Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend☆

Takaaki Hori*,[a], Zhuo Chen[a,b], Hakan Erdogan[a,c], John R. Hershey[a], Jonathan Le Roux[a], Vikramjit Mitra[d], Shinji Watanabe[a]

[a] *Mitsubishi Electric Research Laboratories, Cambridge, MA, USA*
[b] *Columbia University, New York, NY, USA*
[c] *Sabanci University, Istanbul, Turkey*
[d] *SRI International, Menlo Park, CA, USA*

Received 11 April 2016; received in revised form 8 December 2016; accepted 19 January 2017

Abstract

This paper gives an in-depth presentation of the multi-microphone speech recognition system we submitted to the 3rd CHiME speech separation and recognition challenge (CHiME-3) and its extension. The proposed system takes advantage of recurrent neural networks (RNNs) throughout the model from the front-end speech enhancement to the language modeling. Three different types of beamforming are used to combine multi-microphone signals to obtain a single higher-quality signal. The beamformed signal is further processed by a single-channel long short-term memory (LSTM) enhancement network, which is used to extract stacked mel-frequency cepstral coefficients (MFCC) features. In addition, the beamformed signal is processed by two proposed noise-robust feature extraction methods. All features are used for decoding in speech recognition systems with deep neural network (DNN) based acoustic models and large-scale RNN language models to achieve high recognition accuracy in noisy environments. Our training methodology includes multi-channel noisy data training and speaker adaptive training, whereas at test time model combination is used to improve generalization. Results on the CHiME-3 benchmark show that the full set of techniques substantially reduced the word error rate (WER). Combining hypotheses from different beamforming and robust-feature systems ultimately achieved 5.05% WER for the real-test data, an 84.7% reduction relative to the baseline of 32.99% WER and a 44.5% reduction from our official CHiME-3 challenge result of 9.1% WER. Furthermore, this final result is better than the best result (5.8% WER) reported in the CHiME-3 challenge.

Keywords: CHiME-3; Robust speech recognition; Beamforming; Noise robust feature; System combination,

1. Introduction

With the wide-spread availability of portable devices equipped with automatic speech recognition (ASR), there is increasing demand for accurate ASR in noisy environments. Although great strides have been made in the

☆ This paper has been recommended for acceptance by Roger Moore
* Corresponding author.
E-mail address: thori@merl.com (T. Hori).

http://dx.doi.org/10.1016/j.csl.2017.01.013

advancement of recognition accuracy, background noise and reverberation continue to pose problems even for the best systems. The presence of highly non-stationary noise is typical of public areas such as cafés, streets, or airports, and tends to significantly degrade recognition accuracy in such situations. Such noises can be challenging to model and estimate due to their diverse and unpredictable spectral characteristics. Therefore, robust speech recognition in noisy environments has attracted increasing attention in ASR research and development.

Several challenge-based workshops focusing on related tasks have been recently held (Barker et al., 2013; Vincent et al., 2013; Kinoshita et al., 2013) to provide common data and benchmarks suitable for comparing and contrasting the performance of different methods. The 3rd CHiME speech separation and recognition challenge (CHiME-3) (Barker et al., 2015) is a new challenge task, which was designed around the well-studied Wall Street Journal corpus. In contrast with the previous CHiME challenges (Barker et al., 2013; Vincent et al., 2013), the CHiME-3 scenario focuses on typical use cases of portable devices. It features speakers talking in challenging noisy environments (cafés, street junctions, public transports and pedestrian areas), recorded using a 6-channel tablet-mounted microphone array.

The CHiME-3 challenge has successfully finished in December 2015. 26 systems were submitted and various strategies that improved the recognition accuracy were proposed and discussed (Barker et al., 2015). We built the MERL/SRI system for CHiME-3 and achieved the 2nd best result among the 26 systems (Hori et al., 2015). The goal of the study was to create an advanced system by determining the best combination of the leading methods on development data and testing their generalization to the evaluation data. Although our system achieved a good level of performance for the challenge task, it was not yet complete as regards to exploiting all the component technologies in the best combination. In this paper, we further extend our previous work and present the complete system and new evaluation results, eventually achieving a better word error rate to that of the best system in the CHiME-3 challenge.

A noteworthy aspect of our system is the pervasive use of deep neural networks (DNNs) and recurrent neural networks (RNNs) at multiple levels throughout the system: the front-end speech enhancement based on long short-term memory (LSTM) RNNs, DNNs for acoustic modeling, and LSTM RNNs for language modeling. Furthermore, we apply noisy data training of DNN acoustic models, which improves the recognition accuracy as reported in Seltzer et al. (2013), Narayanan and Wang (2014) and Delcroix et al. (2015).

For the CHiME-3 task, our system relies on the following key technologies: (1) beamforming to enhance the target speech from the multi-channel signals; (2) noise-robust feature extraction, either directly from the beamformed signal, or from the output of LSTM-based single-channel speech enhancement after beamforming; (3) DNN acoustic models, and large-scale LSTM RNN language models; and (4) system combination of different beamforming/robust-feature systems. Through a series of experiments with different combinations of these techniques, we investigate the relative contributions of the methods, and show that in combination they are surprisingly effective for the CHiME-3 task, ultimately achieving 5.05% WER for the real-test data, an 81.8% reduction relative to the baseline of 32.99% WER and a 44.5% reduction from our official CHiME-3 challenge result of 9.1% WER.

2. Proposed system

2.1. System overview

Fig. 1 describes our proposed system, which is separated into training and recognition stages. In the training stage, 6-channel microphone array signals $\{y_1, \cdots, y_6\}$ are processed independently by three feature extraction modules for mel-frequency cepstral coefficients (MFCCs), damped oscillator coefficient cepstrum (DOCC) and modulation of medium duration speech amplitudes (MMeDuSA), and converted to feature vector sequences $\{\mathbf{x}_1, \cdots, \mathbf{x}_6\}$, $\{\mathbf{x}_1^D, \cdots, \mathbf{x}_6^D\}$ and $\{\mathbf{x}_1^M, \cdots, \mathbf{x}_6^M\}$, respectively, where DOCC and MMeDuSA are noise robust features described in Section 2.4. After that, a DNN acoustic model for each feature extraction method is created by cross-entropy (CE) training followed by state-level Minimum Bayes Risk (sMBR) training. In the training phase, we do not use any speech enhancement techniques based on microphone arrays, and simply deal with the 6-channel signals independently to obtain a larger data set which is 6 times larger than the actually spoken data. The advantage of this architecture is demonstrated in Section 3.

In the recognition stage, we use three types of beamforming, a weighted delay-and-sum (WDAS) beamformer, a minimum variance distortionless response (MVDR) beamformer and a generalized eigenvector (GEV) beamformer to extract enhanced signals $\hat{y}$, $\hat{y}'$ and $\hat{y}''$ from 6-channel microphone array signals $\{y_1, \cdots, y_6\}$, as described in

Please cite this article as: T. Hori et al., Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.01.013