SUBSPACE MIXTURE MODEL FOR LOW-RESOURCE SPEECH RECOGNITION IN CROSS-LINGUAL SETTINGS

Yajie Miao, Florian Metze, Alex Waibel

Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{ymiao, fmetze, ahw}@cs.cmu.edu

ABSTRACT

The subspace Gaussian mixture model (SGMM) has been exploited for cross-lingual speech recognition. The general motivation is that the subspace parameters can be estimated on multiple source languages and then transferred to the target language. In this work, we investigate an extension to SGMM, referred to as subspace mixture model (SMM), in which subspace parameters on the target language are casted as a linear mixture of the subspaces derived from source languages. This approach reduces the number of SGMM model parameters, while retaining the flexibility of subspace learning on the target language. Experiments show that the proposed SMM method outperforms SGMM significantly when the target language has limited training data.

Index Terms— Subspace models, acoustic modeling, cross-lingual speech recognition

1. INTRODUCTION

State-of-the-art speech recognition systems rely on a large amount of transcribed data to robustly estimate HMM-GMM acoustic models. These data can be easily obtained for rich-resource languages such as English and Mandarin. However, for languages with only limited speech and resources, the acquisition of large data collections becomes both expensive and challenging. The subspace Gaussian mixture model (SGMM) [1, 2] has been proposed to deal with this data sparseness. Instead of estimating GMM parameters directly, SGMM learns lower-dimension subspaces which are able to capture the main phonetic and speaker variability. The subspaces can be shared across HMM states, which results in a more compact representation and helps to shrink the size of model parameters. Previous studies have shown that SGMM can outperform the conventional GMM consistently, especially when we have limited speech data for acoustic modeling.

There has been considerable interest in using SGMM for multilingual and cross-lingual acoustic modeling [3, 4, 5, 6]. In multilingual scenarios, the SGMM subspace parameters are estimated based on combined statistics from multiple languages [3]. This effectively gives us more data for subspace learning and improves the robustness of parameter estimation. Another application of SGMM is in cross-lingual speech recognition where a group of source languages are available. The general motivation is to improve the performance on the target language which may be under-resourced. On this aspect, Lu et al. proposed to estimate the globally shared subspace parameters on the source languages and transfer them to the target language [5]. However, since the subspace parameters are fixed for the target system, this method may suffer from mismatch between the source and target languages. To alleviate this effect, a maximum a posterior (MAP) adaptation approach was introduced in [6], where the source subspace serves as prior for the subspace of the target language.

In this paper, we further develop this line of ideas and propose the subspace mixture model (SMM) for cross-lingual acoustic modeling. Specifically, for the target language, SMM learns its subspace parameters through a linear combination of the subspaces from source languages. Cross-lingual speech recognition in this manner is attractive since incorporating or eliminating new source languages is easy to achieve. In contrast, the direct transfer [5] and MAP [6] methods have to re-estimate the shared subspace parameters on the source languages from scratch. Moreover, compared with [5], SMM is a more flexible framework in that the target language can tune its subspace, or equivalently mixture coefficients, on the available training data. We present maximum likelihood estimation (MLE), as well as implementation considerations, for SMM. Our focus is on the scenario where the target language has highly limited training data. On the GlobalPhone corpus [7], we experimented with two low-resource conditions with 1.2 hours and 3 hours of target language data respectively. The proposed SMM method achieves consistent reduction on WER compared with the SGMM baseline.

2. REVIEW OF SGMM

In conventional HMM-GMM acoustic models, the emission probability of each state is modeled with a Gaussian mixture model (GMM). In SGMM, each state or substate is also represented with a GMM. The major difference is that GMM means are derived from phonetic and speaker

subspaces while mixture weights are from a set of weight projections [1]. In addition, we have I covariance matrices which are normally full rather than diagonal. Formally, on each Gaussian index i, the phonetic and speaker subspaces are $\mathbf{M}_i$ and $\mathbf{N}_i$, the weight projection is $\mathbf{w}_i$, the covariance matrix is $\mathbf{\Sigma}_i$. Then, the SGMM model for state j on speaker s can be expressed as

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

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)} \qquad w_{jmi} = \frac{\exp(\mathbf{w}_i^T \mathbf{v}_{jm})}{\sum_{i'=1}^{I} \exp(\mathbf{w}_{i'}^T \mathbf{v}_{jm})} \qquad (1)$$

where $\mathbf{x}(t) \in \mathbb{R}^D$ is a D-dimension feature vector, m is a substate of j, and i is a Gaussian component. The subspace parameters, weight projections and covariance matrices are shared by all the HMM states. SGMM has state-specific parameters including substate vectors $\mathbf{v}_{jm} \in \mathbb{R}^S$ and substate mixtures c_{jm} , where S is the dimension of the phonetic subspace. The GMM means can be obtained from phonetic subspace and substate vectors. To model speaker variability, there is a speaker-specific offset to the mean vector, derived from the speaker subspace $\mathbf{N}_i$ and speaker vector $\mathbf{v}^{(s)} \in \mathbb{R}^T$ where T denotes the dimension of the speaker subspace. In this work, we do not consider speaker adaptive training for SGMM, and thus exclude the speaker subspace from the model. It can be seen that the SGMM parameters in fact span a subspace of the entire GMM parameter space. The subspace parameters can be shared and collaboratively estimated over multiple languages or domains. Therefore, SGMM has been studied for multilingual and cross-lingual speech recognition.

3. THE SUBSPACE MIXTURE MODEL

The general idea of cross-lingual SGMM [5] is shown in Figure 1(a). The subspace parameters are estimated on the source languages using multilingual SGMM [3], and then used in acoustic modeling on the target language. The parallel arrays indicate that the subspace of the source languages and the subspace of the target language are equivalent. In spite of promising results, this approach has limitations. For instance, since the subspace is fixed for the target language, we lose the flexibility of tuning subspace parameters on the target language. Also, this manner overemphasizes the generality of the source subspace, but ignores the potential mismatch (channel conditions, noise levels, speaking styles, etc) between the source and target languages. More importantly, if we want to add or remove source languages, we need to estimate the new source subspace, which is expensive especially when there is sufficient training data on the source languages.

Figure 1(b) depicts the motivation behind our SMM method. Instead of learning one single source subspace,

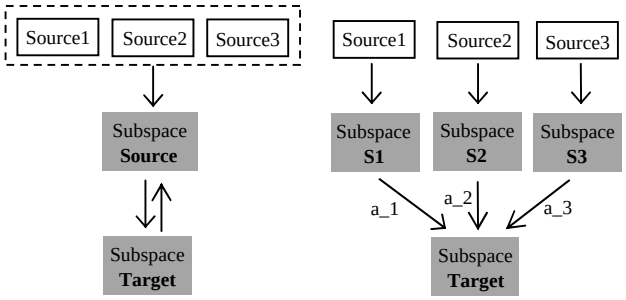

Figure 1: Illustrations for (a) cross-lingual SGMM and (b) SMM. The markers on the arrays in (b), i.e., a_1, a_2 and a_3, represent mixture weights for source languages.

SMM learns the subspace separately for each source language. Then, the target subspace is formed by a linear combination of the source subspaces. The multiple source subspaces bear information regarding individual source languages. The target language "assembles" its subspace with the mixture weights based on its own acoustic characteristics. Compared with cross-lingual SGMM, SMM has a more flexible manner of subspace learning, while at the same time reducing model parameters dramatically.

We examine the phonetic subspace firstly. For each source language g, we model its *i-th* phonetic subspace matrix $\mathbf{M}_i^{(g)}$ with the SGMM model. On the target language, SMM derives the corresponding subspace $\mathbf{M}_i$ by a linear mixture of the source languages:

$$\mathbf{M}_i = \sum_{g=1}^{L} a_i^{(g)} \mathbf{M}_i^{(g)} \qquad (2)$$

where $a_i^{(g)}$ is the mixture weight for $\mathbf{M}_i^{(g)}$, L is the total number of source languages.

In SMM, state-specific parameters $\mathbf{v}_{jm}$ and c_{jm} are estimated similarly as in SGMM. Our goal here is on the estimation of the mixture weights $\mathbf{a}_i$ which has the dimension of L. In the context of SGMM, ML estimation is presented in [1] where the auxiliary function involving the phonetic subspace is

$$Q_{SGMM}(\mathbf{M}_i) = \sum_{t,j,m} ¥_{jmi}^{(t)} \Big[\mathbf{x}(t)^T \mathbf{\Sigma}_i^{-1} \mathbf{M}_i \mathbf{v}_{jm} - \frac{1}{2} \mathbf{v}_{jm}^T \mathbf{M}_i^T \mathbf{\Sigma}_i^{-1} \mathbf{M}_i \mathbf{v}_{jm} \Big] \qquad (3)$$

where $¥_{jmi}^{(t)}$ is the occupancy of state j, substate m and Gaussian i conditioned on the observation $\mathbf{x}(t)$ from the target language, i.e., $¥_{jmi}^{(t)} = p(j,m,i \mid \mathbf{x}(t))$, the superscript T represents matrix transpose.

By substituting Eq. (2) into (3), we get the SMM auxiliary function w.r.t. the mixture weights:

$$Q_{SMM}(\mathbf{a}_i) = \mathbf{a}_i^T \mathbf{b}_i - \frac{1}{2} \mathbf{a}_i^T \mathbf{\Phi}_i \mathbf{a}_i \qquad (4)$$

where the $L \times 1$ vector $\mathbf{a}_i$ contains the mixture weights for $\mathbf{M}_i$, the $L \times 1$ vector $\mathbf{b}_i$ is the first-order statistics collected with state occupancy and its *g-th* element is

$$\mathrm{b}_i^{(g)} = \sum_{t,j,m} ¥_{jmi}^{(t)} \mathbf{x}(t)^T \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i^{(g)} \mathbf{v}_{jm} \qquad (5)$$

The $L \times L$ symmetric matrix $\boldsymbol{\Phi}_i$ is the second-order statistics where the (l, g) element is:

$$\boldsymbol{\Phi}_i^{(l,g)} = \sum_{j,m} ¥_{jmi} \mathbf{v}_{jm}^T \mathbf{M}_i^{(l)T} \boldsymbol{\Sigma}_i^{-1} \mathbf{M}_i^{(g)} \mathbf{v}_{jm} \qquad (6)$$

where $¥_{jmi}$ is the aggregated occupancy $¥_{jmi} = \sum_t ¥_{jmi}^{(t)}$. If $\boldsymbol{\Phi}_i$ is non-singular, we can take the derivative of $Q_{SMM}(\mathbf{a}_i)$ to be 0 and obtain $\mathbf{a}_i$ as follows:

$$\hat{\mathbf{a}}_i = \arg\max_{\mathbf{a}_i} Q_{SMM}(\mathbf{a}_i) = \boldsymbol{\Phi}_i^{-1} \mathbf{b}_i \qquad (7)$$

However, in practice, $\boldsymbol{\Phi}_i$ may be singular or have poor condition. Thus, we adopt a generalized solution for this optimization problem, i.e., the `solve_vec` function described in [1]. ML estimation of $\mathbf{a}_i$ based on expectation maximization (EM) can be summarized as follows.

1) Initialize all the mixture weights to be $1/L$
2) Collect statistics $\boldsymbol{\Phi}_i$ and $\mathbf{b}_i$ according to Eq. (5) and (6), using the source subspaces and current model parameters.
3) Update mixture weights $\mathbf{a}_i$ via `solve_vec` and get the new phonetic subspace based on Eq. (2)
4) Go to step 2 until converged.

In principle, we can apply SMM to the speaker subspace and weight projections as well. But in low-resource cases, the phonetic subspace takes a large proportion of SGMM model parameters (see Section 4). Therefore, SMM is only applied to the phonetic subspace in this work. During optimization, we impose no constraints on $\mathbf{a}_i$. In fact, $\mathbf{a}_i$ can be constrained to be mixture probabilities, meaning that its elements should sum up to 1. Also, to further improve robustness, we can add to Eq. (4) a regularization factor, such as l_1 norm, on $\mathbf{a}_i$. We leave the investigation of these constraints to our future work.

4. EXPERIMENTS

The performance of SMM is evaluated on the GlobalPhone corpus [7] which has been widely used in multilingual speech recognition. This corpus contains up to 19 languages. We take German (GE) as the target language, and Spanish (SP), Portuguese (PO), Swedish (SW) as the source languages. Table 1 summarizes their statistics in terms of the amount of training and development data.

Table 1. Dataset statistics and monolingual SGMM models.

	SP	PO	SW	GE
training (h)	17.6	22.7	17.4	14.9
dev (h)	2.0	1.6	2.0	2.0
# of states	2493	2310	2400	2527
# of substates	20k	20k	20k	20k
WER of SGMM	28.6	22.8	35.2	21.4

We firstly build monolingual SGMM models on individual source languages to get the source subspaces. On all the languages, we use a 13-dimensional MFCC front-end including the C0 energy and its first, second derivatives with per-speaker mean normalization. An LDA transform reduces the feature dimension to 40, on which MLLT is applied. On top of the LDA+MLLT context-dependent model, we construct the ML-SAT system using fMLLR [8]. In the speaker-adapted feature space, a universal background model (UBM) is trained over all the source languages, by clustering diagonal Gaussians in their ML-SAT acoustic models. In our experiments, it is observed that a shared UBM is critical for the performance of SMM, since this guarantees the phonetic subspace to be aligned across source languages. Starting from this common UBM, SGMM is trained separately on each source language, with the fMLLR transforms fixed. We set the number of Gaussians I = 400, and the phonetic subspace dimension S = 40. For recognition, fMLLR adaptation with respect to ML-SAT is performed on the testing data, i.e., development set. The resulting testing speaker transforms are used in SGMM decoding, together with trigram language models. The performance of monolingual SGMMs, as well as the number of tied states and substates, are shown in Table 1. An SGMM model is also trained on the complete German data. Because of speaker adaptation and more language model training materials, we obtain better WER than the results in [5].

4.1. Cross-lingual Experiments with 1.2 Hour data

On the target language German, two levels of data sparseness are investigated. In the first case, 1.2 hour data is selected from the complete training set. The ML-SAT model and fMLLR transforms are estimated on this subset. The baseline SGMM totally has 695 triphone tied states. During SMM training, the same clustering decision tree from SGMM is used. The phonetic subspace parameters are estimated according to Section 3, while the others are updated similarly with SGMM. Table 2 gives a comparison between SGMM and SMM regarding the number of model parameters when we have 4k substates. In SGMM, the phonetic subspace takes almost half of all the parameters. After applying SMM, parameters attached with the phonetic subspace are reduced significantly, which helps to shrink the total size of model parameters.

The recognition performance of SGMM and cross-lingual SMM on the German development set is shown in Figure 2. The SMM model exploiting multiple source languages

Table 2. Comparison of model parameters sizes between SGMM and SMM. The number of substates is 4k.

	SGMM	SMM
all the parameters	13×10^6	6.6×10^6
phonetic subspace	6.4×10^6	1200

results in considerably lower WER compared with the SGMM baseline. The only exception is when the number of substates is 1k. This is because the relatively small number of SMM parameters cannot fully capture speech variability. As we increase the number of substates, the best WER achieved by SMM is 28.3%, while the best WER of SGMM is 30.2%. Also, we observe from Figure 2 that removing one of the source languages degrades the performance of SMM. This indicates that the three source languages are complementary with each other in constructing the phonetic subspace of the target language.

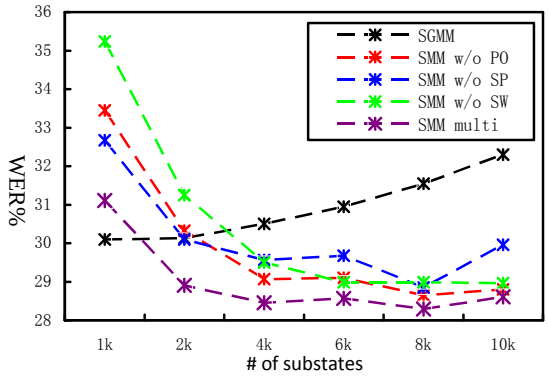

Figure 2. Performance of SGMM and SMM on 1.2 hour training data. Here "multi" means all the source languages are used and "w/o SP" means Spanish is excluded from the source languages.

4.2. Cross-lingual Experiments with 3 Hour data

In the second case, we increase the amount of training data to 3 hours, on which the SGMM and SMM models are trained. Both models share the same decision tree and have 1114 tied states. The phonetic subspace dimension S is still 40. Figure 3 shows the comparison between SGMM and SMM in terms of WER. Again, the SMM model with all the source languages ("SMM multi") achieves the best WER (24.0%), while the best WER of SGMM is 25.1%. However, the gains of SMM over SGMM become less significant compared with on the 1.2 hour training set. This is because in SGMM the phonetic subspace now takes less than 2/5 of the whole set of parameters. Moreover, excluding one of the source languages degrades the performance of SMM, which is consistent with what we observe on the 1.2 hour data.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose the subspace mixture model (SMM) for cross-lingual speech recognition. This SMM approach can effectively reduce the number of model parameters in SGMM, which makes it suitable for under-resourced target languages. Experiments with the Global-Phone corpus show that SMM outperforms the SGMM baseline significantly under two conditions of data sparseness. As mentioned in Section 3, for future work, we

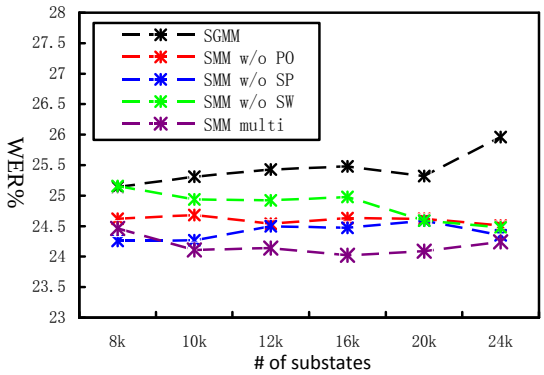

Figure 3. Performance of SGMM and SMM on 3 hour training data as the number of substates increases.

will extend SMM to speaker subspace and weight projections, to further reduce the model size. Another focus will be on investigating various forms of regularization on the source language mixture weights. This is expected to add sparsity to the mixture weights and enhance the robustness of SMM estimation.

6. ACKNOWLEDGMENTS

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0015. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

7. REFERENCES

[1] D. Povey, L. Burget, M. Agarwal, et al, "The subspace Gaussian mixture model-a structured model for speech recognition," *Computer Speech and Language*, vol. 25, issue 2, pp. 404-439, 2011.

[2] D. Povey, L. Burget, M. Agarwal, et al, "Subspace Gaussian mixture model for speech recognition," in *Proc. ICASSP*, pp. 4330-4333, 2010.

[3] L. Burget, P. Schwarz, M. Agarwal, et al, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. ICASSP*, pp. 4334-4337, 2010.

[4] A. Mohan, S. H. Ghalehjegh, and R. C. Rose, "Dealing with acoustic mismatch for training multilingual subspace Gaussian mixture models for speech recognition," in *Proc. ICASSP*, pp. 4893-4896, 2012.

[5] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. ASRU*, pp. 365-370, 2011.

[6] L. Lu, A. Ghoshal, and S. Renals, “Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition,” in *Proc. ICASSP*, pp. 4877-4880, 2012.

[7] T. Schultz, “GlobalPhone: a multilingual speech and text database developed at Karlsruhe University,” in *Proc. ICLSP*, pp. 345–348, 2002.

[8] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” *Computer Speech and Language*, vol. 12, issue 2, pp. 75–98, 1998.