appropriate spatial and temporal processing methods.

3. Speech enhancement scheme

The proposed speech enhancement scheme is illustrated by Fig. 1.

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. A supplementary spatial pre-processing step may consist of a source localization unit if the interfering source location is unknown and sufficient pairs of microphones are available. In the following the subsequent stages of the scheme are explained in detail.

3.1. Blind source separation of interfering point sources

In recent years a number of signal processing algorithms have emerged implementing blind

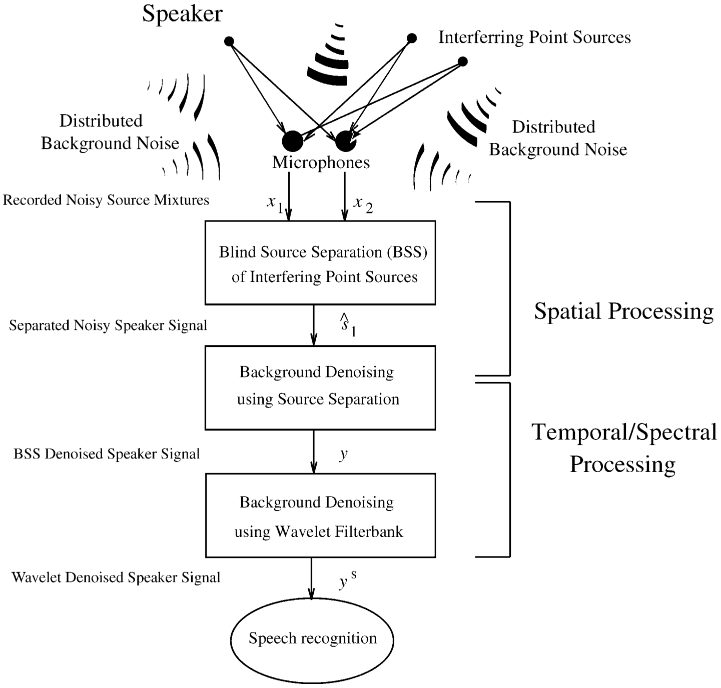
Speaker
Interferring Point Sources
Distributed Background Noise
Distributed Background Noise
Microphones
Recorded Noisy Source Mixtures
x_1
x_2
Blind Source Separation (BSS) of Interfering Point Sources
Separated Noisy Speaker Signal
$\hat{s}_1$
Background Denoising using Source Separation
BSS Denoised Speaker Signal
y
Background Denoising using Wavelet Filterbank
Wavelet Denoised Speaker Signal
y^s
Speech recognition
Spatial Processing
Temporal/Spectral Processing

Fig. 1. Proposed speech enhancement scheme.

source separation (BSS) of mixture signals into its components by decorrelating their higher-order statistics. No a priori information about the sources is assumed other than general statistical properties of their probability density distribution. These BSS methods are also referred to as independent component analysis (ICA). The common approach consists in finding an unmixing matrix $\boldsymbol{W}(t)$

$$\hat{\boldsymbol{s}}(t) = \boldsymbol{W}(t)\boldsymbol{A}(t)\boldsymbol{s}(t)$$

such that higher (larger than second) order correlation between separated sources $\hat{\boldsymbol{s}}$ is minimized. A number of ICA algorithms (Bell and Sejnowski, 1995; Hyvaerinen and Oja, 1997) have been developed for the instantaneous case in the time domain and for the convolved case in the frequency domain (Cardoso and Souloumiac, 1993). A number of successful applications have been reported for biomedical applications (Makeig et al., 1995). Blind separation of reverberated speech sources has been investigated in (Lee et al., 1997, 1998).

However, the second-order decorrelation approach presented in (Parra and Spence, 2000) yielded the most consistent separation performance in a number of experiments (see Section 4 below). The multiple adaptive decorrelation (MAD) algorithm outlined in (Parra and Spence, 2000) is designed for separation of non-stationary convolved signal mixtures. In MAD, it is assumed that m original sources $\boldsymbol{s}(t) = [s_1(t)s_2(t)\cdots s_m(t)]$ can be recovered from m recorded mixtures $\boldsymbol{x}(t) = \sum_{\tau=0}^{P}\boldsymbol{A}(\tau)\boldsymbol{s}(t-\tau) = [x_1(t)x_2(t)\cdots x_m(t)]$ by finding a sequence of $m \times m$ unmixing filter matrices $\boldsymbol{W}(\tau)$ such that

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

Q being the filter length. The unmixing filter computation is executed in the frequency domain where

$$\boldsymbol{X}(\omega,t) \simeq \boldsymbol{A}(\omega)\boldsymbol{S}(\omega,t),$$

$X(\omega,t)$ being the spectrogram obtained by consecutively computing the short time Fourier transform of length T (where $T \gg P$, the convolution order), of $x(t)$ at each time instant t in an overlap-shift fashion (Parra and Spence, 2000).

If the cross-correlation of the measurements is denoted by $\hat{R}_x(\omega,t) = \mathsf{E}[\boldsymbol{X}(\omega,t)\boldsymbol{X}^H(\omega,t)]$ [1] and that of the sources by $\hat{\Lambda}_s(\omega,t) = \mathsf{E}[\boldsymbol{S}(\omega,t)\boldsymbol{S}^H(\omega,t)]$, $\boldsymbol{W}(\omega)$ is found by minimizing

$$\hat{\boldsymbol{W}},\hat{\Lambda}_s = \underset{\hat{\boldsymbol{W}},\hat{\Lambda}_s}{\operatorname{argmin}} \sum_t \sum_{\omega=1}^{T} \| \boldsymbol{W}\hat{R}_x(\omega,t)\boldsymbol{W}^H - \Lambda_s(\omega,t)\|^2, \quad (3)$$

$$\text{s.t. } \boldsymbol{W}(\tau) = 0 \ \forall \tau > Q,\ Q \ll T, \quad (4)$$

$$\boldsymbol{W}_{ii}(\omega) = 1. \quad (5)$$

Since the source correlation is updated as $\hat{\Lambda}_s(\omega,t) = \operatorname{diag}[\boldsymbol{W}(\omega)\hat{R}_x(\omega,t)\boldsymbol{W}^H]$, the cost basically minimizes the off-diagonal elements of the cross-correlation matrix $\hat{R}_x(\omega,t)$. Constraint (4) imposes that the filter length Q be much smaller than T to solve the frequency permutation problem (Parra and Spence, 2000). Also scaling issues are solved by fixing the diagonal elements of the filter matrices to unity (constraint (5)). One finally obtains the learning rule (Parra and Spence, 2000)

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

where μ is the learning rate and $E(\omega,t) = \boldsymbol{W}\hat{R}_x(\omega,t)\boldsymbol{W}^H - \hat{\Lambda}_s(\omega,t)$. The filter $\boldsymbol{W}(\omega)$ is learned over the complete available mixture data by solving problem (3) with a moving window of $T/2$. At every reoptimization, the cross-correlations of the mixtures are updated using $\hat{R}_x(\omega,t) = (1-\gamma)\hat{R}_x(\omega,t) + \boldsymbol{X}(\omega,t)\boldsymbol{X}^H(\omega,t)$ where γ is a forgetting factor. In (Parra and Spence, 2000), an additional learning rule to remove additive distributed noise sources parallel to removing point sources is presented. This requires however a number of microphones much larger than the number of point sources and corresponding prohibitive computational load.

The algorithm assumes no a priori information about the mixed sources and therefore implements so-called "blind" source separation. Note that dereverberation of the desired speaker signal and separation from interfering sources are performed at the same time. The approach has shown robust and near real time performance in a number of applications and will be discussed further below. Moreover, the diagonal filter element equality constraints (5) in optimization problem (3) ensure that the dominant speaker voice will be separated at the microphone position at which its amplitude is highest for most of the duration of the signal. Since the unmixing filters are computed using data over the whole signal time range available, no temporarily dominating disturbance can perturbate the output order of the separated sources. This makes an additional algorithmic step to determine which of the separated sources is the speaker voice unnecessary.

3.2. Denoising of distributed noise signals

This previous BSS step will remove highly interfering point sources from the recorded mixtures but leaves background noise distributed in space unaffected. In (Parra and Spence, 2000), a learning rule for removing the distributed background noise $n(t)$ simultaneously to point sources is presented. This would however require a number of microphones much larger than the number of point sources. Therefore an alternative procedure for reducing the noise level using the same BSS algorithm in an additional step is proposed next.

3.2.1. Background denoising using source separation

In this denoising step, the separated speaker's source signal from the previous BSS stage and an artificially generated 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. This enhancement stage thus combines spatial and temporal processing.

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 noise-only intervals in the speaker's signal and taking the complete phase information of the speaker's signal. Using the estimated noise power spectrum in these intervals and phase

[1] The index H denotes the complex hermitian transpose.

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.

This new noise signal is constructed in the following manner. Consider the separated speaker's signal $\hat{s}_1(t)$, $t \in [t_o, t_f]$, and noise-only signal $s_n(t) = \hat{s}_1(t)$, $t \in [t_o, t_1]$, where $T_r = t_1 - t_o$. The Fourier transform of $s_n(t)$ yields

$$S_n(\omega) = |S_n(\omega)| \mathrm{e}^{\mathrm{i}\Phi_{S_n}(\omega)}. \tag{6}$$

By consecutively computing the Fourier transform of length T_r of $\hat{s}_1(t)$

$$S_{T_r}(\omega) = |S_{T_r}(\omega)| \mathrm{e}^{\mathrm{i}\Phi_{S_{T_r}}(\omega)} \tag{7}$$

and replacing its magnitude spectrum by the noise magnitude yielding

$$S^a_{T_r}(\omega) = |S_n(\omega)| \mathrm{e}^{\mathrm{i}\Phi_{S_{T_r}}(\omega)}, \tag{8}$$

a noise signal $S^a(\omega)$ is constructed having the noise power characteristics but the speaker's signal's phase. The time-domain version $s^a(t)$ of $S^a(\omega)$ 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 (Parra and Spence, 2000), i.e. taking Fourier/inverse Fourier transforms of window length T_r and shifting the window by $T_r/2$. Using $s^a(t)$ and $\hat{s}_1(t)$ as inputs to the MAD algorithm will denoise $\hat{s}_1(t)$ yielding the BSS denoised speaker signal $y(t)$. As opposed to spectral subtraction, musical noise artifacts resulting from power over-subtraction (Ephraim and Malah, 1984) are minimized since the generated artifactual spectrum would increase the correlation between signals again. The extension of the formulation to the case where instationary noise power is estimated in a time-varying manner along $\hat{s}_1(t)$ from speech-absent time-intervals as indicated before is straightforward.

3.2.2. Background denoising using wavelet filterbank

The previous denoising step may not remove all distributed noise since its power may be underestimated on short, noise-only sample intervals especially in the low frequency subbands. Therefore a wavelet filterbank exploiting the statistical difference of speech and noise is added to perform complementary denoising using a technique referred to as wavelet coefficient shrinkage (Donoho et al., 1995).

The statistical distribution of background noise can in general be approximated by a Gaussian distribution by invoking the central limit theorem (Rabiner and Juang, 1993). This is a realistic assumption especially when the noise sources appear at the same time. 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 will be small. Statistically speaking, the latter have a high probability of being zero and can thus be eliminated by applying a coring or shrinkage function (Donoho et al., 1995).

The continuous wavelet transform provides such a signal mapping into sparse subspaces using linear filters (Vetterli and Kovacevic, 1995). Studies have shown that wavelet coefficients are naturally sparse (Buccigrossi and Simoncelli, 1997). The transform is efficiently implemented by using an oversampled, shift-invariant multi-resolution filter bank (Holschneider et al., 1989). The filterbank decomposes each signal into orthogonal subbands, each subband having a constant subband frequency width-to-center frequency ratio in analogy to the mel scale used in speech recognition (Rabiner and Juang, 1993). Therefore a sparse representation and physiologically intuitive frequency subdivision are obtained at the same time.

Fig. 2 illustrates the wavelet filterbank based denoising approach.

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^s_i for every i such that the joint probability $P(y_i - y^s_i|\sigma_i)P(y^s_i)$ of the independent noise $(y_i - y^s_i)$ and speech (y^s_i) distributions is maximized, σ_i being the noise level.

The priors are given by the Gaussian $P(y_i - y^s_i|\sigma_i) \simeq \mathrm{e}^{-\left((y_i - y^s_i)^2/2\sigma^2_i\right)}$ and the Laplacian distribution $P(y^s_i) \simeq \mathrm{e}^{-|y^s_i|}$. When the log-likelihood is considered, one obtains (Donoho et al., 1995):

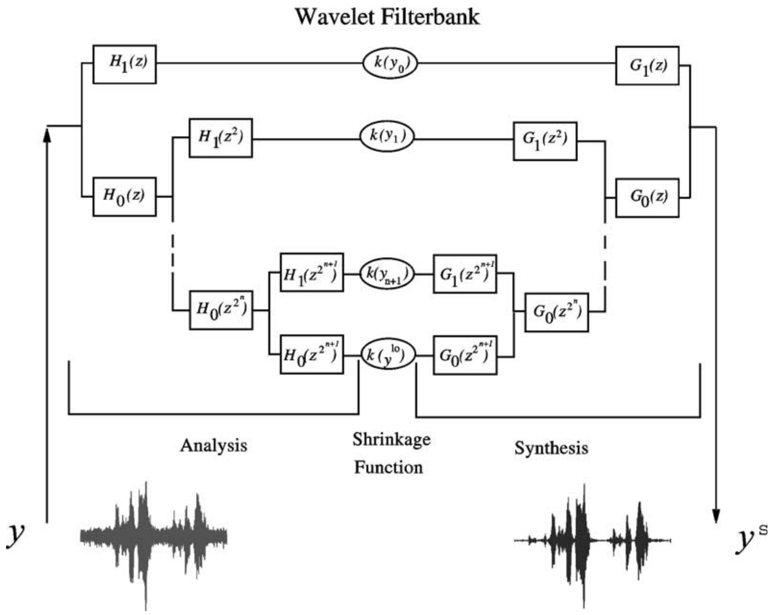

Fig. 2. Denoising using a wavelet filterbank: signal analysis is implemented using upsampled versions of orthogonal low- and high-pass filters H_0 and H_1, respectively. A shrinkage function k is applied. Signal reconstruction is done using upsampled versions of synthesis filters G_0 and G_1 (Vetterli and Kovacevic, 1995).

$$y_i^s = \arg\min_{y_i^s}\left(\sum_j \frac{(y_i(j)-y_i^s(j))^2}{2\sigma_i^2} + \sum_j |y_i^s(j)|\right)$$

whose approximate analytical solution is given by the shrinkage function k (Donoho et al., 1995):

$$y_i^s = k(y_i) = \mathrm{sign}(y_i)\max\left([(|y_i| - \sqrt{2}\sigma_i)\ 0]\right). \tag{9}$$

An accurate estimation of the noise level σ_i is essential and can be supplied by conventional techniques (Mokbel and Chollet, 1995; Ephraim and Malah, 1984). The resulting frequency subdivision as well as the subband wavelet coefficients before and after denoising are illustrated by Fig. 3.

4. Real recording and recognition experiments

To formally quantify the performance of the proposed speech enhancement scheme in comparison to commonly used multi-microphone techniques, speech recognition experiments were carried out on speech data recorded in a real noisy car environment. The experiments were specifically setup to generate a worst case scenario of combined interfering point source and diffuse background noise to illustrate the scheme's robustness in a complex, real-life setting. The quantitative speech enhancement of the individual spatial and temporal processing stages of the proposed scheme can also be adequately measured in this way and compared to standard enhancement techniques. Moreover, since the clean desired speaker signals are unknown, speech enhancement cannot be quantified in terms of SNR improvement so speech recognition accuracy was used instead. The design was thus not intended to represent an average noise situation encountered in car environments nor are the obtained speech recognition accuracies meant to be final achievable standards for car applications.

Also, no stereo, noisy benchmark database for speech recognition experiments involving multiple microphone setups is available in the speech community. Hence a multi-channel test database had to be specifically recorded for this application. On the other hand a single microphone, connected digits database is given by the AURORA 2

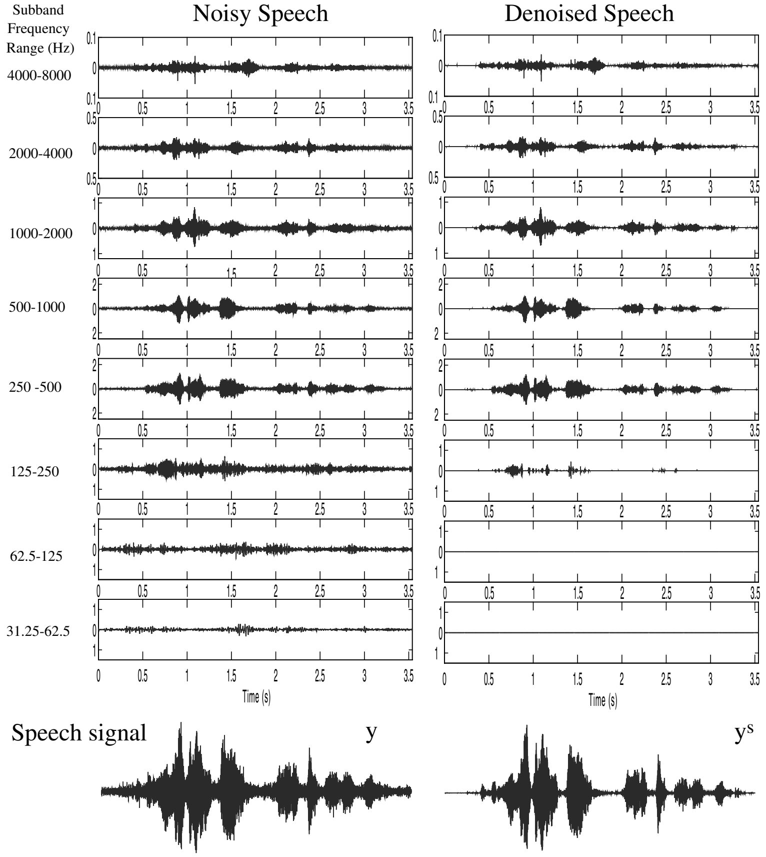

Fig. 3. Subbands No. 0, ..., 6 and lo (=lowest frequency channel) (from top to bottom) with sampling frequency 16 kHz.

benchmark dataset (Hirsch and Pearce, 2000) which considers a large variety of noise scenarios. Since the final step in a speech recognizer is performed on single channel speech data, the AURORA 2 package can still be used in continuous digit recognition experiments for robustly training the HMM of the recognition engine (Rabiner and Juang, 1993).

4.1. Experimental setup

The following recordings were carried out inside a Toyota Corolla (CE, model 2000) automobile by two male speakers. The driver was uttering digit sequences while the passenger seated next to the driver was talking simultaneously on his cell phone while driving at 40 mph with open windows and fan turned on. The car radio (factory-built-in, cassette module) played pop-music on an FM station at low to medium volume (knob position 7–11 o'clock) through four loudspeakers (150 W) inside the car (two speakers in the front doors and two in the back of the car). Two uni-directional, stereo microphones (SONY Stereo/Zoom ECM-ZS90) were attached on each side of the rear view mirror to record the speech mixtures. The distance between microphones was 15 cm and the recorded speech data was sampled at 8 kHz. The digit sequences had a maximum length of 5 digits per sentence with varying silence intervals between digits. The task at hand is thus continuous word recognition encountered in continuous digit dialing.

In technical terms, the driver's instructions are perturbated by a highly interfering passenger's signal and a noisy background. To determine the signal-to-noise ratio (SNR), the resulting noise signal variance was estimated from the difference between the variance of the noisy recorded digit utterances and the signal variance of clean digits recorded with the same microphone setup inside the silent car environment, averaged over 1 s frames (average digit duration). While the SNR of the mixture recorded by the microphone on the driver's side ranged from 2 to 5 dB due to the passenger's interference alone, its SNR due to combined interfering signal and background noise varied from 0 to 4 dB. Before each test utterance, both the driver and passenger were silent for 1 s to estimate the noise background constituted by fan, wind and road noise as well as music from the radio.

Two microphones were considered sufficient since the driver's position is known and fixed and no additional microphones are required to estimate the desired speech source location. Also the most interfering source originates from the front passenger seat so the microphone setup behind the rear window mirror is ideally positioned to capture sufficient spatial information to perform separation.

4.2. Standard reference method

The performance of the proposed speech enhancement scheme is compared to beamforming followed by spectral subtraction.

Rather than trying to achieve explicit source *separation*, interfering sources can be *suppressed* by aligning and summing up source mixtures recorded at different locations in a microphone array (Johnson and Dudgeon, 1993). The conceptual idea behind this common spatial processing technique called beamforming is that the desired signal is emphasized by in phase summation while the interfering sources are gradually reduced because of out-of-phase summation (Johnson and Dudgeon, 1993). If this delay and sum procedure is carried out in the frequency domain, reverberation of source signals is minimized as well (Brandstein and Silverman, 1997). A number of techniques exist to estimate the correct delay (Carter, 1993; Chow and Schultheiss, 1981; Fertner and Sjolund, 1986; Friedlander and Porat, 1984; Knapp and Carter, 1976). In our experiments, a delay of 3–4 taps could be estimated from the fixed driver's location and the microphone positions. Also, since only one pair of microphones was available, the beamforming procedure was basically reduced to realigning the recorded mixtures with the estimated time delay and taking the arithmetic average. It is noted that better beamformers exist using more microphones or sharper microphone directivity as well as a priori knowledge about the desired speaker and interfering sources. However the emphasis in this study was on comparing speech enhancement methods using as little a priori information as possible and only two microphones.

The standard temporal processing method for removing additive, distributed background noise is spectral subtraction (Ephraim and Malah, 1984; Dembo and Zeitouni, 1988; Boll, 1979; Berouti et al., 1979). We applied a basic linear spectral subtraction method in which the noise power was estimated from the noise-only, pre-speech 1 s intervals and subtracted from the noisy speech signal power (Boll, 1979). The denoised signal power was used together with the noisy speech phase to obtain the denoised driver's signal.

4.3. Implementation of speech enhancement scheme

Several BSS approaches for tackling the convolved mixture case have been mentioned. Although the time-domain ICA methods were easier to implement and faster to execute, the frequency domain algorithm presented in (Parra and Spence, 2000) consistently yielded better performance in our experiments. The frequency-domain version of the MAD algorithm has been implemented in C code. It is executable in 2 times slower than real-time on a 550 MHz PC but its processing speed can be optimized for real-time applications (Parra and Spence, 2000). The algorithm successfully separated the passenger's voice from the driver's voice.

However, as illustrated in Fig. 4, both separated source files still contain diffuse noise originating from the open window, vibrations of the car, fan or music from the radio. It should be noted that if the driver's side signal is fed into the MAD algorithm as the first input file and the passenger's side recorded signal as the second one, the order of the MAD output files will correspond to the driver followed by passenger files also. This is due to

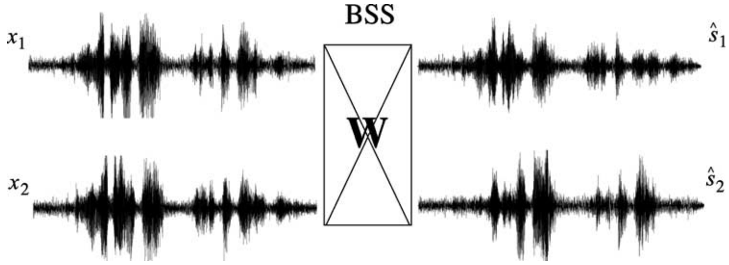

Fig. 4. Blind source separation (BSS) of interfering point sources: Input recorded noisy source mixtures (left) and output separated noisy sources (right).

constraint (5) in problem (3) which will preserve the highest amplitude signal on the corresponding car side. Since the driver's voice is stronger in the driver's side recorded signal and correspondingly for the passenger's, the output order of the BSS step is the same than the input order and no additional methods such as driver's speech detection or speaker recognition are necessary.

In order to remove the noise background, denoising using BSS as outlined in Section 3.2.1 was performed in a next stage. The noise power for the new artificial noise signal was estimated from noise-only, 1 s pre-speech intervals like in the spectral subtraction procedure. The separated driver's signal from the previous speech enhancement stage and the latter artificial noise signal were then used as new inputs to the MAD algorithm (Parra and Spence, 2000). Background noise is removed and the output driver's signal has a higher SNR. The procedure is illustrated by Fig. 5.

Additional denoising was performed with a wavelet filterbank (illustrated by Fig. 3). Music from the radio 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 (Rabiner and Juang, 1993). Wavelet coefficient shrinkage was only done in the three lowest frequency bands since the non-linear operation of thresholding the coefficients causes important phase distortions in high frequency bands. The BSS denoising and wavelet shrinkage steps are thus complementary in frequency space. Twelve tap Daubechies wavelet filters were chosen for H_0, H_1, G_0 and G_1 (Vetterli and Kovacevic, 1995). Wavelet coefficients were shrinked by applying (9). The noise threshold levels σ_i was determined from the maximum absolute values of the wavelet coefficients in the noise-only, pre-speech onset time interval.

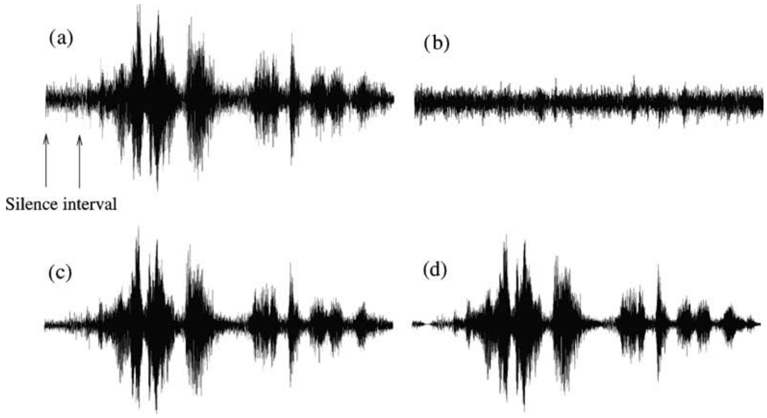

Fig. 5. Illustration of BSS denoising: (a) BSS separated driver's signal; (b) artificial noise-only signal; (c) BSS denoised and separated driver's signal; (d) denoised signal obtained by spectral subtraction applied to (a) (noise power estimate from silence interval).

Fig. 6 compactly illustrates the obtained speech enhancement at each stage on another example.

Case (a) and (b) clearly illustrate that the highly interfering passenger's voice is removed from the recorded mixture leading to a significant SIR improvement. Further enhancement is achieved in the denoising stages using BSS and wavelet shrinkage as indicated by case (d) and (e). Case (c) and (d) compare spectral subtraction and denoising using BSS. Although no clear difference can be seen from the time-domain signals, the speech recognition results in the next section clarify the quantitative performance comparison.

4.4. Speech recognition results

As indicated before, the speech recognizer as well as a multiple noise condition database for training the HMM was provided by the AURORA 2 benchmark dataset (Hirsch and Pearce, 2000). The speech feature extraction front-end FE_v2_0 (Hirsch and Pearce, 2000) 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 digit sequences with a total number of 147 digits. Six different speech recognition cases were considered, using the MFCC's extracted from the

- *Case 1:* recorded driver's side signal $(x_1(t))$,
- *Case 2:* driver's signal processed by beamforming followed by spectral subtraction,
- *Case 3:* separated driver's signal using BSS $(\hat{s}_1(t))$,
- *Case 4:* denoised, separated driver's signal using BSS followed by spectral subtraction,

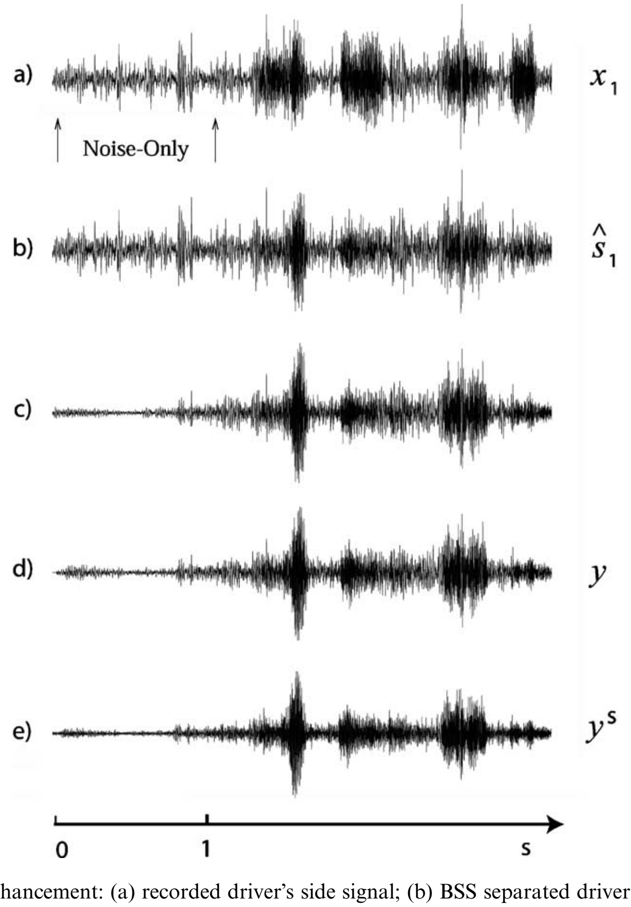

Fig. 6. Illustration of speech enhancement: (a) recorded driver's side signal; (b) BSS separated driver's signal; (c) spectral subtracted, BSS separated driver's signal; (d) BSS denoised and separated driver's signal; (e) wavelet-BSS denoised, separated driver's signal.

- *Case 5:* denoised, separated driver's signal using BSS for separation followed by an additional BSS stage for denoising ($y(t)$) and
- *Case 6:* denoised, separated driver's signal using BSS for separation, BSS for denoising and additional denoising using wavelet coefficient shrinkage ($y^s(t)$).

The recognition results are as follows. The digit recognition accuracy dropped to 46.9% (Case 1) when applying the recognizer engine to the mixture signal recorded by the microphone on the driver's side. The interfering passenger's voice as well as high background noise SPL contribute to this unacceptable performance. The common 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% (Case 2). This observed performance reflects insufficient removal of the highly interfering passenger's voice and severe low frequency car vibrations and fan noise. By comparison, the BSS step of our speech enhancement scheme alone achieved an accuracy of 72.1% (Case 3). While applying spectral subtraction at this point yields an accuracy of 72.8% (Case 4), denoising using BSS achieves an accuracy of 74.8% (Case 5). Hence, although the noise power was estimated in both cases on the same noise-only, 1 s 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% (Case 6) was obtained by denoising the separated driver's signal further with wavelet coefficient shrinkage. This latter accuracy indicates the necessity of low frequency filtering with the wavelet filterbank. Most of the noise power is contained in the low frequency bands (car vibrations) and underestimated by the denoising procedure using BSS or spectral subtraction since their noise power estimate relies on short, 1 s intervals only. The latter denoising methods therefore essentially remove high frequency noise. The best reference accuracy of 92.5% was determined in recognition experiments with a clean recorded digit dataset of similar size than the noisy test dataset. Digits were uttered alternatively by the two male speakers in the driver's seat and recorded with the same microphone set in the noise-free car environment. This reference accuracy quantifies the effects of microphones, speakers and recording environment different from the ones used in the AURORA 2 training database for the HMM.

To summarize, more than 30% recognition accuracy improvement was obtained with the proposed speech enhancement scheme compared to 10% with standard techniques. The bulk of the performance improvement was achieved by applying a BSS algorithm processing spatial information recorded by only two microphones.

4.5. Discussion

The results show significantly improved speech enhancement compared to standard techniques. The commonly used spatial processing approach of beamforming yields significantly worse performance than BSS when only a few microphones are available. Microphone arrays as well as more a priori knowledge about the desired and interfering signal covariances are necessary for adaptive beamforming algorithms to handle reverberation and interference efficiently. Temporal processing via spectral subtraction was shown to be inferior to combined BSS denoising and wavelet coefficient shrinkage.

The ASR results on this specific dataset were primarily intended to quantify the gradual *performance improvement* at each enhancement step and illustrate the scheme's superiority over commonly used multi-microphone techniques in a complex real recording situation. Its performance in more controlled settings (i.e. noise scenarios with only one interfering point source or only background noise) can be readily inferred from the experimental results shown above. If the digit utterance is masked by an interfering point source only, the first BSS stage separates it from the driver's side recorded signal while the denoising stages have no effect with no noise left in the silence interval. Hence, while the scheme yields at least an enhancement observed from Case 1 to Case 3 (∼25%), a considerably smaller improvement from Case 1 to Case 2 (∼10%) is found for beamforming followed by spectral subtraction.

The mixed (interfering point source—background noise) scenario in Case 2 can indeed be regarded either as a single point source or background noise situation since the reference methods used are rather insensitive to the noise's spatial properties: the beamforming method emphasizes the digit signal based on knowledge of the driver's position only and spectral subtraction is a single channel denoising method. If only background noise is considered, the first BSS stage leaves the recorded signal's SNR on the driver's side basically unchanged while removing the driver's utterance from the passenger's side recorded mixture. The proposed scheme's performance in this noise scenario is thus similar to the improvement from Case 3 to Case 6 (~8%), with the residual interfering point source present in Case 3 considered part of the background noise. Using the same arguments than in the single interfering point source scenario, this performance is comparable to the standard denoising improvement observed from Case 1 to Case 2, which is of the same order of magnitude (~10%). To summarize, a considerable advantage of the proposed scheme over the studied reference methods stems from the BSS algorithm's ability to exploit spatial information for efficiently separating interfering point sources. If only distributed background noise is considered, significant improvements over standard schemes are not expected in general as noise power estimation is limited to the pre-speech onset, silence interval, which may not be sufficient to characterize highly non-stationary noise.

Several optimizations to the present scheme are considered to improve the final achievable accuracy. Firstly the BSS denoising and wavelet denoising step can yield better performance if the time-varying nature of the background noise is taken into account. This can be done by segmenting the noisy speech signal into noise-only—mixed speech/noise intervals using voice detection algorithms (Kim et al., 1998) and adapting the noise power estimate along the noisy speech signal.

Moreover additional microphone pairs should be included if interfering point sources originate elsewhere than the passenger's seat and the present microphone setup is not optimally positioned. If specific noise sources are known to cause significant speech degradation in a particular application, much benefit is expected from measuring these signals by placing additional microphones in an ad hoc manner and feeding them forward to the speech enhancement system for direct compensation. Also, the accuracy of location estimation is limited by room reverberation; some interesting papers in this context are (Champagne et al., 1996; Li and Hoffman, 1999; Marro et al., 1998). The techniques are centered around time-delay estimation between microphone pairs and subsequent triangulation of the estimated delays. Additional microphones as well as the accompanying source estimation techniques will robustify the performance of the BSS step in the enhancement scheme. The final number of microphones to use for a specific application will depend on a trade-off between the performance enhancement they allow and the computational as well as economic cost to implement them.

The present speech enhancement scheme is almost completely speech or noise model independent. As pointed out before, the number of noise scenarios in a real-life environment may be virtually unlimited. However as the noise is preprocessed by the "blind" stages of the present scheme to levels which can be handled by a priori models, subsequent denoising fine-tuning with speech and noise models suitably identified for a particular scenario will improve overall performance. We therefore consider including a number of off-line learned speech and noise models in our scheme and perform additional Wiener filtering with statistically weighted covariance models in a final enhancement step.

5. Conclusions

A new speech enhancement scheme for robust speech recognition in noisy car environments has been presented. The approach uses a spatio-temporal signal processing strategy to remove noise originating from spatially localized point sources and spatially distributed background noise in subsequent enhancement steps. The scheme provides a framework to address reverberation, highly interfering sources and background noise in car environments without the need of a technically demanding multiple microphone array infrastructure nor any prior speech or noise source models involved. The method works with standard hardware and software requirements and requires only two microphones. It was shown to outperform common speech enhancement techniques based on beamforming combined with spectral subtraction.

Future improvements to the scheme include incorporation of additional microphones, source location estimation techniques at the front-end to optimize separation of interfering sources, and a multi-model based, Wiener filtering stage at the back-end to perform additional denoising fine-tuning. The scheme is expected to enhance recognition accuracies in very noisy situations and be applicable to a large number of real-life environments.

References

Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S., 2002. QUALCOMM-ICSI-OGI features for ASR. In: Proc. ICSLP2002, pp. 21–24.

Atal, B.S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55, 1304–1312.

Attias, H., Platt, J.C., Acero, A., Deng, L., 2001. Speech denoising and dereverberation using probabilistic models. In: Leen, T. (Ed.), Advances in Neural Information Processing Systems, vol. 13. MIT Press, Cambridge, MA.

Bell, A.J., Sejnowski, T.J., 1995. An information-maximisation approach to blind separation and blind deconvolution. Neural Comput. 7 (6), 1004–1034.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, pp. 208–211.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-29, 113–120.

Brandstein, M., Silverman, H., 1997. A practical methodology for speech source localization with microphone arrays. Comput. Speech Lang. 11 (2), 91–126.

Buccigrossi, R.W., Simoncelli, E.P., 1997. Progressive wavelet image coding based on a conditional probability model. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing, Munich.

Cardoso, J.-F., Souloumiac, A., 1993. Blind beamforming for non gaussian signals. IEEE Proc. F 140, 362–370.

Carter, G.C. (Ed.), 1993. Coherence and Time-Delay Estimation. IEEE Press book.

Champagne, B., Bédard, S., Stéphenne, A., 1996. Performance of time-delay estimation in the presence of room reverberation. IEEE Trans. Speech Audio Process 4 (2), 148–152.

Chow, S.-K., Schultheiss, P.M., 1981. Delay estimation using narrow-band processes. IEEE Trans. Acoust. Speech Signal Process.

Dahl, M., Claesson, I., 1999. Acoustic noise and echo cancelling with microphone array. IEEE Trans. Veh. Technol. 48 (5), 1518–1526.

Dembo, A., Zeitouni, O., 1988. Maximum a posteriori estimation of time-varying ARMA processes from noisy observations. IEEE Trans. Acoust. Speech Signal Process. 36 (4), 471–476.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1995. Wavelet shrinkage: Asymptopia? J. Roy. Statist. Soc. Ser. B 57, 301–337.

Droppo, J., Deng, L., Acero, A., 2001. Evaluation of the SPLICE algorithm on the AURORA 2 database. In: Proc. Eurospeech 2001, Aalborg, Denmark, pp. 217–220.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32, 1109–1121.

Fertner, A., Sjolund, A., 1986. Comparison of various time delay estimation methods by computer simulation. IEEE Trans. Acoust. Speech Signal Process.

Fischer, S., Simmer, K.U., 1996. Beamforming microphone arrays for speech acquisition in noisy environments. Speech Comm. 20 (3–4), 215–227.

Friedlander, B., Porat, B., 1984. A parametric technique for time delay estimation. IEEE Trans. Aerospace Electron. Syst. 1 (November).

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, ISCA ITRW ASR2000 "Challenges for the New Millennium", Paris, September.

Holschneider, M., Kronland-Martinet, R., Morlet, J., Tchamitchian, Ph., 1989. A real-time algorithm for signal analysis with the help of the wavelet transform. In: Wavelets, Time-Frequency Methods and Phase Space. Springer-Verlag., pp. 289–297.

Hyvaerinen, A., Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Comput. 9, 1483–1492.

Johnson, D.H., Dudgeon, D.E., 1993. Array Signal Processing: Concepts and Techniques. Prentice Hall, Englewood Cliffs.

Kim, D.Y., Un, C.K., Kim, N.S., 1998. Speech recognition in noisy environments using first-order vector Taylor series. Speech Comm. 24, 39–49.

Knapp, C.H., Carter, G.C., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24 (4), 320–326.

Lee, T.-W., Bell, A., Lambert, R.H., 1997. Blind separation of delayed and convolved sources. In: Advances in Neural Information Processing Systems, vol. 9. Cambridge, MA, pp. 758–764.

Lee, T.-W., Ziehe, A., Orglmeister, R., Sejnowski, T., 1998. Combining time-delayed decorrelation and ICA: towards solving the cocktail party problem. In: Proc. ICASSP, Seattle, WA, pp. 1249–1252.

Li, Z., Hoffman, M.W., 1999. Evaluation of microphone arrays for enhancing noisy and reverberant speech for coding. IEEE Trans. Speech Audio Process. 7 (1), 91–95.

Lieb, M., Fischer, A., 2001. Experiments with the Philips continuous ASR system on the AURORA noisy digits database. In: Proc. Eurospeech 2001, Aalborg, Denmark, pp. 625–628.

Macho, D., Mauuary, L., No, B., Cheng, Y.M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of a noise-robust DSR front-end on AURORA databases. In: Proc. ICSLP 2002, pp. 17–20.

Mahieux, Y., Le Tourneur, G., Saliou, A., 1996. A microphone array for multimedia workstations. J. Audio Eng. Soc. 44 (5), 365–372.

Makeig, S., Bell, A.J., Jung, T.-P., Sejnowski, T.J., 1995. Independent component analysis of electroencephalographic data. In: Mozer, M. et al. (Eds.), Advances in Neural Information Processing Systems, vol. 8. MIT Press, Cambridge, MA.

Marro, C., Mahieux, Y., Simmer, K.U., 1998. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. IEEE Trans. Speech Audio Process. 6 (3), 240–259.

Mokbel, C.E., Chollet, G.F.A., 1995. Automatic word recognition in cars. IEEE Trans. Speech Audio Process. 3 (5), 346–356.

Parra, L., Spence, C., 2000. Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Process. 8, 320–327.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice Hall, NJ.

Silverman, H.F., Patterson, W.R., Flanagan, J.L., Rabinkin, D., 1997. A digital processing system for source location and sound capture by large microphone arrays. In: 1997 IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, Munich, Germany, vol. 1, pp. 251–254.

Vetterli, M., Kovacevic, J., 1995. Wavelets and Subband Coding. Prentice-Hall, NJ.

Ward, D.B., Williamson, R.C., Kennedy, R.A., 1998. Broadband microphone arrays for speech acquisition. Acoust. Aust. 26 (1), 17–20.

Zhu, Q., Cui, X., Iseli, M., Alwan, A., 2001. Noise robust feature extraction for ASR using the AURORA 2 database. In: Proc. Eurospeech 2001, Aalborg, Denmark, vol. 1, pp. 185–188.