Pattern Recognition Letters 75 (2016) 36–40

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Vector Taylor series based model adaptation using noisy speech trained hidden Markov models☆

Yongjoo Chung*

Department of Electronics Engineering, Keimyung University, Shindang-Dong 1000, Daegu, South Korea

ARTICLE INFO

Article history:
Received 19 January 2015
Available online 10 March 2016

Keywords:
Noisy speech recognition
Vector taylor series
Model parameter adaptation

ABSTRACT

Conventionally, in vector Taylor series (VTS) based compensation for noise-robust speech recognition, hidden Markov models (HMMs) are usually trained with clean speech. However, it is known that better performance is generally obtained by training the HMM with noisy speech rather than clean speech. From this viewpoint, we propose a novel VTS-based HMM adaptation method for the noisy speech trained HMM. We derive a mathematical relation between the training and test noisy speech in the cepstrum-domain using VTS and the mean and covariance of the noisy speech trained HMM are adapted to the test noisy speech in an iterative expectation-maximization (EM) algorithm. In the experimental results on the Aurora 2 database, we could obtain about 10–25% relative improvements in word error rates (WERs) over multi-condition training (MTR) method depending on speech front-ends and the HMM complexity.

1. Introduction

Speech recognition in noisy environments still remains a difficult problem despite many technical advances that have been made in this field. The techniques which try to mitigate the mismatch between test noisy speech and clean speech HMM can be generally categorized as noise-robust speech feature extraction, speech enhancement, and feature compensation and model parameter adaptation [1–6].

Among these methods, we focus on the HMM parameter adaptation methods because they are very powerful in improving noisy speech recognition accuracy compared with other approaches and many new research efforts are still actively going on in these areas. Maximum likelihood linear regression (MLLR) is one of the earliest and most popular model adaptation methods [6]. A set of linear transformations is used to adapt the means and covariance matrices of the HMM. Some variants of MLLR have been used to cope with the large amount of adaptation data required to estimate the transformation matrices [7]. Originally, developed for speaker adaptation, MLLR can be also effective for noise robust speech recognition. However, since it does not take into account actual noise corruption process, its performance is generally lower than the model adaptation methods such as parallel model combination (PMC) and VTS adaptation [3,4] that consider explicitly the relationship between clean and noisy speech signal.

Besides their superior performance, PMC and VTS require quite less adaptation data compared with MLLR. Especially, VTS adaptation is performed at runtime just using the test noisy speech and its performance is known to be better than PMC since the nonlinear approximation based on a first-order Taylor series expansion is more accurate than the lognormal approximation used in PMC. Joint uncertainty decoding (JUD) adapts HMM parameters by explicitly modeling the joint probability density function (PDF) of noisy and clean speech signal [8]. Compared with VTS and PMC, JUD requires less computation time by considering the joint PDF for a small set of regression classes rather than for each mixture component of the HMM at a small decrease in performance. The joint PDF in JUD can be obtained either by using stereo data consisting of noisy and clean speech signal or by applying the conventional model adaptation methods like PMC/VTS.

Although the aforementioned model adaptation methods have shown very successful results in noisy speech recognition, they use HMMs trained with clean speech as the baseline, which have some limitation in improving noisy speech recognition accuracy due to the inevitable mismatch between the observed test noisy speech and the adapted HMM parameters. This comes from the inaccuracies in the assumed noise corruption model and the mathematical approximations such as the first-order Taylor series expansion in the VTS adaptation.

As a different point of view from the model adaptation approaches for noise robust speech recognition, training HMMs

☆ This paper has been recommended for acceptance by Crocco Marco.

* Corresponding author. Tel.: +82 1035558767.
E-mail address: yjjung@kmu.ac.kr

http://dx.doi.org/10.1016/j.patrec.2016.02.013

directly with noisy speech has been proposed to show very promising results in noisy speech recognition [9–11]. For example, in the multi-condition training (MTR) method, noisy speech signals under various noise conditions are collected to train one set of HMMs [9]. Also, multiple HMM sets corresponding to various noise types and signal-to-noise ratio (SNR) values are constructed during training in a multiple-model based speech recognition (MMSR) framework [10]. The resulting noisy speech trained HMM significantly reduces the mismatch with the test noisy speech and performs much better than the clean speech HMM.

Some research efforts have been proposed to incorporate the conventional feature and model parameter compensation methods to the noisy speech trained HMM. This is motivated by the idea that the noisy speech trained HMM is more advantageous to performance improvement than the clean speech HMM. One of the earliest approaches to adapt the parameters of the noisy speech trained HMM is Jacobian adaptation (JA) [12]. Since it is based on a simple linear approximation of the nonlinear cepstral distortion, it seems to have difficulty in accurately reflecting the changing noise conditions into the HMM parameters. A noise-type dependent minimum mean square error (MMSE) estimation of the feature vectors has been applied successfully in the MMSR framework [13]. In our previous study, the nonlinear relation between the test and training noisy speech in the log-spectrum domain was used to re-estimate the test noisy speech feature vector to make it match better with the noisy speech trained HMM [14]. In recent studies, VTS and JUD based approaches have been popularly used in the feature and model compensation for the noisy speech trained HMM [15–17]. For example, in [15], an MTR trained HMM is transformed into a pseudo-clean HMM during training by using VTS. Then, the pseudo-clean HMM is used for recognition instead of the clean speech HMM to reduce environmental variations due to the noise leading to successful recognition results.

In this letter, we propose a new approach to adapt the parameters of the noisy speech trained HMM to the test noisy speech at recognition. It is based on a novel relation between the training and test noisy speech in the cepstrum domain. Compared with the recent model adaptation methods which adapt the pseudo clean speech HMM, the proposed algorithm is relatively simple in its implementation since it does not require estimate the parameters of the pseudo clean speech HMM [15]. In addition, the proposed method improves performance by adapting directly the MTR trained HMM, which makes it possible to take advantage of the inherent noise robustness of the noisy speech trained HMM.

2. VTS-based model adaptation

In this section, we derive an adaptation formula for the mean vectors and covariance matrices of the noisy speech trained HMM using VTS approximation. Contrary to the conventional VTS algorithms which require clean speech HMM, the noisy speech HMM trained by MTR method is used in the proposed adaptation algorithm. A novel relation between the training and test noisy speech is first derived in the cepstrum-domain. The non-linear relation is approximated using VTS to obtain the mean vectors and covariance matrices corresponding to the test noisy speech assuming that the additive and channel noises are known. An iterative EM algorithm is employed to update both the noise and HMM parameters.

2.1. Adaptation of HMM parameters

It is generally assumed that the clean speech $\mathbf{x}$ and the test noisy speech $\mathbf{y}$ contaminated with additive and channel noise $\mathbf{n}, \mathbf{h}$ is related in the cepstrum domain as follows:

$$\mathbf{y} = \mathbf{x} + \mathbf{h} + \mathbf{C}\log(\mathbf{i} + \exp(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{x} - \mathbf{h})) \tag{1}$$

where $\mathbf{i}$ is a unity vector and $\mathbf{C}$ and $\mathbf{C}^{-1}$ represent the discrete cosine transformation (DCT) and its inverse, respectively.

Assuming there is no channel noise in the training noisy speech$\mathbf{y}_{Tr}$, it can be expressed as follows:

$$\mathbf{y}_{Tr} = \mathbf{x} + \mathbf{C}\log(\mathbf{i} + \exp(\mathbf{C}^{-1}(\mathbf{n}_{Tr} - \mathbf{x})) \tag{2}$$

where $\mathbf{n}_{Tr}$ represents the additive noise in the training speech and is determined during training.

We can derive the following equation by taking inverse DCT and exponents on both sides of (2)

$$\exp(\mathbf{C}^{-1}\mathbf{x}) = \exp(\mathbf{C}^{-1}\mathbf{y}_{Tr}) - \exp(\mathbf{C}^{-1}\mathbf{n}_{Tr}) \tag{3}$$

Substituting (3) into (1) and (2), the relation between the test noisy speech $\mathbf{y}$ and training noisy speech $\mathbf{y}_{Tr}$ can be expressed as follows:

$$\mathbf{y} = \mathbf{y}_{Tr} + \mathbf{h} + \mathbf{g}(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}) \tag{4}$$

$$\begin{aligned}\mathbf{g}(\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}, \mathbf{n}_{Tr}) \equiv \mathbf{C}\log(\mathbf{i} + \exp(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{h} - \mathbf{y}_{Tr}))\\ - \exp(\mathbf{C}^{-1}(\mathbf{n}_{Tr} - \mathbf{y}_{Tr})))\end{aligned} \tag{5}$$

Eq. (4) is expanded using a first-order VTS [18] around the initial value $\{\mu_{y_{Tr},0}, \mu_{\mathbf{n},0}, \mathbf{h}_0\}$ of $\{\mathbf{y}_{Tr}, \mathbf{n}, \mathbf{h}\}$ as follows:

$$\begin{aligned}\mathbf{y} \approx \mu_{y_{Tr},0} + \mathbf{h}_0 + \mathbf{g}(\mu_{y_{Tr},0}, \mu_{\mathbf{n},0}, \mathbf{h}_0, \mathbf{n}_{Tr})\\ + \mathbf{G}_{y_{Tr}}(\mathbf{y}_{Tr} - \mu_{y_{Tr},0}) + \mathbf{G}_{\mathrm{n}}(\mathbf{n} - \mu_{\mathbf{n},0}) + \mathbf{G}_{\mathrm{h}}(\mathbf{h} - \mathbf{h}_0)\end{aligned} \tag{6}$$

where $\mathbf{G}_{y_{Tr}}$, $\mathbf{G}_h$ and $\mathbf{G}_n$ are the Jacobians of (4) with respect to $\mathbf{y}_{Tr}$, $\mathbf{h}$ and $\mathbf{n}$, respectively, which will be explained in more detail in the following.

Let $\mu_{y_{Tr},sm}$ and $\mathbf{\Sigma}_{y_{Tr},sm}$ denote the mean vector and diagonal covariance matrix of the mth Gaussian component in the sth state of the MTR trained noisy speech HMM. The additive noise $\mathbf{n}$ of the test noisy speech is assumed to be Gaussian with mean μ_n and covariance $\mathbf{\Sigma}_n$. Also, the alignment between speech frame and the corresponding Gaussian component of the HMM does not alter due to the change in noise conditions.

By taking the expected value of the terms in (6), the mean vector $\mu_{y,sm}$ and covariance matrix $\mathbf{\Sigma}_{y,sm}$ of the adapted HMM can be estimated as follows:

$$\begin{aligned}\mu_{y,sm} = \mu_{y_{Tr},sm} + \mathbf{h}_0 + \mathbf{g}(\mu_{y_{Tr},sm}, \mu_{n,0}, \mathbf{h}_0, \mu_{n_{Tr}})\\ + G_{n,sm}(\mu_n - \mu_{n,0}) + \mathbf{G}_{h,sm}(\mathbf{h} - \mathbf{h}_0)\end{aligned} \tag{7}$$

$$\mathbf{\Sigma}_{y,sm} \approx \mathbf{G}_{y_{Tr},sm}\,\mathbf{\Sigma}_{y_{Tr},sm}\mathbf{G}^T_{\mathbf{y}_{Tr},sm} + \mathbf{G}_{n,sm}\mathbf{\Sigma}_n\mathbf{G}^T_{n,sm} \tag{8}$$

$$[\mathbf{G}_{y_{Tr},sm}]_{il} = \sum_k C_{ik}\frac{C^{-1}_{kl}}{1 + B_k - A_k} \tag{9}$$

$$\left[\mathbf{G}_{h,sm}\right]_{il} = \sum_k C_{ik}\frac{C^{-1}_{kl}\,(1 - A_k)}{1 + B_k - A_k} \tag{10}$$

$$[\mathbf{G}_{n,sm}]_{il} = \sum_k C_{ik}\frac{C^{-1}_{kl}\,B_k}{1 + B_k - A_k} \tag{11}$$

$$A_k = \exp\left(\sum_j C^{-1}_{kj}(\mu_{n_{Tr,j}} - \mu_{y_{Tr},sm,j})\right) \tag{12}$$

$$B_k = \exp\left(\sum_j C^{-1}_{kj}(\mu_{\mathbf{n},0,j} - \mathbf{h}_{0,j} - \mu_{y_{Tr},sm,j})\right) \tag{13}$$

$$C_{ik} = [\mathbf{C}]_{ik}, C^{-1}_{ik} = \left[\mathbf{C}^{-1}\right]_{ik} \tag{14}$$