Data-driven UBM Generation via Tied Gaussians for GMM-Supervector Based Accent Identification

Rong Zheng, Ce Zhang, Bo Xu

Digital Content Technology Research Center, Institute of Automation
Chinese Academy of Sciences, Beijing 100190, China
{rzheng,czhang,xubo}@hitic.ia.ac.cn

10.21437/Interspeech.2011-325

Abstract

This paper presents a new approach to exploit data-driven universal background model (UBM) generation using tied Gaussians for accent identification (AID). The motivation of the proposed algorithm is to potentially utilize broad phonetic-specific accent characteristics by Gaussian mixture model (GMM) and examine data-driven phonetically-inspired UBM creation for GMM-supervector based accent classification. In this work, we discuss the issues involved in applying cumulative posterior probability based Gaussian selection and tree structure based UBM parameter estimation. Derivation and validation of the UBM refined by tied Gaussians are reported in this paper. Performance evaluations comparing our system with other well-known techniques for AID are also provided. Better performance is further achieved by fusing these acoustic-based accent classifiers. Comparison experiments conducted on the CSLU foreign-accented English (FAE) dataset show the effectiveness of the proposed method.

Index Terms: accent identification, GMM-supervector, tied Gaussians, Gaussian selection, UBM generation

1. Introduction

Accent can be defined as a pattern of pronunciation, grammar and vocabulary of a language used by the community of nonnative speakers belonging to some geographical area. Accent classification (especially foreign accent identification) is the task of automatically detecting the accent of a foreign speaker from a speech utterance. In this paper, we focus on accent classification in foreign-accented English.

Although language identification (LID) and dialect identification (DID) are extensively studied, less works have been done in the area of AID. Some techniques widely used for LID have been gradually extended for AID [1-6]. However, accent classification is more challenging because of two aspects. Firstly, some linguistic knowledge, such as syllable structure, may be relatively consistent for different accents of the same language. Secondly, compared with LID task, the capacities of the popular LID approaches, especially phonotactic techniques, may be limited to capture subtle phonotactic differences [6].

Techniques used for LID may be applicable to AID if they only consider acoustic information. In [3], detection rate of 13.2% was reported for 23 accents using naive Bayes accent classification on the FAE database (refer to section 3.1 for details). Choueiter et al. obtained 32.7% detection accuracy based on cumulative application of HLDA, MMI and Gaussian Tokenization [4]. The authors in [5] demonstrated improvements in AID using an automatic speech recognition (ASR) dependent phonetically-inspired approach. When combined with ASR front-end (many ASR-dependent techniques involved, i.e. MLLT, FMLLR, FMPE) and a novel supervector representation, 56.1% accuracy was achieved.

In this paper, we propose a data-driven tied Gaussians based UBM generation algorithm for GMM-supervector (GSV) accent identification. Generally, each individual component Gaussian is interpreted to represent some broad acoustic classes (underlying broad phonetic sounds) [7], which are used to describe potentially phonetic-specific accent characteristics here. A tree structure is built to model the whole acoustic space via tied Gaussians. Maximum likelihood (ML)-trained UBM in conventional GSV is then refined by cumulative posterior probability based Gaussian selection and parameter estimation.

The remainder of this paper is organized as follows. Section 2 describes the proposed algorithm of data-driven UBM creation via tied Gaussians for GSV based AID. In section 3, the corpus and different AID techniques are presented. The experimental results and performance comparisons are shown in section 4. Finally, some conclusions are given in section 5.

2. Description of the proposed method

The application of support vector machine (SVM) in GMM-supervector space has provided interesting results [8]. In conventional GSV system, Kullback-Leibler (KL) divergence is used to measure the distance of two GMMs. Using some approximation to the KL divergence for two GMM probability distributions, the KL kernel function that satisfies the Mercer condition can be given as follows,

$$K_{KL}(X,Y)=\sum_{c=1}^{C}[\sqrt{\omega_c}\,\Sigma_c^{-\frac{1}{2}}(\mu_c^{X\_adapt}-\mu_c^{ubm})]^t[\sqrt{\omega_c}\,\Sigma_c^{-\frac{1}{2}}(\mu_c^{Y\_adapt}-\mu_c^{ubm})] \quad (1)$$

where ω_c, μ_c^{ubm} and Σ_c are the c-th mixture weights, mean and diagonal covariance matrix of UBM, $\mu_c^{X\_adapt}$ and $\mu_c^{Y\_adapt}$ are the *maximum a posteriori* (MAP) adapted means corresponding to the component c for two speech utterances, X and Y.

In the GMM-supervector modeling for AID task, a UBM is trained using speech utterances from a large group of accent utterances to represent the characteristics of all different accents. Each utterance is represented by a high dimensional normalized mean-concatenated vector through MAP adaptation. It is known that Gaussian components can be considered to model some underlying broad phonetic sounds. Many of the accent

Supported by National Natural Science Foundation of China (No. 90820303)

variations are conditioned on phonetic classes, phonetic sequences, and syllabic classes [5]. However, ML-trained UBM in conventional GSV may be not enough to capture all the accent characteristics and there is some accent-dependent phonetic information loss. In this section, we propose a data-driven phonetically-inspired UBM generation, which considers using tied Gaussians for UBM refinement to capture the potentially broad phonetic-specific accent information.

Fig.1 shows a general block diagram of the proposed phonetically-inspired UBM creation method. The details of the algorithm's functional blocks and related issues are discussed in the next paragraphs.

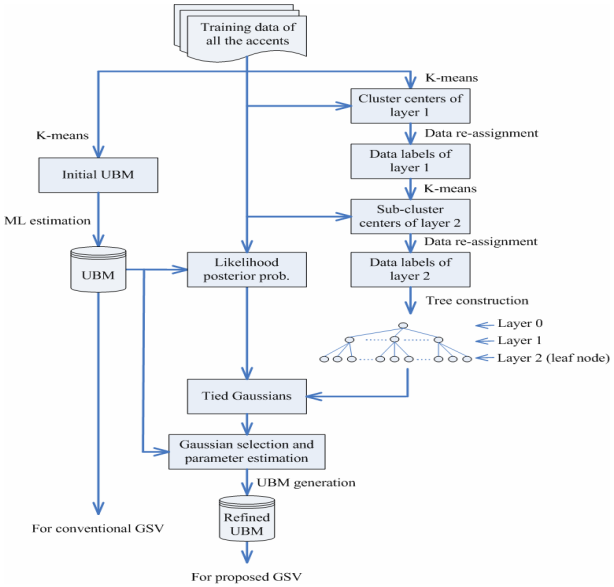

Figure 1: *Block diagram of the proposed tied Gaussians based phonetically-inspired UBM generation.*

2.1. Data-driven tied Gaussians

For the training data of all accents, a two-layer tree, as shown in Fig.1, can be constructed to model the acoustic space hierachically. Firstly, the cluster centroids of layer-1 are estimated directly from the acoustic features by K-means clustering. All of the acoustic features are re-assigned to certain cluster via Euclidean distance between the cluster centroid and feature vector, so data labels for the input features are obtained. Secondly, based on the data labels of layer-1, similar procedures are performed for layer-2. Thirdly, for a sequence of cluster-dependent feature vectors, the sufficient statistics of individual Gaussian components are calculated on the ML-trained UBM.

$$N_{(n,c)} = \sum_{x_t \in n(l_1,l_2)} \Pr(c / x_t) = \sum_{x_t \in n(l_1,l_2)} \frac{\omega_c p_c(x_t)}{\sum_{j=1}^{C} \omega_j p_j(x_t)} \quad (2)$$

where $N_{(n,c)}$ is the zeroth order *Baum-Welch* statistics for leaf node n (i.e. index l_1 in layer-1 and index l_2 in layer-2) and mixture c in the ML-trained UBM. $\Pr(c / x_t)$ is the c -th component likelihood posterior probability of feature vector x_t .

For each leaf node n , tied Gaussians can be associated by posterior probability $q_{(n,c)}$, $q_{(n,c)} = N_{(n,c)} / \sum_{j=1}^{C} N_{(n,j)}$, in the descending order for further use. Each leaf node of the tree structure represents a cluster of Gaussian components tied in ML-trained UBM and could be modeled by a single Gaussian probability distribution function (PDF) for the refined UBM.

2.2. UBM generation using Gaussian selection and parameter estimation

The tree structure models the acoustic space with different Gaussian-dependent information, that is, posterior probabilities are included to build the tree. Underlying broad phonetic information captured by tied Gaussians could be utilized to generate new UBM for GSV classifier. It is known that each phone is expected to be modeled by several mixtures, so before creating the new UBM model, two problems should be solved: (1)which Gaussians should be selected for each leaf node to retain phonetically discriminative regions and suppress common accent acoustic scope? (2)what are the parameter estimation equations for Gaussian PDF? For the first question, we use the cumulative posterior probability to filter out unimportant tied Gaussians. For each leaf node, if the cumulative posterior probability of top tied Gaussians is larger than the preset threshold, the corresponding Gaussian components of the ML-trained UBM are selected. For the second question, two kinds of parameter estimation methods are provided as follows.

By taking into account of the posterior probabilities, the mixture weights, mean and covariance of a set of selected tied Gaussians, R , the ***ML-approximated*** parameter estimation equations for a leaf node n with D -dimensional vector are given,

$$\hat{\omega}_n = \frac{\sum_{c \in R} q_{(n,c)} \omega_c}{\sum_{c \in R} q_{(n,c)}} \quad (3)$$

$$\hat{\mu}_n(d) = \frac{\sum_{c \in R} q_{(n,c)} \omega_c \mu_c(d)}{\sum_{c \in R} q_{(n,c)} \omega_c} \quad (4)$$

$$\hat{\sigma}_n^2(d) = \frac{\sum_{c \in R} q_{(n,c)} \omega_c (\sigma_c^2(d) + \mu_c^2(d))}{\sum_{c \in R} q_{(n,c)} \omega_c} - \hat{\mu}_n^2(d) \quad (5)$$

where $\hat{\omega}_n$, $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are the weight, mean and covariance of the n -th mixture component of the generated UBM.

In [9], a KL-divergence based approach for cluster centroid estimation was employed. By including cumulative posterior probability based Gaussian selection and approximations to the KL divergence, the PDF parameters of ***KL-approximated*** estimation are derived (c.f. [9,5] for details).

There is some similar work reported by scientific literature in [5]. They combined ASR-dependent front-end and phonetically-inspired UBM for AID. Note that the authors in [5] applied many ASR-dependent techniques and clustered the ASR acoustic model, which consisted of 250K Gaussian components to achieve the final supervector representation. It's not difficult to discriminate our work from [5]. First of all, only low-level acoustic-based system is performed in this work, that is, no ASR system is needed. Secondly, all of the models used in this work (including intersession compensation matrices, feature transformation model, et.) are only trained on the training set of FAE corpus. Finally, the procedures of UBM generation are substantially different and the acoustic resolution of generated UBM could be flexibly adjusted due to tree structure, which will be explained in section 4.

3. Data and AID systems

3.1. Database

Experiments are conducted on the CSLU foreign-accented English (FAE) dataset [10]. The corpus consists of 4925 telephone-quality utterances spoken by 23 accents, which involve variable-length speech utterances (the average duration is 17.9 seconds in length). For the FAE task, we use the same dataset configuration in [4], that is, 70% for training, 15% for development and 15% for test respectively.

3.2. Feature extraction

All acoustic systems use the common feature set. The 7-1-3-7 configuration of the shifted delta cepstra (SDC) concatenated with 7 static PLP coefficients are calculated to produce 56 dimensional feature vectors. The feature streams are processed through energy-based speech activity detection to eliminate low-energy acoustic vectors. Mean/variance normalization is used to remove the linear channel effects. Feature warping and feature domain intersession compensation (FDIC, not used in section 3.3.1) are then applied to reduce session variability [11].

3.3 Experimental setup

3.3.1. TVM

Total variability modeling (TVM, also referred as I-vector) outlined by Dehak et al. [12] simultaneously models the speaker and channel variability in only one space. Let F and C be the acoustic feature dimension and the total number of Gaussian mixture components, the GMM-supervector is a CF - dimensional vector which is formed by concatenating the means of each mixture component. Given an utterance, the speaker- and channel- dependent GMM-supervector, M , is written as,

$$M = m + Tw \quad (6)$$

where m is speaker- and channel- independent supervector representing the center of the full parameter space, T is a rectangular matrix of low rank and w is a standard normally distributed random vector. T and w are commonly referred as total variability space and total factor, separately.

Here we introduce TVM for AID. The total factors are extracted regardless of accent variability and channel variability. And therefore, we perform linear discriminant analysis (LDA) prior to within-class covariance normalization (WCCN) [13] for intersession compensation in the total factor space, which was successfully applied for speaker recognition [12].

In this work, the dimension of the total factor is experimentally set to 400. Due to limited accent classes, i.e. 23 accents in FAE corpus, we split the multi-session training set of each accent to multiple sub-accent classes according to the utterance similarity, which totally produces 221 sub-accent classes. Then, LDA and WCCN are applied with reduced dimension of 200. However, TVM and TVM-LDA-WCCN give similar performances on the FAE dataset.

3.3.2. GLDS-SVM

The GLDS-SVM (generalized linear discriminant sequence) system utilizes the polynomial-based sequence kernel as presented in [14]. Three-degree of polynomial expansion is used to form the input supervector (32509 dimensions) of SVM training, which is built with LIBSVM library.

3.3.3. MMI

The GMM-MMI system is based on the work described in [15]. GMM models with 512 Gaussians per accent are trained with maximum mutual information (MMI) criterion.

3.3.4. GSV

The GSV system uses a kernel based on an approximation to the KL divergence [8]. Each input utterance is represented by the difference of mean vectors of the UBM model and MAP-adapted model normalized by the corresponding standard deviation. Except where otherwise indicated, 512 Gaussian mixtures are used for GSV-based AID in this paper.

3.3.5. The proposed method

The proposed algorithm mentioned in section 2 is investigated. We report the experimental results and analyze the effects of cumulative posterior probability based threshold, two kinds of parameters estimation, UBM order, as well as the number of nodes in the tree structure.

4. Experiments and results

4.1. Experiments on comparative systems

Table 1 lists the detection rates of comparative AID systems on the development and test set, respectively.

Table 1. *Comparisons of detection rates of comparative systems on the development and test set.*

Systems	Development set	Test set
TVM	24.59%	25.00%
GLDS-SVM	24.86%	25.27%
MMI	34.92%	35.46%
GSV	34.24%	35.33%

As shown in Table 1, we remark that: (1) the AID performances of MMI and GSV are very similar, which are slightly higher than 32.7% detection rate reported in [4]; (2) the classifier detection rates on TVM and GLDS-SVM are worse. It seems that TVM for AID is not as powerful as in speaker recognition task. This discrepancy can be explained to some extent by less observable variation and generally small amount of training data for accent and intersession variability modeling. In future, an in-depth study should be investigated further.

4.2. Gaussian selection and parameter estimation

To provide some perspective on data-driven tie Gaussians based GSV AID system, experiments are conducted as a function of the cumulative posterior probability based threshold. The ML-approximated and KL-approximated parameter estimation of refined UBM are also compared. The performances on development set and test set are illustrated in Fig.2.

The cumulative posterior probability varies from 40% to 95% (If the largest posterior probability of tied Gaussian is higher than the preset threshold, one tied Gaussian component

at least is used). As depicted in Fig.2, the proposed system mostly benefits from the neighborhood of 60%, where the average number of Gaussian components selected from the tied Gaussians is 6.2 for each leaf node, which seems to be reasonable to the assumption that each phone is expected to be modeled by several mixtures. In view of the overall situation, the accuracy rates of ML-approximated parameter estimation are better than that of KL-approximated. So 60% for cumulative posterior probability threshold and ML-approximated parameter estimation are used for the next experiments.

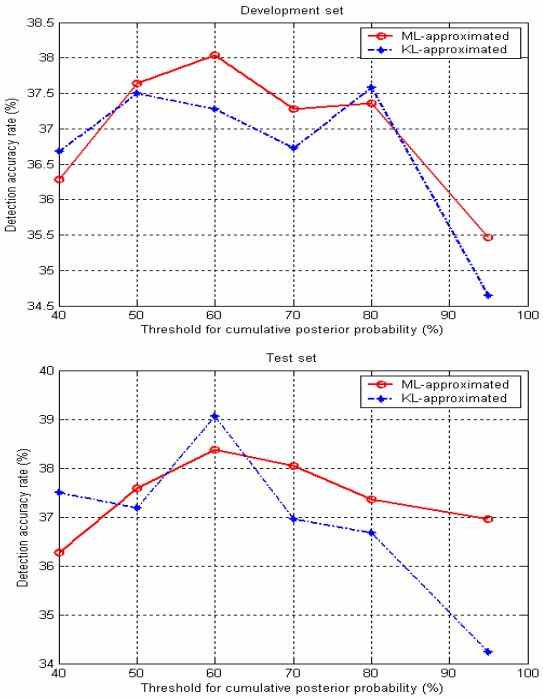

Figure 2: *Detection rate on the development and test set as a function of cumulative posterior probability based threshold and parameter estimation of UBM generation.*

4.3. UBM order

UBM of orders 512, 1024 and 2048 are compared for conventional GSV and our proposed algorithm. The detailed results are provided in Table 2. For experimental simplicity in our 2-layer tree construction, the generated UBM order of 512 consists of 32 nodes in layer-1 and 16 nodes as its descendants (i.e. 512=32x16, then 1024=32x32, 2048=64x32 are used here).

Table 2. *Comparisons of detection rates as a function of UBM order for conventional GSV and the proposed method.*

Systems		Development set	Test set
512Mix	GSV	34.24%	35.33%
	Proposed	38.04%	38.38%
1024Mix	GSV	36.41%	36.14%
	Proposed	39.53%	40.08%
2048Mix	GSV	37.50%	36.82%
	Proposed	41.58%	43.07%

The results indicate that the performance of proposed method is stably better than the conventional GSV system.

4.4. System fusion

The output scores of individual systems are fused with linear combination. The weighting factors for fusion are optimized on the development set and then directly applied for the test set. UBM of order 2048 is used for GSV and the proposed method.

Table 3. *Summary of detection rates for system fusion on the development and test set.*

Systems	Development set	Test set
Proposed	41.58%	43.07%
Proposed+GLDS-SVM+MMI+GSV	44.70%	44.97%
Proposed+GLDS-SVM+MMI+GSV+TVM	44.83%	45.10%

As reported in Table 3, the system performance can be further enhanced by fusing our proposed algorithm with GLDS-SVM, MMI, GSV and TVM, which achieves 44.83% detection accuracy rate on the development set and 45.10% on the test set of 23-way FAE corpus.

5. Conclusions

In this paper we investigate a data-driven UBM generation via tied Gaussians for GSV based accent classification over FAE dataset. We propose a new method of incorporating posterior probability of tied Gaussians into UBM refinement, and then compare it with other modeling techniques for AID, i.e. TVM, GLDS-SVM, MMI and the conventional GSV. The resulting system consistently leads to considerable performance gains. Comparison experiments show that UBM generation using tied Gaussians performs better and it is efficient to enhance the GSV system in the accent identification task.

6. References

[1] Wu, T.Y., Duchateau, J. et al.: Feature subset selection for improved native accent identification. Speech Communication, Vol.52, pp.83-98, 2010

[2] Torres-Carrasquillo, P.A., Sturim, D.E. et al.: Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition. Proc. of InterSpeech, pp.723-726, 2008

[3] Macias-Guarasa, J.: Acoustic adaptation and accent identification in the ICSI MR and FAE corpora. ICSI Meeting slides, 2003

[4] Choueiter, G., Zweig, G., Nguyen, P.: An empirical study of automatic accent classification. Proc. of ICASSP, pp.4265-4268, 2008

[5] Omar, M.K., Pelecanos, J.: A novel approach to detecting non-native speakers and their native language. Proc. of ICASSP, pp.4398-4401, 2010

[6] Liao, Y.F., Yeh, S.C., et al.: Latent prosody model-assisted mandarin accent identification. Proc. of ROCLING, 2009

[7] Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. SAP, pp.72-83, 1995

[8] Campbell, W.M. et al.: Support vector machines using GMM supervectors for speaker verification. IEEE Sign. Process. Lett. 13(5), pp.308-311, 2006

[9] Xiang, B., Berger, T.: Efficient text-independent speaker verification with structural Gaussian mixture models and neural network. IEEE Trans. SAP, 2003

[10] CSLU.: Foreign-accented English corpus. Available: http://www.cslu.ogi.edu/corpora/fae/

[11] Castaldo, F. et al.: Compensation of nuisance factors for speaker and language recognition. IEEE Trans. ASLP, Vol.15, pp.1969-1978, 2007

[12] Dehak, N., Kenny, P. et al.: Front-end factor analysis for speaker verification. IEEE Trans. ASLP, Vol.19, pp.788-798, 2011

[13] Hatch, A.O. et al.: Within-class covariance normalization for SVM-based speaker recognition. Proc. of InterSpeech, pp.1471-1474, 2006

[14] Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. Proc. of ICASSP, pp.161-164, 2002

[15] Matejka, P. et al.: Brno university of technology system for NIST 2005 language recognition evaluation. Proc. of IEEE Speaker Odyssey, 2006