CSA-BF: NOVEL CONSTRAINED SWITCHED ADAPTIVE BEAMFORMING FOR SPEECH ENHANCEMENT & RECOGNITION IN REAL CAR ENVIRONMENTS

Xianxian Zhang and John H. L. Hansen

Robust Speech Processing Group, Center for Spoken Language Research
University of Colorado at Boulder, USA
{zhang, jhlh}@cslr.colorado.edu http://cslr.colorado.edu

ABSTRACT

While a number of studies have investigated various speech enhancement and processing schemes for in-vehicle speech systems, little research has been performed using actual voice data collected in noisy car environments. In this paper, we propose a new constrained switched adaptive beamforming algorithm (CSA-BF) for speech enhancement and recognition in real moving car environments. The proposed algorithm consists of a speech/noise constraint section, a speech adaptive beamformer, and a noise adaptive beamformer. We investigate CSA-BF performance with a comparison to classic delay-and-sum beamforming (DASB) in realistic car environments using a large quantity of data recorded in various car noise environments from across the United States. After analyzing the experimental results and considering the range of complex noise situations in the car environment using the CU-Move corpus, we formulate the CSA-BF algorithm. This method is shown to decrease WER for speech recognition by up to 31% and improve speech quality via the SEGSNR by up to 5.5dB on the average simultaneously.

1. INTRODUCTION

The increased use of mobile telephones and voiced controlled features in cars has created a greater demand for hands-free, in-car installations. Many countries now restrict handheld cellular technology while operating a vehicle. As such there is a greater need to have reliable voice capture within automobile environments. 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) [1]. This method is simple and robust if we know the direction of the speech source. However, if the source location changes during operation, this method is less effective due to the mismatch in delay estimation between microphones. Another practical problem of DASB is that the theoretical maximum noise attenuation $10\log_{10}(M+1)$ is not easy to obtain in car noise environments with small microphone arrays. Nordholm, et. al [2] considered a built-in calibration procedure for data collection instrumentation in the car environment. Compernolle [3] presented an approach using switching adaptive filters, with no a priori knowledge about the speech source. While this was an important contribution, it was evaluated only in a reverberant laboratory setting, and not in a noisy moving car environment. Oh, et. al [4] applied a Griffiths-Jim beamformer in a car environment with a 7-channel microphone array. Their general recommendations were that the generalized side-lobe canceller (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 will make the GSC unstable. Shinde et. al. [5] presented a multichannel method for noisy speech recognition which estimates the log speech spectrum for a close-talking microphone based on a multiple regression of the log spectra (MRLS) of noisy signals captured by the distributed microphones. Visser et. al. [6], presented a speech enhancement scheme, which combined a spatial and temporal processing strategy to handle reverberation, highly interfering sources and background noise. While a number of studies have investigated various speech enhancement and processing schemes for in-vehicle speech systems, the vast majority are conducted under controlled simulated conditions inside a room or with pre-recorded car noise. Little research has been performed using actual voice data collected in the car with associated environmental noise conditions.

In this paper, we investigate and propose several potential speech processing solutions for in-vehicle speech systems. The performance of a classic DASB technique in a realistic car environment is first considered. After analyzing experimental results, we propose a constrained switched adaptive beamforming (CSA-BF) method.

2. CSA-BF: CONSTRAINED SWITCHED ADAPTIVE EBAMFORMING

The proposed CSA-BF algorithm consists of four parts: a constraint section (CS), a speech adaptive beamformer (SA-BF), a noise adaptive beamformer (NA-BF) and a switch. Fig. 1 shows the detailed structure of CSA-BF, where we assume a 5-microphone array. 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, 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. The combination of SA-BF and NA-BF processing results in a framework that achieves noise cancellation for interference in

This work was supported by DARPA through SPAWAR under Grant No. N66001-8906

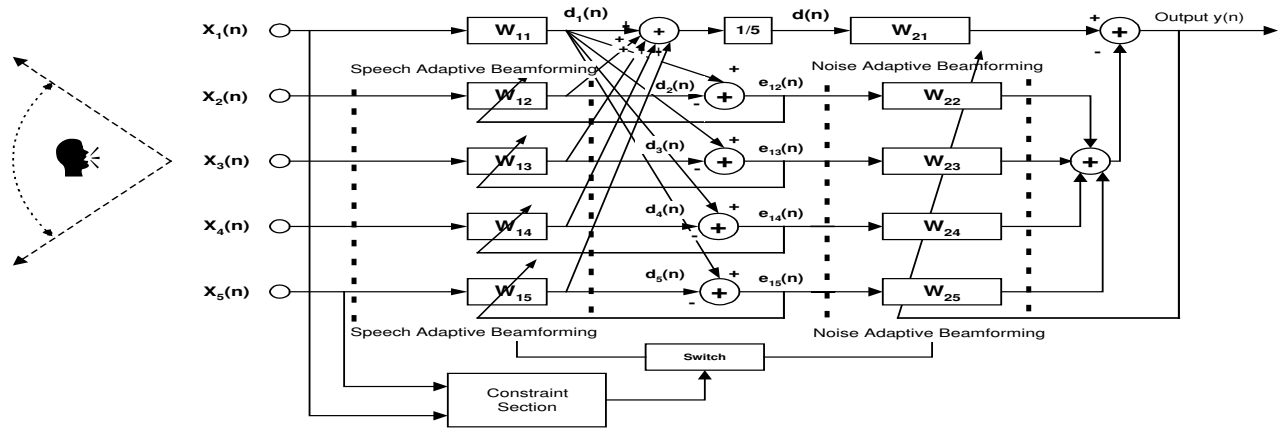

Figure 1: Structure of the Proposed Constrained Switched Adaptive Beamforming (CSA -BF)

both time and spatial orientation. Next, we consider each processing stage of the proposed CSA-BF scheme.

2.1. Constraint Section

Many source localization methods have been presented and report effective performance with large microphone arrays in conference rooms or large auditoriums. Their ability to perform well in changing noisy car conditions has not been documented to the same degree, but is expected to be poor. Here, we propose three practical constraints which can be used to separate speech and noise sources with high accuracy. The constraints introduced here are more effective than switching filters [3] in dealing with impulsive noise sources, as well as locating sources in the car.

It is known that the input microphone signal can be one or any combination of the following sources:

(i.) A desired speech signal (i.e., driver's voice);
(ii.) An unwanted speech signal (i.e., 2nd passenger);
(iii.) Various environmental car noises (vibration, turn signal, car passing, radio, air conditioner, etc).

Here, we view (ii) and (iii) as sources of interference.

2.1.1. *Criterion 1*

It is assumed that the microphone array is positioned on the windshield near the sun visor in front of the driver who is the assumed 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. To measure the speech energy, we employ the nonlinear Teager Energy Operator (TEO) [7]. Thus, our first criterion is based on the average TEO energy as follows:

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

Here, $\overline{E}_{signal}$ denotes the energy of the current signal analysis window, $\overline{E}_{speech}$ denotes the speech energy threshold, and $\overline{E}_{noise}$ denotes the noise energy threshold, where

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

In order to track the changing environmental noise and speech conditions, we also update the speech and noise thresholds according to the following rules:

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

$$\overline{E}_{speech}^{new} = \alpha \times (\overline{E}_{speech}^{old}) + (1-\alpha) \times \overline{E}_{signal} \quad (2)$$

$$\overline{E}_{speech} = \rho_{speech} \times \overline{E}_{speech}^{new} \quad (3)$$

(ii.) when the current analysis window is a noise candidate:

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

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

with $0 < \alpha, \beta < 1$ and ρ_{speech} and ρ_{noise} are the constants which control the level of speech and noise threshold respectively. Fig. 2 shows the average TEO energy and corresponding thresholds for a portion of noisy speech from a speaker in the CU-Move database.

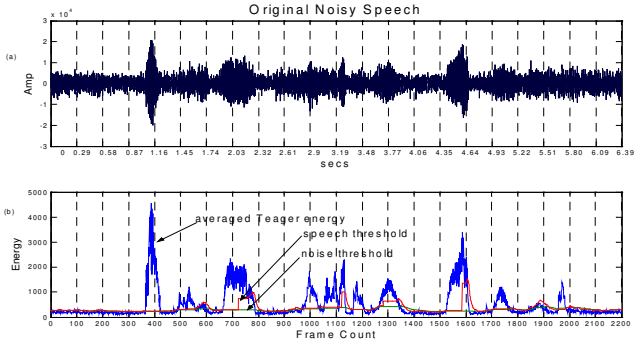

Figure 2: Averaged TEO. Energy versus corresponding thresholds
(a) Noisy speech waveform from car environment;
(b) TEO profile and resulting speech and noise thresholds.

2.1.2. *Criterion 2*

Independent of how the driver positions his head while speaking, the direction of his speech will be significantly different to that of another person sitting in the car. Therefore, in order to separate front-seat driver and passenger, we need a criterion to decide the direction of speech. We choose the adaptive LMS filter [8] method. In our case, we insert a delay that corresponds to the peak of the filter weight. According to the geometric structure of the microphone array and the arriving incident sound wave, we are able to locate the source from this delay. Fig. 3 shows this relationship. Obviously, if we take the axis between the center of the desired microphone (mic1) and reference microphone (mic5) as the standard axis, the desired source should be located within some symmetric area $\alpha \le \alpha_{thres}$ from both sides of this axis. α_{thres} can be fixed, or variable to obtain further noise suppression. In order to simulate this, we delayed the desired signal by $L/2$, for which the corresponding delay will be a positive or negative number as shown in Fig. 3.

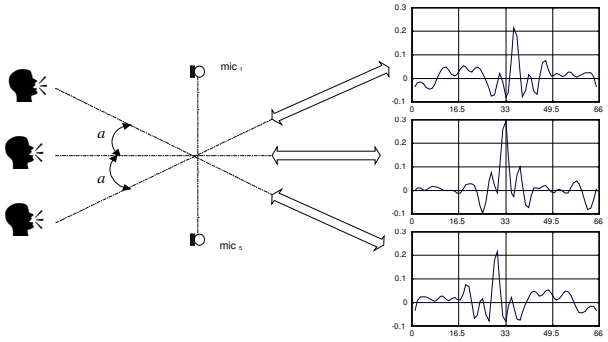

Figure 3: Relation. between speaker position and weight of LMS filter

2.1.3. Criterion 3

This final criterion is employed as a special case for road impulse/bump noise. Bump noise has a high-energy content, is typically impulsive in nature, and does not arrive from a particular direction. Fortunately, impulsive bump noise has obvious high-energy characteristics versus time, and thus the average TEO energy response will be higher than noisy speech and other noise types. Therefore, we can set a bump noise threshold E_{bump} during our implementation to avoid instability in the filtering process**.**

Finally, we note that the signal is labeled as speech if and only if all three criteria are satisfied.

2.2. Speech Adaptive Beamformer (SA-BF)

The function of SA-BF is to form an appropriate beam pattern to enhance the speech signal. Since adaptive filters are used to perform the beam steering, we can change beam pattern with a movement of the source. The degree of adaptation steering speed is decided by the convergence behavior of the adaptive filters. In our implementation, we select microphone 1 as the primary microphone, and build an adaptive filter between it and each of the other four microphones. These filters compensate for the different transfer functions between the speaker and the microphone array. A normalized LMS algorithm is used to update filter coefficients only when the current signal is detected as speech. There are two kinds of output from the SA-BF: namely the enhanced speech $d(n)$ and noise signal $e_{1i}(n)$, which are given as follows,

$$d(n) = \frac{1}{5}\sum_{i=1}^{5} \mathbf{w}_{1i}^{T}(n)\mathbf{x}_{1i}(n) \tag{6}$$

$$e_{1i}(n) = \mathbf{w}_{11}^{T}(n)\mathbf{x}_{1}(n) - \mathbf{w}_{1i}^{T}(n)\mathbf{x}_{i}(n) \tag{7}$$

$$\mathbf{w}_{1i}(n+1) = \mathbf{w}_{1i}(n) + \frac{2\mu}{\mathbf{x}_{i}^{T}(n)\mathbf{x}_{i}(n)} e_{1i}(n)\mathbf{x}_{i}(n) \tag{8}$$

for mic. channels $i = 2, 3, 4, 5$, where $\mathbf{w}_{11}(n)$ is a fixed filter.

2.3. Noise Adaptive Beamformer (NA-BF)

The NA-BF processor operates in a scheme like a multiple noise canceller, in which both the reference speech signal of the noise canceller and the speech free noise references are provided by the output of the SA-BF. Since the filter coefficients w_{2i} are updated only when the current signal is detected as noise, they form a beam that is directed towards the noise, thus the reason to name it a noise adaptive beamformer (NA-BF). The output response is given as,

$$y(n) = \mathbf{d}\ (n)\mathbf{w}_{21}^{T}(n) - \sum_{i=2}^{5} \mathbf{w}_{2i}^{T}(n)\mathbf{e}_{2i}(n) \tag{9}$$

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

for microphone channels $i = 2, 3, 4, 5$.

3. PERFORMANCE EVALUATION

3.1. CU-Move Corpus

The CU-Move [10] database include 5 parts: command and control words, digit strings of telephone and credit card numbers, street names and addresses, phonetically balanced sentences, and Wizard of Oz interactive navigation conversation. A total of 500 speakers, balanced across gender and age, produced over 600GB of data during a six-month collection effort across the United States. The database and noise conditions are discussed in detail in [9]. We point out that the noise conditions are changing with time and are quite different in terms of SNR, stationarity and spectral structure. In this study, we use the digits portion that includes speech under a range of varying complex car noise environments and contains approximately 40 words from approximately 100 speakers in Minn., MN (i.e., Release 1.1a).

3.2. Experiment Establishment

In the CSA-BF algorithm, there are a number of adaptive filters which are parameter dependent, such as, step-size of each adaptive filter, the speech/noise threshold, and the definition for the desired speech range. In addition to these parameters, the accuracy of the speech/noise decision in the CS is also important. Thus, in order to evaluate the performance of the DASB and CSA-BF algorithms in car noisy environments, we designed the following two experiments:

Exp #1: Establish algorithm setting using small speaker set;

Exp #2: Establish performance over large speaker group.

In Exp #1, we select ten speakers from the CU-Move database that are balanced across gender and age. Each speaker was processed using the DASB and CSA-BF algorithms. CSA-BF is evaluated for a range of parameter setting for these ten speakers, and the best parameter set was selected for use in open test Exp #2. In order to compare the result of CSA-BF with that of DASB thoroughly, we also investigated the enhanced speech output from SA-BF. In Exp #2, we process all available speakers in release 1.1a [9] of the CU-Move corpus using DASB and CSA-BF algorithms. This release consists of 153 speakers, of which 117 were from the Minneapolis, MN area. We selected 67 of these speakers that include 28 males and 39 females, which reflect 8 hours of data. In processing with CSA-BF, we used the parameter settings established for the ten speakers in Exp #1, except the speech range definition. We choose two different speech ranges for the speakers in Exp #2, since it is not practical to restrict all 67 speakers to speech from the same direction.

3.3. Evaluations

For evaluation, we consider two different performance measures using CU-Move data. One measure is the Segmental Signal-to-Noise Ratio (SEGSNR)[12] 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 Sonic Recognizer [13] is used to investigate speech recognition performance. For the processed data used in Exp #1, the size of the set is not large enough for recognizer evaluation, therefore, we adopted the cross-validation method [14]. For the processed data in Exp #2, we used 49 speakers (23 male, 26 female) as the training set, and 18 speakers (13 male, 5 female) as the test set.

3.4. Experiment Results

Fig. 4 shows SEGSNR results for Exp #1. Table 1 shows average SEGSNR improvement, average WER, CORR (word correct rate), SUB (Word Substitution Rate), DEL (Word Deletion Rate) and INS (Word Insertion Rate) for the speakers in Exp #1.

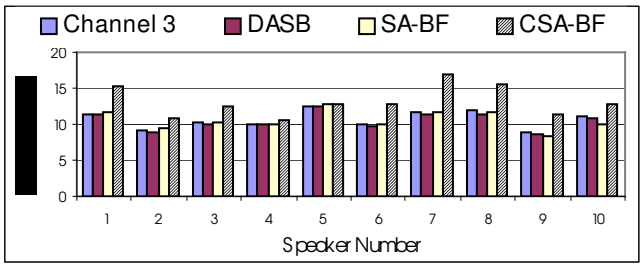

Figure 4: SEGSNR Performance for Ref. 3 Microphone and beamforming Scenarios in Exp #1

method / measure	chan3	DASB	SA-BF	CSA-BF
Ave. (dB) SEGSNR	10.67	10.48	10.60	13.16
WER	11.31	9.66	9.68	7.85
SUB	5.09	4.29	4.038	3.83
DEL	3.7	1.64	1.53	1.63
INS	2.51	3.76	4.13	2.41
CORR	91.22	94.09	94.58	94.58

Table 1: Average SEGSNR, WER, CORR, SUB, DEL and INS for Ref. 3 Microphone and beamforming Scenarios in Exp #1

Fig. 5& 6 illustrate average SEGSNR improvement and WER speech recognition performance results form Exp #1 and #2 respectively. The average SEGSNR results are indicated by the bars using the left-side vertical scale (dB), and the WER improvement is the solid line using the right-side scale (%).

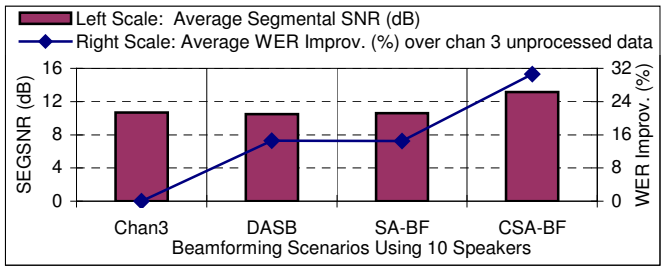

Figure 5: SEGSNR and WER Results for Ref. 3 Microphone and beamforming Scenarios in Exp #1 using 10 speakers

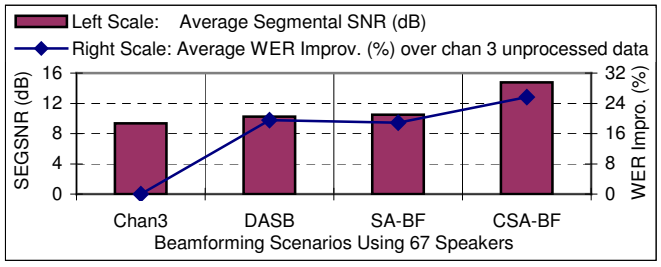

Figure 6: SEGSNR and WER Results for Ref. 3 Microphone and beamforming Scenarios in Exp #2 using 67 speakers

From these results, we draw the following points:
(i.) Employing delay-and-sum beamforming (DASB) or the proposed speech adaptive beamforming (SA-BF), increases SEGSNR slightly, but some variability exists across speakers. These two methods are able to improve WER for speech recognition by more than 14% and 19% for Exp #1 & 2.
(ii.) There is a measurable increase in SEGSNR and decrease in WER when noise cancellation processing is activated (CSA-BF). With CSA-BF, SEGSNR improvement is 2.5dB in Exp #1 and 5.5dB in Exp #2, and also provides WER improvement by 30.6% in Exp #1 and 26% in Exp #2.
(iii.) If the optimal parameter settings for CSA-BF are altered slightly, the SEGSNR improvement is not affected. However, the WER degrades slightly because of speech leakage.

4. CONCLUSIONS

In this paper, we have proposed a novel constrained switched adaptive beamforming (CSA-BF) for speech enhancement and recognition in real car environments based on experiments using a large quantity of voice data recorded in moving car environments. We demonstrated that the proposed CSA-BF processor can improve voice communications quality as reflected in a +5.5dB increase in SEGSNR, and speech recognition performace improvement by decreasing WER by 26-30.6% using CU-Move in-vehicle speech data. We have also shown that the CSA-BF solution outperforms a single channel microphone (channel 3) and traditional delay-and-sum beamforming. Finally, CSA-BF requires neither calibration signal nor a priori knowledge of speech or noise sources.

5. REFERENCE

[1] Brandstein and Ward, *Microphone Arrays*, Springer, 2001.
[2] Nordholm, Claesson and Dahl, "Adaptive Microphone Array Employing Calibration Signals: Analytical Evaluation", *IEEE Trans. on SAP,* May 1999.
[3] Compernolle,"Switching Adaptive Filters for Enhancing Noisy and Reverberant Speech from Microphone Array Recordings", *in ICASSP'*1990.
[4] Oh, Viswanathan and Panamichalis, "Hands-free Voice Communication in an Automobile with a Microphone Array", *in ICASSP'*1992.
[5] Shinde, Takeda and Itakura, "Multiple Regression of Log-spectra for In-Car Speech Recognition", *in ICSLP'2002.*.
[6] Visser et. al., "A Spatio-Temporal Speech Enhancement Scheme for Robust Speech Recognition", *in ICSLP'2002.*
[7] Kasier, "On a Simple Algorithm to Calculate the 'Energy' of a Signal", *ICASSP-90.*
[8] Reed, Feintuch, Bershad, "Time delay Estimation Using the LMS Adaptive Filter-Static Behavior", *IEEE Trans. On Acoustics, Speech and Signal Processing*, June 1981.
[9] Hansen, et.al., "CU-Move": Analysis & Corpus Develop. for Interactive In-vehicle Speech Systems", *Eurospeech'01.*
[10] http://cumove.colorado.edu/
[11] Deller, Hansen and Proakis, "*Discrete-Time Processing of Speech Signals*," *IEEE Press*, Chapter 8, 2000.
[12] http://www.nist.gov
[13] Pellom, "Sonic: The University of Colorado Continuous Speech Recognizer", *University of Colorado, Technical Report #TR-CSLR-2001-01*, Boulder, Colorado, March, 2001.
[14] Rabiner and Juang, "*Fundamentals of Speech Recognition*", Englewood Cliffs, NJ: Prentice-Hall, 1993.