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,