Multi-microphone speech recognition in everyday environments

Jon Barker, Ricard Marxer, Emmanuel Vincent, Shinji Watanabe

Preface

Special issue on multi-microphone speech recognition in everyday environments

---

Multi-microphone signal processing techniques have the potential to greatly improve the robustness of speech recognition (ASR) in distant microphone settings. However, in everyday environments, typified by complex non-stationary noise backgrounds, designing effective multi-microphone speech recognition systems is non trivial. In particular, optimal performance requires the tight integration of the front-end signal processing and the back-end statistical speech and noise source modelling. The best way to achieve this in a modern deep learning speech recognition framework remains unclear. Further, variability in microphone array design — and consequent lack of real training data for any particular configuration — may mean that systems have to be able to generalise from audio captured using mismatched microphone geometries or produced using simulation.

This special issue presents 14 papers focused on multi-microphone speech recognition. The backbone is formed by the CHiME-3 Speech Separation and Recognition Challenge and the ensuing special session held at the 2015 IEEE Automatic Speech Recognition and Understanding Workshop. The first paper by Barker et al. presents the design and outcomes of the challenge, which featured speech recorded in real noisy environments using a 6-channel tablet based microphone array. By comparing the results across recording sessions, utterances, and the 26 submitted systems, the most successful techniques and the signal properties which correlate most with the word error rate (WER) are identified. The second paper by Vincent et al. provides a complementary analysis of the mismatches between training and test data. With one notable exception, environment, microphone, and data simulation mismatches are shown to have a minor impact on the WER. The authors then introduce the CHiME-4 Challenge, which revisits CHiME-3 by reducing the number of microphones available for testing.

The following four papers focus on enhancing the recorded signals using front-end processing techniques. Moore et al. evaluate established techniques, namely spectral subtraction, delay-and-sum (DS) beamforming, weighted prediction error (WPE) based inverse filtering, and combinations of these techniques. Heymann et al. introduce a new approach to beamforming based on esti-

mating a spectral mask using a deep neural network (DNN) and deriving a generalised eigenvalue beamformer. This approach does not require knowledge of the microphone array geometry. Barfuss et al. propose to apply a post-filter based on the spatial coherence of the multichannel signal to the beamformer output to further reduce diffuse noise. Various estimators of the coherent-to-diffuse ratio, which may or not depend on the estimated direction of arrival of the speaker, are assessed in this context. Cho et al. build a feature-domain enhancement system by bringing together independent vector analysis and a model of reverberation to estimate the log-power spectrum of clean speech.

Besides enhancing the recorded signals, another approach to robust ASR consists of adapting the ASR back-end to the test conditions. Falavigna et al. adapt the DNN acoustic model by gradient descent and control the deviation from the original unadapted model by using the Kullback-Leibler divergence as a regularization term. They obtain significant WER improvement by using a system to automatically predict the quality of ASR hypotheses to select the best sentences for adaptation. Lin et al. train the DNN acoustic model on a multi-condition training set, which consists of the original noisy training data augmented with data generated using a denoising autoencoder.

Several papers describe complete robust ASR systems that exploit both front-end enhancement and back-end adaptation. Sivasankaran et al. conduct a series of experiments to assess the combined WER impact of dereverberation based on DS, WPE or DNN and acoustic model adaptation based on learning hidden unit contributions or i-vectors in an extensive range of reverberation conditions. Hori et al. combine DS and minimum variance distortionless response (MVDR) beamforming, recurrent neural network (RNN) based single-channel post-filtering, robust feature extraction by damped oscillator coefficients or modulation of medium duration speech amplitudes, and RNN language modeling. Moritz et al. combine MVDR beamforming or multichannel nonnegative matrix factorization (NMF) with amplitude modulation filter bank features, also together with RNN language modeling. Tu et al. present a principled approach to designing a DNN acoustic model that combines early fusion of multiple enhanced speech features, speaker-related features, and other auxiliary features concatenated as the inputs to various subnets and late fusion of different subnet outputs.

The last two papers address practical considerations arising in real application scenarios. Takeda et al. propose three complementary methods to reduce the computational footprint of DNN acoustic models by adaptively quantizing the DNN parameters, reducing memory usage within the CPU cache, and adaptively pruning neurons. Finally, Rodomagoulakis et al. present a full pipeline for always-listening, far-field voice command in smart homes equipped with multiple microphone arrays, which consists of room-dependent speech activity detection, wake-up word detection, multi-condition acoustic model training,

channel adaptation, and fusion of single-channel ASR outputs.

Jon Barker
j.barker@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

Ricard Marxer
r.marxer@dcs.shef.ac.uk
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

Emmanuel Vincent
emmanuel.vincent@inria.fr
Inria, F-54600 Villers-lès-Nancy, France

Shinji Watanabe
watanabe@merl.com
Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA