Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language xxx (2017) xxx-xxx

www.elsevier.com/locate/csl

# A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions[☆]

Q1 Sunit Sivasankaran[*,a,b,c], Emmanuel Vincent[a,b,c], Irina Illina[a,b,c]

[a] *Inria, Villers-lès-Nancy, F-54600, France*
[b] *Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*
[c] *CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*

Received 18 April 2016; received in revised form 3 February 2017; accepted 7 February 2017

**Abstract**

Robustness to reverberation is a key concern for distant-microphone ASR. Various approaches have been proposed, including single-channel or multichannel dereverberation, robust feature extraction, alternative acoustic models, and acoustic model adaptation. However, to the best of our knowledge, a detailed study of these techniques in varied reverberation conditions is still missing in the literature. In this paper, we conduct a series of experiments to assess the impact of various dereverberation and acoustic model adaptation approaches on the ASR performance in the range of reverberation conditions found in real domestic environments. We consider both established approaches such as WPE and newer approaches such as learning hidden unit contribution (LHUC) adaptations, whose performance has not been reported before in this context, and we employ them in combination. Our results indicate that performing weighted prediction error (WPE) dereverberation on a reverberated test speech utterance and decoding using a deep neural network (DNN) acoustic model trained with multi-condition reverberated speech with feature-space maximum likelihood linear regression (fMLLR) transformed features, outperforms more recent approaches and helps significantly reduce the word error rate (WER).

© 2017 Elsevier Ltd. All rights reserved.

*Keywords:* Robust ASR; Dereverberation; Acoustic model adaptation; Evaluation

## 1. Introduction

Many automatic speech recognition systems (ASR) such as the Amazon Echo do not require the user to stay close to the microphone while interacting with the device. This raises two main problems, namely noise and reverberation. The performance of ASR systems is known to degrade under these external influences (Brutti and Matassoni, 2016; Petrick et al., 2014; Wölfel and McDonough, 2009; Yoshioka et al., 2012).

Various methods have been proposed to tackle ASR under noisy conditions (Li et al., 2013). Evaluation challenges have also been conducted on this specific issue (Barker et al., 2015). The problem can be addressed at different stages of ASR. At the signal level, single-channel and multichannel speech enhancement techniques have been applied (Virtanen et al., 2012; Sivasankaran et al., 2015). At the feature level, cepstral mean and variance

UNCORRECTED PROOF

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

Q2 [*] Corresponding author at: Inria, Villers-lès-Nancy, F-54600, France.
*E-mail address:* sunit.sivasankaran@inria.fr (S. Sivasankaran).

http://dx.doi.org/10.1016/j.csl.2017.02.003
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

ARTICLE IN PRESS

normalization (CMVN) (Viikki et al., 1998), histogram equalization (Molau et al., 2003) and subband histogram normalization (Joshi et al., 2015) techniques have been employed. At the model level, multi-condition training and model adaptation techniques have been used. Another approach is to model the uncertainties created by the noise and compensate them across the ASR system in order to achieve noise robustness (Abdelaziz et al., 2015). An overview of noise robust techniques for ASR is described in Li et al. (2013).

Similarly, ASR for reverberated speech has previously been studied (Kinoshita et al., 2013; Brutti and Matassoni, 2016; Yoshioka et al., 2012; Brutti and Matassoni, 2014). Room acoustics induces three components, namely: direct sound, early reflections and late reverberation. The boundary between early reflections and late reverberation is generally considered to be around 50 ms after the direct signal. The energy ratio of direct sound and early reflections with respect to the late reverberation is measured by the clarity index, $C_{50}$ (Kuttruff, 2009; Nishiura et al., 2007). This quantity is also sometimes referred to as the early-to-late ratio (ELR) and it is known to influence the ASR performance (Wölfel and McDonough, 2009). Another parameter which is commonly associated with reverberant signals is the reverberation time (RT) which is defined as the time taken for the audio signal to decay by 60 dB. Early reflections can be considered as a convolution of the speech signal with a stationary channel response, which can be handled by CMVN, and they have been reported to improve ASR performance under certain conditions (Petrick et al., 2014; Brutti and Matassoni, 2016). Late reverberation, by contrast, is considered to be uncorrelated to the speech signal and it contributes most to the degradation of the ASR performance.

Reverberation has been tackled across multiple fronts in ASR systems. Single-channel and multichannel dereverberation techniques have been investigated (Naylor and Gaubitch, 2010). Features which are more robust to reverberation have been proposed (Hermansky and Morgan, 1994; Kingsbury et al., 1998; Falk and Chan, 2010; Ganapathy et al., 2011). Time delay neural networks (Peddinti et al., 2015b; 2015a) and polyphone (instead of triphone) acoustic models (Yoshioka et al., 2012), which take into account the longer term characteristics of speech, have shown promising results at the expense of a larger computational cost. I-vector based model adaptation (Peddinti et al., 2015b) and ROVER based system combination (Fiscus, 1997) have also been proposed as solutions. Evaluation campaigns such as REVERB (Kinoshita et al., 2013) and ASpIRE (Harper, 2015) have recently contributed to popularizing this research topic.

To the best of our knowledge there are only three studies in the literature which attempt to evaluate in detail the performance of various techniques for reverberant ASR. Brutti and Matassoni (2016) have analyzed the effect of reverberation on ASR and correlated its performance with the characteristics of room impulse responses (RIR). The RIRs used were synthetically generated and the results were validated using real RIRs. They showed the dependence of ASR accuracy on the features derived from the structure of early reflections and late reverberation. However, the experiments were conducted using multi-condition training only: no dereverberation or acoustic model adaptation was applied. Another work of interest is the summary of the REVERB challenge by Kinoshita et al. (2016). Many techniques were tried as part of this challenge. However, the test material covered a small range of ELR and RT conditions, which questions the validity of the results in the larger range of conditions found in real environments. Delcroix et al. (2015) detailed different techniques to counter reverberation using the REVERB dataset such as dereverberation using weighted prediction error (WPE) and minimum variance distortionless response (MVDR) beamforming. The effect on the performance using unsupervised adaptation and increased dataset size was also studied. This detailed work however lacks a comparison of the older WPE and MVDR methods to newer DNN based dereverberation and adaptation methods. Performance comparison using different features such as fMLLR are also missing. FMLLR features have however been used in Peddinti et al. (2015b) to obtain state of the art results as part of the ASpIRE challenge. Nevertheless a comparison of techniques and their combinations are missing in the literature.

In this work, we build upon these studies. We conduct a series of experiments to assess the impact of various dereverberation and acoustic model adaptation approaches on the ASR performance. We evaluate the validity of these approaches in varied ELR and RT reverberation conditions on data simulated from real RIRs. The best performing methods on the simulated data are then applied on the real data. We consider both established approaches and newer approaches, whose performance across different reverberation conditions has not been reported before, and we employ them in combination. As a side contribution, we evaluate the performance of DNN based dereverberation using fMLLR features, which hasnt been reported earlier to the best of our knowledge.

The rest of the paper is organized as follows. Section 2 details the dataset creation and sets up the baseline for our experiments. Section 3 explains how clean alignments for training are obtained and shows the relevant results. Section 5 details the different dereverberation techniques tried and Section 6 explains the model adaptation techniques.

UNCORRECTED PROOF

Please cite this article as: S. Sivasankaran et al., A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.003

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات