Digital Signal Processing 32 (2014) 109–123

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

A new framework for robust speech recognition in complex channel environments ☆

Yongjun He [a,*], Jiqing Han [b], Tieran Zheng [b], Guanglu Sun [a]

[a] School of Computer Science and Technology, Harbin University of Science and Technology, No. 52, Xuefu Road, Harbin, Heilongjiang, China
[b] Harbin Institute of Technology, No. 92, West Da-Zhi Street, Harbin, Heilongjiang, China

ARTICLE INFO

Article history:
Available online 11 June 2014

Keywords:
Channel distortion
Expectation-maximization
Spectrum missing
Automatic speech recognition

ABSTRACT

Channel distortion is one of the major factors which degrade the performances of automatic speech recognition (ASR) systems. Current compensation methods are generally based on the assumption that the channel distortion is a constant or slowly varying bias in an utterance or globally. However, this assumption is not sustained in a more complex circumstance, when the speech records being recognized are from many different unknown channels and have parts of the spectrum completely removed (e.g. band-limited speech). On the one hand, different channels may cause different distortions; on the other, the distortion caused by a given channel varies over the speech frames when parts of the speech spectrum are removed completely. As a result, the performance of the current methods is limited in complex environments. To solve this problem, we propose a unified framework in which the channel distortion is first divided into two subproblems, namely, *spectrum missing* and *magnitude changing*. Next, the two types of distortions are compensated with different techniques in two steps. In the first step, the speech bandwidth is detected for each utterance and the acoustic models are synthesized with clean models to compensate for *spectrum missing*. In the second step, the constant term of the distortion is estimated via the expectation-maximization (EM) algorithm and subtracted from the means of the synthesized model to further compensate for *magnitude changing*. Several databases are chosen to evaluate the proposed framework. The speech in these databases is recorded in different channels, including various microphones and band-limited channels. Moreover, to simulate more types of *spectrum missing*, various low-pass and band-pass filters are used to process the speech from the chosen databases. Although these databases and their filtered versions make the channel conditions more challenging for recognition, experimental results show that the proposed framework can substantially improve the performance of ASR systems in complex channel environments.

1. Introduction

Although extensive research on automatic speech recognition (ASR) in adverse environments has been carried out for many years, it remains a challenge because of various possible types of environmental distortion. As one of the major factors that degrade the performance of ASR systems, channel distortion is inevitable. Various conditions, such as different microphones and transmitting channels or encoders, can cause channel distortion. Within the last few decades, many approaches have been proposed to address this problem. These approaches can be broadly divided into two main categories: feature enhancement and model adaptation [1].

Feature enhancement attempts to extract features that are less sensitive to channel distortion. One class of feature enhancement is feature normalization; some examples are cepstral mean normalization (CMN) [2], cepstral mean and variance normalization (CMVN) [3], and relative spectra (RASTA) [4]. These methods rely on the assumption that the channel distortion is present in the constant or slowly varying components of speech features. Hence, these methods normalize features by removing such components. CMN removes the mean vector from the features of an utterance; CMVN extends CMN by adjusting both the mean and variance of features; and RASTA designs a filter in the cepstral domain to remove the slowly varying components.

☆ This research is partly supported by the National Natural Science Foundation of China under grant Nos. 61305001, 91120303 and 91220301, the Heilongjiang Postdoctoral Fund under grant No. LBH-Z13099, the China Postdoctoral Science Foundation under grant No. 2013M531042, the Scientific Research Fund of Heilongjiang Provincial Education Department under grant No. 12511096, the Natural Science Foundation of Heilongjiang Province under grant No. F200936.

* Corresponding author.
E-mail addresses: heyongjun@hit.edu.cn (Y. He), jqhan@hit.edu.cn (J. Han), zhengtieran@hit.edu.cn (T. Zheng), sunguanglu@hrbust.edu.cn (G. Sun).

http://dx.doi.org/10.1016/j.dsp.2014.06.002
1051-2004/

Another class of feature enhancement methods aims to estimate pseudo-undistorted features from distorted speech. These methods operate in the log-spectrum or Mel-frequency cepstral coefficient (MFCC) domains, in which channel distortion is modeled as an additive term. In addition, clean speech can be modeled with a Gaussian, a Gaussian mixture model (GMM) or a hidden Markov model (HMM). The channel distortion is then estimated under a Bayesian framework, and the pseudo-undistorted features are obtained by subtracting the channel distortion from the distorted features [5]. When additive noise and channel distortions are present simultaneously, the relationship between the clean and distorted speech is highly nonlinear in the MFCC domain [6,23]. To solve this problem, the vector Taylor series (VTS) is introduced to linearize the distortion model. Next, the channel distortion is estimated with the expectation-maximization (EM) algorithm [6,7]. In addition, the ETSI's advanced frontend [8] adopts the least mean square (LMS) algorithm [9] to compensate for the channel distortion in the feature domain. Recently, a new feature named robust compressive gammachirp filterbank cepstral coefficient is proposed [10]. This feature is 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.

Feature enhancement methods are attractive because they operate in the frontend and can be easily implemented into an ASR system without modifying the recognizer. Moreover, these methods are typically simpler computationally, satisfying the demand of a real-time application. However, feature enhancement methods are shown to only achieve a medium-level distortion reduction [26], and errors in feature estimations can cause further mismatches between the features and the acoustic models, resulting in degraded performance [1].

On the contrary, model adaptation methods work in the backend to compensate by modifying or even retraining the acoustic models. The most straightforward method is to train models from the distorted speech, namely, matched training. This method provides an upper limit for the performance of an ASR system under a given condition, so it is often used for experimental comparison. However, it is impractical to use this method in real-life applications because it needs a large amount of distorted speech under each possible environment. Multi-style training [11] is designed to train models with different types of distorted speech, each of which would be reasonable to expect in deployment. When the likely running environments of the recognizer are included in the training data, this method can achieve good performance; otherwise, the performance will be degraded. In the model domain, the widely adopted strategy is to adapt the models trained with clean speech to distorted environments.

When adaptation data from the new environment are available, speaker adaptation methods, such as the constrained maximum likelihood linear regression (CMLLR) [12], the maximum likelihood linear regression (MLLR) [13–15] and maximum *a-posteriori* (MAP) adaptation [16], can also be used for environment adaptation. These methods do not make any assumption about the nature of the corrupting process and instead adopt a data-driven style. Their performance approaches those of the matched training with an increase in the amount of adaptation data. However, large amounts of adaptation data are required to achieve good performance, especially under severe distortion.

Unlike speaker adaptation methods, another type of model adaptation methods learn environment characteristic from adaptation data but instead takes advantage of the known relationship between the clean and distorted acoustic models. Typically, the signal bias removal (SBR) [17] models the channel distortion as a constant term added to the Gaussian means of the clean HMMs. This method estimates the channel term in a maximum likelihood estimation (MLE) manner and removes this term from the means in the HMMs. The parallel model combination (PMC) [18–20] exploits the mismatch function to combine the clean models with the noise models in the log-spectrum domain. This method is proposed to compensate for additive noise only in [18], and extended to address both additive noise and channel distortion in [19,20] by modifying the mismatch function. The PCA–CMS based PMC combines robust feature and PMC to improve the robustness of speech recognition systems [21]. This algorithm utilizes cepstral mean subtraction (CMS) normalization ability and principal component analysis (PCA) compression and de-correlation capability in the combination with PMC model transformation method. The VTS adaptation method [22–26] linearizes the distortion model [6] with the VTS approximation. Next, it estimates the additive noise and channel distortion in an EM algorithm framework and modifies the clean acoustic models to match the distorted speech. In addition, current methods can further improve their performance by adopting uncertainty decoding [27].

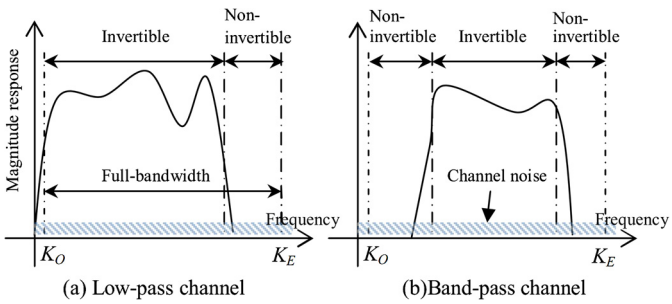

Fig. 1. Illustration of the invertible and non-invertible regions of band-limited channels. The full-bandwidth is K_O–K_E. (a) A low-pass channel is non-invertible in the high-frequency band, (b) a band-pass channel is non-invertible in both the low- and high-frequency bands.

Missing feature techniques [53] have achieved success in compensation for additive noise, and they can also be combined with traditional channel compensation methods to deal with both additive noise and channel distortion. Segbroeck and Hamme [54] propose a method which treat channel distortion as a constant term in the log-spectrum domain and then estimate it by maximizing the log-likelihood of the optimal state sequence of an observation sequence. Palomaki et al. [55] combine spectral features and cepstral features within the missing data framework to handle convolutional distortion and additive noise. They also propose a method for handling reverberated speech which attempts to identify time-frequency regions that are not badly contaminated by reverberation [56].

Although current methods have achieved improvements in simple channel environments, their performance are still limited in complex channel environments. In this paper, we consider two complex situations. The first case is that the channel changes from one to another frequently. Under this condition, it is impractical for data-driven methods to compensate for the channel distortion that varies over time. The second case is that the speech is band-limited in comparison with the training speech. Under this situation, the speech spectrum in a stop-band is removed completely and the effect of the channels is non-invertible. As shown in Fig. 1, the bandwidth of the training speech, i.e., the full-bandwidth is K_O–K_E. If speech passes through a low-pass (Fig. 1(a)) or band-pass (Fig. 1(b)) channel, the spectrum within stop-bands is heavily attenuated to be channel noise [29]. This type of speech is also called band-limited speech. It is shown in Section 3 that the channel distortion varies over speech frames when the channel has non-invertible bands (stop-bands). Therefore, the methods based on the assumption that the channel distortion is a constant bias