Accepted Manuscript

IRM Estimation Based on Data Field of Cochleagram for Speech Enhancement

Xianyun Wang , Feng Bao , Changchun Bao

PII: S0167-6393(17)30265-0
DOI: 10.1016/j.specom.2017.12.014
Reference: SPECOM 2521

To appear in: *Speech Communication*

Received date: 12 July 2017
Revised date: 26 December 2017
Accepted date: 29 December 2017

Please cite this article as: Xianyun Wang , Feng Bao , Changchun Bao , IRM Estimation Based on Data Field of Cochleagram for Speech Enhancement, *Speech Communication* (2018), doi: 10.1016/j.specom.2017.12.014

IRM Estimation Based on Data Field of Cochleagram for Speech Enhancement

Xianyun Wang[1], Feng Bao[2] and Changchun Bao[1]

[1]*Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China*

[2]*Department of Electrical and Computer Engineering, The University of Auckland, Auckland 1142, New Zealand*

E-mail: b201402001@emails.bjut.edu.cn, fbao026@aucklanduni.ac.nz, baochch@bjut.edu.cn

(Corresponding author: Changchun Bao Tel: 8610-67391635 Fax: 8610-67391625 E-mail: baochch@bjut.edu.cn Pos: Beijing University of Technology, Beijing 100124, China)

Abstract

When computational auditory scene analysis (CASA) is used for the speech enhancement, it can mask noise effectively by an accurate mask estimation approach. In this paper, we attempt to apply the ideal ratio mask (IRM) estimation based on the spectral dependency into the speech cochleagram for enhancing speech. To achieve the spectral dependency, the concept of data field (DF) is introduced to model the time-frequency (T-F) relationship of the cochleagram so that the obtained results (termed as the potentials) with the adjacent spectral information are used eventually to estimate the IRM. In the estimation framework, we firstly use a pre-processed module to obtain initial T-F values of noise and speech. Then, given initial estimations of noise and speech, we can employ DF model to obtain the forms of speech and noise potentials, which are viewed as the energy with the information of its neighbors. Subsequently, based on the forms of speech and noise potentials, their optimal potentials that reflect their respective optimal distribution are obtained by the optimal influence factors. Finally, we attempt to obtain the masking value using the potentials of speech and noise for restoring clean target speech signal. Our algorithm is evaluated and compared with the reference methods, and it can yield an effective improvement in speech quality.

Key Words: Speech enhancement, Data field, CASA, Ratio mask

1. Introduction

Speech enhancement is generally used to improve the quality of speech deteriorated by the noise and increase the intelligibility of noisy speech in many speech communication systems. It has a wide range of applications, such as mobile phone disturbed from background noise, the front-end of speech recognition technology, military communications, hearing aids and so on.

For the speech signal collected by one microphone, the intrinsic properties of speech and noise are often considered. Various algorithms have been proposed for single-channel speech enhancement, and they are generally based on some characteristic analysis of speech or noise and subsequently perform speech amplification or noise reduction. For example, some methods have been proposed to obtain speech by estimating the short-time spectrum of noise [1]-[3] or to extract speech based on speech modeling [4]-[6], which have demonstrated their abilities under stationary noise conditions, but they do not perform well when the non-stationary noise is introduced. The main reason is that these algorithms are designed to be dependent on the accuracy of the estimation of noise power spectrum. We know that it is relative easy to estimate noise spectral level for many noise estimation methods, such as minimum statistics [2][3], when the noise is stationary. However, it becomes very difficult in following the rapid changes of non-stationary noise energy. For example, the noise power spectrum in [3] was determined by following the minima in a buffer of past noisy spectra. And generally, large enough buffer needs to be given to ensure that the buffer contains the minima. However, larger buffer usually makes it difficult to track time-varying noise with quickly energy changes. Thus, the performance of the noise reduction methods becomes unsatisfactory when the noise becomes non-stationary.