Combined Front-End Signal Processing for In-Vehicle Speech Systems

Jay P. Plucienkowski, John H. L. Hansen, Pongtep Angkititrakul

CSLR: Center for Spoken Language Research; RSPL: Robust Speech Processing Lab
University of Colorado at Boulder, Boulder, Colorado 80309-0594, USA
{plucienk, jhlh, angkitit}@cslr.colorado.edu, http://cslr.colorado.edu

Abstract
In this paper, we investigate the integration of two processing methods to improve speech quality for in-vehicle speech systems: multi-sensor beamforming and constrained iterative (Auto-LSP) speech enhancement. The intent is to establish an intelligent microphone array processing scheme in high noise environments by considering the effectiveness of a multi-sensor beamformer method and the Auto-LSP single channel speech enhancement method. The goal therefore is to design a system where the strengths of one method help compensate any potential weaknesses of the other. The noise cancellation method is an acoustic beamformer designed and constructed using a linear microphone array. The speech enhancement method is the constrained iterative Auto-LSP approach, previously considered for single channel enhancement. After establishing the combined processing scheme, evaluations are performed using speech and acoustic noise data collected in vehicles. Noise suppression levels by the beamformer is established for different road noise conditions. Quality improvement from the enhancement scheme is assessed using objective speech quality measures over a test speech corpus using TIMIT data. The results show that while beamforming alone can suppress background noise levels, the combination of beamforming and constrained enhancement can provide as much as a 63% improvement in objective quality, suggesting a potential single comprehensive solution for in-vehicle speech systems.

1. Introduction
The CU-Move system is a proposed in-vehicle interactive system for route planning and navigation. The goal is to develop a real-time naturally spoken dialog system in which a driver can obtain navigation and route planning in a hands-free automobile environment [1]. In order to obtain a high degree of performance in the relatively harsh acoustic environment of the automobile, a front-end noise cancellation and speech enhancement system will be employed. This front-end system uses array processing for noise cancellation and constrained iterative techniques for speech enhancement. A five-channel microphone array and data recording hardware were designed and constructed, and acoustic data and measurements were recorded in a number of different automobiles along a 15-mile test route. Front-end processing and performance tests were conducted in a non-real-time environment after downloading the acoustic data to a PC.

This paper presents the current performance results of the front-end processing system. It is organized as follows. In Section 2, background and research issues are discussed. In Section 3, the noise cancellation system and speech enhancement system are presented in detail. Section 4 gives the analysis methods and results in terms of noise suppression and objective quality measures. Section 5 concludes with the integration of the two sub-systems and other future considerations for the proposed front-end of the CU-Move system.

2. Background
Effective speech processing and recognition in automobile environments offers a potentially wide range of useful applications that might include: (i) hands-free dialing for cellular communications, (ii) command and control of automobile functions, and the most interesting and challenging (iii) hands-free dialog for route navigation and wireless information access. Though the rewards of such a system are great, given the harsh acoustic environment of the automobile, implementation remains a formidable task. Existing systems to date exhibit far from desirable performance. Attempts at speech recognition result in high word-error-rates which typically range from 30-65 percent [1], and almost exclusively focus on isolated-word, command-and-control, or limited structured dialog prompting.

A major factor in the poor performance of these systems is the multitude and variability of noise conditions experienced in the automobile environment. There are noise conditions that may be viewed as somewhat stationary and car dependent, such as engine noise at a given speed. There are noises that are impulsive in nature, such as turn signals, wiper blades, or car horns; and noise sources that are clearly not vehicle dependent and random nature such as a truck passing the vehicle with windows open at 65mph. The unpredictable nature of the noise experienced in the

1 This work was supported by DARPA through SPAWAR under Grant No. N66001-002-8906.
automobile makes training and adapting recognition models a difficult process. Therefore, it is evident that the potential for considerable improvement in these speech systems lies in the ability of an intelligent front-end signal processing framework. This being said, an effective front-end must make an attempt to compensate for the wide range of noise characteristics experienced in the automobile. The system should both suppress acoustic noise where possible, but potentially also provide information to back-end recognition model adaptation methods when noise sources are determined to be stationary over a given time window. The system proposed here is a combination of a noise cancellation method and a speech enhancement method. These methods were chosen as a pair in order to mutually compensate for any weakness one or the other might possess. The goal is to provide a system that performs well in the multitude of noise conditions experienced in the automobile.

This system is composed of a microphone array beamformer and a constrained iterative enhancement method connected in series. The beamformer was chosen because of its effectiveness in canceling the highly directional noise located in the upper range of the frequency spectrum. The constrained iterative approach performs very well at enhancing speech in the presence of broadband or low-frequency, slowly varying noise (an advantage for in-vehicle systems, since some car dependent noise types tend to dominate below 1kHz). Combined, these two systems offer the potential to markedly improve the quality of speech input to a recognizer.

3. System Overview

3.1 Noise Cancellation: Microphone Array Beamformer

In many applications of processing signals that propagate in space, it is desirable to extract signal components of desired temporal and spatial characteristics. For example, in a noisy environment with many sources, we may wish to extract a single frequency band propagating from one source and cancel those propagating from other directions. One widely used method of achieving this is by filtering in the spatial domain with the use of a multi-sensor array beamformer. By using a response that is pure delay in time and a simple weight in space, we sum the signals received at the array sensors and construct the well-known weighted delay-and-sum beamformer. The beamformer response is thus given as:

$$bf(t) = \frac{1}{N} \sum_{i=0}^{N-1} w_i r_i(t - \tau)$$

(1)

where N is the number of sensors in the array, $r_i(t)$ is the signal received at the $i^{th}$ sensor, $\tau_i$ is the sensor delay, and $w_i$ is the sensor weight[2]. The delays, $\tau_i$, are chosen to compensate for time differences of arrival of the desired signal at each sensor. Thus, the desired signal components are added in phase, while signals propagating from other directions are added out of phase, canceling each other. For non-vehicle dependent noise sources, such as exterior or time varying traffic noise when automobile windows are open, the directionality of the beamformer can be useful in suppressing this interference. Sensor weights are chosen to trade off the width of the main beam with sidelobe power. We use the coefficients of a Chebyshev polynomial[3] to effectively achieve this task.

To realize the weighted delay-and-sum beamformer for speech acquisition in the automobile, a microphone array was constructed. In order to prevent the appearance of grating lobes, in which high frequency signals of an undesirable propagation direction are passed, the spacing between microphones in the array must be less than half the shortest wavelength in the signal [4]. Since we are dealing with a bandwidth of 4kHz, we chose our microphones to have a spacing of 4.25 cm (note that for the data recording used for in-vehicle processing, a 44kHz sample rate is used to obtain more accurate delay offsets before downsampling to subsequent processing stages). Another consideration in constructing the array is effective beamwidth. Effective beamwidth of an array is inversely proportional to the length of the array in wavelengths [4]. To avoid grating lobes while at the same time maintaining a tight beamwidth, it is apparent that a large array of many tightly spaced sensors is needed. However, in a confined automobile environment, this is simply not feasible. We chose to construct a 5 microphone array in order to give reasonably good performance, while minimizing the size of the array. The theoretical spatial response of the beamformer is shown in Fig. 1 for frequencies of 1, 3, and 4kHz.
some time varying noise sources such as high frequency sirens, horns, etc. are difficult to address using typical speech enhancement methods or model adaptation methods, and therefore beamforming offers an attractive means of suppressing such interference for speech recognition. Performance of the constructed array was tested using a function tone generator and loudspeaker, rotated 180° about the array. The results are shown in Fig. 2, and it is evident that the performance is very close to the ideal response from Fig. 1.

Fig. 2: Actual Beamformer Response (1,3,4kHz)

3.2 Speech Enhancement: NA-AutoLSP

The speech enhancement algorithm used, is a noise adapted Auto-LSP (NA-AutoLSP), which has been the subject of previous research studies\[5,6\]. It is essentially a constrained iterative noise-adaptive Wiener filtering technique that employs speech production spectral constraints between sequential MAP estimation. An estimate of the power spectral density of the acoustic noise is generated in a portion of the data in which no speech activity is detected. LPC spectral estimates of the speech plus noise are generated on a frame-by-frame basis and are transformed to a line-spectral-pair representation. Constraints across iteration and across time are then placed on these line-spectral-pairs to reduce jitter an ensure a more effective speech estimate representation, thereby subsequently reducing noise. These constrained LSP’s are then transformed back to an LPC representation and a Weiner filter is constructed based upon the constrained LPC based spectrum and the initial estimate of the noise spectrum. The data is filtered using the constrained Wiener filter, and the procedure is repeated iteratively. Fig. 3 shows a sample speech spectrograms of the TIMIT sentence “She had your dark suit in greasy wash water all year.” Male speaker, 8kHz sample rate. (a) original clean; (b) degraded with “Wind” noise from SUV vehicle, windows open ~2 inches traveling 65mph; (c) beamformer output; (d) beamformer plus 6 iterations of NA-Auto-LSP enhancement.

It is apparent that the enhancement method does an excellent job of removing the low frequency, slowly varying noise, while maintaining the spectral structure of the speech signal. It is the effectiveness of the method at enhancing low frequencies that makes it an ideal choice for complementing the beamformer.

4. Analysis and Results

The beamformer and enhancement processing front-end was evaluated separately and together using data recorded in vehicles on a 15-mile course through the city of Boulder, CO (i.e., this road course includes city and highway driving, along with other noise conditions such as acceleration, turn signal noise, wiper blades, and a variety of exterior road noise sources). Noise suppression abilities of the beamformer were tested in the presence of four different noise conditions in the automobile (extracted from our field data collection). These four conditions were chosen as a representative set of the wide range of noise conditions present in the typical automobile environment. The four conditions are as follows: (1) windows rolled up and air-conditioning on high, (2) windows partly rolled down and a truck passing by, (3) windows rolled up and turn signal on, and (4) windows partly rolled down (1-2 inches) at a speed of 45mph. These conditions are referred to herein as AC, Truck, Turn, and Wind, respectively. Examples of these conditions were extracted from the data set and passed through the beamformer. Fig. 4 shows the effectiveness of the beamformer at attenuating these four noise conditions. The log power in the upper 2500Hz of the spectrum
(1.4-4kHz) is plotted vs. the log power in the lower 1500 Hz. The circled data points represent the log power contained in the signals after beamforming to the driver’s location. The dependence of attenuation on frequency is apparent in the figure, as attenuation of high frequencies ranges from about 5 to 7 dB, whereas the lower frequency attenuated is between 2 to 3 dB.

Next, the system was tested objective speech quality measures, namely the Itakura-Saito (IS) log-likelihood measure. Ten clean sentences spoken by ten different speakers were taken from the TIMIT database and degraded with the previously mentioned four noise conditions recorded in the automobile. Signal-to-Noise ratios of +5, 0, and -5 dB were used in the evaluation. Quality measures of the degraded, AutoLSP enhanced, and the beamformer plus AutoLSP enhanced sentences were calculated with respect to the clean TIMIT sentences and averaged. Overall quality measure results are shown in Fig. 5 (each entry represents an average over 1715 frames of processed data). Quality improvement was obtained in all noise conditions with enhancement engaged. Further improvements were seen when the beamformer was used in conjunction with enhancement in all cases except the Air Conditioning noise. When compared to the original degraded, improvements are obtained in all cases with the noise cancellation and speech enhancement systems used together.

5. Conclusions

Considerable improvements in speech quality can be realized with the concurrent use of noise cancellation and speech enhancement in the presence of the varying types of automobile noise. The two methods proposed here exhibit good performance in even relatively harsh amounts of noise (-5dB SNR). Noise attenuation of 5 to 7 dB was obtained with the beamformer at high frequencies, and the enhancement method proves to be effective in the lower frequency ranges. Combined, these two offer a competitive solution to improving speech quality for in-vehicle speech systems in the presence of many types of noise. The performance gains of such a system undoubtedly result in higher-quality speech input to a recognizer, as evidenced in the quality measure results presented here.

<table>
<thead>
<tr>
<th>SNR +5dB</th>
<th>AC</th>
<th>Truck</th>
<th>Turn</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Degraded</td>
<td>2.875</td>
<td>4.997</td>
<td>6.614</td>
<td>4.346</td>
</tr>
<tr>
<td>NA-AutoLSP</td>
<td>1.945</td>
<td>2.662</td>
<td>5.145</td>
<td>2.293</td>
</tr>
<tr>
<td>Beamformer and NA-AutoLSP</td>
<td>2.201</td>
<td>2.303</td>
<td>4.684</td>
<td>1.582</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR 0dB</th>
<th>AC</th>
<th>Truck</th>
<th>Turn</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Degraded</td>
<td>3.807</td>
<td>6.702</td>
<td>7.359</td>
<td>4.834</td>
</tr>
<tr>
<td>NA-AutoLSP</td>
<td>2.938</td>
<td>3.831</td>
<td>5.385</td>
<td>3.563</td>
</tr>
<tr>
<td>Beamformer and NA-AutoLSP</td>
<td>3.188</td>
<td>3.277</td>
<td>5.157</td>
<td>2.590</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNR -5dB</th>
<th>AC</th>
<th>Truck</th>
<th>Turn</th>
<th>Wind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Degraded</td>
<td>4.535</td>
<td>7.217</td>
<td>9.008</td>
<td>5.515</td>
</tr>
<tr>
<td>NA-AutoLSP</td>
<td>4.343</td>
<td>4.953</td>
<td>6.744</td>
<td>5.242</td>
</tr>
<tr>
<td>Beamformer and NA-AutoLSP</td>
<td>4.500</td>
<td>4.370</td>
<td>6.610</td>
<td>3.869</td>
</tr>
</tbody>
</table>

Fig. 5: Overall IS Quality Measures averaged across TIMIT sentences in the presence of four different automobile noise sources at -5, 0, and 5 dB SNR.

6. References