Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language 00 (2016) 1–23
www.elsevier.com/locate/csl

An analysis of environment, microphone and data simulation mismatches in robust speech recognition

Emmanuel Vincent[a,*], Shinji Watanabe[b], Aditya Arie Nugraha[a], Jon Barker[c], Ricard Marxer[c]

[a] *Inria, 54600 Villers-lès-Nancy, France*
[b] *Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA*
[c] *Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK*

Received 25 April 2016; received in revised form 12 October 2016; accepted 18 November 2016

Abstract

Speech enhancement and automatic speech recognition (ASR) are most often evaluated in matched (or multi-condition) settings where the acoustic conditions of the training data match (or cover) those of the test data. Few studies have systematically assessed the impact of acoustic mismatches between training and test data, especially concerning recent speech enhancement and state-of-the-art ASR techniques. In this article, we study this issue in the context of the CHiME-3 dataset, which consists of sentences spoken by talkers situated in challenging noisy environments recorded using a 6-channel tablet based microphone array. We provide a critical analysis of the results published on this dataset for various signal enhancement, feature extraction, and ASR backend techniques and perform a number of new experiments in order to separately assess the impact of different noise environments, different numbers and positions of microphones, or simulated vs. real data on speech enhancement and ASR performance. We show that, with the exception of minimum variance distortionless response (MVDR) beamforming, most algorithms perform consistently on real and simulated data and can benefit from training on simulated data. We also find that training on different noise environments and different microphones barely affects the ASR performance, especially when several environments are present in the training data: only the number of microphones has a significant impact. Based on these results, we introduce the CHiME-4 Speech Separation and Recognition Challenge, which revisits the CHiME-3 dataset and makes it more challenging by reducing the number of microphones available for testing.

Keywords: Robust ASR; Speech enhancement; Train/test mismatch; Microphone array

1. Introduction

Speech enhancement and automatic speech recognition (ASR) in the presence of reverberation and nonstationary noise are still challenging tasks today (Baker et al., 2009; Wölfel and McDonough, 2009; Virtanen et al., 2012; Li et al., 2015). Research in this field has made great progress thanks to real speech corpora collected for various application scenarios such as voice command for cars (Hansen et al., 2001), smart homes (Ravanelli et al., 2015), or

* Corresponding author.
E-mail address: emmanuel.vincent@inria.fr (E. Vincent).

http://dx.doi.org/10.1016/j.csl.2016.11.005

tablets (Barker et al., 2015), and automatic transcription of lectures (Lamel et al., 1994), meetings (Renals et al., 2008), conversations (Harper, 2015), dialogues (Stupakov et al., 2011), game sessions (Fox et al., 2013), or broadcast media (Bell et al., 2015). In most corpora, the training speakers differ from the test speakers. This is widely recognized as good practice and many solutions are available to improve robustness to this mismatch (Gales, 1998; Shinoda, 2011; Karafiát et al., 2011; Swietojanski and Renals, 2014). By contrast, the acoustic conditions of the training data often match (or cover) those of the test data. While this allows for significant performance improvement by multi-condition training, one may wonder how the reported performance would generalize to mismatched acoustic conditions. This question is of tantamount importance for the deployment of robust speech processing technology in new environments. In that situation, the test data may differ from the training data in terms of reverberation time (RT60), direct-to-reverberant ratio (DRR), signal-to-noise ratio (SNR), or noise characteristics. In a multichannel setting, the number of microphones, their spatial positions and their frequency response also matter.

Regarding multichannel speech enhancement, the impact of the number of microphones and the microphone distance on the enhancement performance has been largely studied in the microphone array literature (Cohen et al., 2010). The impact of imprecise knowledge of the microphone positions and frequency responses has also been addressed (Cox et al., 1987; Doclo and Moonen, 2007; Anderson et al., 2015). For traditional speech enhancement techniques, which require either no training or training on the noise context preceding each test utterance (Cohen et al., 2010; Hurmalainen et al., 2013), the issue of mismatched noise conditions did not arise. This recently became a concern with the emergence of speech enhancement techniques based on deep neural networks (DNNs) (Wang et al., 2014; Xu et al., 2014; Weninger et al., 2015), which require a larger amount of training data not limited to the immediate context. Chen et al. (2015) and Kim and Smaragdis (2015) considered the problem of adapting DNN based enhancement to unseen test conditions, but their experiments were conducted on small, simulated datasets and evaluated in terms of enhancement metrics.

Regarding ASR, the variation of the word error rate (WER) as a function of the SNR was studied in several evaluation challenges, e.g., Hirsch and Pearce (2000) and Barker et al. (2013). The adaptation of DNN acoustic models to specific acoustic conditions has been investigated, e.g., Seltzer et al. (2013) and Karanasou et al. (2014), however it has been evaluated in multi-condition settings rather than actual mismatched conditions. The impact of the number of microphones on the WER obtained after enhancing reverberated speech was evaluated in the REVERB challenge (Kinoshita et al., 2013), but the impact of microphone distance was not considered and no such large-scale experiment was performed with noisy speech. To our knowledge, a study of the impact of mismatched noise environments on the resulting ASR performance is also missing.

Besides mismatches of reverberation and noise characteristics, the mismatch between real and simulated data is also of timely interest. In the era of DNNs, there is an incentive for augmenting the available real training data by perturbing these data or simulating additional training data with similar acoustic characteristics. Simulation might also allow for rough assessment of a given technique in a new environment before real data collected in that environment become available. Suspicion about simulated data is common in the speech processing community, due for instance to the misleadingly high performance of direction-of-arrival based adaptive beamformers on simulated data compared to real data (Kumatani et al., 2012). Fortunately, this case against simulation does not arise for all techniques: most modern enhancement and ASR techniques can benefit from data augmentation and simulation (Kanda et al., 2013; Brutti and Matassoni, 2016). Few existing datasets involve both real and simulated data. In the REVERB dataset (Kinoshita et al., 2013), the speaker distances for real and simulated data differ, which does not allow fair comparison. The CHiME-3 dataset (Barker et al., 2015) provides a data simulation tool which aims to reproduce the characteristics of real data for training and twinned real and simulated data pairs for development and testing. This makes it possible to evaluate the improvement brought by training on simulated data in addition to real data and to compare the performance on simulated vs. real test data for various techniques.

In this article, we study the above mismatches in the context of the CHiME-3 dataset. Our analysis differs from the one of Barker et al. (2016), which focuses on the speaker characteristics and the noise characteristics of each environment and compares the achieved ASR performance with the intelligibility predicted using perceptual models. Instead, we focus on mismatched noise environments, different microphones, and simulated vs. real data. We provide a critical analysis of the CHiME-3 results in that light and perform a number of new experiments in order to separately assess the impact of these mismatches on speech enhancement and ASR performance. Based on these results, we conclude that, except for a few techniques, these mismatches generally have little impact on the ASR performance compared to, e.g., reducing the number of microphones. We introduce the CHiME-4 Speech Separation and

Please cite this article as: E. Vincent et al., An analysis of environment, microphone and data simulation mismatches in robust speech recognition, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.11.005