Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language 00 (2017) 1–16

www.elsevier.com/locate/csl

Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition

Niko Moritz[a,d,*], Kamil Adiloğlu[b,d], Jörn Anemüller[c,d], Stefan Goetze[a,d], Birger Kollmeier[a,b,c,d]

[a] *Fraunhofer IDMT, Project Group for Hearing, Speech, and Audio Technology, Marie-Curie-Str. 2, 26129 Oldenburg, Germany*
[b] *Hörtech gGmbH, Marie-Curie-Str. 2, 26129 Oldenburg, Germany*
[c] *University of Oldenburg, Medizinische Physik, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany*
[d] *Cluster of Excellence 'Hearing4all', 26129 Oldenburg, Germany*

Received 8 April 2016; received in revised form 30 September 2016; accepted 21 November 2016

Abstract

The paper describes a system for automatic speech recognition (ASR) that is benchmarked with data of the 3rd CHiME challenge, a dataset comprising distant microphone recordings of noisy acoustic scenes in public environments. The proposed ASR system employs various methods to increase recognition accuracy and noise robustness. Two different multi-channel speech enhancement techniques are used to eliminate interfering sounds in the audio stream. One speech enhancement method aims at separating the target speaker's voice from background sources based on non-negative matrix factorization (NMF) using variational Bayesian (VB) inference to estimate NMF parameters. The second technique is based on a time-varying minimum variance distortionless response (MVDR) beamformer that uses spatial information to suppress sound signals not arriving from a desired direction. Prior to speech enhancement, a microphone channel failure detector is applied that is based on cross-comparing channels using a modulation-spectral representation of the speech signal. ASR feature extraction employs the amplitude modulation filter bank (AMFB) that implicates prior information of speech to analyze its temporal dynamics. AMFBs outperform the commonly used frame splicing technique of filter bank features in conjunction with a deep neural network (DNN) based ASR system, which denotes an equivalent data-driven approach to extract modulation-spectral information. In addition, features are speaker adapted, a recurrent neural network (RNN) is employed for language modeling, and hypotheses of different ASR systems are combined to further enhance the recognition accuracy. The proposed ASR system achieves an absolute word error rate (WER) of 5.67% on the real evaluation test data, which is 0.16% lower compared to the best score reported within the 3rd CHiME challenge.

Keywords: Speech enhancement; Non-negative matrix factorization; Feature extraction; Modulation frequency analysis; CHiME; Amplitude modulation filter bank

[*] Corresponding author at: Fraunhofer IDMT, Project Group for Hearing, Speech, and Audio Technology, Marie-Curie-Str. 2, 26129 Oldenburg, Germany.

E-mail address: niko.moritz@idmt.fraunhofer.de (N. Moritz).

http://dx.doi.org/10.1016/j.csl.2016.11.004
0885-2308/ 2017 Published by Elsevier Ltd.

1. Introduction

Robust automatic speech recognition (ASR) with distant microphones in noisy environments still constitutes a major technological challenge and simultaneously a sought-after application scenario, e.g., when used in consumer electronics. The 3rd CHiME speech separation and recognition challenge provides a platform for the development and comparison of different noise robust front- and back-end technologies targeting the use case of a tablet device that is equipped with multiple microphones (Barker et al., 2015). One important part of robust ASR systems with distant microphones are multi-channel signal processing techniques that can emphasize spatial information to attenuate interfering sounds not arriving from a desired direction. The ASR system proposed here employs various techniques to enhance noise robustness, e.g., multi-channel speech enhancement, extraction of acoustic features by analyzing temporal dynamics of speech, speaker adaptation, and deep neural network (DNN) based acoustic modeling. In addition, due to occasional malfunctions of single microphones, a microphone channel failure detection algorithm is developed and employed to identify and exclude corrupted recordings prior to speech enhancement and ASR. In order to further reduce word error rates (WERs), sophisticated language modeling based on a recurrent neural network (RNN) and a minimum Bayes risk (MBR) system combination method are applied.

Two different speech enhancement methods are used and compared in the present study. The first method aims at separating the target speaker's voice from environmental noise sources. Many source separation algorithms have been discussed in the literature, ranging from early algorithms like independent component analysis (ICA), binaural cue clustering, and sparse component analysis (SCA) to more recent algorithms such as nonnegative matrix factorization (NMF), nonnegative tensor factorization (NTF), and factorial hidden Markov models (HMM) (Comon and Jutten, 2010; Vincent et al., 2014). The proposed source separation method is based on NMF and uses a general variational Bayes (VB) algorithm to estimate model parameters (Adiloğlu and Vincent, 2012; Ozerov et al., 2012). NMF results are compared with a second speech enhancement method, which is the time-varying minimum variance distortionless response (MVDR) beamformer (Bitzer and Simmer, 2001; Mestre and Lagunas, 2003). The direction of arrival (DOA) estimation is based on the steered response power (SRP) source localization technique that is computed from the phase transform (PHAT) (DiBiase et al., 2001).

The enhanced audio signal is forwarded to the acoustic feature extraction that analyzes temporal amplitude fluctuations within spectral sub-bands by employing the amplitude modulation filter bank (AMFB) (Moritz et al. 2015a). The development of the AMFB is inspired by two major observations from psychoacoustics and psychophysics, respectively: (1) Amplitude modulation (AM) frequencies ranging from approx. 2 to 16 Hz (Kollmeier and Koch, 1994; Shannon et al., 1995) and the modulation phase (Greenberg and Arai, 2001; Moritz et al., 2015b) are essential acoustic cues used by the human auditory system to recognize speech; (2) The processing of AM frequencies in the inferior colliculus (IC), a neuronal structure of the auditory midbrain with approximately orthogonal tonotopic and periodotopic organization (Langner et al., 1997), shows similarities with a filter bank operating on sub-band temporal envelopes. The assignment of AMFB center frequency (CF) and bandwidth (BW) settings is inspired by the design criteria of auditory filters, e.g., the Mel filter bank that employs a scale to determine the bandwidth as a function of the acoustic frequency. In Moritz et al. (2016), different polynomial functions are evaluated in terms of word error rates (WERs) to find a scale that describes the AM frequency to bandwidth characteristics best. The thus obtained optimal AMFB parameters are also applied in the present study. We show that the AMFB with its preselected and fixed temporal patterns for the extraction of AM information outperforms a comparable data-driven temporal processing method, i.e., the frame splicing technique in combination with a deep neural network (DNN). In addition, acoustic features of the presented ASR system are speaker adapted using the feature-space maximum likelihood linear regression (fMLLR) approach (Li et al., 2002; Gales, 1998).

The detection of erroneous microphone channels is attained by cross-correlating signal representations of each microphone to identify outliers and failures, respectively. The computation of signal representations employs the AMFB to separate speech and noise information in the modulation frequency domain in order to reduce the influence of noise for the failure detection.

2. System components

The main parts of the proposed ASR system, which is applied to the 3rd CHiME speech separation and recognition challenge data (Barker et al., 2015), are the microphone channel failure detection, the NMF-based speech