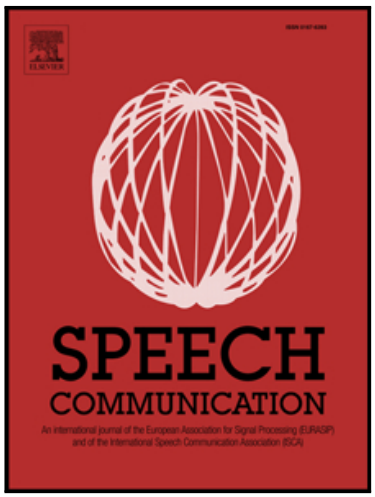

A Unified DNN Approach to Speaker-dependent Simultaneous Speech Enhancement and Speech Separation in Low SNR Environments

Tian Gao, Jun Du, Li-Rong Dai, Chin-Hui Lee

PII: S0167-6393(16)30384-3
DOI: 10.1016/j.specom.2017.10.003
Reference: SPECOM 2495

To appear in: *Speech Communication*

Received date: 30 December 2016
Revised date: 14 October 2017
Accepted date: 16 October 2017

Please cite this article as: Tian Gao, Jun Du, Li-Rong Dai, Chin-Hui Lee, A Unified DNN Approach to Speaker-dependent Simultaneous Speech Enhancement and Speech Separation in Low SNR Environments, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.10.003

A Unified DNN Approach to Speaker-dependent Simultaneous Speech Enhancement and Speech Separation in Low SNR Environments

Tian Gao[a,*], Jun Du[a,*], Li-Rong Dai[a], Chin-Hui Lee[b]

[a] *National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, China*
[b] *Georgia Institute of Technology, Atlanta, Georgia, United States*

Abstract

We propose a unified speech enhancement framework to jointly handle both background noise and interfering speech in a speaker-dependent scenario based on deep neural networks (DNNs). We first explore speaker-dependent speech enhancement that can significantly improve system performance over speaker-independent systems. Next, we consider interfering speech as one noise type, thus a speaker-dependent DNN system can be adopted for both speech enhancement and separation. Experimental results demonstrate that the proposed unified system can achieve comparable performances to specific systems where only noise or speech interference is present. Furthermore, much better results can be obtained over individual enhancement or separation systems in mixed background noise and interfering speech scenarios. The training data for the two specific tasks are also found to be complementary. Finally, an ensemble learning-based framework is employed to further improve the system performance in low signal-to-noise ratio (SNR) environments. A voice activity detection (VAD) DNN and an ideal ratio mask (IRM) DNN are investigated to provide prior information to integrate two sub-modules at frame level and time-frequency level, respectively. The results demonstrate the effectiveness of the ensemble method in low SNR environments.

Keywords: Speaker-dependent speech processing; speech enhancement; speech separation; deep neural network; low SNR

1. Introduction

Speech enhancement [1] and speech separation [2] are important front-ends of speech processing systems aimed at noise reduction and segregating speech from mixed speakers, respectively. Background noise and human voice interference can reduce both the quality and intelligibility of the speech signals and cause performance degradations in real-world applications, including speech communication, hearing aids and speech and speaker recognition. A key goal of speech enhancement [3] is to improve quality and intelligibility in the presence of interfering noise. On the other hand, speech separation [2, 4] aims to separate the voice of a target speaker when multiple speakers talk simultaneously.

Numerous methods were developed over the past several decades for speech enhancement and speech separation. For enhancement, the conventional methods include a wide range of approaches, such as spectral subtraction [5], Wiener filtering [6] and statistical-model-based algorithms [7]. Spectral subtraction is one of the first algorithms proposed for noise reduction. However, the resulting enhanced speech often suffers from an annoying artifact called musical noise [8]. The Wiener algorithm, minimum mean squared error (MMSE) estimation [9, 10] and optimally modified log-spectral amplitude (OM-LSA) speech estimator [11] all exist in a statistical estimation framework that attempts to find a linear (or nonlinear) estimator of the parameters of interest. OM-LSA utilizes a minima controlled recursive averaging (MCRA) noise estimation approach to avoid the musical residual noise phenomena. One limitation of the conventional speech enhancement algorithms is that they can't improve speech intelligibility effectively. Supervised and unsupervised nonnegative matrix factorization (NMF) methods were investigated in [12, 13]. The basic idea is to decompose the training data into bases and weight matrices for speech and noise, respectively.

For separation, one broad class is the so-called computational auditory scene analysis (CASA) [14], usually in an unsupervised mode. CASA-based approaches [15, 16, 17, 18, 19], use the psychoacoustic cues such as pitch, onset/offset, temporal continuity, harmonic structure and modulation correlation, and segregate a voice of interest by masking the interfering sources. For example, in [18], pitch and amplitude modulation are adopted to separate the voiced portions of cochannel speech. In [19], unsupervised clustering is used to separate speech re-

*Corresponding author
Email addresses: gtian09@mail.ustc.edu.cn (Tian Gao), jundu@ustc.edu.cn (Jun Du), lrdai@ustc.edu.cn (Li-Rong Dai), chl@ece.gatech.edu (Chin-Hui Lee)