In-Vehicle Based Speech Processing for Hearing Impaired Listeners

Xiaoxian Zhang\textsuperscript{1}, John H.L. Hansen\textsuperscript{2}, Kathryn Arehart\textsuperscript{2}, Jessica Rossi-Katz\textsuperscript{2}

\textsuperscript{1}Robust Speech Processing Group, Center for Spoken Language Research, University of Colorado, Boulder, Colorado 80309 U.S.A.
\textsuperscript{2}Department of Speech, Language & Hearing Science, University of Colorado, Boulder, Colorado 80309 U.S.A.

Abstract

Noisy cars are very difficult listening environments for persons with hearing loss. While there have been numerous studies in the field of speech enhancement for car noise environments, the majority of these studies have focused on noise reduction for normal hearing individuals. In this paper, we present recent results in the development of more effective speech capture and enhancement processing for wireless voice interaction for persons with hearing loss in real car environments. We first present a data collection experiment for a proposed FM wireless transmission scenario using a 5-channel microphone array in the car, followed by several alternative speech enhancement algorithms. After formulating 6 different processing methods, we evaluate the performance by SegSNR improvement using data recorded in a moving car environment. Among the 6 processing configurations, the combined fixed/adaptive beamforming (CFA-BF) obtains the highest level of SegSNR improvement by up to 2.65 dB.

1. Introduction

There are estimated 28 million people in the U.S. who are hearing impaired\textsuperscript{[1]}. Noisy cars are very difficult listening environments for persons with hearing loss. It is important to optimize communication for hearing-impaired persons in cars. There are numerous areas where it is necessary to enhance the quality of speech degraded by background noise. Examples of environments include in-vehicle hands-free voice communications, mobile phone use in public noisy environments, hearing impaired persons in large classrooms or meeting halls, and others. In addition, a number of speech enhancement algorithms have been proposed in the past and a survey can be found in \cite{2}. However, the majority of these studies have focused on noise reduction for normal-hearing individuals. Listeners with cochlear hearing loss have much more difficulty understanding speech than do normal-hearing (NH) listeners. This increased difficulty is especially pronounced in noisy environments \cite{3, 4}. While wireless FM devices exist and can be used in applications such as student-teacher settings in classrooms to help hearing-impaired students, the highly time-varying noise conditions within car environments make such solutions less attractive for human-to-human communications. Therefore, human voice interaction within car environments becomes a challenging task for listeners with hearing loss.

In this study, we present recent results in the development of more effective speech capture and enhancement processing for wireless voice interaction for listeners with hearing loss in real car environments. We first present a data collection experiment for a proposed FM wireless transmission scenario using a 5-channel microphone array in the car, followed by several alternative speech enhancement algorithms. These algorithms are tested in the context of noise reduction for hearing aid applications. The in-vehicle data collection also simulates the signal capture expected by a driver fitted with a behind-the-seat hearing aid. In our evaluations, two speakers (one male, one female) were used to collect in excess of 800 phonetically balanced sentences using a dashboard mounted computer display. Our goals were as follows: (1) determine acoustic environment for typical listener situation in real car environment, (2) evaluate effectiveness of different processing strategies for improving SegSNR in the different noise environments, (3) determine whether the effectiveness of the algorithm varies between normal-hearing and hearing-impaired individuals, (3) find a better solution to increase hearing impaired listeners’ abilities to understand speech in noisy car environments.

2. In-Vehicle Data Collection

In the proposed collection framework, we assume that the driver is a hearing-impaired person, and the speaker is the passenger who sits beside him. Our goal of this study is to investigate the driver’s ability to understand speech while the passenger is talking for different processing configurations under car noise situations. Therefore, our data collection focuses on collecting acoustic noise from a variety of driving conditions. Fig. 1 shows images of the data collection setup for the multi-channel microphone array, close talking microphone, and constructed behind-the-seat microphone. A fixed 10 mile route through Boulder, CO was used for data collection. The route included stop-and-go traffic, and prescribed locations where driver/passenger windows, turn signals, wiper blades, and air conditioning were operated. The data collection run lasted approximately 45 minutes. A selection of phonetically balanced 450 IEEE sentences were produced along the same route locations.

\footnotesize† This project was supported by a grant from Whitaker Foundation.
3. Algorithms development

Given the challenging range of car noise environments, we considered a range of multi-channel and single-channel algorithms for this project: (i) fixed single channel unprocessed data, (ii) delay and sum beamforming, (iii) Modified GMMSE-AMT-ERB, (iv) Improved DASB with GMMSE-AMT-ERB, (v) CFA-BF: Combined Fixed/Adaptive Beamforming. Next, we consider each of the processing configurations.

3.1. Single Channel Unprocessed data --- Baseline

In our data collection, we used a 5 channel linear microphone array. Therefore, we chose a single channel from the center of the microphone array as our baseline.

3.2. Delay-and-sum Beamforming --- DASB

One of the simplest beamforming solutions is weighted delay and sum beamforming (DASB) \[5, 6\]. 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. The beamformer output \( y(n) \) is formed by averaging weighted and delayed versions of the microphone signals as follows

\[
y(n) = \frac{1}{N} \sum_{i=1}^{N} W_i x_i(n - \tau_i)
\]

(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 \( i \)-th microphone are given as \( W_i \) and \( \tau_i \). The first portion of 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 \( \tau_i = (i-1)d/T_s \).

3.3. Constructed Behind-the-ear Microphone

In order to simulate the real conversation situation inside a moving car, we also placed a microphone behind the ear of the driver. The microphone was constructed using Knowles omnidirectional microphones. During our data collection, the speaker is the passenger who sits besides the driver, and the listener is the driver. We studied the case without additional speech enhancement processing to determine how the listener perceives the conversation.

3.4. Modified GMMSE-AMT-ERB

A recently developed noise suppression algorithm uses the auditory-masked (AMT) in conjunction with a modified generalized minimum mean squared error estimation. This GMMSE-AMT-ERB \[7\] speech enhancement processor was used for single channel speech enhancement. The reason we chose this algorithm is that it employs an auditory masked threshold with equivalent rectangular bandwidth based filter partitioning, bandwidth spreading and elevated threshold representation for hearing-impaired listeners. We want to extend this method to car noise environments. In the original implementation, the speech is assumed to be degraded with additive noise and the speech and noise segments are uncorrelated as in Eqn (2):

\[
y(n) = x(n) + n(n)
\]

(2)

The short term power spectrum is calculated by applying a Hamming window to a frame of speech. Under this assumed model, one can obtain a family of MMSE speech spectral estimation as:

\[
\hat{X}_p = (E[X_p^a Y_p])^{1/2}
\]

(3)

Here, \( P_{ik} \) be the noise power spectrum for the \( k^\text{th} \) subband, and \( P_{ik} \) be the noisy speech power spectrum for the \( k^\text{th} \) subband. The values of \( P_{ik} \) and \( P_{ik} \) are calculated as follows:

\[
P_{ik}[n] = \rho P_{ik}[n-1] + \frac{1 - \lambda}{1 - \beta} (P_{ik}[n] - \beta P_{ik}[n-1])
\]

(4)

\[
P_{ik}[n] = \alpha P_{ik}[n-1] + (1 - \alpha) Y_i[n] F_i
\]

(5)

In the implementation, the first ten frames of noisy speech, which consists of only noise, is taken as the estimation of the noise for the entire noisy speech sentence. This assumption is valid if the noise does not change. However, once the noise spectrum changes, enhancement performance will decrease, resulting in either under or over noise suppression. Therefore, in our experiments, we modified this algorithm and added a front-end noise estimation stage. Once a noise change is detected, noise spectrum updating is performed. We do not update the noise spectrum on a frame by frame basis, since we believe this will increase speech distortion. The difference between original GMMSE-AMT-ERB and modified GMMSE-AMT-ERB is the presence (e.g., with/without) of the noise classification and update stage in the modified algorithm.
3.5. Improved DASB with GMMSE-AMT-ERB

We also employed an improved DASB processor with subsequent GMMSE-AMT-ERB processing that employs improved speech/noise detection using a Target Energy Operator [8] based scheme to detect the presence of the speech. Since beamforming algorithms (delay-and-sum beamforming or adaptive beamforming) obtain the enhanced signal by selecting the appropriate delays (fixed or adaptive) between each microphone and summing the delayed signals in phase for direction angle $\theta$, we will have destructive interference for noise signals arriving from other angles. Therefore, in this processing method, we use the output of the delay and sum beamforming instead of the single microphone. Fig. 2 illustrates the configuration of this method.

![Fig. 2: Configuration of Improved DASB with GMMSE-AMT-ERB](image)

3.6. CFA-BF: Combined Fixed/Adaptive Beamforming

The fifth algorithm is a recently developed CFA-BF (combined fixed and adaptive beamforming) algorithm developed for speech enhancement for car noise environments [9]. The reason we selected this algorithm is to determine if it works equally well for hearing impaired persons. The working scheme of CFA-BF consists of two steps: source location calibration and target signal enhancement. As is well known, an adaptive algorithm such as the normalized Least Mean Square algorithm (NULMS) can more easily reach its convergence behavior in quiet or stationary noise environments, than under non-stationary noise environments (e.g., car noise environments). Therefore, the first step of this method is to pre-record the transfer functions between speaker and microphone array from different potential source positions using adaptive beamforming under quiet environments. The second step is to use this pre-recorded information to enhance the desired speech when the car is operating on the road.

4. Performance Evaluation

For our evaluations, we report objective measure performance using SegSNR sentence level distributions. Fig. 3 shows the SegSNR histogram of close-talking and the constructed behind-the-ear microphone (both from 454 sentences). Here, the close talking microphone is attached to the mouth of the speaker, and the constructed behind-the-ear microphone is attached to the ear of listener.

![Fig. 3: SegSNR Histograms for Microphone](image)

Fig. 4 shows the SegSNR histograms with Mean/Standard deviation across 454 sentences for each processing configuration.

![Figure 4: SegSNR Histogram for Processing Configuration](image)

Fig. 5 shows the rank ordered SegSNR improvement across all 454 sentences compared with the baseline system for all the processing configurations. From these results, we make the following observations:

i. Reliable noise suppression in car environments is challenging since the difference between average SegSNR generated between the speaker and perceived by the listener is 26.7 dB.
Among the 6 processing configurations, the combined fixed/adaptive beamforming algorithm (CFA-BF) obtained the highest averaged SegSNR improvement of 2.65 dB.

However, across the entire set of test sentences, the single channel modified GMMS-E-AMT-ERB has the highest SegSNR improvement at an individual sentence level of 10.25 dB.

5. Conclusions and Future Work

In this study, we have considered a range of speech processing configurations to improve quality for human-to-human communication within car environments when one person has a hearing loss (i.e., the driver fitted with a behind-the-ear microphone). We report objective measure performance for 6 different processing configurations using SegSNR sentence level distributions each across 454 sentences recorded in real moving car environments. The algorithms included (i) a baseline single channel microphone with no processing, (ii) traditional delay-and-sum 5 microphone array processing, (iii) a constructed behind-the-ear microphone to simulate a digital hearing aid worn by the driver, (iv) a single channel modified GMMS-E-AMT-ERB scheme that includes auditory masking, (v) DASB with GMMS-E-AMT-ERB, (vi) and a combined fixed/adaptive 5-mike array processing scheme (CFA-BF). We have shown that effective speech enhancement and noise suppression in real car noise environments is a challenging task. We also demonstrated that the experimental results do show improvement in SegSNR. We found that among all the processing configurations, the combined fixed/adaptive beamforming (CFA-BF) approach can obtain the highest averaged SegSNR improvement 2.65 dB.

However, there remain some issues to address for future work:

- Performing a formal listener evaluation using both hearing-impaired and normal-hearing listeners.
- Determining if this improvement is consistent for a larger number of test speakers; and if competing speaker noise in the car can be addressed with the proposed methods.
- Combining multi-channel array with constrained iterative speech enhancement algorithm (Auto-SP) [10] to improve SegSNR further.

5. References