Digital Signal Processing 29 (2014) 147–157

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Robust feature extraction based on an asymmetric level-dependent auditory filterbank and a subband spectrum enhancement technique

Md Jahangir Alam [a,b,*], Patrick Kenny [b], Douglas O'Shaughnessy [a]

[a] INRS-EMT, University of Quebec, Montreal, Quebec, Canada
[b] CRIM, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Available online 19 March 2014

Keywords:
Speech recognition
Compressive gammachirp
Auditory spectrum enhancement
Feature normalization

ABSTRACT

In this paper we introduce a robust feature extractor, dubbed as robust compressive gammachirp filterbank cepstral coefficients (RCGCC), based on an asymmetric and level-dependent compressive gammachirp filterbank and a sigmoid shape weighting rule for the enhancement of speech spectra in the auditory domain. The goal of this work is to improve the robustness of speech recognition systems in additive noise and real-time reverberant environments. As a post processing scheme we employ a short-time feature normalization technique called short-time cepstral mean and scale normalization (STCMSN), which, by adjusting the scale and mean of cepstral features, reduces the difference of cepstra between the training and test environments. For performance evaluation, in the context of speech recognition, of the proposed feature extractor we use the standard noisy AURORA-2 connected digit corpus, the meeting recorder digits (MRDs) subset of the AURORA-5 corpus, and the AURORA-4 LVCSR corpus, which represent additive noise, reverberant acoustic conditions and additive noise as well as different microphone channel conditions, respectively. The ETSI advanced front-end (ETSI-AFE), the recently proposed power normalized cepstral coefficients (PNCC), conventional MFCC and PLP features are used for comparison purposes. Experimental speech recognition results demonstrate that the proposed method is robust against both additive and reverberant environments. The proposed method provides comparable results to that of the ETSI-AFE and PNCC on the AURORA-2 as well as AURORA-4 corpora and provides considerable improvements with respect to the other feature extractors on the AURORA-5 corpus.

1. Introduction

Speech intelligibility as well as the performance of speech recognition systems degrades in practical environments due to a variety of signal variabilities. Additive noise and reverberation are the important causes of signal variabilities. Additive noise from interfering noise sources and convolution noise arising from acoustic environments mainly cause a reduction of speech recognition performance.

The acoustic features most commonly used in speech recognition systems are Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2] features. Both MFCC and PLP front-ends perform well in matched environments, where speech data are collected from reasonably clean environments. However, their performance degrades severely when the testing environment is different from the training environment. Degradation of performance due to mismatched environments has been a barrier for deployment of speech recognition technologies. Various sources give rise to this mismatch, such as background noise, channel/handset distortion, room reverberation. Most of the sources of speech variability produce additive distortion (e.g., background noise) and/or convolutional distortion (e.g., channel/handset mismatch) in the speech signal. Among all different additive noises, a multi-speaker or babble noise environment, where the interference is speech from speakers in the vicinity, is one of the most challenging noise conditions. Therefore, it is necessary to address this problem, i.e., performance degradation due to mismatch environments, to enable the deployment of recognition systems in real world conditions.

Various research has been reported in the literature to improve the robustness of speech recognition systems under additive noise and reverberation. The methods to compensate for the effects of environmental mismatch can be implemented at the front-end (feature domain techniques) or at the back-end (model domain techniques) or both. The model domain methods adapt each acoustic model to make it fit better to the mismatched acoustic environment so that the adapted models will be able to classify the mismatched speech features collected in the testing environment.

* Corresponding author.
E-mail addresses: alam@emt.inrs.ca (M.J. Alam), Patrick.kenny@crim.ca (P. Kenny), dougo@emt.inrs.ca (D. O'Shaughnessy).

http://dx.doi.org/10.1016/j.dsp.2014.03.001
1051-2004/

The typical examples of this category include the well-known noise masking [20], speech and noise decomposition (SND) [21], vector tailor series (VTS) [22], maximum a posteriori (MAP) [23,24], maximum likelihood linear regression (MLLR) [25], model-based stochastic matching [26], statistical reestimation (STAR) [27–29], parallel model combination (PMC) [30], optimal likelihood weighting based on the criteria of minimum classification error (MCE) and maximum mutual information (MMI) [31], etc.

The objective of feature domain techniques is to make features more consistent in diverse environmental conditions. Feature domain methods can be classified into two subgroups. One subgroup of feature domain techniques aims at modifying the test features and making those features match the acoustic conditions better for the trained models: speech enhancement methods [32–37], codeword dependent cepstral normalization (CDCN) [38], feature-based stochastic mapping [26], multivariate Gaussian based cepstral normalization (RATZ) [39], feature normalization methods, such as cepstral mean normalization (CMN) [7,11,40], the stereo-based piecewise linear compensation for environment (SPLICE) [41,42]. The other subgroup of feature domain techniques, on the other hand, aims at making a special robust speech feature representation, which is used for both training and testing, to reduce the sensitivity to the various acoustic conditions. This paper deals with the latter subgroup. Robust feature extractors are usually obtained either by appending a pre-processing step, like speech enhancement [8,9,14,15], or by incorporating algorithms in an MFCC or PLP computation framework such as PNCC [3], amplitude modulation-based cepstral features [43,44], frequency masking [9], or by adding a post-processing step, like feature normalization techniques [7,11] (e.g., cepstral mean normalization (CMN)) or by combining any two or all of the above mentioned steps [3,4]. Most of the front-ends use, in addition to other techniques for environmental mismatch compensation, a feature normalization technique, at the least CMN, as a post-processing scheme.

Additive noise reduction approaches usually have a tradeoff between the amount of noise reduction and speech distortion induced due to processing of a speech signal. At very low SNR the intensity of this induced distortion is high, thereby deteriorating the performance of the speech recognition systems. Compensation of reverberant noise is usually done by dereverberation, which can be obtained by inverse filtering the impulse response of the room [12,13]. However, room impulse response is dependent on the distance between the speaker and the microphone and on the conditions of the room. Therefore, extracting a common set of robust features, which can perform well at low SNRs and also can handle various room impulse responses, is a difficult and challenging task.

To deal with additive noise distortion various speech enhancement methods have been proposed in [8,9,14,15]. In [10,12,13] several approaches have been proposed for handling convolutional noise distortions. The ETSI-AFE, described in [4], uses a two-stage Wiener filter and a blind equalization technique, which is based on the comparison to a flat spectrum and the application of the LMS algorithm, for improving robustness of ASR systems against additive noise distortions and channel effects. The PNCC technique, proposed in [3], includes the use of a gammatone filter-bank (GTFB) and a power law nonlinearity in place of the Mel filter-bank and log nonlinearity, used in conventional MFCCs framework, a medium duration power bias subtraction technique, for noise reduction, based on the arithmetic mean (AM)–geometric mean (GM) ratio and cepstral mean normalization as a post-processing scheme for DC offset removal, for robust feature extraction.

In this work, for robust features extraction, we propose to enhance the speech auditory spectrum using a weighting rule based on the subband *a posteriori* signal-to-noise ratio (SNR). In order to allow a realistic and controllable frequency-domain asymmetry and to model most of the level dependency observed in basilar membrane (BM) filtering, the proposed method includes the use of a compressive gammachirp filter-bank (cGCFB) [16] for auditory spectral analysis. We use a power function nonlinearity as it has been found in [3] that it is more robust than the logarithmic nonlinearity used in a conventional MFCC framework. As a post-processing scheme for the normalization of the features, we use the short-time cepstral mean and scale normalization (STCMSN) technique, proposed in [7]. Feature normalization is normally performed over the whole utterance with the assumption that the channel effect is constant over the entire utterance, such as CMN (or CMVN). Also, normalizing a feature vector over the entire utterance is not a feasible solution in real-time applications as it causes an unnecessarily long processing delay. To relax this assumption and to reduce the processing delay, cepstral features in the proposed method are normalized over a sliding window of 1.5 s duration. In this paper, we denote this proposed feature extractor as the Robust Compressive Gammachirp filterbank Cepstral Coefficients (RCGCC). Replacing cGCFB with the GTFB in the proposed RCGCC feature extraction framework, we also present Robust Gammatone Filterbank Cepstral Coefficient (RGFCC) features to show the effectiveness of using cGCFB for the auditory spectral analysis in the proposed front-end. Experimental recognition results on the AURORA-2, AURORA-4, and AURORA-5 corpora demonstrate that the proposed RCGCC feature extractor outperforms the MFCC and PLP front-ends and provides comparable (and sometimes better) results to most state-of-the-art front-ends used in this work.

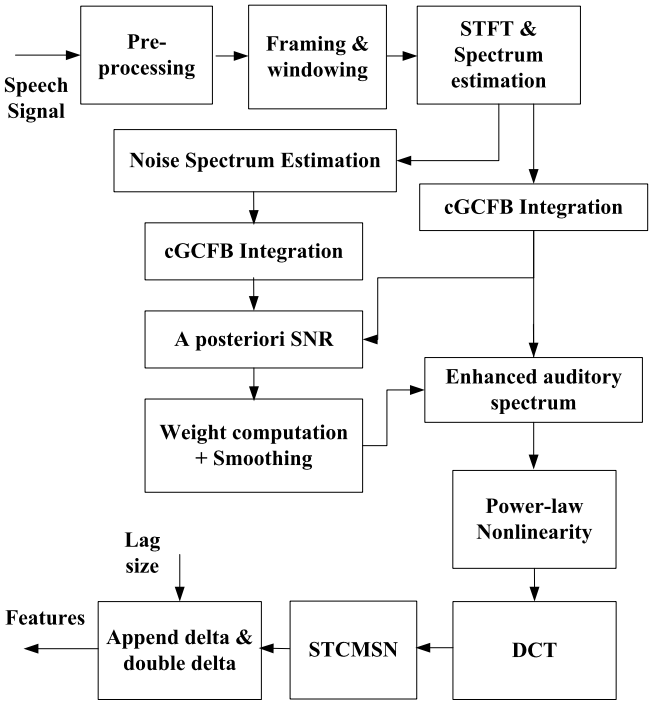

Fig. 1. Block diagram showing various stages of the proposed robust compressive gammachirp filterbank cepstral coefficient (RCGCC) feature extractor.

2. Proposed feature extractor

The complete block diagram of the proposed robust compressive gammachirp filterbank cepstral coefficient (RCGCC) feature extractor for robust speech recognition is shown in Fig. 1. In the RCGCC feature extractor, processing of a speech signal begins with pre-processing (including DC removal and pre-emphasis, typically using a first-order high-pass filter). Short-time Fourier Transform (STFT) analysis is performed using a finite duration