Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 73 (2015) 28–46

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Regularized minimum variance distortionless response-based cepstral features for robust continuous speech recognition

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

Received 10 June 2013; received in revised form 8 June 2015; accepted 22 July 2015
Available online 29 July 2015

Abstract

In this paper, we present robust feature extractors that incorporate a regularized minimum variance distortionless response (RMVDR) spectrum estimator instead of the discrete Fourier transform-based direct spectrum estimator, used in many front-ends including the conventional MFCC, to estimate the speech power spectrum. Direct spectrum estimators, e.g., single tapered periodogram, have high variance and they perform poorly under noisy and adverse conditions. To reduce this performance drop we propose to increase the robustness of speech recognition systems by extracting features that are more robust based on the regularized MVDR technique. The RMVDR spectrum estimator has low spectral variance and is robust to mismatch conditions. Based on the RMVDR spectrum estimator, robust acoustic front-ends, namely, are regularized MVDR-based cepstral coefficients (RMCC), robust RMVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC). In addition to the RMVDR spectrum estimator, RRMCC and NRMCC also utilize auditory domain spectrum enhancement methods, auditory spectrum enhancement (ASE) and medium duration power bias subtraction (MDPBS) techniques, respectively, to improve the robustness of the feature extraction method. Experimental speech recognition results are conducted on the AURORA-4 large vocabulary continuous speech recognition corpus and performances are compared with the Mel frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP), MVDR spectrum estimator-based MFCC, perceptual MVDR (PMVDR), cochlear filterbank cepstral coefficients (CFCC), power normalized cepstral coefficients (PNCC), ETSI advancement front-end (ETSI-AFE), and the robust feature extractor (RFE) of Alam et al. (2012). Experimental results demonstrate that the proposed robust feature extractors outperformed the other robust front-ends in terms of percentage word error rate on the AURORA-4 large vocabulary continuous speech recognition (LVCSR) task under clean and multi-condition training conditions. In clean training conditions, on average, the RRMCC and NRMCC provide significant reductions in word error rate over the rest of the front-ends. In multi-condition training, the RMCC, RRMCC, and NRMCC perform slightly better in terms of the average word error rate than the rest of the front-ends used in this work.

Keywords: Speech recognition; Robust feature extraction; Regularized MVDR; ASE; Feature normalization; Multi-condition training

1. Introduction

Mel-frequency cepstral coefficients (MFCC) Davis and Mermelstein (1980), which have proven to be one of the most effective feature sets for speech and speaker recognition tasks, are frequently used as a low-dimensional set of features to represent short-time speech signals. MFCC are usually computed by integrating a triangular-shaped

* Corresponding author at: INRS-EMT, University of Quebec, 800, de La Gauchetière West, Suite 6900, Montréal, Québec H5A 1K6, Canada. Tel.: +1 (514) 840 1235x3344; fax: +1 (514) 840 1244.
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.specom.2015.07.007
0167-6393/

Mel-scaled filterbank (MelFB) either to the DFT-based short-time spectrum or to the linear predictive coding (LPC)-based spectrum. MFCC and perceptual linear prediction (PLP) Hermansky (1990)-based speech recognizers perform well under matched training/test conditions but the performance gap between automatic speech recognizers (ASRs) and human listeners in real world settings is significant (Huang et al., 2001; O'Shaughnessy, 2000). Different operating conditions during signal acquisition (e.g., channel response, handset type, additive background noise, reverberation, etc.) lead to feature mismatch across training and testing and thereby degrade the performance of MFCC (and PLP)-based speech recognition systems. To tackle this problem, various robust feature extractors are employed in speech recognition tasks, such as the ETSI advanced front-end (ETSI-AFE) (ETSI ES 202 050, 2003), power normalized cepstral coefficients (PNCC) (Kim and Stern, 2010), and the robust feature extractors proposed in Alam et al. (2012, 2013a, 2014b), van Hout and Alwan (2012), Mitra et al. (2012), Chiu et al. (2012), etc. In MFCC (Davis and Mermelstein, 1980) and PLP (Hermansky, 1990) front-ends, and in most of the robust feature extractors the features are computed from a windowed (e.g., Hamming) direct spectrum estimate (the squared magnitude of the Fourier transform of the short-time windowed observed signal) that has a high spectral variance. The variances of these features are greatly influenced by the variances of the spectral estimates of the observed speech signal. Variance in the feature vectors has a direct bearing to the variance of Gaussians modeling the speech classes. Reduction in the variance of the feature vector increases class separability and improved class separability can potentially increase recognition accuracy and decrease search speed (Dharanipragada and Rao, 2001). Although direct spectrum estimators (also known as non-parametric spectrum estimators) are entirely independent of data and therefore do not suffer from problems arising from modeling deficiencies, these methods are not robust to noise and hence they perform poorly under mismatched training/test conditions. Among the parametric spectrum estimators, the linear predictive coding (LPC) based all-pole spectrum estimator is most widely used (Capon, 1969). It has been noted in speech modeling literature that the LP-based all-pole models do not provide good models of the spectral envelope for medium and high pitch voiced speech (Dharanipragada and Rao, 2001). Also, the LP-based cepstra are known to be very sensitive to noise. They tend to overestimate or overemphasize sparsely spaced harmonic peaks (Wolfel et al., 2009). The standard feature extractors used for speech recognition are based on either DFT, e.g., MFCC or linear prediction, e.g., PLP. The MFCC feature extractor is not robust and therefore shows poor performance under noisy and adverse conditions. On the other hand, the PLP front-end is ill-suited for reliable estimation of the spectra of speech signals, which is true for all methods using linear prediction envelopes (Wolfel et al., 2009). In order to overcome the problems associated with linear prediction, namely, over-estimation of spectral power at the harmonics of voiced speech, the MVDR method was proposed in Murthi and Rao (2000). It is also known as Capon's method (Capon, 1969) for all pole modeling of speech.

In this paper, we propose to incorporate a regularized minimum variance distortion-less response (RMVDR) spectrum estimator, in place of the DFT-based direct spectrum estimator, into the traditionally used feature extraction framework, e.g., MFCC, for speech recognition task. Based on RMVDR spectrum estimation method we also propose robust feature extractors, dubbed as robust regularized MVDR cepstral coefficients (RRMCC) and normalized RMVDR cepstral coefficients (NRMCC), that include the use of sigmoid-shape auditory domain spectrum enhancement (ASE) (Alam et al., 2012) and medium duration power bias subtraction (Kim and Stern, 2010) techniques, respectively, to improve the robustness of speech recognition systems in adverse conditions while having little performance reduction in matched train/test conditions. The advantages of a RMVDR spectrum estimator are:

(a) It overcomes the problems apparent in linear prediction spectral estimation.
(b) The regularization parameter helps to penalize rapid changes in all-pole spectral envelopes thereby producing smooth spectra without affecting the formant positions (Murthi and Kleijn, 2000; Hanilci et al., 2012).
(c) It provides robust spectral estimates under noisy environments (Alam et al., 2013b, 2013c, 2013d).

The MVDR spectral estimator has already been applied in speech recognition (Dharanipragada and Rao, 2001) and speaker identification (Wolfel et al., 2009) tasks. An extension of the MVDR method was proposed in Wolfel and McDonough (2005) by warping the frequency axis with the bilinear transformation prior to MVDR spectral estimation. In Yapanel and Dharanipragada (2003), a perceptual MVDR-based cepstral coefficients (PMCC) approach is proposed where perceptual information is directly incorporated into the spectrum estimation. The perceptually motivated MVDR (PMVDR) front-end, proposed in Yapanel and Hansen (2008), completely eliminates the auditory filterbank processing step and directly performs warping on the DFT power spectrum.

In order to compare the performance of the proposed front-ends, the following conventional and robust front-ends were chosen: MFCC (Davis and Mermelstein, 1980), PLP (Hermansky, 1990), MVDR-based MFCC (Dharanipragada and Rao, 2001), PMVDR (Yapanel and Hansen, 2008), ETSI-AFE (ETSI ES 202 050, 2003), power normalized cepstral coefficients (PNCC) (Kim and Stern, 2010), cochlear filterbank cepstral coefficients (CFCC) (Li and Huang, 2010), and the robust feature extractor (RFE) proposed in Alam et al. (2012). The ETSI-AFE,