ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Two-stage Speaker Adaptation Approach for Subspace Gaussian Mixture Model based Nonnative Speech Recognition

Bo, LI and Khe Chai, SIM

School of Computing, National University of Singapore, Singapore
{li-bo,simkc}@comp.nus.edu.sg

10.21437/Interspeech.2012-483

Abstract

Nonnative speech recognition is becoming more and more important as many speech applications are deployed world wide. Meanwhile, due to the large population of nonnative speakers, speaker adaptation remains the most practical way for providing high performance speech services. Subspace Gaussian Mixture Model (SGMM) has recently been shown to yield superior performance on various native speech recognition tasks. In this paper, we investigated different speaker adaptation techniques of SGMM for nonnative speech recognition. A two-stage direct model adaptation approach has been proposed based on the analysis of SGMM model parameter functionalities. Our initial experiments have also verified that the proposed approach is much more effective than the traditional feature-space Maximum Likelihood Linear Regression(MLLR) on SGMM based nonnative speaker adaptation tasks.

Index Terms: Speaker Adaptation, Nonnative Speech Recognition, Subspace Gaussian Mixture Model

1. Introduction

More and more automatic speech recognition (ASR) based voice activated applications have been deployed into market during recent years, especially with the launch of Siri on iPhone 4S. However, due to the large amount of time and efforts required to build a robust ASR system for a particular language, only a small set of major languages are supported in some applications and most only provide English services. For these applications to reach out to the global customers, the vast variations (including both regional and personal speaking characteristics) in the realization of English phonemes between nonnative users and the native training speakers pose a major challenge.

Building speaker dependent (SD) systems or even just region dependent systems is still impractical in most services for general public due to the requirement of amassing sufficient amount of data from each speaker or region. Speaker adaptation techniques, which eliminate training and testing mismatches by adapting the speaker independent (SI) model's parameters to the new individual user's voice with a small amount of data from that user, could bridge up the gap between the SI and SD performances.

On account of the small amount of data available and the large number of model parameters to estimate, the essence of speaker adaptation techniques is thus to constrain the degree of parameter updates. Generally speaking, there are two major categories of constraints. Transformation based constraints are rather popular, which usually estimate linear transforms of either Gaussian means or variances or both from the adaptation data, such as Maximum Likelihood Linear Regression (MLLR) [1] for Gaussian Mixture Model (GMM) based Hidden Markov Model (HMM). This approach and its variations are commonly adopted because of their simplicity and robustness. However, the assumption of linear transformations for modeling speaker characteristics is hard to validate and may limit the adaptation capability. An exponential transform has been recently shown to be promising for speaker adaptation [2]. The other kind of approaches interpolate parameter estimations between training and adaptation processes such as Maximum a Posterior (MAP). Due to the inherent likelihood based estimation, model parameters without sufficient observations in the adaptation data are not updated. We could hence deem MAP as constraining the updates by first selecting only a subset of model parameters and then interpolate them with training parameters. The major concern for MAP like approaches is how to ensure a robust parameter selection and interpolation based estimation.

Subspace Gaussian Mixture Model (SGMM) [3] has recently attracted large interests for ASR. It constrains the whole parameter space into a much smaller subspace. The compact Gaussian sharing and subspace modeling merits give SGMM much more robustness against variations and lead to promising results compared to the conventional GMM based systems. It would thus be interesting to apply SGMM to nonnative speech recognition. This paper investigated the application of SGMM for nonnative speech recognition and proposed a two-stage direct model adaptation method for SGMM. Experimental results have shown that SGMM consistently outperforms GMM with large improvements and further error reductions could be achieved using the two-stage model adaptation method which has also been shown to be robust with a small amount of nonnative data.

The rest of the paper is organized as follows: The SGMM is detailed in Section 2. Section 3 presents both the existing adaptation methods and our proposed two-stage direct model adaptation approach. Experimental results are reported in Section 4 and Section 5 discusses the possible future research and concludes the paper.

2. Subspace Gaussian Mixture Model

The sub-state version of the basic Subspace Gaussian Mixture Model (SGMM) [3, 4] is adopted in this work. In SGMM-HMM, the distribution of features $\mathbf{x}$ in the HMM state j is a mixture of sub-state distributions, each of which itself is a mixture of Gaussians:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i) \qquad (1)$$

where c_{jm} is the sub-state mixture weight and the number of sub-states M_j for each state j is independent and proportional to a small power of the state data count. The same number of

mixture components I is used for all the sub-states. The covariances $\mathbf{\Sigma}_i$ are shared across sub-states, also across states. Unlike in a conventional GMM, the mean vectors μ_{jmi} and mixture weights w_{jmi} are not directly estimated as parameters of the model. Instead, a particular sub-state jm is associated with a vector $\mathbf{v}_{jm}$ which determines the means and weights. The mean vectors are derived as:

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} \tag{2}$$

where the parameters $\mathbf{M}_i$ are shared across the state distributions and define the subspace in which the GMM mean vectors can live. The mixture component weights are derived from the vector $\mathbf{v}_{jm}$ using a log-linear model

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}, \tag{3}$$

with a globally shared parameter $\mathbf{w}_i$ determining the mapping. To simplify following elaborations, the SGMM parameters are tabulated in Table 1.

Table 1: *Summary of SGMM model parameters.*

Parameters	Explanations
$\mathbf{\Sigma}$	Gaussian full covariance
$\mathbf{M}$	mean projection matrix
$\mathbf{w}$	weight projection vector
$\mathbf{v}$	sub-state vector
$\mathbf{c}$	sub-state coefficient vector

3. Speaker Adaptation for SGMM

3.1. Feature-space MLLR

Feature-space MLLR (fMLLR), also known as Constrained MLLR [5], uses an affine transform for the feature vectors,

$$\mathbf{x}^{(s)} = \mathbf{A}^{(s)} \mathbf{x} + \mathbf{b}^{(s)}, \tag{4}$$

where the transform parameters $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ are estimated to maximize the likelihood of the transformed speaker-specific data under the acoustic model. Commonly a more compact notation is adopted,

$$\mathbf{x}^{(s)} = \mathbf{W}^{(s)} \mathbf{x}^{+}, \tag{5}$$

where $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)}, \mathbf{b}^{(s)}]$, and $\mathbf{x}^{+} = [\mathbf{x}^T, 1]^T$ denotes a vector $\mathbf{x}$ extended with a 1.

For each speaker transform $\mathbf{W}^{(s)}$, there are totally $d^2 + d$ parameters, where d is the dimensionality of the feature vectors. When small amounts of per speaker adaptation data are present, these estimations may not be reliable. In such cases, a basis representation of fMLLR transforms is proposed in [6]. $\mathbf{W}^{(s)}$ is thus represented as a linear combination of a set of basis transforms $\mathbf{W}_b$ ($1 \leq b \leq B$),

$$\mathbf{W}^{(s)} = \mathbf{W}_0 + \sum_{b=1}^{B} \alpha_b^{(s)} \mathbf{W}_b, \tag{6}$$

where $\mathbf{W}_0 = [\mathbf{I}; \mathbf{0}]$ which is included to serve as the default transform. The bases $\mathbf{W}_b$ are estimated from the entire training set, and essentially define a subspace for the fMLLR transforms. For each speaker, only the coefficients α_b need to be estimated, where typically $B \ll d(d+1)$.

When the underlying probability model has a diagonal covariance structure, the fMLLR transform can be estimated by iteratively updating the matrix rows [5]; a similar row-by-row update has also been proposed for the full covariance case [7]. While in this work, we adopted the recently proposed gradient based approach, which is designed to be easily combined with subspace techniques and is as efficient or more efficient than standard approaches[6]. It directly maximizes the likelihood function by repeated line search in the direction of the gradient and the estimation is done in a pre-transformed parameter space such that an approximation to the expected Hessian is proportional to the unit matrix.

3.2. Two-stage direct model adaptation

According to SGMM parameters' phonetic dependencies, we could group them into three categories:

- The phonetically independent parameters: $\mathbf{\Sigma}$. These Gaussian covariance parameters, which are shared among all the states, model the generic speech acoustic variations of the training dataset. They have no awareness of the high level phonetic information. Those language and region related information is thus believed to be captured in these parameters.
- The partially phonetically dependent parameters: $\mathbf{M}$ and $\mathbf{w}$. These parameters transform each sub-state vector to a state specific mean vector and mixture weight. In other words, they project the large state specific parameter space to a smaller subspace. Hence, the estimation of these parameters has the awareness of the state information. However, these parameters themselves are shared and are not specific to any state. We could thus deem them as intermediate parameters linking the acoustic and phonetic information.
- The state dependent parameters: $\mathbf{v}$ and $\mathbf{c}$. These parameters directly model the states and thus define the phonetic distributions.

As for nonnative speech recognition, the largest mismatch comes from the language differences between native and nonnative speakers' mother tongues and the speaker specific characteristics sever as the second major difference. Based on SGMM parameters' functionalities and nonnative speech characteristics, we thus proposed a two-stage speaker adaptation method for nonnative speech recognition. Firstly, we pool the observations for all the different states together to robustly adapt the SI model's phonetically independent parameters $\mathbf{\Sigma}$. Due to the limited availability of adaptation data, there are usually few observations for each state. It would be better to pool them together for robust estimations to reduce the dramatic language/region mismatches. Secondly, all the shared Gaussian parameters, namely $\mathbf{M}$, $\mathbf{w}$ and $\mathbf{\Sigma}$ are jointly adapted in a maximum likelihood manner to refine the language/region adapted model to further capture speaker specific variations.

4. Experiments

This section presents the initial experimental results of speech recognition on the nonnative English from a Mandarin speaker. Due to the unavailability of nonnative speech corpus at hand, the author recorded his own speech using prompts from the standard WSJ0 5K closed vocabulary task's test and development sets. A summary of the dataset used in this work is

Table 2: *Summary of the dataset used in this work.*

Dataset		Amount (minutes)	# of Utterances
Test		26.4	179
Adaptation	adapt1	10.4	70
	adapt2	32.8	210

tabulated in Table 2. There are totally 26.4 minutes of testing data and 43.2 minutes of adaptation data. A randomly selected 10-minute subset "adapt1" is used as the main adaptation data for following experiments and the others are only used to evaluate the effect of varying the adaptation data amount. A GMM system (2350 tied triphone states and 33.7k Gaussians) and a SGMM system (2386 tied triphone states, 8k substates and 400 shared full covariance Gaussians per sub-state) are trained on the SI-84 WSJ0 native corpus using commonly adopted 39D MFCC features (MFCC_E_D_A_Z). In SGMM, the dimension of the subspace vector is 40 leading to an almost square mean projection matrix $\mathbf{M}_i$. For word recognition, the test data are decoded with the WSJ0 5K trigram language model (LM) and Word Error Rates (WER) are reported; while for phoneme recognition, the TIMIT bigram phoneme language model is adopted and Phoneme Error Rates (PER) are stated.

4.1. Speaker adaptation techniques

Table 3: *WER (%) on test data for both GMM and SGMM on nonnative speech recognition.*

Methods	GMM	SGMM
SI	63.51	49.69
fMLLR	32.03	38.83
fMLLR_regtree	**27.04**	-
fMLLR_basis	-	37.08
mean	65.93	19.24
full	71.07	**17.32**

Existing speaker adaptation techniques are firstly investigated to adapt the native GMM and SGMM, which have WERs of 4.33% and 3.18% respectively for the native 5K WSJ0 task. Nonnative recognition results are presented in Table 3. Comparing the two baseline SI models, SGMM has an absolute 13.82% lower WER than GMM, indicating its better generalization capability brought by its compact Gaussian sharing structure; however, the WER is still quite high compared with native speech recognition.

Feature-space Maximum Likelihood Linear Regression (fMLLR) is one of the most popular speaker adaptation methods for HMM systems. It reduces almost half of the errors for the GMM model, from 63.51% to 32.03% with a single global transform, and with 9 regression classes (fMLLR_regtree in Table 3), which is the optimal for our case, it further improves the performance to 27.04% WER. However, the gain from fMLLR on SGMM is relatively small, only 10.86% absolute WER reduction. Even with 200 bases estimated on the native training corpus, the fMLLR_basis approach only yields another 1.75% absolute improvement. This may be explained by the compact Gaussian sharing structure. Unlike GMM-HMM, all the Gaussians are independent, linear transforms such as fMLLR, especially regression tree based fMLLR, have more freedom to refine the model parameters towards nonnative speech.

The robustness of SGMM itself and the ineffectiveness of fMLLR for SGMM lead us to the idea of directly updating the model parameters. Two sets of experiments are carried out: updating the Gaussian means only and re-estimating all the parameters with the small amount of adaptation data, which are indicated as "mean" and "full" in Table 3 respectively. In SGMM, the mean projection matrix is updated to represent the "mean" update. The more the amount of parameters updated, the larger the degradation of GMM is. By contrast, with all the parameters updated, SGMM could achieve a 17.32% WER, a relative 65% improvement over the SI SGMM. This result has demonstrated that the compact Gaussian sharing structure in SGMM makes the model estimation much more robust even with a small amount of adaptation data.

Table 4: *WER(%) of adapting different SGMM parameters.*

	$\mathbf{cv}$	$\mathbf{\Sigma Mw}$	$\mathbf{Mw}$	$\mathbf{\Sigma}$	$\mathbf{M}$	$\mathbf{w}$
WER	22.85	**15.12**	18.59	26.60	19.24	46.67

Table 5: *Comparison between different two-stage parameter adaptations for SGMM.*

Settings	Stage1 WER(%)	Stage2 WER(%)
$\mathbf{\Sigma Mw_\Sigma Mw}$	15.12	15.43
$\mathbf{Mw_\Sigma Mw}$	18.59	17.25
$\mathbf{M_\Sigma Mw}$	19.24	16.56
$\mathbf{\Sigma w_\Sigma Mw}$	23.47	14.05
$\mathbf{\Sigma_\Sigma Mw}$	26.60	**13.99**

4.2. Adaptation towards speaker dependent SGMM

To find the optimal set of parameters that leads to the best performance in previous experiments (i.e. "full" in Table 3), further investigations are carried out and the results are reported in Table 4. Firstly, the shared parameters $\mathbf{\Sigma Mw}$ jointly yield a 7.73% lower WER than the state specific ones $\mathbf{cv}$. Furthermore, the phoneme aware parameters $\mathbf{Mw}$ perform better than the phonetically independent $\mathbf{\Sigma}$, but is worse than the joint adaptation, which clearly shows the effectiveness of the second stage joint update in our proposed approach. To further verify the proposed two-stage approach, we compared several different two-stage settings in Table 5. It can be found that re-estimating the mean parameters $\mathbf{M}$ is more effective in the first stage(i.e. $\mathbf{Mw_\Sigma Mw}$ and $\mathbf{M_\Sigma Mw}$). But with Gaussian covariance parameters updated first and then a joint update of $\mathbf{\Sigma Mw}$, the $\mathbf{\Sigma_\Sigma Mw}$ setup, yields the best performance.

In addition, all the adaptation data (i.e. both "adapt1" and "adapt2") are pooled together to evaluate the robustness of the proposed approach against the amount of data, which is illustrated in Figure 1. Both word recognition and phoneme recognition are carried out for this experiment. With no more than 10 minutes of adaptation data, both the WER and the PER decrease quickly with the increase of data amount; but the improvement slows down using even more data. Meanwhile, with only 1 minute of adaptation data, our approach has a WER of 30.10% which is already better than the 37.08% WER got from the fMLLR_basis adaptation for SGMM with 10 minutes of data.

4.3. Generalization on IDEA dataset

To further justify the proposed approach for nonnative speech recognition, we experimented with nonnative recordings from the International Dialects of English Archive (IDEA). The

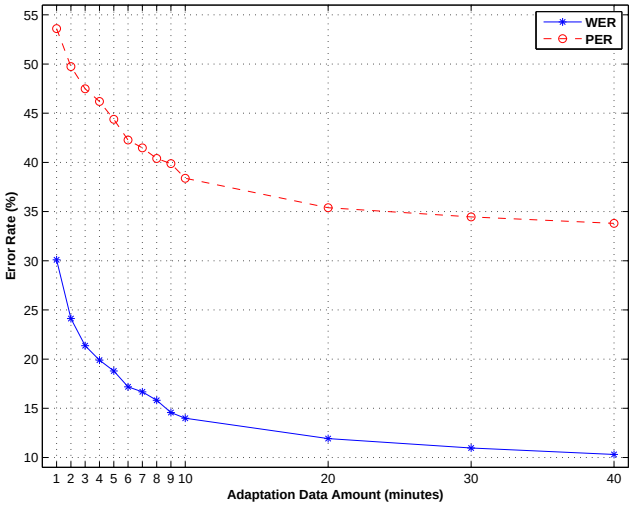


Figure 1: *WER(%) and PER(%) on test data with different amounts of adaptation data.*

Table 6: *WER (%) and PER (%) on the Jiang Su, China subset of the IDEA dataset.*

Settings			Average	spkA	spkB
WER	GMM	SI	82.24	54.91	93.68
		fMLLR_spk	73.99	44.51	90.80
	SGMM	SI	**76.45**	51.45	89.08
		fMLLR_spk	73.02	48.55	90.23
		$\mathbf{\Sigma}$_$\mathbf{\Sigma M w}$_spk	**66.83**	43.93	81.03
		fMLLR_set	74.40	52.60	90.80
		$\mathbf{\Sigma}$_$\mathbf{\Sigma M w}$_set	**60.30**	**39.31**	**71.84**
PER	GMM	SI	59.93	48.69	64.27
		fMLLR_spk	54.27	41.31	60.77
	SGMM	SI	56.88	42.79	63.94
		fMLLR_spk	54.24	40.33	64.44
		$\mathbf{\Sigma}$_$\mathbf{\Sigma M w}$_spk	49.45	39.67	52.25
		fMLLR_set	55.49	42.79	62.27
		$\mathbf{\Sigma}$_$\mathbf{\Sigma M w}$_set	**45.93**	**34.75**	**49.25**

IDEA consists of various audio recordings of nonnative speakers worldwide. For Mandarin speakers, each recording consists of three short parts: an English reading passage, some unscripted speech and a Chinese reading text. In this study, only the English reading part is used. Among all the 26 speakers from Jiang Su, China, 3 recordings with background noises are excluded. For each speaker, there are roughly 2.68 minutes of data, which is equally split into adaptation and test sets. Due to the domain mismatch, in this experiment we adopted WSJ0 20K trigram LM for word recognition. Even though, the LM still has a 605.53 perplexity and 3.82% OOV rate. While for phoneme recognition, the same TIMIT bigram LM is used. The average performance is tabulated in Table 6 under the column "Average". Although the WER is relatively high, the PER is in the similar level as previous experiments. Same trends can also be observed from both the WER and PER results. SI SGMM performs better than SI GMM and for speaker adaptation (indicated by "_spk"), $\mathbf{\Sigma}$_$\mathbf{\Sigma M w}$ has a relative 12.58% WER improvement compared with fMLLR's 4.49% over the SI SGMM. Additionally, the less than 2 minutes of adaptation data per speaker may not be sufficient to capture all the acoustic variations. We thus pooled all the 23 speakers' adaptation data together and did a test set adaptation (indicated as "_set" in Table 6). Another 6.53% absolute WER reduction is achieved. Moreover, beside the nonnative accents, other nonnative characteristics such as insertions of unnecessary pauses, repetitions and so on could also cause recognition errors. We hence compared the recognition performance per speaker and the performance of the best, spkA, and the worst, spkB, are listed in the last two columns of Table 6. The WERs have a much larger difference than the PERs and human listening suggests that spkA is much more fluent than spkB in speaking. This implies that, besides the accent, other nonnative aspects such as prosody should be accounted for a better word recognition performance.

5. Conclusions

In this study, we have investigated speaker adaptation techniques for the SGMM based nonnative speech recognition. It shows that due to SGMM's inherent Gaussian sharing structure and subspace modelling nature, the model parameter estimation is much more robust to small amounts of data. We proposed a two-stage direct model adaptation approach to adapt the SGMM to nonnative speakers according to the theoretical analysis of the parameter functionalities. Initial experimental results using both self-recorded data and the IDEA nonnative data have verified the effectiveness of the proposed approach over conventional fMLLR for nonnative speaker adaptation in SGMM. Although further investigation and verification of the proposed approach on large nonnative corpus is needed, this work still has its good practical implications that before turning to complicated adaptation methods for nonnative speaker adaptation of SGMM, apply our proposed two-stage approach first to adapt the SI model due to its simplicity. In future, we will continue the investigation of SGMM for nonnative speech recognition and especially on how to incorporate other nonnative characteristics such as disfluencies into the recognition process.

6. Acknowledgements

The authors would like to thank Professor Paul Meier, the Founder and Director of International Dialects of English Archive (IDEA), for his permission to the use of the Chinese subset of the IDEA recordings. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

7. References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer speech and language*, vol. 9, no. 2, 1995.

[2] D. Povey, G. Zweig, and A. Acero, "Speaker adaptation with an exponential transform," in *Proc. ASRU*, 2011.

[3] D. Povey, L. Burget *et al.*, "Subspace gaussian mixture models for speech recognition," in *Proc. ICASSP*, 2010.

[4] D. Povey, "A tutorial-style introduction to subspace gaussian mixture models for speech recognition," *Microsoft Research*, 2009.

[5] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech and language*, vol. 12, no. 2, 1998.

[6] A. Ghoshal, D. Povey *et al.*, "A novel estimation of feature-space MLLR for full-covariance models," in *Proc. ICASSP*, 2010.

[7] K. C. Sim and M. J. F. Gales, "Adaptation of precision matrix models on large vocabulary continuous speech recognition," in *Proc. ICASSP*, 2005.