Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 129–142

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A wavelet-based thresholding approach to reconstructing unreliable spectrogram components ☆

Shirin Badiezadegan [1], Richard C. Rose *

Electrical and Computer Engineering Department, McGill University, Montreal, Quebec, Canada

Received 18 September 2013; received in revised form 22 October 2014; accepted 21 November 2014
Available online 8 December 2014

Abstract

Data imputation approaches for robust automatic speech recognition reconstruct noise corrupted spectral information by exploiting prior knowledge of the relationship between target speech and background through the use of spectrographic masks. Most of these approaches are model-based techniques that can only provide accurate estimates of the underlying clean speech when the characteristics of the noise corrupted features do not deviate from those of the model. Discrete wavelet transform (DWT) based de-noising methods can also be used for re-estimating the underlying clean speech from a noise corrupted signal, but often require that the background noise is stationary and modeled by a Gaussian distribution. A novel approach is presented here for incorporating the information derived from spectrographic masks in a DWT-based de-noising method. The spectrographic masks are used for deriving thresholds for de-noising wavelet domain coefficients making DWT based de-noising more suitable for non-stationary noise conditions. The results of an experimental study are presented to demonstrate the performance of DWT based data imputation relative to other established techniques on the Aurora 2 noisy speech recognition task. It will be shown that the proposed approach reduces the impact of model mismatch associated with parametric approaches and exploits the robustness of non-parametric wavelet de-noising approach.

Keywords: Spectrogram reconstruction; Data imputation; Spectrographic mask; Discrete wavelet transform; Thresholding; De-noising

1. Introduction

The goal of data imputation based missing feature approaches is to reconstruct spectral components derived from noise corrupted speech to improve ASR performance. Most existing implementations are model based (Raj, 2000; Raj and Stern, 2005; Raj and Singh, 2005; Badiezadegan and Rose, 2010; Stern and Raj, 1997; Borgstrom and Alwan, 2010; Barker et al., 2000; Barker et al., 2000; El-Maliki and Drygajlo, 2001). These techniques can provide accurate estimates of the features as long as the characteristics of the corrupted features do not deviate significantly from those of the uncorrupted features used for model training. This mismatch, however, in low signal to noise ratios and non-stationary noise backgrounds can degrade the performance of the model-based missing feature techniques.

Wavelet-based thresholding and de-noising techniques have been widely used in many signal and image processing applications and have been shown to perform well when the corrupting noise is stationary. In this case, identifying the clean and noisy features in the wavelet domain is quite straight-forward when a universal threshold estimate is used (Donoho, 1995). The challenge, however, is to identify the threshold value when the corrupting noise is non-stationary and cannot be assumed to have a Gaussian distribution.

☆ This work was sponsored by Google, NSERC, and MDEIE.

* Corresponding author.
E-mail addresses: shirin.badiezadegan@mail.mcgill.ca (S. Badiezadegan), rose@ece.mcgill.ca (R.C. Rose).

[1] Principle corresponding author.

http://dx.doi.org/10.1016/j.specom.2014.11.005
0167-6393/© 2014 Elsevier B.V. All rights reserved.

This paper presents an alternative approach to data imputation on mel log-spectral features which incorporates the spectrographic masks described in Badiezadegan and Rose (2010) to serve as the "oracle" for labeling the wavelet domain noisy (missing/unreliable) and clean (reliable) features. Using the information provided by the spectrographic masks, we can tackle the threshold estimation challenge of the wavelet thresholding techniques when dealing with non-stationary noise, and at the same time exploit the strength of this non-parametric re-estimation approach. The spectrographic masks can provide accurate estimates of the probability of speech presence for mel log-spectral features in stationary and non-stationary noise environments. Employing the information provided by the spectrographic masks in a wavelet-based thresholding technique, not only enables the system to track the variations in the background noise characteristics and to deal with the non-stationary noise types for thresholding the wavelet coefficients, but also makes it possible to identify the noise-dominant approximation features and to perform an additional processing step on the approximation coefficients in the wavelet domain to better suppress the noise.

The method proposed in this paper propagates the feature level probabilities provided by the spectrographic mask through a discrete wavelet transform filter bank to identify the noisy and clean wavelet domain features. This approach generates the "oracle" information needed for wavelet thresholding. Next, the noisy features are processed according to a hard thresholding scheme to recover the underlying clean features. The spectrographic masks are generated using the process presented in Badiezadegan and Rose (2010) which tracks speech presence uncertainty following a procedure originally developed for speech enhancement in Malah et al. (1999). This approach is briefly described as speech presence probability (SPP) estimation in Section 2.2. A brief introduction to the practice of propagating mask information from the spectral domain to the wavelet domain so that noise suppression can be performed in the wavelet domain was given in Badiezadegan and Rose (2011). This paper expands on that discussion by providing a detailed formulation of the mask propagation approach. Furthermore, the sensitivity of this approach to the choice of several free parameters in this approach is investigated as part of the experimental study presented in this paper.

Results of the experimental study presented in Section 4 demonstrate the ability of the DWT based spectrogram reconstruction, coupled with spectrographic mask estimation, to provide improved ASR word accuracy (WAC) in non-stationary noise environments. The rest of the paper is organized as follows: Section 2 provides an introduction to data imputation based missing feature techniques for spectrogram reconstruction. Section 3 introduces the proposed wavelet-based spectrogram reconstruction technique. The experimental study is presented in Section 4 followed by conclusions and future work in Section 5.

2. Background and motivation

This section includes a brief review of missing feature based data imputation, also known as missing feature based spectrogram reconstruction in Section 2.1. One of the most important steps in any spectrogram reconstruction technique is the spectrographic mask estimation, since it is the spectrographic mask which identifies the reliable and unreliable spectrogram components. The SPP masks used throughout this work are presented in Section 2.2. Next, minimum mean squared error (MMSE) based spectrogram reconstruction technique is briefly reviewed in Section 2.3. This is commonly used as a reference in the experimental study presented in Section 4. Finally, in Section 2.4, the basics of the discrete wavelet transform and the DWT-based signal de-noising in its general form are presented.

2.1. Missing feature based spectrogram reconstruction

Acoustic robustness in automatic speech recognition (ASR) implies the ability to maintain a high level of recognition accuracy in the presence of multiple sources of variability, especially when the recognizer has been trained under noise-free conditions (Raj, 2000; Stern and Raj, 1997). There have been a large number of feature compensation and model adaptation approaches that have demonstrated good performance in environments where the main source of acoustical variability is stationary noise with moderate SNR (Cooke et al., 2001). However, most scenarios of general interest, like, for example, ASR services designed for mobile users and voice control of devices from far field microphones, can be characterized by rapidly varying non-stationary noise conditions.

Missing feature techniques have been widely used for reconstructing speech features corrupted by stationary or non-stationary background noise. This class of techniques first segment sound sources in the time–frequency domain into "reliable" regions which are dominated by the target speech and "missing" or "unreliable" regions which are dominated by background noise (Raj et al., 2004). This segregation is represented by a spectro-temporal mask which assigns a probability of speech presence to individual time–frequency spectral components. In the data imputation approach to missing feature based ASR, an estimate is obtained for the spectral components labeled as unreliable, ASR features are computed from the reconstructed spectral representation, and these features are used in an ASR system which has been trained in a noise-free environment. In this work, a minimum mean squared-error based data imputation approach is used to re-estimate the unreliable spectrogram components (Raj, 2000; Raj and Singh, 2005). This technique is used as a reference to evaluate the performance of the DWT-based approach to the state of the art. A more detailed description of the spectrographic masks and the MMSE-based spectrogram reconstruction technique is presented in Sections 2.2 and 2.3.