Applied Soft Computing 19 (2014) 198–214

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Stochastic feature compensation methods for speaker verification in noisy environments

Sourjya Sarkar*, K. Sreenivasa Rao

School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

ARTICLE INFO

Article history:
Received 31 July 2013
Received in revised form 8 February 2014
Accepted 17 February 2014
Available online 26 February 2014

Keywords:
Speaker verification
Noisy environment
Minimum mean squared error
Maximum likelihood estimate
Expectation Maximization algorithm
Gaussian Mixture Models

ABSTRACT

This paper explores the significance of stereo-based stochastic feature compensation (SFC) methods for robust speaker verification (SV) in mismatched training and test environments. Gaussian Mixture Model (GMM)-based SFC methods developed in past has been solely restricted for speech recognition tasks. Application of these algorithms in a SV framework for background noise compensation is proposed in this paper. A priori knowledge about the test environment and availability of stereo training data is assumed. During the training phase, Mel frequency cepstral coefficient (MFCC) features extracted from a speaker's noisy and clean speech utterance (stereo data) are used to build front end GMMs. During the evaluation phase, noisy test utterances are transformed on the basis of a minimum mean squared error (MMSE) or maximum likelihood (MLE) estimate, using the target speaker GMMs. Experiments conducted on the NIST-2003-SRE database with clean speech utterances artificially degraded with different types of additive noises reveal that the proposed SV systems strictly outperform baseline SV systems in mismatched conditions across all noisy background environments.

1. Introduction

Speaker verification (SV) is the process of validating the claimed identity of an individual using his or her speech. The task is achieved by classifying a claimant's utterance as true (authentic) or false (impostor) based on its statistical similarities with an enrolled (claimed) speaker's utterance. Acoustic modeling is used in the training stage of SV to effectively capture the distribution of features unique to an enrolled speaker. In the standard Gaussian Mixture Model (GMM)-based SV systems [1,2], acoustic speaker models are GMMs obtained by *Maximum a Posteriori* (MAP) adaptation [3] of a Universal Background Model (UBM) [2]. During evaluation, given a test speech segment, the log-likelihood ratio of scores obtained from the MAP-adapted GMMs and UBM is compared with an empirically determined threshold for final classification decision. Though the simple GMM-based SV systems perform quite well for clean speech, the performance is severely degraded in the presence of environmental noise [4]. A major challenge in the field of speaker recognition is to make the SV system robust towards its acoustic environment. Apart from channel distortions, additive background noise has been identified as a prominent factor for degraded SV performance [4]. The loss of performance accuracy can be mainly attributed to the mismatch occurring due to the differences in training and recognition environment. Since accurate estimation of noise is infeasible in nature, traditional approaches aim to compensate for environmental noise. Noise compensation can be broadly categorized in two domains i.e., acoustic model level and feature level.

Acoustic model adaptation techniques alter the statistical model parameters learned during the training/enrollment phase to reflect the acoustic environment of testing/recognition phase [5]. Popular data-driven model adaptation techniques like *Maximum aPosteriori* (MAP) [3] and Maximum Likelihood Linear Regression (MLLR) [6] use various amounts of adaptation data to achieve this task. State-of-the-art model compensation techniques like Parallel Model Compensation (PMC) [7] and Vector Taylor Series (VTS) [8] use an analytical relationship of the clean and noisy environment. These methods exploit prior knowledge about the test environment in the form of a statistical model of the noise or reliable estimates of the noise distribution. The model adaptation techniques are usually superior to their feature-level counterparts because they can appropriately capture the uncertainty caused by noise statistics [9]. However, besides depending on available clean speaker models, these methods are computationally intensive and often require high amount of training data [10].

* Corresponding author. +91 9883350943.
E-mail addresses: sourjyasarkar@gmail.com (S. Sarkar), ksrao@iitkgp.ac.in (K. Sreenivasa Rao).

http://dx.doi.org/10.1016/j.asoc.2014.02.016
1568-4946/

Feature compensation techniques map feature vectors extracted during the recognition phase to reflect the acoustic environment of the training/enrollment phase. The wide range of methods explored in this domain can be viewed in three groups. The first group of methods include high-pass filtering techniques like Cepstal Mean Subtraction (CMS) [11,12] and RelAtive SpecTral Amplitude (RASTA) [13]. Despite limited performance improvement, these techniques along with various feature transformation methods like feature warping [14] and nonlinear spectral magnitude normalization [15] find generic application in most speech related tasks. The second group of feature compensation techniques are noise model-based. They assume prior knowledge of the noise spectrum. An estimate of the clean speech parameters is made using either a noise model or representation of the effects of noise in speech. The parameters of the noise model are estimated and applied to the appropriate inverse operation to compensate the recognition signal. Examples include Spectral Subtraction (SS) [16], Codeword Dependent Cepstral Normalization (CDCN) [17], Kalman Filtering [18], and feature-level VTS [19].

The third group of feature compensation techniques are entirely data-driven and are stochastic in nature. They are 'blind' towards the nature of the corrupting process and are based on empirical compensation methods that use direct spectral comparison. Prior work shows that they often outperform the previous two approaches for feature enhancement [20]. During the training phase, some transformations are estimated by computing the frame-by-frame differences between the vectors representing speech in the clean and noisy environments (stereo data). The differences between clean and noisy feature vectors are modeled by training additive bias vectors on the mean and covariance of either of the two (clean or noisy) probability distributions. During the evaluation phase, the bias vectors are used to transform noisy test feature vectors to their clean feature equivalent based on a minimum mean squared error (MMSE) estimate. Earlier methods like CDCN [17,21] used vector quantization (VQ) codebooks to represent the distribution of clean feature vectors. Due to their quantization-based framework, these algorithms were unable to learn the variance of a distribution and were later replaced by the more flexible Gaussian Mixture Model(GMM)-based normalization techniques e.g., multivaRiate GAussian-based cepsTral normaliZation (RATZ) [22]. Although the RATZ family of algorithms approximated the normalized features, the posterior probability of clean GMM components with respect to the noisy test feature vectors were usually distorted causing poor MMSE estimates. To suppress these distortions, the Stereo-based Piecewise LInear CompEnsation for Environments (SPLICE) algorithm [23] modeled the noisy feature space using GMMs instead. This produced significantly better result in robust speech recognition tasks compared to its predecessors [24]. The effectiveness of SPLICE framework has since then encouraged it's extended applications e.g., speech recognition in non-stationary noisy environments within cars using the Multi Environment Model-based LInear Normalization (MEMLIN) algorithm [25] and word recognition using Noise Adaptive Training [24]. The more recently proposed Stereo-based Stochastic Mapping (SSM) [26] is principally a more accurate version of SPLICE based on joint probability modeling of the noisy and clean feature spaces using GMMs.

In the intermediate stages of the MMSE estimation, algorithms like RATZ, SPLICE, MMCN rely on approximations of the conditional distribution of clean and noisy features since its closed form solution is hard to estimate. Though SSM overcomes this limitation by deriving an exact conditional distribution from the joint probability model, a fundamental drawback still exists. Individual frames of an utterance are treated independent of each other during feature transformation. This often resulted in inappropriate dynamic characteristics. Addressing this problem, feature enhancement based on mapping sequence of frames (cepstral trajectory) was proposed in [27]. The motivation was based on successful applications of GMM-based trajectory mapping techniques for voice conversion [28].

The family of stochastic feature compensation algorithms till date remains a preferable choice for robust speech recognition tasks due to their relatively lower computational cost and reasonably good performance. An added advantage is their independence of any structural assumption about the nature of noise degradation. However, to the best of the authors' knowledge, the application of these techniques has not been studied extensively for robust SV tasks. In this paper we propose application of standard stochastic feature compensation methods in a SV framework. Through a comparative study of these methods, we highlight their significance for speaker verification in noisy environment.

The rest of the paper is organized as follows. Section 2 provides a brief introduction to stochastic feature compensation. Algorithmic descriptions of stereo-based feature compensation techniques used for a comparative study in the present work, are given in Section 3 and Section 4. The experiments conducted are discussed in Section 5, results and discussion in Section 6 followed by a brief summary and conclusion of the work in Section 7.

2. Stochastic feature compensation

Since accurate enumeration of the environmental effects on speech is a non-trivial task, a simplified version of speech signal degradation based on additive and convolutional channel noise is used in practice. It is assumed that the noise is statistically independent of speech while the convolutive channel distortions are linear time-invariant. In the cepstral domain a noisy mel frequency cepstral coefficient (MFCC) vector y_t is represented in terms of MFCC vectors of clean speech x_t, additive background noise n_t and channel noise h_t as follows

$$y_t = x_t + h_t + Clog(1 + \exp(C^{-1}(n_t - x_t - h_t))) \quad (1)$$

where t, C and C^{-1} are the time frame index, Discrete Cosine Transform (DCT) matrix and the inverse DCT matrix respectively. Due to the random nature of noise, a given clean feature vector can generate different noisy feature vectors, and vice-versa, which causes an uncertainty. Conventionally, Gaussian Mixture Models (GMMs) are used to represent the MFCC distribution. The additive noise in general alters the distribution by reducing the variance of each Gaussian component while the convolutional noise shifts the mean vectors. State-of-the-art model compensation techniques like PMC and VTS use the analytical relation in Eq. (1) and an available GMM of noise to adapt model parameters of noisy speech. In uncontrolled environments where Eq. (1) is not necessarily valid, faulty adapted parameters can be generated.

Stochastic feature compensation (SFC) methods are independent of any mathematical structure of noise degradation. They model stereo training data using GMMs. The effect of noise is represented as additive terms to the mean vectors and covariance matrices of the clean speech GMMs. Given a noisy test feature vector y_t, a minimum mean squared error (MMSE) criterion is used to estimate a clean vector $\hat{x}_t$ as follows

$$\hat{x}_t = E[x|y_t] = \int_X xp(x|y_t)dx \quad (2)$$

where x is a random variable representing clean feature vectors and $p(x|y_t)$ is the conditional probability distribution function (pdf) of x given y_t. Depending on the nature of the feature compensation algorithm, the two broad approaches of deriving $p(x|y_t)$ can be categorized as (i) independent probability modeling and (ii) joint probability modeling. The independent probability modeling methods construct individual GMMs for clean and noisy data.