Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 1–23

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Bounded cepstral marginalization of missing data for robust speech recognition☆

Kian Ebrahim Kafoori, Seyed Mohammad Ahadi*

Department of Electrical Engineering, Amirkabir University of Technology, Tehran, Iran

Received 2 April 2015; received in revised form 10 July 2015; accepted 29 July 2015
Available online 8 August 2015

Abstract

Spectral imputation and classifier modification can be counted as the two main missing data approaches for robust automatic speech recognition (ASR). Despite their potentials, little attention has been paid to the classifier modification techniques. In this paper, we show that transferring bounded marginalization, which is a classifier modification method, from spectral to cepstral domain would be beneficial for robust ASR. We also propose improved solutions on this transfer toward a better performance. Two such techniques are presented. The first approach still does not need training of any extra model. It benefits from an observed characteristic of cepstral features and raises accuracy of previously proposed method to a comparable level with that of a classic imputation method. The second technique combines our originally proposed method with an imputation technique but replaces spectral reconstruction with a simpler and faster possible range estimation of missing components. We show that the resulting method improves the accuracies of either of the two combined methods. The proposed techniques also show good robustness when implemented with an inaccurate spectrographic mask.

Keywords: Automatic speech recognition; Missing data theory; Noise robustness; Cepstral analysis

1. Introduction

Achieving practical ways to Human-Machine spoken communication has been an ultimate goal in the past few decades. One of the most challenging parts of this task is Automatic Speech Recognition (ASR) (Jokinen and McTear, 2010). Currently, systems that exploit statistical modeling of speech – especially Hidden Markov Modeling (HMM) – are almost dominating the ASR field, because they have gained excellent recognition accuracies under unpolluted acoustic conditions (Benesty et al., 2008). However, their performances fall drastically when acoustic conditions in which they are tested differ from which they were trained. Additive environmental acoustic noise – which is common in practice – is one of the main contributors to the training-testing mismatch. It is customary to either train recognizer merely by information derived under clean conditions, which is known as clean training, or train it by information derived under various additively polluted conditions, which is known as multi-condition training. Extensive efforts

☆ This paper has been recommended for acceptance by Hugo Van hamme.

* Corresponding author. Tel.: +98 21 6454 3336; fax: +98 21 6640 6469.
E-mail addresses: kafoori@aut.ac.ir (K. Ebrahim Kafoori), sma@aut.ac.ir (S.M. Ahadi).

http://dx.doi.org/10.1016/j.csl.2015.07.005
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

have been carried out to make the recognizer more robust against additive environmental acoustic noise (Virtanen et al., 2012).

Methods employed for robust ASR can be divided into two general categories of model compensation and feature compensation (Virtanen et al., 2012). In model compensation, the ultimate goal is to shift clean-trained statistical model parameters in such a way as if they were trained under present noisy condition. Parallel Model Combination (PMC) (Gales, 1993) and vector Taylor series (VTS) noise adaptation (Moreno et al., 1996) are among the most successful methods in this category. In contrast, in signal/feature compensation methods, models trained under the clean condition remain intact, while the efforts are concentrated on either extracting speech features that contain more information from speech than from noise, or remove the noise impairments from them. Cepstral mean and variance normalization (CMVN) (Furui, 1981), RASTA filtering (Hermansky and Morgan, 1994), histogram equalization (De la Torre et al., 2005), autocorrelation based feature extraction (Farahani et al., 2007) and missing data approaches (Cooke et al., 2001; Raj and Stern, 2005) are good examples in this category. Our proposed methods belong to the latter category.

Recently, missing data robustness approach, which is a feature compensation approach based on the missing data theory, has shown promising results toward robust ASR. The fact that even in severe noisy conditions some regions of the log-spectrogram representation of speech remain almost intact is the main initiative in this approach. Hence, if one could detect these regions and somehow perform decoding only using them, destructive effects of noise could be mostly removed. First step to reach this goal is to identify these useful regions from the missing ones, which is known as mask estimation. Rich literature on mask estimation is now available (Virtanen et al., 2012), most of them like SNR estimation approaches, Bayesian classifiers (Seltzer et al., 2004) and perceptual (like Zhao et al., 2012) and binaural approaches (Harding et al., 2006) perform estimation as a preliminary phase to decoding. However, recently, there has been a trend to perform mask estimation and decoding simultaneously (Barker et al., 2010; Ma et al., 2012; Narayanan and Wang, 2013). After mask estimation, decoding should be carried out using incomplete log-spectrogram. The approaches proposed to achieve this goal can generally be divided into two groups, Classifier Modification (CM) and Spectral Imputation (SI) (Raj and Stern, 2005). In CM, the classifier is modified in a way to make it capable of classifying incomplete features. In other words, decoding is made merely using the reliable data. In SI approaches, the missing components are estimated and restored based on a trained prior knowledge. Decoding is then performed in a conventional complete-feature manner. While covariance-based (Raj et al., 2004), cluster-based (Raj et al., 2004) and sparse (Gemmeke et al., 2010; Gemmeke et al., 2011; Ahmadi et al., 2014; Yilmaz et al., 2014) imputation are among the most successful approaches in this category, the most important CM approach is bounded marginalization (Cooke et al., 1999). Furthermore, class-conditional imputation (Josifovski et al., 1999) is a combined technique of SI and CM. Uncertainty decoding (Droppo et al., 2003) also modifies the classifier in order to incorporate the uncertainty of features in decoding process. Since complete spectrogram reconstruction allows for the straightforward employment of common robustness procedures, such as cepstral analysis or feature normalization, SI approaches could be easily merged with these kinds of methods and consequently, majority of research in the field of missing data ASR have been carried out in the SI category (González et al., 2013; Wang et al., 2013; Remes et al., 2011; Goodarzi et al., 2010; Borgström and Alwan, 2010; Keronen et al., 2013; Badiezadegan and Rose, 2015). However, they also suffer certain drawbacks. Regardless of any imputation method employed, either a distinct statistical spectral estimation model or an ensemble dictionary should be trained and missing components of the log-spectrogram should be estimated at the decoding phase as well. These two processes are often computationally expensive, and if higher estimation accuracies were needed, algorithms became more complex and even more computationally demanding. Furthermore, even the best estimation algorithm is not perfect and adds an additional noise to the features (Virtanen et al., 2012; Raj and Stern, 2005; Hartmann et al., 2013).

On the other hand, CM approaches need neither separate model training nor an estimation procedure, as they only use the reliable parts of data without any estimation (Virtanen et al., 2012; Raj and Stern, 2005). Nevertheless, due to certain obstacles, minor attention has been paid to them compared to SI approaches (Virtanen et al., 2012; Wang et al., 2013). Since modification of the classifier by marginalization of missing components is more straightforward in spectral domain, CM approaches have been usually carried out in this domain (Raj and Stern, 2005; May et al., 2012). This means that the structure of ASR classifier, learning process, and feature extraction should be modified to adapt to the spectral domain. According to a pervasive belief, efficiency of cepstral features is higher than the spectral ones, as their lower statistically correlated elements allow for a better acoustic modeling with diagonal covariance matrices. Furthermore, cepstral features have shown more robustness to the additive noise, while their normalization has led to a more effective compensation against channel distortion (Virtanen et al., 2012).