Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 134–149
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Analysis of CFA-BF: Novel combined fixed/adaptive beamforming for robust speech recognition in real car environments

John H.L. Hansen *, Xianxian Zhang

CRSS: Center for Robust Speech Systems, Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Electrical Engineering, EC33, P.O. Box 830688, Richardson, Texas 75083-0688, USA

Received 16 November 2006; received in revised form 26 March 2009; accepted 4 September 2009

Abstract

Among a number of studies which have investigated various speech enhancement and processing schemes for in-vehicle speech systems, the delay-and-sum beamforming (DASB) and adaptive beamforming are two typical methods that both have their advantages and disadvantages. In this paper, we propose a novel combined fixed/adaptive beamforming solution (CFA-BF) based on previous work for speech enhancement and recognition in real moving car environments, which seeks to take advantage of both methods. The working scheme of CFA-BF consists of two steps: source location calibration and target signal enhancement. The first step is to pre-record the transfer functions between the speaker and microphone array from different potential source positions using adaptive beamforming under quiet environments; and the second step is to use this pre-recorded information to enhance the desired speech when the car is running on the road. An evaluation using extensive actual car speech data from the CU-Move Corpus shows that the method can decrease WER for speech recognition by up to 30% over a single channel scenario and improve speech quality via the SEGSNR measure by up to 1 dB on the average.

Keywords: Array processing; Robust speech recognition; In-vehicle speech systems; Beamforming

1. Introduction

The increased use of mobile telephones in cars has created a greater demand for hands-free, in-car installations. Many countries now restrict the use of hand-held cellular technology while operating a vehicle (Komarow, 2000). As such, there is a greater need to have reliable voice capture within automobile environments. However, the distance between a hands-free car microphone and the speaker will cause a severe loss in speech quality due to changing acoustic environments. Therefore, the topic of capturing clean and distortion-free speech under distant talker conditions in noisy car environments has attracted much attention. Earlier studies on signal channel speech enhancement offer one viable path for signal quality improvement (Deller et al., 2000) and speech recognition advancements (Hansen and Clements, 1991; Hansen, 1994; Pellom and Hansen, 1998; Jensen and Hansen, 2002) in the car environment. Dual-channel methods can also improve speech quality as well including such methods as ACE-1, ACE-2 (auditory constrained iterative speech enhancement (Nandkumar and Hansen, 1995; Hansen and Nandkumar, 1995), but multi-microphone array solutions have a greater potential to track speakers and time varying background noise. Microphone array processing and beamforming is one promising area which can yield effective performance.

The classic array beamforming method is delay-and-sum beamforming (DASB), and is based on applying time shifts to a set of microphone array signals to compensate for the

* Corresponding author. Tel.: +1 972 883 2910; fax: +1 972 883 2710.
E-mail address: John.Hansen@utdallas.edu (J.H.L. Hansen).
URL: http://crss.utdallas.edu (J.H.L. Hansen).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.09.001

propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. This method is very simple and robust if we know the direction of the speech source and the number of microphones and microphone spacing is selected appropriately. A simple DASB approach has been shown to be effective for real in-vehicle systems by Plucienkowski et al. (2001). However, if the source location changes during operation, this method will be less effective due to the mismatch in estimating the delays between the microphones. Another practical problem of DASB is that the theoretical maximum noise attenuation $10\log_{10} M$ (Haykin et al., 1985) (where M is the number of the microphones in the array) is not easy to obtain in car noise environments due to the small microphone array, since car noise is not entirely uncorrelated and traditional beamforming technique with small standard arrays do not provide substantial improvement in signal to noise ratio as compared to single omni-directional microphones (Galanenko et al., 2001). In the study by Nordholm et al. (1999), they formulate a simple built-in calibration procedure for data collection instrumentation in the car environment. Their working scheme is to find the transfer function among the speaker, jammer signal, and microphone array in a quiet setting, and assume this function does not change when the car is moving on the road. This algorithm is one of several typical beamforming algorithms that have been used in car environments. However, it should be noted that microphone array calibration does have a problem, since it is not easy to keep a human being steady during operation, and most of the movement of his/her head will change the source position, which will change the transfer function. In another study, Compernolle (1990) presented an approach using switching adaptive filters, with no *a priori* knowledge about the speech source. The filters have two sections, where the first section implements an adaptive look direction and cues in on the desired speech, while the second section acts as a multichannel adaptive noise canceler. This method is able to simultaneously track the movement of the speaker source and compensate for the transfer function between the microphone array and speaker in real-time. While this was an important contribution, it was evaluated only in a reverberant laboratory setting (Compernolle et al., 1990), and not in a noisy moving car environment. Another study by Oh et al. (1992) applied a Griffiths–Jim beamformer (Griffiths and Jim, 1982) in a car environment with a 7-channel microphone array. They evaluated Signal-to-Noise (SNR) and word error rate (WER) improvement of their algorithm, and compared this to the case when only a DASB was used. Their general recommendations were that the generalized side-lobe canceler (GSC) was relatively stable and robust. However, from our analysis using real car data we collected, we found that noise signals with high frequency energy, such as road bump noise, which routinely happens for road surface repairs of potholes or expansion joints across bridges, will make the GSC unstable. This phenomenon is also observed and mentioned by Korompis et al. (1995). In a study by Zhang and Hansen (2003a), a method to identify this kind of noise is proposed and thereby allows the adaptive filters to work more robustly. In the study by Shinde et al. (2002), they presented a multichannel method for noisy speech recognition which estimates the log spectrum of speech for a close-talking microphone based on a multiple regression of the log spectra (MRLS) of noisy signals captured by the distributed microphones. This method was reported to improve speech recognition performance by up to 20%. In a later study by Li et al. (2005), an improved version of this method has been implemented by automatically adapted the regression weights for different noise environments, and 58.5% word error rate (WER) was reported. However, the MRLS based method requires a specific microphone arrangement in the car. It should also be noted that the noise signals captured by distributed microphones within the car are not necessarily the real noise that reaches the close-talking microphone. Hoshuyama et al. (1999) considered an adaptive beamforming solution for microphone arrays with a blocking matrix using constrained adaptive filters. Abut (2002), Wahab et al. (1997) and Wahab et al. (1998) presented a speech enhancement framework using a DCT-based (discrete cosine transform) Generalized Amplitude Spectral Estimator (ASE), which can be used for a stereo microphone noise cancellation system in the car. Visser et al. (2002), presented a speech enhancement scheme, which combined a 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. Meyer and Simmer (1997) considered the diffuse noise field in cars, and presented a multichannel-algorithm for speech enhancement. It consists of a delay-and-sum beamformer (DASB), a spectral subtraction algorithm for low frequency and a Wiener filter for high frequency. These methods were reported to have good performance under a single controlled driving condition (i.e., windows closed traveling at a given speed). Wallace and Goubran (1992) proposed a sub-banded two-stage beamforming multi-reference adaptive noise canceler with sub-banded second stage for noise suppression in car noise environments. This method was shown to have a significant noise reduction during non-speech segments but the performance during speech segments is degraded. In another study by Haan et al. (2003), a method for the design of over-sampled uniform DFT-filter banks aiming at minimal source signal degradation at the microphone array output was proposed. Their method consists of two steps. In the first step the analysis filter bank was designed in such a way that the aliasing terms in each sub-band were minimized individually, contributing to minimal aliasing at the output without aliasing cancellation. In the second step the synthesis filter bank was designed to match the analysis filter bank where the analysis-synthesis response was optimized while all aliasing terms in the output signal were individually suppressed,

rather than aiming at aliasing cancellation. They evaluated their method in an automobile environment observing that the background noise is suppressed by about 15 dB and that the interference signal is suppressed by about 17 dB. However, the desired signal distortion is still audible. In a study by Goulding and Bird (1990), they investigated the properties of the noise field in an automobile and proposed a delay equalized near-field beamformer, which attempts to enhance the speech, rather than reduce the noise. In another study by Grenier (1992), a 8-channel non-uniform microphone array was configured for car environments, and both Frost adaptive beamformer (Frost et al., 1972) and Griffith–Jim adaptive beamformer (Griffiths and Jim, 1982) were implemented and evaluated using a limited recorded database. This is a valuable work, as the performance of two important beamformers were compared. Gazor and Grenier (1995) and Gazor and Grenier (1994) also investigated the optimal positions and numbers of sensors for a microphone array, however we can not summarize their important conclusions here as this topic is out of the scope of this paper.

While a number of studies have investigated various speech enhancement and processing schemes for in-vehicle speech systems, the vast majority of results are conducted under controlled simulated conditions inside a room or summing pre-recorded car noise with clean speech. Little research has been performed using actual voice data collected in the car with associated environmental noise conditions. Because of the variety of simulated in-vehicle evaluation scenarios, it is difficult to compare performance between studies, and to predict if simulated performance will hold for actual, in-vehicle conditions. In our previous work (Zhang and Hansen, 2003a,b), an analysis was performed on data recorded in various car noise environments from across the United States, and the performance of traditional Delay and Sum beamforming (DASB) has been benchmark using the collected multi-channel microphone array data. There, we also proposed a constrained switched adaptive beamforming algorithm (CSA-BF), which detects the head movement of the driver and adjusts the time delay between microphones automatically. That method was shown to decrease WER (word error rate) for speech recognition by up to 31% and improve speech quality by up to 5.5 dB on the average simultaneously using the CU-Move corpus (Cu-Move, 2004).

In this paper, a novel combined fixed/adaptive beamforming (CFA-BF) scheme is proposed which is designed specially for robust speech recognition in car noise environments.[1] The proposed method is based on our previous work and analysis results for the potential driver movements during voice interaction by selecting 10 speakers from the CU-Move corpus (Hansen et al., 2001; Cu-Move, 2004). Our proposed method combines fixed and adaptive beamformers and also applies source localization techniques. Therefore, it has several novel advantages:

(i) low computational complexity with robustness;
(ii) target signal distortion reduction by omitting the parameter adjustment in adaptive filters;
(iii) automatically tracking driver movement, and no speech range definition is needed;
(iv) directional sources can be suppressed;
(v) especially suitable for use in car noise environments.

For the formulation of a microphone array front-end system for in-vehicle automatic speech recognition and navigation purpose, the low complexity and robustness of the multi-channel noise suppression algorithms are critical because of the limitation of resources. Several microphone array processing systems have been proposed and report good performance by some companies. For example, Andrea Electronics' microphone array, which was purchased by BMW of North America for use with hands-free digital phone application in BMW Z8 sports car in 2000[2], and selected by Delphi automotive systems for inclusion in demonstration vehicle at the international motor Show Passenger Cars in 2001[3]. A Singapore based company BITwave presented their array solutions for Motorcycle Helmet and handsfree array car kit for effective suppressing environmental noise and provides crystal clear speech for automobile speech communication[4].

For traditional array processing methods, the procedure used to separate the desired speech and sources of interference is also critical. Several source localization methods have been proposed in the literature (Brandstein and Ward, 2001) and report good performance using experimental data. Source localization methods are more effective when using larger microphone arrays in situations such as conference rooms or large auditoriums. Their ability to perform well in changing noisy conditions such as the car has not been documented to the same degree, but it is clearly expected to be poor compared to applications such as conference rooms. In our previous study (Zhang and Hansen, 2003a,b), we proposed three practical constraints which can be used in separating the desired speech and sources of interference with high accuracy, and will continue to use them for the present study.

Since in-vehicle speech systems could focus on hands-free wireless cell phone communications, as well as automatic speech recognition, our performance criteria for algorithm evaluation are segmental speech-to-noise ratio (SEGSNR) and word error rate (WER) using a speech recognition platform. This paper is organized as follows. In Section 2, we

[1] An earlier version of the CFA-BF: combined fixed/adaptive beamforming algorithm was presented at Interspeech-2003 (Zhang and Hansen, 2003b), which received Best Paper Award for Interspeech-2003.

[2] http://www.andreaelectronics.com/PressReleases/2000/2000_12_11.htm.

[3] http://www.andreaelectronics.com/PressReleases/2001/2001_09_17.htm.

[4] http://www.bitwave.com.sg.

introduce the CU-Move in-vehicle speech database collected for development of in-vehicle route navigation. Next, we briefly introduce our previous work the constrained switched adaptive beamforming (CSA-BF) algorithm and present the combined fixed/adaptive beamforming (CFA-BF) in Section 4. An extensive series of evaluations are then performed and presented in Section 5. Finally, we draw conclusions and discuss future work in Section 6.

2. CU-Move: in-vehicle speech corpus for interactive speech systems

The goal of the University of Colorado CU-Move project (Cu-Move, 2004) is to develop algorithms and technology for robust access to information via spoken dialog systems in mobile, hands-free environments. This requires reliable speech recognition access across changing acoustic conditions. However, the various noises in the car environment degrade speech signals and speech recognition system performance. In order to solve this problem, we formulate a new microphone array and multi-channel noise suppression front-end to provide high quality speech for in-vehicle speech systems. The microphone array that we constructed for the CU-Move project is a linear five-channel array, with microphone spacing between consecutive microphones of 4.25 cm to avoid spatial aliasing, given a frequency bandwidth of 4 kHz. Since the resolution on the beamforming delays is one sample period, which determines the resolution of the angle of the main beam, we employ a 44.1 kHz sample rate to obtain the highest resolution possible for our system.

The CU-Move corpus includes five parts:

1. NAVIGATION Direction Phrases section: a collection of phrases which are determined to be useful for In-vehicle navigation interaction [prompts fixed for all speakers];
2. DIGITS prompts section: strings of digits for the speaker to say [prompts randomized];
3. STREETS/Address/Route locations section: street names or locations within the city; some street names will be spelled, some just spoken. [prompts randomized];
4. SENTENCES – General Phonetically Balanced Sentences section: collection of phonetically balanced sentences for the speaker to produce [prompts randomized];
5. DIALOG Wizard - of - Oz Collection: Wizard of Oz interactive navigation conversation.

The driver performs a fixed route similar in structure to what was done for Phase I data collection (CU-Move, 2004; Hansen et al., 2001) that includes a combination of driving conditions (city, highway, traffic noise, etc.) for each speaker. A total of 500 speakers, balanced across gender and age, produced over 600 GB of data during a six month collection effort across the United States. The database and noise conditions are discussed in detail in (Hansen et al., 2001, 2000; Yapanel et al., 2002). We note that the noise conditions are changing with time and are quite different in terms of SNR, stationarity, and spectral structure. In this study, we chose 10 speakers from approximately 100 speakers in Minn., MN (i.e., Release 1.1A) and use the digits portion that includes speech under a range of varying complex car noise environments and contains approximately 40 words (i.e., Release 1.1a).

3. Prior beamforming algorithms

3.1. Delay and sum beamforming principle

Traditional multi-microphone array processing has focused on beamforming where a high quality speech signal is acquired by forming a directive pattern sensitive to the propagating direction. This section briefly considers the principle of delay-and-sum beamforming.

One of the simplest beamforming solutions is the weighted delay-and-sum beamformer (DASB) (Brandstein and Ward, 2001; Pillai, 1989; Johnson et al., 1993). The beamformer output $y(n)$ is formed by averaging weighted and delayed versions of the microphone signals as follows:

$$y(n) = \frac{1}{N}\sum_{i=1}^{N} w_i x_i(n-\tau_i). \quad (1)$$

Here, $x_i(n)$ represents a noisy speech sample from microphone "i" at time location $t = nT_s$, where T_s is the sample period. The weight and relative delay for the ith microphone are given as w_i and τ_i. In particular, the delays τ_i are selected so as to center the beamformer's passband along some particular angle of orientation. Essentially, we can formulate a (θ, r) space that reflects possible angles and distances for the speaker within the car environment as shown in Fig. 1. Plane sound waves approaching from the perpendicular direction will be added together in phase while those approaching from other directions will be added with different phases and will tend to be attenuated.

Fig. 2 shows the delay-and-sum beamformer consisting of a summed set of outputs from a number of microphones (i.e., $N = 5$) which have been delayed to "steer" the beam. Here, the spacing between each microphone is the same distance "d", therefore the delay terms are $\tau_i = (i-1) * \tau$.

In general, we assume the speech to be a plane wave arriving from direction θ to the axis of the microphone array which is composed of N microphones set up linearly and each separated by a distance of d. Thus, the delay is found as $\tau = f_{sam} * \left(\frac{d\sin\theta}{c}\right)$, where c is the sound propagation speed and f_{sam} is the sample frequency. Clearly, by selecting the appropriate delays between each microphone, we obtain signals that are summed in phase for direction angle θ, with destructive interference for signals arriving from other angles.

3.2. Constrained switched adaptive beamforming

Fig. 3 illustrates a block diagram of the CSA-BF algorithm. The CSA-BF consists of a speech/noise constraint

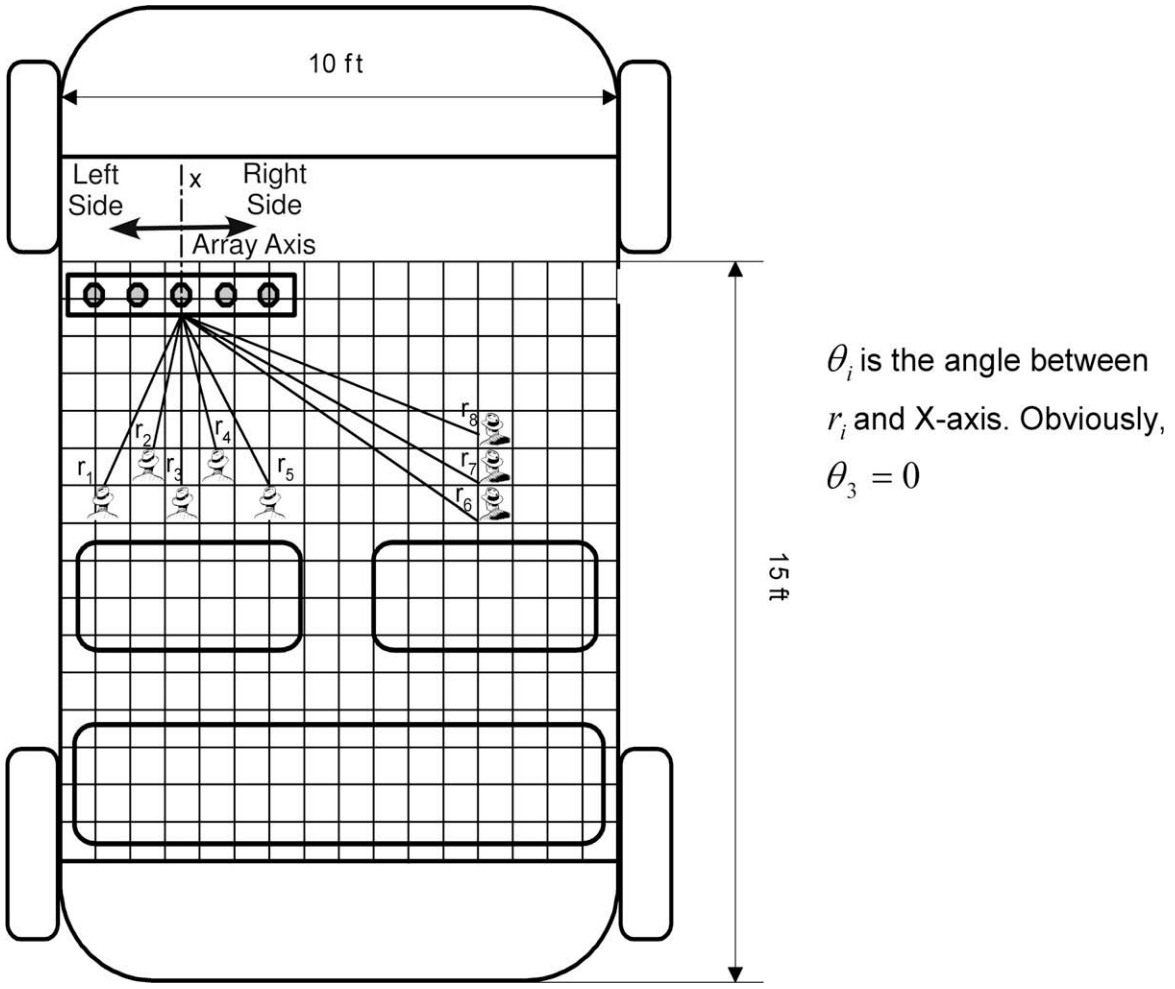


Fig. 1. Possible angles versus distances for speaker within the car environment.

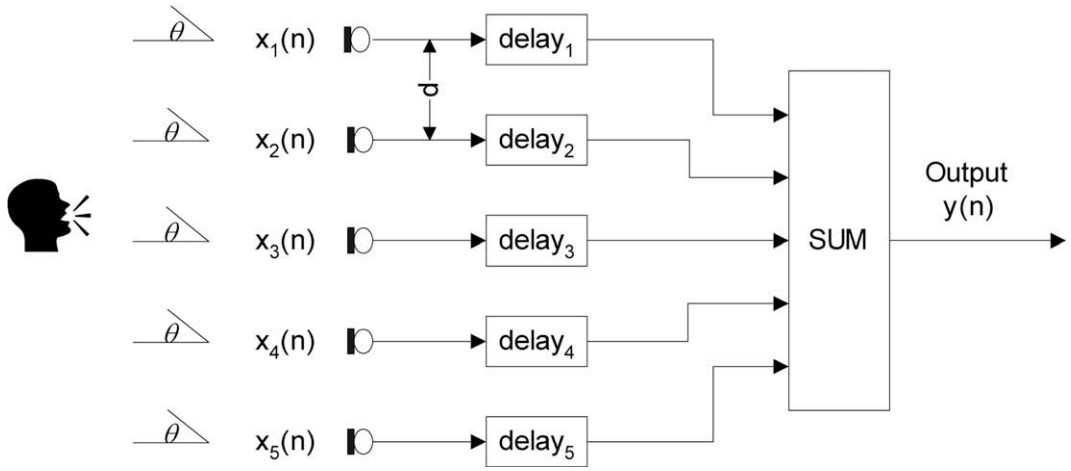


Fig. 2. Block diagram of a 5-element delay-and-sum beamforming.

section (CS), a speech adaptive beamformer (SA-BF), and a noise adaptive beamformer (NA-BF). The CS is designed to identify potential speech and noise locations. If a speech source is detected, the switch will activate SA-BF to adjust the beam pattern and enhance the desired speech. At the same time, the NA-BF is disabled to avoid speech leakage. If however, a noise source is detected, the switch will activate NA-BF to adjust the beam pattern for noise and switch off SA-BF processing to avoid the speech beam pattern from being altered by the noise. A set of adaptive filters are used to perform the beam steering. Also, a normalized LMS algorithm is used to update the filter

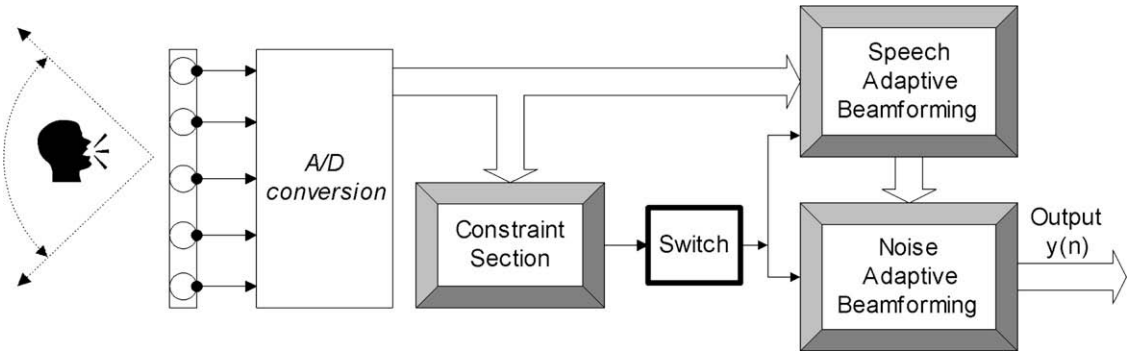


Fig. 3. Block diagram of the constrained switched adaptive beamforming (CSA-BF).

coefficients. The combination of SA-BF and NA-BF processing results in a framework that achieves noise cancelation for interference in both time and spatial orientation.

3.3. Motivations for DASB and CSA-BF

The advantage for selecting DASB for in-vehicle speech systems is that it is simple and robust, especially when the goal is to formulate effective speech input capture for real-time implementation. An alternative to employing a fixed delay based DASB is to dynamically adapt the delay values to steer the beam in the appropriate direction. Compared with adaptive beamforming which requires a number of adaptive filters, the computation complexity of DASB is quite low. If the direction of the desired source is known, then parameter adjustment is not needed. These advantages suggest an attractive solution (CU-Move, 2004; Hansen et al., 2001, 2000), since in car environments the driver's head position is restricted based on the seat position. With the microphone array positioned appropriately in the car, it is possible to maintain the speaker position (i.e., the head of the driver) at a consistent direction under most situations.

The most novel advantage of CSA-BF method is that source movement can be tracked and directional sources can be suppressed with reduced target signal distortion. However, the adaptive filters used in CSA-BF increase the computation complexity greatly, which limits the implementation of CSA-BF algorithm in real-time. Another disadvantage of CSA-BF is the sensitivity of parameter setting for the adaptive filters. From the experiment results in (Nordholm et al., 1999), we know that if the optimal parameter settings for CSA-BF are altered slightly, the WER will degrade slightly because of speech leakage.

4. Source location analysis in real car environments

4.1. Source location techniques

In this section, the techniques employed for locating speaker position are briefly introduced. In the implementation used here, the Teager Energy Operator (TEO) technique is first applied to decide the speech period for the selected speakers and find speech candidates based on maximum averaged energy. Since speech arriving from the driver's direction will have on average the highest intensity of all sources present in the vehicle, the proposed method here assumes the peak to be the arrival direction for the speaker. In order to more accurately track the energy peak in the arrival direction, the average signal TEO (Kaiser, 1993) energy is calculated on a frame by frame. If this energy value is greater than a given threshold, the current signal frame is marked as a speech candidate. Next, the adaptive LMS filter technique is applied on the speech candidate according to the geometric structure of the microphone array to locate the source (i.e., the position of the head of each speaker).

Next, the two techniques for source location will be discussed in detail.

4.1.1. Technique 1 (Teager Energy Operator (TEO))

It is known that when the microphone array is used in the car, it is generally positioned on the windshield near the sun visor in front of the driver who is assumed to be the speaker. Therefore, the driver to microphone array distance will be shorter than for other passengers in the vehicle. Therefore, speech from the driver's direction will have on average the highest intensity of all sources present in the car. Thus, the first proposed technique is based on energy averages as follows:

1. if $E_{signal} > E_{speech}$, then the current signal will be a speech candidate;
2. if $E_{signal} < E_{noise}$, then the current signal will be a noise candidate.

Here, E_{signal} denotes the present signal energy, E_{speech} denotes speech energy threshold, and E_{noise} denotes the noise energy threshold. To measure the speech energy, the nonlinear energy operator developed by Teager, and described by Kaiser (1993) and Zhou et al. (2001) are employed as follows:

$$\psi[x(n)] = x^2(n) - x(n+1)x(n-1). \quad (2)$$

Here, $\psi[\cdot]$ is referred to as the TEO, and $x(n)$ is the sampled speech signal. The TEO was used as an energy measurement method over a traditional window based energy scheme since the TEO method is capable of estimating

the instantaneous energy over a small sample window, while traditional energy measurement is generally obtained from an average mean square energy estimate which represents a smoothed response. In order to overcome instances of impulsive high-energy interference in the proposed implementation, an analysis window consisting of 256 samples is used instead of the 3 sample window needed to compute the average Teager energy. Assume the analysis window size is N, then the average Teager energy of this window is given as follows:

$$\overline{E}_{signal} = \frac{1}{N}\sum_{i=1}^{N}[x^2(n) - x(n+1)x(n-1)]. \quad (3)$$

With this representation, the following terms are used for developing frame selection criteria. First, $\overline{E}_{signal}$ is defined as the present energy estimate; $\overline{E}_{speech}$ is the present speech energy threshold; $\overline{E}_{noise}$ is the present noise energy thredhold. Also, the terms $\overline{E}^{new}_{speech}$ and $\overline{E}^{old}_{speech}$) represent the previous and updated speech energy thresholds, and $\overline{E}^{new}_{noise}$) and $\overline{E}^{old}_{noise}$) represent the previous and updated noise energy thresholds. Therefore, the first criterion becomes,

(1) if $\overline{E}_{signal} > \overline{E}_{speech}$, then the current signal analysis window will be a speech candidate;
(2) if $\overline{E}_{signal} < \overline{E}_{noise}$, then the current signal analysis window will be a noise candidate.

In order to track the changing environmental noise and speech conditions within the vehicle, the algorithm also updates the speech and noise thresholds according to the following rules:

(1) when the current analysis window is a speech candidate:

$$\overline{E}^{new}_{speech} = \alpha \times (\overline{E}^{old}_{speech}) + (1-\alpha) \times \overline{E}_{signal}, \quad (4)$$

$$\overline{E}_{speech} = \rho_{speech} \times \overline{E}^{new}_{speech}. \quad (5)$$

(2) when the current analysis window signal is a noise candidate:

$$\overline{E}^{new}_{noise} = \beta \times (\overline{E}^{old}_{noise}) + (1-\beta) \times \overline{E}_{signal}, \quad (6)$$

$$\overline{E}_{noise} = \rho_{noise} \times \overline{E}^{new}_{noise}, \quad (7)$$

where $0 < \alpha, \beta < 1$, ρ_{speech} and ρ_{noise} are constants which control the levels of speech and noise threshold respectively. Fig. 4 shows the averaged Teager energy and the corresponding thresholds for a portion of noisy speech from a speaker in the CU-Move database. It was previously shown in Zhang and Hansen (2003a,b) that for most cases, this technique is able to maintain high accuracy in separating speech and noise. In the scenario for speech in Fig. 4, the driver spoke during fixed periods, and background noise was present throughout most of the recording. In the next section, a method is introduced on how to find the optimal weights associated with a source position using an LMS adaptive filter technique.

4.1.2. Technique 2 (adaptive LMS filter)

A number of source localization methods have been proposed in the past in array processing (Yamada et al., 2002; Omologo and Svaizer, 1994, 1996; Giuliani et al., 1996;

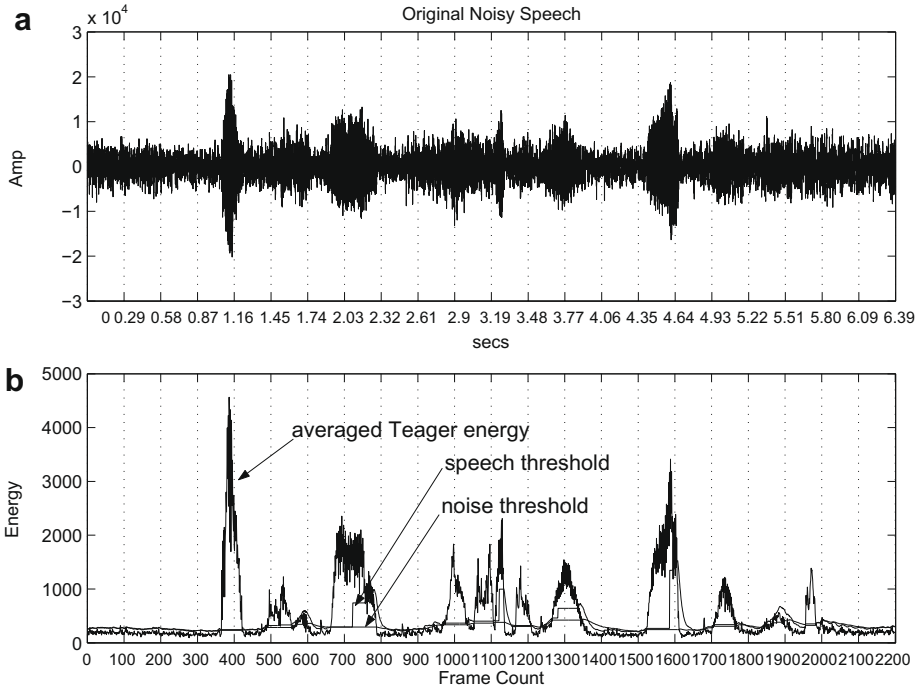


Fig. 4. Averaged TEO Energy profile and corresponding speech and noise thresholds: (a) noisy speech waveform from car environment; (b) TEO profile and resulting speech and noise thresholds.

Svaizer et al., 1997; Capon, 1969; Lang and McClellan, 1980; Knapp and Carter, 1976; Senadji and Grenier, 1993; Reed et al., 1981). Among these methods, the adaptive LMS filter (Reed et al., 1981) method is selected which is suggested to be the most suitable for a confined car environment (It is noted that alternative source localization methods could also be explored in future studies, but the adaptive LMS filter produced reliable performance for purposes needed here to track the speaker head position in the car environment). It is known that the peak of the weight coefficients in the LMS algorithm corresponds to the best delay between the reference signal $s(t)$ and desired signal $s_d(t)$. It is noted that in discrete time, $t = nT_s$ will be denoted with $s(n)$ and $s_d(n)$. Further details on traditional LMS adaptive filtering can be found in Reed et al. (1981).

The algorithm adapts the FIR filter to insert a delay equal and opposite to that existing between the two signals. In an ideal situation, the filter weight corresponding to the true delay would be unity and all other weights would be zero (Knapp and Carter, 1976; Reed et al., 1981; Widrow, 1985). In the present case, which is not an ideal situation, mic1 is selected in Fig. 5 (or Fig. 10) as the desired microphone, and mic2 in Fig. 5 (or mic5 in Fig. 10) as the reference microphone. Next, a delay is inserted that corresponds to the peak of the filter weight. According to the geometric structure of the microphone array and the arriving incident sound wave, the proposed method is able to locate the source from this delay. Fig. 5 shows this relationship. In order to simulate this, the desired signal is delayed by $L/2$, for which the corresponding delay will be a positive or negative number. A positive number means that the speech is coming from the right side of the array axis, and a negative value means that the speech is coming from the left side of the array axis.

In the experiments performed here, a sample rate of 44.1 kHz was used for each microphone signal, with an FIR filter length of 65, and therefore the center point of the filter coefficients will be located at the 33rd tap. If the angle between the speech direction and the axis of the two microphones is α, then the difference between the position of the maximum value and the center point in the coefficients is given roughly as $x = round(\eta * (\alpha * 33)/90)$. Here, η is comprised of several factors, including the distance between the speaker and the microphone array and the distance between the microphones. However, after constructing the microphone array system which is fixed, the impact of these factors does not change in the above relationship. Therefore, we can set aside η and employ the following equation:

$$\alpha = \pm 30^\circ \leftrightarrow x = \pm 11. \tag{8}$$

In the experiments in this study, x is taken as the head position number associated with the source direction.

4.2. Source location analysis

In order to analyze the movement of the driver's head in the car during voice interaction, a subset of 10 speakers was selected from the CU-Move database (managed by CRSS-UTD (CU-Move, 2004; Hansen et al., 2001)) representing a balance across gender and age. Next, the TEO energy operator technique was employed to determine the

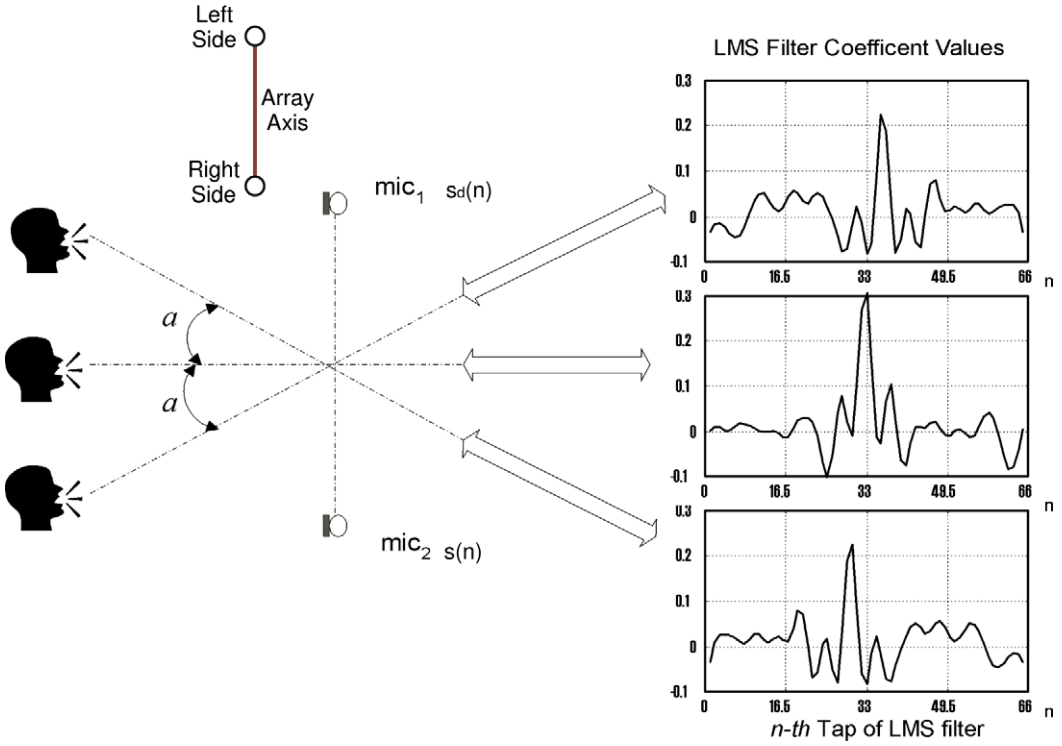


Fig. 5. Relationship between speaker position and weight of LMS filter (plots show filter response of 65 taps from 1 to 65).

speech periods for each of the 10 speakers, followed by the application of the adaptive LMS filter technique to locate the position of the head of each speaker(source). Table 1 summarizes the entire recording time for each speaker and the distribution percentage (%) of time each speaker's head stays in a certain rotational position.

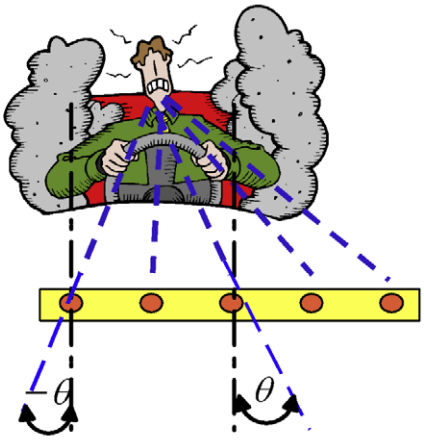
Fig. 6 shows the position number in Table 1 corresponding to the source rotational angle to the axis of the microphone array during the recording. From this table, it is seen that each driver will change their head position often during their up to 9 minutes of voice recording. The reason there some driver cases with a percentage of unknown positions is that the source location technique employed cannot at times make a reliable decision as to the current source location. This may happen when the noise level is very high, the noise changes too fast, and/or the step-size of the filter is too large or too small. This is actually a common situation for in-vehicle systems because of the complex noise situations as well as the limitations of the adaptive LMS filter technique. This also is a motivation for the proposed CFA-BF algorithm.

Fig. 7 shows detailed head movements for some of the speakers. From the figure, it can be seen that speaker #1 spends 39% of his time in head position #0 and 36% in position #1. Speaker #4 spends 66% of his time in head position #7 and 34% time in other head positions. Because of limitations in the source tracking algorithms, there are some unknown positions which imply that it is not possible to determine the actual position of the driver's head.

Table 1 shows the entire recording time for each speaker and the percentage (%) of time each speaker's head stays in a certain position. Fig. 6 shows the position number in Table 1 corresponding to the source angle to the axis of the microphone array during the recording. The table shows that the driver rearranges their head position often during even a short 9 minute period of voice recording.

Table 1
Percentage of time for each speaker source location over CU-Move in-vehicle recording (i.e., Speaker 1 spends 39% of his total 5.6 minutes of speech in digits portion with head position 0 from Fig. 4).

Position Number	Speaker number									
	1	2	3	4	5	6	7	8	9	10
	Amount of recording time (in minutes)									
	5.6	8.2	7.4	8.1	8.2	7.4	6.5	6.1	6.6	6.4
0	**39**%	**57**	1			8	14		2	
1	**36**					**50**	**82**		**80**	
2	4	9			5	17	2		8	**55**
3		0.1			14	18		**67**	3	38
4					0.5	4			0.4	
5		1.6			0.2	0.4			0.5	
6		0.3		1	0.3				0.1	
7				**66**		0.2				
8				10		0.4				
−1	2		**91**					1		
−2			8							
−8				1						
unknown	19	32	0	22	**80**	3	2	32	6	7



position number	corresponding θ
0	0
1	2.8°
2	5.6°
3	8.4°
4	11.2°
5	14°
6	16.8°
7	19.6°
8	22.4°
-1	-2.8°
-2	-5.6°
unknown	unknown

Fig. 6. Relationship between position number and angle of source.

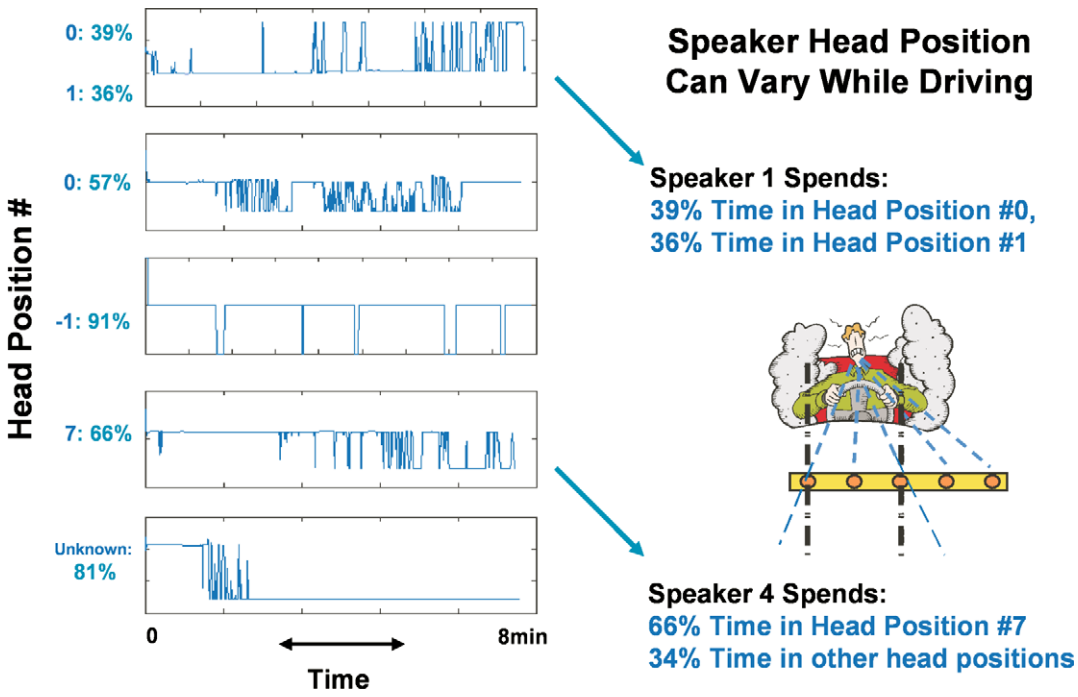


Fig. 7. Time Analysis for speaker head movements during recording. Head positions are estimated from angle of arrival from speech (see table summary in Fig. 6). Individual plots are for Speakers 1–5 from Table 1.

Fortunately, for each speaker it is always possible to find a dominant position.

5. CFA-BF: combined fixed/adaptive beamforming

In this proposed method, it is assumed that if the source position (driver's head) does not change, then the transfer function between the speaker and microphone array in a quiet setting will not change if the car is moving on the road. Therefore, it is proposed to find the transfer function between the speaker and microphone array for different possible source positions when the car is in a quiet environment (for example, a parking plot), and pre-store these for later use when the car is driven on the road. Fig. 8 describes the flowchart of the proposed CFA-BF algorithm.

5.1. Source location calibration – adaptive beamforming

As is well known, an adaptive algorithm such as normalized Least Mean Square algorithm (NLMS) can more easily reach its convergence behavior in quiet or stationary noise environments, than under non-stationary noise environments (for example, changing car noise environments). Also, from source location analysis of the CU-Move corpus in Section 3, it is known that although different drivers will move their heads in different positions, almost all will maintain one position more than 50% of the time while driving. Thus, it is possible to study candidate positions which the driver's head can reach inside a car, and then apply the previous developed CSA-BF (Zhang and Hansen, 2003a) to pre-record the weight coefficients of the adaptive filters for speech adaptive beamforming (SA-BF)

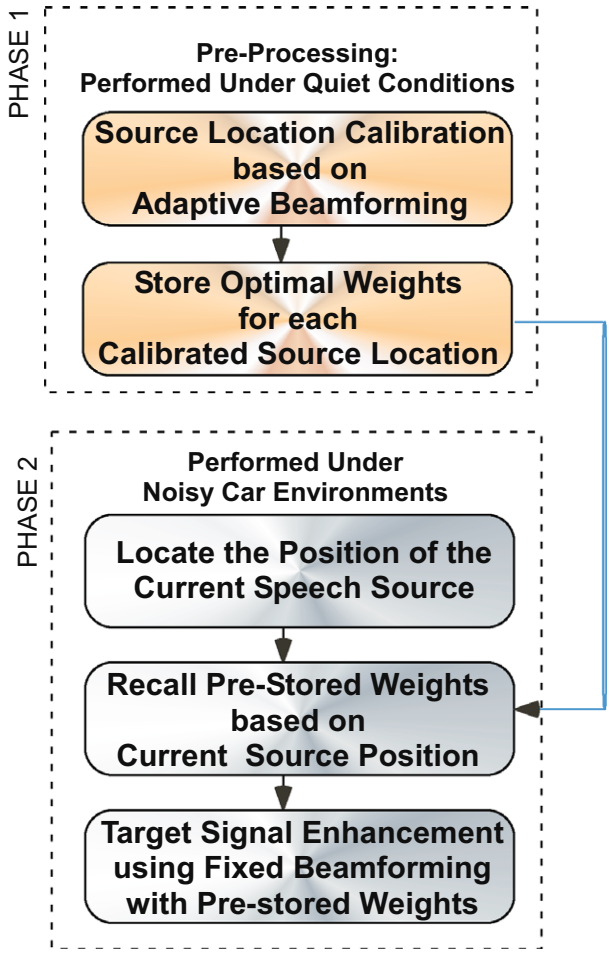


Fig. 8. Flow diagram of the proposed CFA-BF: combined fixed/adaptive beamforming algorithm (shows Phase 1 and Phase 2 processing).

from all the possible source positions in a quiet environment. Fig. 9 is the working scheme of the source calibration procedure. Here, only 3 positions are shown to illustrate the concept, where a peak indicates the estimated source direction. A normalized LMS algorithm is used to update the filter coefficients, and the update equations are given as follows:

$$e_{1i}(n) = \mathbf{x}_1(n - L/2) - \mathbf{w}_{1i}^T(n)\mathbf{x}_i(n). \tag{9}$$

$$\mathbf{w}_{1i}(n+1) = \mathbf{w}_{1i}(n) + \frac{2\mu}{\mathbf{x}_i^T(n)\mathbf{x}_i(n)} e_{1i}(n)\mathbf{x}_i(n). \tag{10}$$

where $\mathbf{x}_i(n) = (x_i(n), x_i(n-1), \ldots, x_i(n-N+1))$ are the current and the past N-1 microphone input signals, $e_{1i}(n)$ is the noise output signal of SA-BF, and w_{1i} the weights of the adaptive filter. It is also noted that channel 1 ($i = 1$) is used as the *primary* microphone, and channels $i = 2, \ldots, 5$ as multiple reference channels. The coefficients stored in the bank of weights implement the transfer functions between the microphone array and the speaker in different positions respectively. These weights also reflect the relative delays between microphones for the array. Fig. 10 shows how the SA-BF operates.

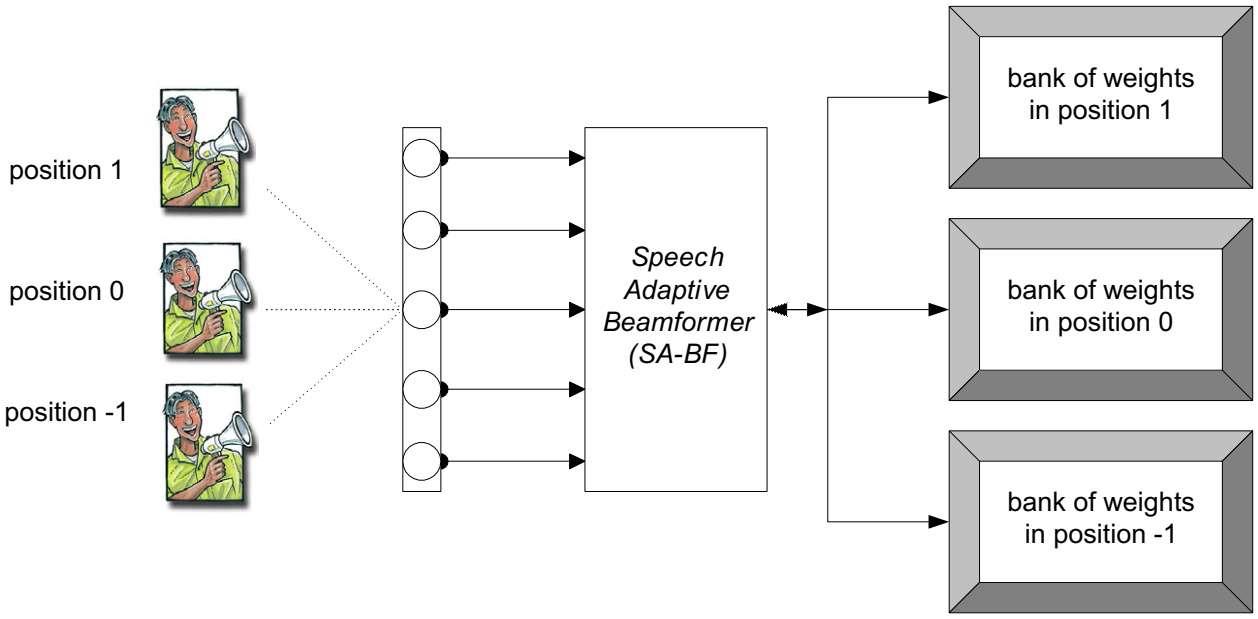


Fig. 9. Working Scheme for the proposed CFA-BF.

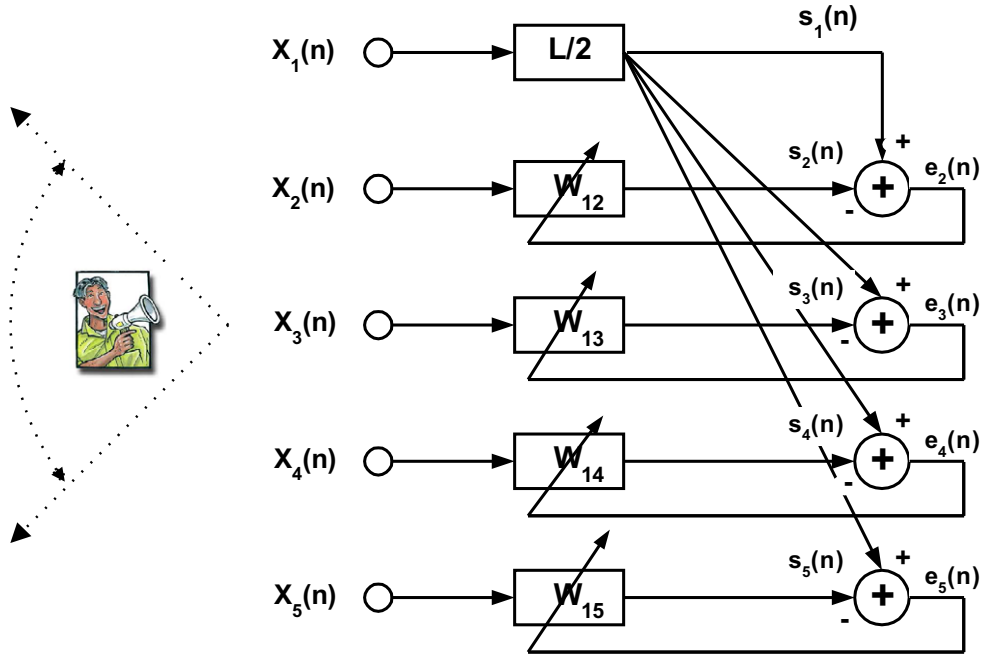


Fig. 10. Structure of speech adaptive beamformer (SA-BF).

5.2. Target signal enhancement – fixed beamforming

Fig. 11 shows the working scheme of the target speech enhancement. At this point, the method has the transfer functions from the speaker in different positions, (i.e., weight coefficients ($[W_{12}^{\theta}, W_{13}^{\theta}, W_{14}^{\theta}, W_{15}^{\theta}]$). With the help of a source localization technique, it is possible to find the source position first and then extract the corresponding weight coefficient bank from the pre-recorded weights and use them in this section. With this procedure, the enhanced speech will be given as follows:

$$s(n) = \sum_{i=1}^{5} \mathbf{w}_{1i}^{T}(n)\mathbf{x}_i(n). \tag{11}$$

where $[W_{12}^{\theta}, W_{13}^{\theta}, W_{14}^{\theta}, W_{15}^{\theta}]$ are functions of the angle between the source and axis of the microphone array θ, and W_{11} is a pure delay which is half of the filter length (i.e., $L/2$).

For the proposed CFA-BF algorithm, careful calibration of the weight coefficients and source location decision will have a significant impact on the performance of the algorithm. Imprecise inter-channel delay estimation generated by imperfect steering may result in serious signal distortion. Fig. 12 shows the effect of calibration procedure for speech enhancement by CFA-BF. Fig. 12a shows waveform of the beamforming output signal using calibrated weights, and Fig. 12b shows the waveform of beamforming output signal without calibration procedure. From this figure, it can be seen that without calibration, the enhanced signal suffers distortion due to imperfect steering. The purpose of this figure is only to illustrate the level of noise suppression in the waveform. In the next section, objective measures are employed to more accurately quantify the degree of noise suppression and speech enhancement.

6. Performance evaluation

6.1. Experiment establishment

In order to evaluate the performance of the CFA-BF under the non-ideal calibration and source location process, a series of experiments are performed as follows:

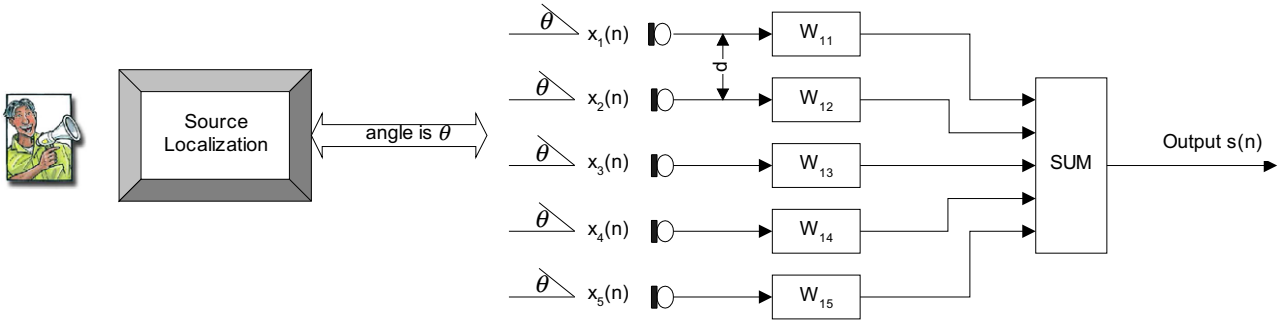


Fig. 11. Structure of fixed beamforming for target signal enhancement.

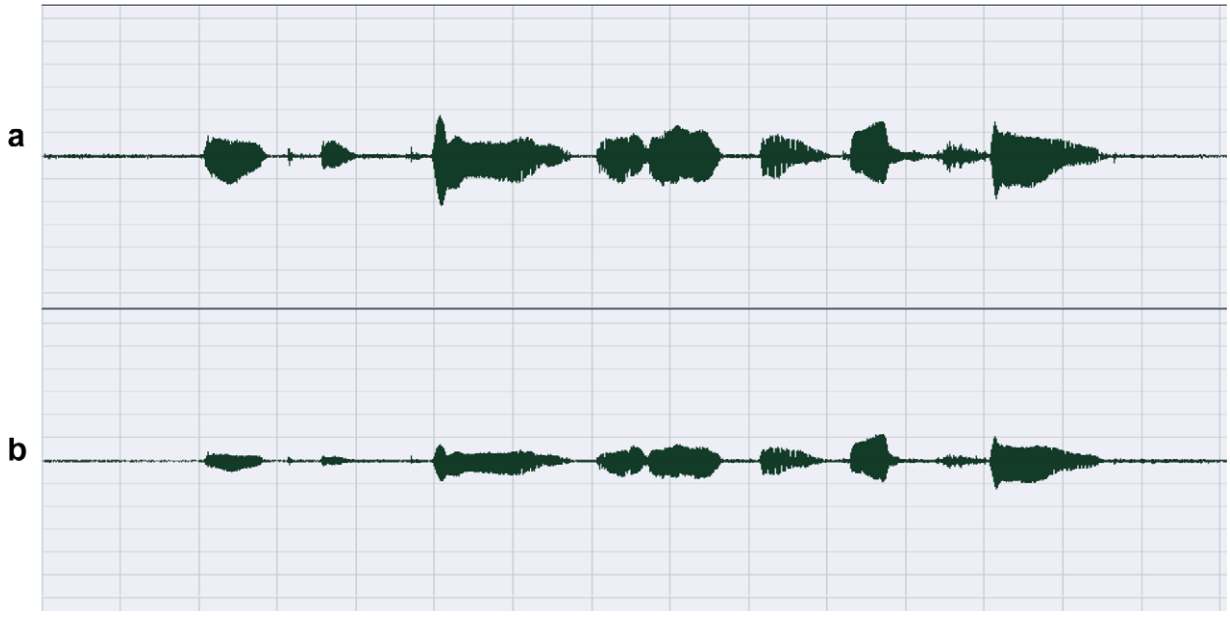


Fig. 12. Waveforms of the beamforming outputs with/without calibration procedure: (a) fixed beamforming output using calibrated weights; (b) adaptive beamforming output (i.e. without calibration procedure).

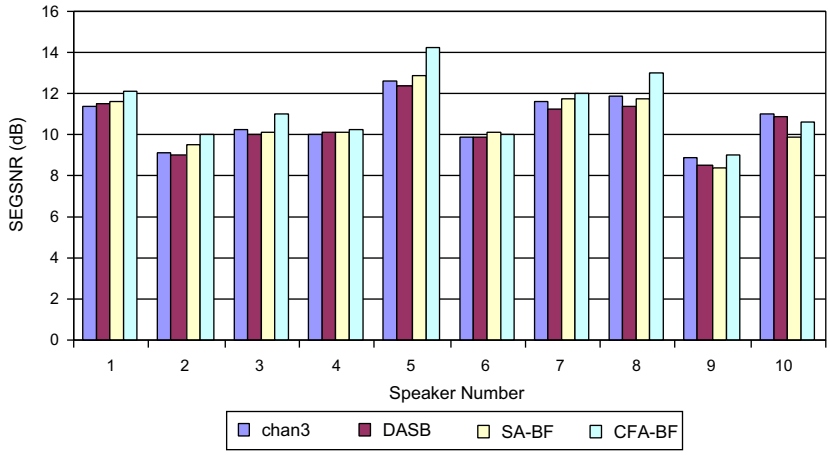


Fig. 13. SEGSNR performance for Ref. 3 microphone (single center microphone from the array) and 4 beamforming scenarios from Experiment #1.

(i) Use the CSA-BF to process each of the ten speakers respectively; the constraint used here is the TEO criterion described in (Kaiser, 1993) only;
(ii) Use the LMS adaptive filter described in (Reed et al., 1981) to identify the dominant source location (i.e., driver's head position);
(iii) Store the weight coefficient set of the speech beamformer (SA-BF) which has the dominant source position, and choose the best from this set as the calibrated weight set for SA-BF for this speaker.

Use the calibrated weight set to re-process the data for this speaker (i.e., delay-and-sum). If CFA-BF can produce better results than DASB and SA-BF under this non-ideal established experiment, it will operate much better in the ideal experimental(calibration) conditions. It has previously been shown in (Zhang and Hansen, 2003a) that with noise cancellation processing activated, both SEGSNR and WER results can be improved compared with SA-BF. In this present study, the cancellation processor is disabled, since if the speech quality (i.e., one of the outputs of SA-BF, which is used as the reference for noise cancellation processor) is improved, the output of GSC will also be improved.

6.2. Evaluation methods

For evaluation, two different performance measures are considered based on experiments with actual in-vehicle speech data from the CU-Move corpus. Recognize that since this data is not laboratory controlled speech data (i.e., artificially added noise to clean speech recordings), a variety of time varying noise conditions can and do exist. One measure of performance is the Segmental Signal-to-Noise Ratio (SEGSNR)[5] which represents a noise reduction criterion for voice communications. The second performance measure is word error rate (WER) reduction, which reflects benefits for speech recognition applications. The traditional HMM based speech recognition engine (Sonic Recognizer Pellom, 2001) is used to investigate speech recognition performance. For the processed data, the size of the set is not large enough for recognizer evaluation, therefore, a standard cross-validation method was adopted (Jelinek and Mercer, 1980; Rabiner and Juang, 1993).

6.3. Experiment results

In order to train the sets of weights, ten speakers were selected from the CU-Move corpus (Cu-Move, 2004) that are balanced across gender and age. Here, the corpus collection consisted of five phases, and for each speaker phase I (i.e., Navigation Direction Phrases section) were used to perform weight calibration. In the CU-Move corpus, phase I speech was collected in the same acoustic environments as the other four parts, but the noise level is much lower than that of others. In most speaker cases, this phase of the speech was collected with the car parked in a parking lot or driven at slow speeds with the windows closed. For the other CU-Move phases, such as part II (i.e., DIGITS prompts section), the speech was collected under a range of varying complex car noise environments (variable speeds, windows open at different positions with wind noise inside the car). In the testing set, the same ten speakers are chosen, but the phase II speech portion is employed to evaluate speech recognition (WER) and noise suppression/speech enhancement (SEGSNR) performance.

Fig. 13 shows the SEGSNR results for reference single channel3, DASB, SA-BF, and proposed CFA-BF. Table 2 shows average SEGSNR improvement, average WER (word error rate), CORR (Word Correct Rate), SUB (Word Substitution Rate), DEL (Word Deletion Rate) and INS (Word Insertion Rate) for the 10 speakers. Fig. 14 illustrates average SEGSNR improvement and WER speech recognition performance results respectively. The average SEGSNR results are indicated by the bars

[5] http://www.nist.gov.

Table 2
Average SEGSNR, WER, CORR, SUB, DEL and INS for Ref. 3 microphone and beamforming scenarios.

Method	Measure			
	chan3	DASB	SA-BF	CFA-BF
Ave. SEGSNR (dB)	10.77	10.48	10.70	**11.34**
WER	10.71	8.28	7.98	**7.51**
SUB	4.76	3.9	3.76	3.51
DEL	4.75	2.35	3.88	2.19
INS	3	3.11	3.16	2.96
CORR	92.28	94.83	95.19	95.46

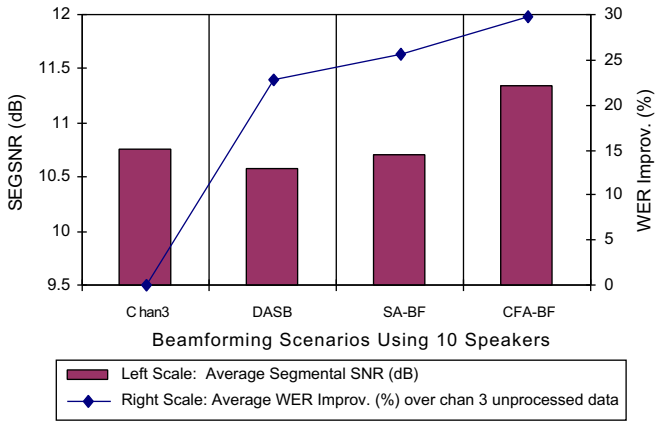


Fig. 14. SEGSNR and WER results for Ref. 3 microphone and beamforming scenarios in experiment #1 using 10 speakers.

using the left-side vertical scale (dB), and the WER improvement is the solid line using the right-side scale (%).

From these results, the following observations can be made:

(i) Employing the proposed combined fixed/adaptive beamforming method, increases SEGSNR slightly, but some variability exists across speakers;
(ii) However, DASB, SA-BF and the proposed method can provide WER improvement by 22.8%, 25.6% and 29.9% respectively over a single microphone (i.e., Channel 3 (Chan 3), the center microphone from the array).

6.4. Analysis

Thus far, the following three algorithms have been investigated to determine the performance of conventional delay-and-sum beamforming (DASB), the constrained switched adaptive beamforming (CSA-BF), and the new combined fixed/adaptive beamforming algorithm (CFA-BF) using a speech corpus collected in real car noise environments. The experimental evidence has demonstrated that a sufficiently high WER improvement can be achieved using the proposed CFA-BF front-end processor. Although CFA-BF processing can improve overall SEGSNR and WER, this does not guarantee it will show improvement across all possible noise conditions in the car (i.e., the portion of the in-vehicle speech data tested was from windows opened at different levels and different vehicle speeds). A separate in depth analysis was performed on specific noise conditions in the car environment, and the following observations summarize these findings:

(1) Bump noise (noise from potholes, concrete joints in the road surface, expansion joint across bridge, etc): This type of noise typically possesses high energy and high frequency content, with very short duration. Adaptive filtering does not perform well on this noise. There are two main reasons for this:
   - The impulsive energy is very high compared with other signals, which makes it difficult to choose a suitable step-size for the adaptive filters.
   - The duration of an impulsive noise signal is very short, and in most cases there is not enough time for the adaptive filter to adjust its coefficients to the optimal noise reduction setting. The proposed CFA-BF is robust to this situation as since the procedure allows for application of the pre-stored weights for the adaptive filters while driving, the bump noises will not affect the filter adaptation anymore.

(2) 20–45 and 65 mph windows closed (road surface, and engine noise in city and highway traffic):
In practice, the most suitable environment for adaptive array processing algorithms is one with stationary noise (e.g., such as an office environment (Zhang et al., 2000)). For an in-vehicle situation where the window is closed traveling 20–45 mph on a smooth road surface, the resulting noise inside the car is relatively stationary. Exterior noise such as other cars passing or vehicle vibration reflects only a small fraction of the noise seen from windows traveling close to city/highway speeds. Adaptive algorithm processing, such as CSA-BF, performs effectively with noticeable SEGSNR improvement. However, when the driving speed increases, the level of SEGSNR improvement will gradually decrease. This occurs because as the speed increases, the effect of vibration noise does not originate from a particular direction, and will vary depending on the exact road surface and weather conditions. Under this situation, the CFA-BF will take all the background noise and interference as noise, and simply place a null to suppress them as long as the source location does not change.

(3) 20–45 mph windows rolled down 2 in. and 65 mph windows rolled down 2 in.:
With windows open traveling on a road surface, the wind and road noise from outside the car dominates the acoustics. If this is city driving speed (20–45 mph), CFA-BF processing can provide a measurable level of improvement in SEGSNR. However, if the

driving speed increases to 65 mph for highway conditions, the SEGSNR improvement is not as significant. Under this situation, it is also more difficult to reliably decide the presence of speech versus noise across time.

7. Conclusions and future work

In this study, a novel combined fixed/adaptive beamforming method (CFA-BF) has been proposed for robust speech recognition in real car environments based on experiments using acoustic data recorded in moving car environments. The CFA-BF was shown to improve SEGSNR slightly, and improve speech recognition performance by decreasing WER by up to 29.3% using CU-Move in-vehicle speech data. It has also been shown that this method outperforms a single channel microphone (channel 3), traditional delay-and-sum beamforming (DASB) and our previous speech adaptive beamformer (SA-BF). However, there remain some issues to address for future work:

(i) Perform source localization calibration in a quiet environment, such as parking plot, and use a larger portion of the CU-Move Corpus to evaluate the performance of CFA-BF;
(ii) Improve the accuracy of source localization by applying alternative source localization techniques, such as CSP (cross-power spectrum technique), and decrease the percentage of unknown positions;
(iii) Activate the GSC noise canceler processor after signal enhancement, and improve the SEGSNR performance without affecting WER improvement.

Acknowledgements

This project was supported by Grants from DARPA through SPAWAR under Grant No. N66001-8906, and by the University of Texas at Dallas under Project EMMITT. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

Abut, H., 2002. IEEE DL Lecture, Japan and Hong Kong. http://akhisar.sdsu.edu/abut/DLLecture-CMAC.pdf.

Brandstein, M., Ward, D. (Eds.), 2001. Microphone Arrays. Springer.

Capon, J., 1969. High resolution frequency-wavenumber spectrum analysis. In: Proc. IEEE, pp. 1408–1418.

Compernolle, D.V., 1990. Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings. In: Proc. IEEE ICASSP'1990, Vol. 2, Alburquerque, NM, USA, pp. 833–836.

Compernolle, D.V., Ma, W., Xie, F., Van Diest, M., 1990. Speech recognition in noisy environments with the aid of microphone arrays. Speech Commun. 9 (5–6), 433–442.

CU-Move, 2004. Originally located at http://cumove.colorado.edu/. Now maintained at http://crss.utdallas.edu/.

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, NY (Chapter 8).

Frost III, O.L., 1972. An algorithm for linearly constrained adaptive array processing. Proc. IEEE 60, 926–935.

Galanenko, V., Kalyuzhny, A., 2001. Investigation of effectiveness of microphone arrays for in car use based on sound field simulation. In: Proc. IEEE ICASSP'2001, Vol. 5, Salt Lake City, Utah, USA, pp. 3017–3020.

Gazor, S., Grenier, Y., 1994. Optimal positioning of sensors for a microphone array. In: Proc. IEEE ICASSP'1994, Vol. 4, Adelaide, Australia, pp. 557–560.

Gazor, S., Grenier, Y., 1995. Criteria for positioning of sensors for a microphone array. IEEE Trans. Speech Audio Process. 3 (4), 94–303.

Giuliani, D., Omologo, M., Svaizer, P., 1996. Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation. In: Proc. IEEE ICSLP'1996, Vol. 3, Philadelphia, PA, USA, pp. 1329–1332.

Goulding, M.M., Bird, J.S., 1990. Speech enhancement for mobile telephony. IEEE Trans. Veh. Technol. 39 (4), 316–326.

Grenier, Y., 1992. A microphone array for car environments. In: Proc. IEEE ICASSP'1992, Vol. 1, San Francisco, CA, USA, pp. 305–308.

Griffiths, L.J., Jim, C.W., 1982. An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antenn. Propag. AP-30, 27–34.

Haan, J.M.D., Grbic, N., Claesson, I., Nordholm, S.E., 2003. Filter bank design for subband adaptive microphone arrays. IEEE Trans. Speech Audio Process. 11 (1), 14–23.

Hansen, J.H.L., 1994. Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect. IEEE Trans. Speech Audio Process. 2 (4), 598–614 (special issue: Robust Speech Recognition).

Hansen, J.H.L., Clements, M., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39 (4), 795–805.

Hansen, J.H.L., Nandkumar, S., 1995. Objective quality assessment and the RPE-LTP vocoder in different noise and language conditions. J. Acoust. Soc. Amer. 97 (1), 609–627.

Hansen, J.H.L., Plucienkowski, J., Gallant, S., Pellom, B.L., Ward, W., 2000. CU-Move: robust speech processing for in-vehicle speech systems. In: Proc. IEEE ICSLP'2000, Vol. 1, Beijing, China, pp. 524–527.

Hansen, J.H.L., Angkititrakul, P., Plucienkowski, J., Gallant, S., Yapanel, U., Pellom, B., Ward, W., Cole, R., 2001. CU-Move: analysis and corpus development for interactive in-vehicle speech systems. In: Interspeech-01/Eurospeech-01, Vol. 3, Aalborg, Denmark, pp. 2023–2026.

Haykin, S., Justice, J.H., Owsley, N.L., Yen, J.L., Kab, A.C., 1985. Array Signal Processing. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Hoshuyama, O., Sugiyama, A., Hirano, A., 1999. A robust adaptive beamforming for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Trans. Signal Process. 47 (10), 2677–2684.

Jelinek, F., Mercer, R.L., 1980. Interpolated estimation of Markov source parameters from sparse data. In: Gelsema, E.S., Kanal, L.N. (Eds.), Pattern Recognition in Practice. North-Holland Pub. Co., Amsterdam, pp. 381–397.

Jensen, J., Hansen, J.H.L., 2002. Speech enhancement using a constrained iterative sinusoidal model. IEEE Trans. Speech Audio Process. 9 (7), 731–740.

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

Kaiser, J.F., 1993. Some useful properties of Teager's energy operator. In: Proc. IEEE ICASSP-1993, Vol. 3, Minneapolis, MN, USA, pp. 149–152.

Knapp, C.H., Carter, G.C., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech Signal Process. ASSP-24, 320–327.

Komarow, S., 2000. Germany targets dialing and driving. USA Today, September 12.

Korompis, D., Wang, A., Yao, K., 1995. Comparison of microphone array designs for hearing aid. In: Proc. IEEE ICASSP'95, Vol. 4, Detroit, MI, USA, pp. 2739–2742.

Lang, S.W., McClellan, J.H., 1980. Frequency estimation with maximum entropy spectral estimators. IEEE Trans. Acoust., Speech Signal Process. ASSP-28, 716–724.

Li, W., Takeda, K., Itakura, F., 2005. Adaptive log-spectra regression for in-car speech recognition using multiple distributed microphones. IEEE Signal Process. Lett. 12 (4), 340–343.

Meyer, J., Simmer, K.U., 1997. Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction. In: Proc. IEEE ICASSP'1997, Vol. 2, Munich, Germany, pp. 1167–1170.

Nandkumar, S., Hansen, J.H.L., 1995. Dual-channel iterative speech enhancement with constraints based on an auditory spectrum. IEEE Trans. Speech Audio Process. 3 (1), 22–34.

Nordholm, S., Claesson, I., Dahl, M., 1999. Adaptive microphone array employing calibration signals: analytical evaluation. IEEE Trans. Speech Audio Process. 7 (3), 241–252.

Oh, S., Viswanathan, V., Panamichalis, P., 1992. Hands-free voice communication in an automobile with a microphone array. In: Proc. IEEE ICASSP'1992, Vol. 1, San Francisco, CA, USA, pp. 281–284.

Omologo, M., Svaizer, P., 1994. Acoustic event localization using a crosspower-spectrum phase based technique. In: Proc. IEEE ICASSP'1994, Vol. 2, Adelaide, Australia, pp. 860–863.

Omologo, M., Svaizer, P., 1996. Acoustic source localization in noisy and reverberant environment using CSP analysis, ICASSP'1996, Vol. 2, Atlanta, Georgia, USA, pp. 921–924.

Pellom, B.L., 2001. Sonic: the University of Colorado – Boulder continuous speech recognizer. University of Colorado, Technical Report #TR-CSLR-2001-01, Boulder, Colorado, USA.

Pellom, B., Hansen, J.H.L., 1998. An improved (Auto:I, LSP:T) constrained iterative speech enhancement algorithm for colored noise environments. IEEE Trans. Speech Audio Process. 6 (6), 573–579.

Pillai, S.U., 1989. Array Signal Processing. Springer-Verlag, New York.

Plucienkowski, J., Hansen, J.H.L., Angkititrakul, P., 2001. Combined front-end signal processing for in-vehicle speech systems. In: Proc. Interspeech-01/Eurospeech-01, Vol. 3, Aalborg, Denmark, pp. 1573–1576.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs, NJ, pp. 447–448.

Reed, F.A., Feintuch, P.L., Bershad, N.J., 1981. Time delay estimation using the LMS adaptive filter-static behavior. IEEE Trans. Acoust., Speech Signal Process. ASSP-29 (3), 561–571.

Senadji, B., Grenier, Y., 1993. Broadband source localization by regularization techniques. In: Proc. IEEE ICASSP'1993, Vol. 1, Minneapolis, MN, USA, pp. 321–324.

Shinde, T., Takeda, K., Itakura, F., 2002. Multiple regression of log-spectra for in-car speech recognition. In: Proc. Interspeech-02/ICSLP-02, Denver, CO, USA.

Svaizer, P., Matassoni, M., Omologo, M., 1997. Acoustic source localization in three-dimensional space using cross-power spectrum phase. In: Proc. IEEE ICASSP'1997, Vol. 1, Munich, Germany, pp. 231–234.

Visser, E., Otsuka, M., Lee, T.W., 2002. A spatio-temporal speech enhancement scheme for robust speech recognition. In: Proc. Interspeech-02/ICSLP-02, Denver, CO, USA.

Wahab, A., Chong, T.E., Abut, H., 1997. Speech enhancement in vehicular environment. In: Proc. 1st Internat. Conf. on Information Communications and Signal Processing, Singapore.

Wahab, A., Chong, T.E., Abut, H., 1998. Noise suppression based on the interior acoustic field of the vehicular chamber. In: Proc. 5th Internat. Conf. on Control, Automation, Robotics and Vision, ICARV'98, Singapore.

Wallace, R.B., Goubran, R.A., 1992. Noise cancellation using parallel adaptive filters. IEEE Trans. Circuits Systems – II: Analog Digital Signal Process. 39 (4), 239–243.

Widrow, B., 1985. Adaptive Signal Processing. Prentice-Hall Inc., Englewood Cliffs.

Yamada, T., Nakamura, S., Shikano, K., 2002. Distant-talking speech recognition based on 3-D viterbi search using a microphone array. IEEE Trans. Speech Audio Process. 10 (2), 48–56.

Yapanel, U., Zhang, X.X., Hansen, J.H.L., 2002. High performance digit recognition in real car environment. In: Proc. Interspeech-02/ICSLP-02, Vol. 2, Denver, CO, USA, pp. 793–796.

Zhang, X.X., Hansen, J.H.L., 2003a. CSA-BF: A constrained switched adaptive beamformer for speech, enhancement,recognition in real car environments. IEEE Trans. Acoust., Speech, Signal Process. 11 (6), 733–745.

Zhang, X.X., Hansen, J.H.L., 2003b. CFA-BF: a novel combined fixed/adaptive beamforming for robust speech recognition in real car environments. In: Proc. Interspeech-03/Eurospeech-03, Geneva, Switzerland.

Zhang, X.X., Ng, B.P., Lim, K.S., 2000. A practical noise suppression method using microphone array. In: Internat. Conf. on Signal Processing and Technology, ICSPAT'2000, Dallas, TX, USA.

Zhou, G., Hansen, J.H.L., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. 9 (2), 201–216.