Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 50 (2008) 797–809

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Combined speech enhancement and auditory modelling for robust distributed speech recognition

Ronan Flynn [a,*], Edward Jones [b]

[a] Department of Electronic Engineering, Athlone Institute of Technology, Ireland
[b] Department of Electronic Engineering, National University of Ireland, Galway, Ireland

Received 3 July 2007; received in revised form 6 May 2008; accepted 13 May 2008

Abstract

The performance of automatic speech recognition (ASR) systems in the presence of noise is an area that has attracted a lot of research interest. Additive noise from interfering noise sources, and convolutional noise arising from transmission channel characteristics both contribute to a degradation of performance in ASR systems. This paper addresses the problem of robustness of speech recognition systems in the first of these conditions, namely additive noise. In particular, the paper examines the use of the auditory model of Li et al. [Li, Q., Soong, F.K., Siohan, O., 2000. A high-performance auditory feature for robust speech recognition. In: Proc. 6th Internat. Conf. on Spoken Language Processing (ICSLP), Vol. III. pp. 51–54] as a front-end for a HMM-based speech recognition system. The choice of this particular auditory model is motivated by the results of a previous study by Flynn and Jones [Flynn, R., Jones, E., 2006. A comparative study of auditory-based front-ends for robust speech recognition using the Aurora 2 database. In: Proc. IET Irish Signals and Systems Conf., Dublin, Ireland. pp. 111–116] in which this auditory model was found to exhibit superior performance for the task of robust speech recognition using the Aurora 2 database [Hirsch, H.G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA ITRW ASR2000, Paris, France. pp. 181–188]. In the speech recognition system described here, the input speech is pre-processed using an algorithm for speech enhancement. A number of different methods for the enhancement of speech, combined with the auditory front-end of Li et al., are evaluated for the purpose of robust connected digit recognition. The ETSI basic [ETSI ES 201 108 Ver. 1.1.3, 2003. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms] and advanced [ETSI ES 202 050 Ver. 1.1.5, 2007. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms] front-ends proposed for DSR are used as a baseline for comparison. In addition to their effects on speech recognition performance, the speech enhancement algorithms are also assessed using perceptual speech quality tests, in order to examine if a correlation exists between perceived speech quality and recognition performance. Results indicate that the combination of speech enhancement pre-processing and the auditory model front-end provides an improvement in recognition performance in noisy conditions over the ETSI front-ends.

Keywords: Speech enhancement; Auditory front-end; Robust speech recognition

1. Introduction

The front-end processor in Automatic Speech Recognition (ASR) systems converts the incoming speech signal into a format that is later used in a classification stage. The front-end extracts a feature from the speech signal that ideally should be independent of the speaker (for speaker independent recognition tasks) and background noise, and distortion introduced by the transmission channel. It is well-known that the presence of noise severely degrades the performance of speech recognition systems, and much research has been devoted to the development

[*] Corresponding author.
E-mail addresses: rflynn@ait.ie (R. Flynn), edward.jones@nuigalway.ie (E. Jones).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.05.004

of techniques to alleviate this effect. One aspect of the robustness of an ASR system is its ability to maintain its recognition accuracy under conditions that are different from the original training conditions. One common approach to improving system performance in noise is to use front-ends that produce robust features. Some of these approaches involve modifications of well established techniques, such as cepstral mean subtraction. Other approaches involve the use of auditory-based front-ends in order to improve robustness (e.g. Ghitza, 1988; Seneff, 1988; Dau et al., 1996).

Another method that has been proposed to improve the robustness of ASR systems is to enhance the speech signal before feature extraction. Enhancement of noisy speech signals is designed to improve the perception of the speech by human listeners or to improve the processing of the speech by ASR systems. It may also have benefits in enhancing robustness in ASR systems. Speech enhancement can be particularly useful in cases where a significant mismatch exists between training and testing conditions, such as where a recognition system is trained with clean speech and then used in noisy conditions. Inclusion of speech enhancement can help to reduce the mismatch.

The enhancement of noisy speech can be described as an estimation problem in which the original clean signal is estimated from a degraded version of the signal. A significant amount of research has been carried out on speech enhancement, and a number of approaches have been well documented in the literature. Ephraim and Cohen (2006) present a survey of a number of approaches to speech enhancement from a single-microphone. Many enhancement techniques are based on the concept of noise spectral estimation coupled with spectral subtraction. The advantage of these methods is a reduction in noise and an improvement in the signal-to-noise ratio. A disadvantage is the introduction of speech distortion and a residual noise called ‘musical noise’.

Two measures that can be used to perceptually evaluate speech are its *quality* and its *intelligibility*. Speech quality is a subjective measure and is dependent on the individual preferences of listeners. It is a measure of how comfortable a listener is when listening to the speech under evaluation. The intelligibility of the speech can be regarded as an objective measure, and is calculated based on the number or percentage of words that can be recognised by listeners. The intelligibility and the quality of speech are not correlated and it is well-known that improving one of the measures can have a detrimental effect on the other one. Speech enhancement algorithms give a trade-off between noise reduction and signal distortion. A reduction in noise can lead to an improvement in the subjective quality of the speech but a decrease in the measured speech intelligibility (Ephraim and Cohen, 2006). The quality and the intelligibility of speech can be evaluated using listening tests. There are however a number of mathematically based tools available that facilitate the evaluation of speech quality and speech intelligibility without the need for listeners. Speech enhancement can have a negative impact on subjective speech intelligibility if the spectral cues and the gross temporal envelope cues in the speech are not adequately preserved by the enhancement algorithm. For example, Hu and Loizou (2007) found that single-microphone speech enhancement algorithms do not improve subjective intelligibility in normal-hearing listeners and that with certain enhancement algorithms the intelligibility was impaired.

When using speech enhancement in an ASR system, the speech is enhanced before feature extraction and recognition processing. The advantage of this is that there is no impact on the computational complexity of the feature extraction or the recognition processes as the enhancement is independent of both. However, every speech enhancement process will introduce some form of signal distortion and it is important that the impact of this distortion on the recognition process is minimised.

Kleinschmidt et al. (2001) combined the model of auditory perception (PEMO) described by Tchorz and Kollmeier (1999), with the noise reduction algorithm proposed by Ephraim and Malah (1984). This noise reduction algorithm is well-known, and has been found to exhibit good performance. Kleinschmidt et al. (2001) compared the performance of this combination with the performance of a front-end based on the standard Mel Frequency Cepstral Coefficient (MFCC) framework, for the task of recognition of an isolated German digit database, and found that the combination of speech enhancement and auditory model resulted in better performance.

This paper extends this paradigm by examining the performance of the auditory model proposed by Li et al. (2000), in combination with a number of different speech enhancement algorithms. Many computational auditory models have been proposed for use in speech recognition systems, often with excellent results, particularly in the presence of noise. In this work, the auditory model of Li et al. (2000) is used. The choice of this auditory front-end is motivated by previous work carried out by Flynn and Jones (2006) where a number of auditory front-ends were investigated in a comparative study of robust speech recognition with the widely-used Aurora 2 database (Hirsch and Pearce, 2000). In that study, there was no pre-processing or enhancement of the speech utterances. The front-ends investigated were Perceptual Linear Prediction (PLP) proposed by Hermansky (1990), the PEMO algorithm proposed by Tchorz and Kollmeier (1999), and the front-end processor proposed by Li et al. (2000). For the task of connected digit recognition using the Aurora 2 database, the front-end proposed by Li et al. gave the best overall recognition results of all the auditory models examined, and with an overall reduction in recognition error compared to the ETSI basic front-end (ETSI ES 201 108 Ver. 1.1.3, 2003) which was used as a baseline for comparison. The ETSI front-ends have been proposed for use in a Distributed Speech Recognition (DSR) paradigm, wherein the front-end would typically be implemented in a mobile handset,