A PUBLICATION OF THE ENGINEERING SCIENCES SOCIETY
The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3-chome, Minato-ku, TOKYO, 105-0011 JAPAN
LETTER
Special Section on Adaptive Signal Processing and Its Applications

A New Speech Enhancement Algorithm for Car Environment Noise Cancellation with MBD and Kalman Filtering

Seungkwon BEACK*1, Seung H. NAM†, and Minsoo HAHN†, Nonmembers

SUMMARY We present a new speech enhancement algorithm in a car environment with two microphones. The car audio signals and other background noises are the target noises to be suppressed. Our algorithm is composed of two main parts, i.e., the spatial and the temporal processes. The multi-channel blind deconvolution (MBD) is applied to the spatial process while the Kalman filter with a second-order high pass filter, for the temporal one. For the fast convergence, the MBD is newly expressed in frequency-domain with a normalization matrix. The final performance evaluated with the severely car noise corrupted speech shows that our algorithm produces noticeably enhanced speech.

key words: speech enhancement, multichannel blind deconvolution, Kalman filter

1. Introduction

Recently, speech recognition algorithms have been applied to the car environments, but inevitably their performances are severely degraded due to car noises. Car noises can generally be categorized into spatial and temporal ones. Spatial noises come from single or multiple point sources such as client’s speech and car audio signals while temporal noises are the ones generated by the engine vibration, the air turbulence, and the tire friction.

The spatial car noises are one of the hardest obstacles in speech enhancement. Recently, as the MBD has been introduced for the spatial noise suppression, the speech quality becomes dramatically improved. However, the simultaneous occurrence case of the spatial and the temporal noises is still a challenging task. In [3], several types of spatial or temporal noises were independently tested with the MBD combined with the sub-band processing and produced rather successful results. The simultaneous cases in car environment were tested with the spatio-temporal enhancement technique [2]. It produced fairly enhanced speech with rather complicated processes.

In this paper, we present a new speech enhancement algorithm basically based on the spatio-temporal process as in [2]. However, the components of our process are newly proposed and the overall structure is relatively simple. Our goal is to design a successful speech enhancement algorithm in a car environment where various car audio and background noises exist simultaneously. We exclude other clients’ interfering speech noises partly because the drivers are usually more often disturbed by the car audio than the client’s speech in a running car.

For the robustness of our algorithm, the frequency-domain block-based MBD (FB-MBD) with a normalization matrix is proposed as a part of the spatial process. The matrix is designed to overcome the intrinsic problems of the time-domain MBD (TD-MBD) such as the whitening effect and the slow convergence [8]. It is derived from the information-maximization-based TD-MBD with natural gradients [1]. The temporal process based on the high pass and the Kalman filtering to suppress the background noises is subsequently applied [7].

We use the following notation throughout the paper. Bold uppercases are used for matrices and vectors, respectively. Normal lowercase are for scalars and vector elements. "n" and "b" are for the time and the block indices, respectively. The superscript f indicates the frequency-domain quantity.

2. Problem Description

The corrupted signal, \( x_j(n) \) from the \( j \)th-microphone can be represented as the convolution sum,

\[
x_j(n) = \sum_{p=0}^{Q} a_{jj,p} s_j(n - p) + \sum_{i=1}^{l} \sum_{n=0}^{Q} a_{ji,p} s_i(n - p) + v(n), \quad j = 1, 2, \ldots, m.
\]

In this equation, \( a_{ji,p} \) is the \((j,i)\) element of the mixing system of \( A(z) = \sum_{p=0}^{Q} A_p z^{-p} \) with the finite impulse response (FIR) \( Q \), and \( s_j(n) \) is the \( j \)th-point source. The second term on the right-hand side indicates the group of spatial noises propagated from the several interfering point sources, where \( s_i(k) \) is the target speech signal, \( v(n) \) is the microphone-independent temporal noise. Similarly, the \( j \)th-unmixing signal can be described as

\[
u_j(n) = \sum_{p=0}^{Q} w_{jj,p}(n)x_j(n - p) + \sum_{i=1}^{l} \sum_{n=0}^{Q} w_{ji,p}(n)x_i(n - p)
+ \sum_{i=1}^{l} w_{ji,p}(n)v(n - p)
\]

(2)