Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 69 (2015) 46–65

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Sub-band based histogram equalization in cepstral domain for speech recognition

Vikas Joshi [a,b,*], Raghvendra Bilgi [b], S. Umesh [b], Luz Garcia [c], Carmen Benitez [c]

[a] IBM India Research Labs, Bangalore, India
[b] Department of Electrical Engineering, Indian Institute of Technology (IIT), Madras, Chennai, Tamil Nadu 630036, India
[c] Department of Signal Theory, Telematics and Communications, University of Granada, Spain

Received 6 September 2014; received in revised form 5 February 2015; accepted 7 February 2015
Available online 19 February 2015

Abstract

This paper describes a novel framework to sub-band based Histogram Equalization (HEQ) applied to robust speech recognition. We propose a frequency band specific equalization to compensate the noise distortion on the individual frequency bands. The proposed equalization framework is a two step process. In the first step, conventional histogram equalization is done. By analyzing the histograms of equalized cepstra, we show that the first stage of conventional HEQ approach does not compensate the sub-band specific noise distortion, even though the overall histogram is normalized. Hence, in the second stage, sub-band specific histogram equalization is done. Every frame of cepstral coefficients is decomposed into low-frequency (LF) cepstra and high-frequency (HF) cepstra. Separate equalization is done on LF and HF cepstra to compensate LF and HF specific noise distortion. The cepstra corresponding to the LF and HF bands are obtained by using simple averaging and differencing filters on the cepstral components *within* a particular frame. The proposed approach is referred to as Sub-band Histogram Equalization (S-HEQ). Using histogram analysis, we show that the S-HEQ approach is able to compensate for the sub-band specific noise distortion. S-HEQ approach shows a consistent improvement over the conventional HEQ approach with a relative improvement of 12% and 22.10% over conventional HEQ in WER on Aurora-2 and Aurora-4 databases respectively. Proposed equalization approach can also be used with the deep neural network based systems and has shown a consistent improvement in the recognition accuracies over conventional HEQ. Finally, the efficacy of the proposed S-HEQ approach for embedded real-time speech applications is shown by comparing the performance and computational complexity trade-off with other state-of-the-art noise compensation methods.

Keywords: Robust speech recognition; Sub-band histogram equalization; S-HEQ; Histogram equalization

1. Introduction

The performance of automatic speech recognition (ASR) systems degrades in noisy environments. The presence of noise in the acoustic signal increases the confusion between phoneme classes and also introduces a mismatch between train and test conditions, thereby degrading the performance of ASR systems. Several techniques have been proposed to compensate such effects of noise (Atal, 1976; Gong, 1995; Moreno et al., 1996; Viikki and Laurila, 1998; Gales, 1998; Kim and Stern, 2014).

Embedded speech recognition systems such as automated voice control systems in automobiles and video games are becoming increasingly popular. Speech recognition systems in such applications should be able to recognize speech under different noisy environments, in real-time. Hence,

* Corresponding author at: IBM India Research Labs, Bangalore, India.
E-mail addresses: joshi.v.vikas@gmail.com (V. Joshi), rrbilgi@gmail.com (R. Bilgi), umeshs@ee.iitm.ac.in (S. Umesh), luzgm@ugr.es (L. Garcia), carmen@ugr.es (C. Benitez).

http://dx.doi.org/10.1016/j.specom.2015.02.005
0167-6393/

noise compensation algorithms deployed in such systems should be able to compensate different types of noise and also be computationally efficient. Cepstral Mean Subtraction (CMS) (Atal, 1976; Liu et al., 1995) is the simplest of the approaches where the mismatch introduced on the mean of training and test features is compensated. CMS, transforms all train and test features into zero mean features, thereby eliminating the effect of noise on the mean of the distribution. Cepstral Mean and Variance Normalization (CMVN) (Furui, 1981; Viikki and Laurila, 1998) compensates the mismatch in mean as well as the variance by transforming train and test features to zero mean and unit variance. Histogram Equalization (HEQ) (Balchandran and Mammone, 1998; Segura et al., 2004; de la Torre et al., 2005; Hilger and Ney, 2006) compensates even the higher order moments, by matching the entire distribution of clean and noisy features. HEQ is shown to compensate different types of noise and is also computationally efficient (Segura et al., 2004). The simplicity and computational efficiency of the above discussed noise compensation methods have made them the de facto noise compensation algorithms in many real-time speech systems. In this work, we extend HEQ to equalize sub-band specific effects of noise and propose a framework to achieve the same.

Distortion introduced by noise is relatively different along the various frequency bands. Moreover, the presence of speech is also not uniform along the frequency bands. Most of the speech energy is often concentrated in low frequency (LF) regions (unlike noise), resulting in low signal-to-noise ratio (SNR) in high frequency (HF) regions. In the past, several works have exploited the non-uniform behavior of speech and noise along the different frequency bands (Fletcher, 1953; Hermansky et al., 1996; Okawa et al., 1998; Cooke et al., 2000). Fletcher's (Fletcher, 1953) work on sub-bands suggests that the linguistic message gets decoded in different frequency sub-bands and the final decision is based on merging of decisions from such different sub-bands. According to Fletcher, the probability of erroneous recognition in the sub-bands multiplies to yield the overall error rate. In missing feature methods (Hermansky et al., 1996), only reliable sub-bands are used for the recombination. In Cooke et al. (2000), the contribution of each sub-band to the overall combination is weighted by its band specific SNR. The above heuristics motivates us to investigate if a sub-band specific noise transformation can be achieved using the popular HEQ technique.

Sub-band specific equalization can be done by applying HEQ on log Mel filter bank (LMFB) features instead of cepstral features. Such an equalization will effectively eliminate the distortion on the individual LMFB coefficient, thereby performing a frequency band specific equalization. However, researchers have argued against equalization in LMFB domain (Obuchi and Stern, 2003; de la Torre et al., 2005) as there is a strong correlation between the components of LMFB feature vector, which is undesirable in the conventional HEQ framework. In conventional HEQ technique, each component of the feature vector is equalized *independent* of the other. Hence, an independent equalization of correlated LMFB features is not appropriate and affects the performance of HEQ. Instead of equalizing each component independently, joint histogram of the entire feature vector can be equalized. However, large amount of data is required to reliably estimate the joint histogram, which may not be available in practice. In most of the literature (Segura et al., 2002; Obuchi and Stern, 2003; de la Torre et al., 2005) equalization is done in the cepstral domain as the cepstral coefficients are approximately uncorrelated due to multiplication of discrete cosine transform of type 2 (DCT-2) (Victoria et al., 1995; Logan, 2000) to LMFB coefficients. Equalization in cepstral domain does not account for a band-specific equalization as the frequency band specific effects of noise get distributed among all the cepstral coefficients by the application of DCT transform to LMFB coefficients. In Section 3, we show that the conventional HEQ technique as applied in the cepstral domain is not effective to compensate the sub-band specific noise distortions.

In this work, we propose a framework to perform sub-band specific equalization *directly* in the cepstral domain, thereby exploiting the property of uncorrelatedness of the cepstral features. We first decompose the conventional Mel-frequency cepstral coefficients (MFCC) into two parts viz., one corresponding to a low frequency band, termed as LF-MFCC and another corresponding to a high frequency band, termed as HF-MFCC. Then a separate histogram equalization is done on LF-MFCC and HF-MFCC features to compensate the sub-band specific noise effects. The approach to decompose a given MFCC frame into corresponding sub-band cepstral coefficients is explained in detail in Section 2. The term "conventional" or "overall" MFCC is used to clearly distinguish them from the cepstra corresponding to low and high frequency bands, i.e., LF-MFCC and HF-MFCC. The proposed S-HEQ features show a significant improvement in recognition results over conventional HEQ for all noise types and all SNR conditions for the Aurora-2 and Aurora-4 databases. S-HEQ features can as well be used in deep neural network (DNN) based ASR systems and they show a consistent improvement in the recognition results over conventional HEQ features. Preliminary work based on this idea was presented in Vikas et al. (2011).

1.1. Related work

Histogram equalization was originally used in digital image processing to correct brightness and contrast levels within an image (Gonzalez and Wintz, 1987; Russ, 1995). Histogram equalization for robust speech recognition was first studied in Balchandran and Mammone (1998) and further explored in detail by several authors (Hilger et al., 2002; de la Torre et al., 2005; Hilger and Ney, 2006). Variants of HEQ have been proposed, where delta and acceleration features are also equalized along with the static cepstra in Obuchi and Stern (2003). Hung et al. (Hung and Fan, 2009) propose to equalize the modulation frequency bands