Applied Acoustics 74 (2013) 640–646

Contents lists available at SciVerse ScienceDirect

Applied Acoustics

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

# Maximum likelihood subband polynomial regression for robust speech recognition

Yong Lü [a,b,*], Zhenyang Wu [b]

[a] College of Computer and Information Engineering, Hohai University, Nanjing 210098, China
[b] School of Information Science and Engineering, Southeast University, Nanjing 210096, China

ARTICLE INFO

*Article history:*
Received 10 January 2012
Received in revised form 2 September 2012
Accepted 20 November 2012
Available online 28 December 2012

*Keywords:*
Model adaptation
Subband polynomial regression
Hidden Markov model
Robust speech recognition

ABSTRACT

In this paper, we propose a model adaptation algorithm based on maximum likelihood subband polynomial regression (MLSPR) for robust speech recognition. In this algorithm, the cepstral mean vectors of prior trained hidden Markov models (HMMs) are converted to the log-spectral domain by the inverse discrete cosine transform (DCT) and each log-spectral mean vector is divided into several subband vectors. The relationship between the training and testing subband vectors is approximated by a polynomial function. The polynomial coefficients are estimated from adaptation data using the expectation–maximization (EM) algorithm under the maximum likelihood (ML) criterion. The experimental results show that the proposed MLSPR algorithm is superior to both the maximum likelihood linear regression (MLLR) adaptation and maximum likelihood subband weighting (MLSW) approach. In the MLSPR adaptation, only a very small amount of adaptation data is required and therefore it is more useful for fast model adaptation.

© 2012 Elsevier Ltd. All rights reserved.

## 1. Introduction

The robustness issue is crucial for speech recognition in real applications because the mismatch between training and testing conditions often degrades the recognition performance significantly. This mismatch is due to the additive background noise, channel distortion (convolutional noise), speaker and other factors. Generally speaking, the methods used to reduce the environmental mismatch can be classified into two major categories: the front-end feature domain methods and the back-end model domain methods.

In the feature domain, spectral subtraction (SS) [1], cepstral mean normalization (CMN) [2] and relative spectra (RASTA) [3] are commonly used to reduce the impact of noise for automatic speech recognition. Besides, model-based feature compensation is also an effective approach to noise robust speech recognition, which is simultaneously proposed by Erell and Weintraub [4] and Acero [5], and further studied in [6–9]. This algorithm typically employs a Gaussian mixture model (GMM) to represent the distribution of the speech feature and uses the minimum mean squared error (MMSE) method to reconstruct the original speech feature. The noise parameters, which are used for transforming the prior trained speech model to the testing condition, are estimated from the noisy speech [6,7] or from the silence duration of the incoming speech [8,9]. In order to obtain the closed-form solution of the noise parameters from noisy speech features, the vector Taylor series (VTS) expansion [6] is proposed to approximate the nonlinear relationship between clean and noisy speech. Another VTS-based noise log-spectral estimation approach is studied in [7] where the maximum a posteriori (MAP) criterion is used instead of the maximum likelihood (ML) criterion. Although the feature domain methods can achieve significant performance improvements in noisy speech recognition, they are not effective in reducing the environmental mismatch resulting from other factors, such as speaker.

In the model domain, the MAP adaptation [10] and maximum likelihood linear regression (MLLR) [11] are two popular adaptation algorithms. The MAP adaptation estimates model parameters by optimally interpolating the associated adaptation data with the prior parameters of hidden Markov models (HMMs). This algorithm only updates the parameters of models which are observed in the adaptation data, and thus fairly large amounts of adaptation data are generally required. MLLR is a transform-based adaptation algorithm, which transforms the overall HMMs by a set of linear regression functions. Compared to MAP, MLLR has better performance when a small amount of adaptation data is available. A model adaptation algorithm using piecewise-linear transformation is studied in [12] where various types of noise are clustered according to their spectral property and signal-to-noise ratio (SNR), and a set of noisy speech HMMs is trained for each cluster. In the recognition phase, the HMM set that best matches the input noisy speech is selected and further adapted using the MLLR method. The maximum likelihood estimation is frequently unreliable in the case of sparse adaptation data. In the maximum a posteriori

* Corresponding author at: College of Computer and Information Engineering, Hohai University, Nanjing 210098, China. Tel.: +86 25 58099120.
*E-mail address:* yonglu@hhu.edu.cn (Y. Lü).

0003-682X/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.apacoust.2012.11.016

linear regression (MAPLR) adaptation [13], the prior knowledge of regression parameters is incorporated into the linear regression estimation to solve the sparseness problem. In theory, MAPLR works better than MLLR when the prior probability density of regression parameters is estimated properly. Besides, some noise adaptation methods, such as parallel model combination (PMC) [14], linear spectral transformation (LST) [15,16] and the VTS adaptation [17], have been proposed for noisy speech recognition. These methods are very effective for noise compensation, but do not take account of other environmental factors such as speaker. Therefore, they cannot be applied to speaker adaptation or minimizing other environmental mismatches.

In recent years, the subband approaches [18–22] have been proposed to improve the robustness of speech recognition against band-limited additive noise. In these approaches, the channels in the full-band filter bank are divided into several subbands, usually of equal partitions, and the subband feature vector is computed from each partition using the discrete cosine transform (DCT). The subband approaches can improve the recognition performance in the presence of narrow-band noise, but may degrade the baseline performance for clean speech because the sub-band features lose the correlation among subbands. In [23], a maximum likelihood subband weighting (MLSW) approach is proposed, where the subband features or subband means are multiplied with weighting factors and then combined and converted to the cepstral domain. The experimental results show that MLSW is more robust than both full-band approaches and conventional subband approaches. However, it achieves higher performance than MLLR only in the case of a very small amount of adaptation data and its performance does not increase obviously with the growth of adaptation data.

In this paper, we propose a model adaptation algorithm based on maximum likelihood subband polynomial regression (MLSPR) for robust speech recognition. Firstly, the channels of Mel filter bank (Mel channels) are divided into several subbands of equal partitions, and the cepstral mean vectors of prior trained HMMs are converted to the log-spectral domain by the inverse DCT. Then each log-spectral mean vector is divided into several subband vectors and each subband vector is transformed to the testing condition by a polynomial function. In other words, all the channels of each subband share a polynomial transformation, which can further improve the robustness of transform-based model adaptation algorithms. The polynomial coefficients are estimated from adaptation data using the expectation–maximization (EM) algorithm [24] under the maximum likelihood criterion. Finally, the estimated subband mean vectors are combined and converted to the cepstral domain.

The rest of this paper is organized as follows. Section 2 describes the MLSPR algorithm. The estimation of polynomial coefficients is given in Section 3. The experimental procedures and results are presented and discussed in Section 4. Section 5 concludes the paper with a summary.

## 2. Subband polynomial regression

In the MAP adaptation [10], the prior trained HMM parameters including the state transition probabilities, mixture weights, mean vectors, and covariance matrices are adapted to the test condition. If enough adaptation data is provided, the MAP estimate is proved to asymptotically approach to the retrained system. Nevertheless, in the test environment, the available adaptation data is often limited. Thus the transform-based adaptation methods, such as MLLR [11], are proposed to improve the robustness of model adaptation, where the data that belong to various Gaussian mixture components are combined to estimate a set of transformation parameters. However, the data sparseness is still a difficult problem to be solved when a small amount of adaptation data is available.

The Mel-frequency cepstral coefficient (MFCC) vector is the most commonly used speech feature for speech recognition. In the cepstral domain, there are weak correlations among the different components of the MFCC vector and thus the different components of the HMM mean vector cannot share the same transformation, while in the log-spectral domain, the adjacent channels of Mel filter bank are overlapped and therefore it can be assumed that the transformation function of one channel is similar to those of its adjacent channels. In this work, the channels of Mel filter bank (Mel channels) are divided into several subbands of equal partitions and all the channels of each subband share a polynomial transformation. There are two reasons why the polynomial regression is used to approximate the relationship between the training and testing subbands. One is that the environmental transformation of each channel is nonlinear and the polynomial regression can approximate any nonlinear function. The other is that there are always some differences among the different channels of one subband and the polynomial regression can cover the different channels better than the linear regression and weighting adaptation.

The proposed subband polynomial regression proceeds as follows:

1. The prior trained cepstral HMM mean vector $\boldsymbol{\mu}$ is converted to the log-spectral domain by the following equation:

$$\boldsymbol{u} = \boldsymbol{C}^{-1}\boldsymbol{\mu} \tag{1}$$

where $\boldsymbol{C}^{-1}$ denotes the inverse DCT matrix and $\boldsymbol{u}$ is the log-spectral mean vector. If the high order coefficients of the MFCC vector are ignored, the full MFCC vector can be extracted from training data and then the MAP adaptation [10] is employed to estimate the full cepstral mean vector for Eq. (1). The original cepstral mean vector can also be padded with zeros to obtain the approximate value of the full mean vector.

2. The log-spectral mean vector $\boldsymbol{u} = [u_1, u_2, \ldots, u_D]^T$ is divided into $K$ subbands:

$$\boldsymbol{u} = [\boldsymbol{m}_1^T, \boldsymbol{m}_2^T, \ldots, \boldsymbol{m}_K^T]^T \tag{2}$$

where $D$ is the number of Mel channels and the superscript $T$ denotes the transpose of the matrix or vector. The mean vector $\boldsymbol{u}$ can also be expressed as the sum of $K$ subband mean vectors,

$$\boldsymbol{u} = \sum_{k=1}^{K} \boldsymbol{u}_k \tag{3}$$

where $\boldsymbol{u}_k = [0, \ldots, 0, \boldsymbol{m}_k^T, 0, \ldots, 0]^T$. The decomposition procedure is illustrated in Fig. 1 where the log-spectral mean vector is decomposed into four subband mean vectors.

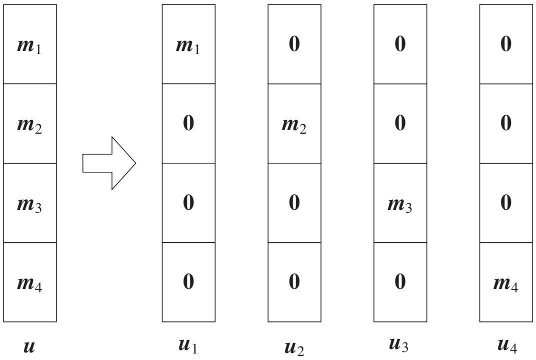

**Fig. 1.** Decomposition of mean vector.

Download English Version:

https://daneshyari.com/en/article/754630

Download Persian Version:

https://daneshyari.com/article/754630

Daneshyari.com