Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 00 (2016) 1–11

www.elsevier.com/locate/csl

# Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures

A.H. Moore*, P. Peso Parada[1], P.A. Naylor

*Department of Electrical and Electronic Engineering, Imperial College London, Exhibition Road, London, United Kingdom*

Received 8 April 2016; received in revised form 28 September 2016; accepted 25 November 2016

**Abstract**

Automatic speech recognition in everyday environments must be robust to significant levels of reverberation and noise. One strategy to achieve such robustness is multi-microphone speech enhancement. In this study, we present results of an evaluation of different speech enhancement pipelines using a state-of-the-art ASR system for a wide range of reverberation and noise conditions. The evaluation exploits the recently released ACE Challenge database which includes measured multichannel acoustic impulse responses from 7 different rooms with reverberation times ranging from 0.33 to 1.34 s. The reverberant speech is mixed with ambient, fan and babble noise recordings made with the same microphone setups in each of the rooms. In the first experiment, performance of the ASR without speech processing is evaluated. Results clearly indicate the deleterious effect of both noise and reverberation. In the second experiment, different speech enhancement pipelines are evaluated with relative word error rate reductions of up to 82%. Finally, the ability of selected instrumental metrics to predict ASR performance improvement is assessed. The best performing metric, Short-Time Objective Intelligibility Measure, is shown to have a Pearson correlation coefficient of 0.79, suggesting that it is a useful predictor of algorithm performance in these tests.

*Keywords:* Automatic speech recognition; Speech enhancement; Dereverberation; Beamforming; Realistic environments; Microphone array signal processing

## 1. Introduction

Real-world applications of Automatic Speech Recognition (ASR), such as meeting transcription and human-robot interaction, demand that the speaker be some distance from the sound capture device. This leads to degradation of the desired signal by both additive noise and reverberation, both of which have a significant degrading effect on ASR accuracy. It has been proposed that robustness in distant-talking ASR be achieved through three approaches, namely enhancement of the audio signal, front-end based approaches which enhance the signal in the feature domain and back-end methods (Haeb-Umbach and Krueger, 2012). Recent challenges such as the REVERB challenge (Kinoshita et al., 2013) and CHiME3 (Barker et al., 2015) have demonstrated the effectiveness of all three approaches.

* Corresponding author.
*E-mail address:* alastair.h.moore@imperial.ac.uk (A.H. Moore).

[1] Present address: Cirrus Logic, Marble Arch House, 66 Seymour St., 1st Floor, London W1H 5BT, United Kingdom.

http://dx.doi.org/10.1016/j.csl.2016.11.003

Our work focusses on speech enhancement algorithms and particularly on the case when the ASR system itself cannot be altered, or it is impractical to alter it. This is relevant in many real-world applications where an ASR engine is supplied by a third-party, either as embedded code or as a cloud-based system. It is also relevant during algorithm development where many alternative parameters or algorithm variations are to be assessed. It is, in general, not practical to process the training data with every algorithm and retrain the ASR system for each. This is particularly true when the range of acoustic conditions that need to be considered is wide, as in everyday environments where a broad range of Reverberation Times (RTs), Direct-to-Reverberant Ratios (DRRs), noise types and Signal-to-Noise Ratios (SNRs) may be encountered.

Speech enhancement for human listening has a long history with a great many algorithms being proposed and refined over the decades which attempt to mitigate both additive and convolutive noise (Naylor and Gaubitch, 2010). By exploiting multiple microphones, the most successful algorithms can improve the perceived audio quality and/or intelligibility. Conducting listening tests using human listeners is time consuming and expensive and so instrumental measures have been developed which attempt to model human responses in order to predict the performance of new algorithms (Taal et al., 2011; ITU-T, 2001).

The ultimate performance measure of speech enhancement algorithms for ASR will always be the Word Error Rate (WER) achieved over a particular set of test data. However, we propose that just as instrumental measures can be useful for modelling human listening, they can also be helpful for evaluating speech enhancement for ASR. To test this assertion, the current study is split into three parts. The first establishes the performance bounds of an “off-the-shelf” recogniser for a broad range of acoustic conditions. The second determines the performance improvements obtained by a range of speech enhancement pipelines for a representative subset of acoustic conditions. The third compares the performance measured using WER and relative WER reduction (rWERR) to that predicted by instrumental metrics.

To summarise, the novel contributions of the current study are (i) results for the CHiME3 (Barker et al., 2015) baseline system under more diverse acoustic conditions than have hitherto been considered; (ii) independent verification of the efficacy of linear prediction-based dereverberation, particularly when combined with beamforming (Yoshioka and Nakatani, 2012; Delcroix et al., 2014, 2015); (iii) evaluation of multi-channel speech enhancement for ASR under reverberant conditions with high levels of noise, and especially babble noise; (iv) a comparison showing strong correlation between ASR performance and instrumental metrics. It is hoped that the results of (i) and (iii), which make use of the Acoustic Characterization of Environments (ACE) Challenge database of Acoustic Impulse Responses (AIRs) and noise, will serve as a baseline for future robust ASR systems and speech enhancement algorithms.

In Section 2, we describe the setup of the experimental framework and then evaluate the ASR system for 210 different acoustic conditions with unprocessed audio. In Section 3, we select a subset of 54 acoustic conditions, each of which is enhanced using 6 different enhancement strategies consisting of different combinations of noise reduction, dereverberation and beamforming. In Section 4, the results from Section 3 are compared to the results of instrumental metrics. The significance of the results is discussed in Section 5 and our conclusions presented in Section 6.

## 2. Experiment 1: effect of acoustic conditions without speech enhancement

The aim of the first experiment is to characterise the performance of a standard state-of-the-art ASR system for a broad range of acoustic conditions using only the received signal. In this way, the effect of reverberation, noise type and SNR can be individually analysed. The results of this experiment also inform the selection of a reduced subset of acoustic conditions for which speech enhancement is used in the second experiment.

### 2.1. ASR system

The ASR employed in this work corresponds to the updated Kaldi CHiME3 baseline recipe available in the Git Kaldi repository.[2] This system is an improvement of the original baseline which includes a feature-space maximum likelihood linear regression transformation on the DNN features and output hypothesis rescoring with a Kneser−Ney smoothed 5-gram model and an RNN language model (Hori et al., 2015). By choosing a system that can be freely downloaded and used, new algorithms can be objectively compared to the baselines presented here.

[2] https://github.com/kaldi-asr/kaldi/tree/master/egs/chime3/s5 commit id: 7e44bc74c3388cd134485672fd1c9687c68073b9.

Please cite this article as: A.H. Moore et al., Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures, Computer Speech & Language (2016), http://dx.doi.org/10.1016/j.csl.2016.11.003