Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 00 (2017) 1–13

www.elsevier.com/locate/csl

# Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments☆

Hendrik Barfuss*, Christian Huemmer, Andreas Schwarz[1], Walter Kellermann

*Multimedia Communications and Signal Processing, Friedrich-Alexander University Erlangen-Nürnberg, Cauerstr. 7, Erlangen 91058, Germany*

Received 8 April 2016; received in revised form 6 February 2017; accepted 10 February 2017

**Abstract**

Speech recognition in adverse real-world environments is highly affected by reverberation and non-stationary background noise. A well-known strategy to reduce such undesired signal components in multi-microphone scenarios is spatial filtering of the microphone signals. In this article, we demonstrate that an additional coherence-based postfilter, which is applied to the beamformer output signal to remove diffuse interference components from the latter, is an effective means to further improve the recognition accuracy of modern deep learning speech recognition systems. To this end, the 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) baseline speech enhancement system is extended by a coherence-based postfilter and the postfilter's impact on the Word Error Rates (WERs) of a state-of-the-art automatic speech recognition system is investigated for the realistic noisy environments provided by CHiME-3. To determine the time- and frequency-dependent postfilter gains, we use Direction-of-Arrival (DOA)-dependent and (DOA)-independent estimators of the coherent-to-diffuse power ratio as an approximation of the short-time signal-to-noise ratio. Our experiments show that incorporating coherence-based postfiltering into the CHiME-3 baseline speech enhancement system leads to a significant reduction of the WERs, with relative improvements of up to 11.31%.


*Keywords:* Robust speech recognition; Postfiltering; Spectral enhancement; Coherence-to-diffuse power ratio; Wiener filter

## 1. Introduction

For a satisfying user experience of human-machine interfaces it is crucial to ensure a high accuracy in automatically recognizing the user's speech. However, as soon as no close-talking microphone is used for capturing the desired speech signal, the recognition accuracy suffers from additional reverberation, background noise and active interfering speakers which are picked up by the microphones (Delcroix et al., 2013; Yoshioka and Gales, 2015). Techniques for robust speech recognition in such reverberant and noisy environments can be categorized into either

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding author.
*E-mail address:* hendrik.barfuss@fau.de (H. Barfuss), christian.huechr.huemmer@fau.de (C. Huemmer), andreas.schwarz@fau.de (A. Schwarz), walter.kellermann@fau.de (W. Kellermann).
[1] Friedrich-Alexander University of Erlangen-Nürnberg while the work has been conducted. He is now with Amazon Development Center, Aachen, Germany.

http://dx.doi.org/10.1016/j.csl.2017.02.005
0885-2308/

front-end (e.g., speech enhancement (Cohen, 2003; Krueger and Haeb-Umbach, 2010; Gales and Wang, 2011)) or back-end (e.g., acoustic-model adaptation (Li and Bilmes, 2006; Liao, 2013; Yu et al., 2013)) processing techniques.

The 3rd CHiME Speech Separation and Recognition Challenge (CHiME-3) (Barker et al., 2015) targets the performance of state-of-the-art Automatic Speech Recognition (ASR) systems in real-world scenarios. The primary goal is to improve the ASR performance of real recorded speech of a person talking to a tablet device in realistic everyday noisy environments by employing front-end and/or back-end signal processing techniques. To this end, a baseline ASR system has been initially provided and updated as follow-up of CHiME-3 to achieve a high recognition accuracy in everyday real-world scenarios. Front-end processing of the updated baseline now employs the BeamformIt toolkit (Anguera et al., 2007) for processing the recorded microphone signals by a Weighted Delay-and-Sum (WDS) beamforming technique. The beamformer output is used as input to the ASR back-end system which contains a Deep Neural Network (DNN)-based acoustic model and a Recurrent Neural Network (RNN)-based language model.

In this article, we extend the updated CHiME-3 baseline system by a low-complexity coherence-based postfilter which is applied to the beamformer output signal to further remove reverberation and non-stationary background noise from the latter. The postfilter is realized as a Wiener filter, where, in contrast to the classical Wiener filter, see, e.g., Diethorn (2000), we use an estimate of the Coherent-to-Diffuse Power Ratio (CDR) as an approximation of the short-time Signal-to-Noise Ratio (SNR) to compute the time- and frequency-dependent Wiener filter gains. The CDR, which is the ratio of the power of direct and diffuse signal components, needs to be estimated from the microphone signals. We compare and evaluate two Direction-of-Arrival (DOA)-independent and two DOA-dependent CDR estimators. Two of the evaluated CDR estimators have been proposed and shown to be very effective for dereverberation by Schwarz and Kellermann (2014, 2015). The remaining two CDR estimators have been proposed by Jeub et al. (2011) and Thiergart et al. (2012a, 2012b) earlier than (Schwarz and Kellermann 2014, 2015), and are evaluated as reference methods. In contrast to the previous work in (Schwarz and Kellermann 2014, 2015), where the dereverberation performance was evaluated using WERs of a Hidden Markov Model (HMM)−Gaussian Mixture Model (GMM)-based ASR system trained on clean speech, we now evaluate the efficacy of the CDR-based Wiener filter realizations with a state-of-the-art HMM−DNN-based ASR system trained on noisy training data from different acoustic environments (provided by CHiME-3 (Barker et al., 2015)). Moreover, the new CDR estimators in (Schwarz and Kellermann 2014, 2015) were proposed and evaluated for a dual-channel microphone array, whereas the recognition task of CHiME-3 involves signal enhancement using a six-channel microphone array. We therefore extend the CDR estimation procedure to a multi-channel (here: six-channel) scenario. To summarize, the contributions of this article are as follows:

1. First-time evaluation of new (Schwarz and Kellermann, 2014, 2015) and previously known (Jeub et al., 2011; Thiergart et al., 2012a, 2012b) CDR estimators with a state-of-the-art HMM−DNN-based ASR system in challenging acoustic scenarios.
2. First-time application of coherence-based dereverberation using the new CDR estimators (Schwarz and Kellermann, 2014, 2015) to a multi-microphone scenario with more than two microphones.

An overview of the signal processing pipeline employed in this work is given in Fig. 1. While the purpose of the beamformer is to spatially focus on the target source, i.e., to reduce the signal components from interfering point sources, the postfilter shall remove diffuse interference components, e.g., reverberation, from the beamformer output signal. The output of the front-end signal enhancement (consisting of beamformer and postfilter) is further processed by the ASR back-end system, which provides an HMM−DNN-based speech recognizer (see Section 3).

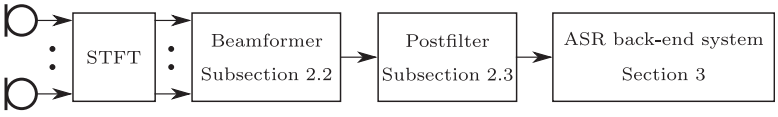


Fig. 1. Overview of the overall signal processing pipeline system with beamformer and postfilter as acoustic front-end signal processing. The acoustic back-end system, including feature extraction/transformation, is equal to the updated baseline acoustic back-end system of CHiME-3 (see Section 3 for more details on the employed ASR system).

Please cite this article as: H. Barfuss et al., Robust coherence-based spectral enhancement for speech recognition in adverse real-world environments, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.005