ACCENT ADAPTATION USING SUBSPACE GAUSSIAN MIXTURE MODELS

*Petr Motlicek, Philip N. Garner**

Idiap Research Institute
Martigny, Switzerland
{motlicek,garner}@idiap.ch

Namhoon Kim, Jeongmi Cho

Samsung Electronics Co. Ltd
Suwon, South Korea
{namhoon.kim,jmcho007}@samsung.com

ABSTRACT

This paper investigates employment of Subspace Gaussian Mixture Models (SGMMs) for acoustic model adaptation towards different accents for English speech recognition. The SGMMs comprise globally-shared and state-specific parameters which can efficiently be employed for various kinds of acoustic parameter tying. Research results indicate that well-defined sharing of acoustic model parameters in SGMMs can significantly outperform adapted systems based on conventional HMM/GMMs. Furthermore, SGMMs rapidly achieve target acoustic models with small amounts of data. Experiments performed with US and UK English versions of the Wall Street Journal (WSJ) corpora indicate that SGMMs lead to approximately 20% and 8% relative improvements with respect to speaker-independent and speaker-adapted acoustic models respectively over conventional HMM/GMMs. Finally, we demonstrate that SGMMs adapted only with 1.5 hours can reach performance of HMM/GMMs trained with 18 hours.

Index Terms— Automatic speech recognition, Acoustic model adaptation, Accented speech, Under-resourced data

1. INTRODUCTION

A major problem in acoustic modeling of dialectical or accented speech is the sparse availability of speech resources. Even in the case of well-resourced languages, acoustic and language model adaptations towards different accents or dialects from a source language (out-of-domain data) require a minimum amount of adaptation (in-domain) data to achieve reasonable performance in the adapted system. Naturally, availability of adaptation data is even more problematic for less viable languages, dialects or infrequent accented speech. Therefore, conventional approaches for developing ASR systems on accented speech are directed by the amount of adaptation data and vary from simply building a recognizer purely using an accented speech to various types of adapting recognizers initially trained on a source language.

*This work was supported by Samsung Electronics Co. Ltd, South Korea, under the project "Domain Adaptation Using Subspace Models."

In this paper, Subspace Gaussian Mixture Models (SGMMs) employed for acoustic model adaptation towards accented speech are investigated. SGMMs have previously been shown to be beneficial in speaker adaptation, where they can be directly compared with conventional HMM/GMMs [1, 2]. SGMMs were also investigated for both multi-lingual and cross-lingual speech recognition tasks, where the globally-shared parameters were estimated by tying across the multiple languages [3]. The work in this paper has commonality with the prior work in that the out-of-domain data is considered as a remedy for training data sparseness.

In the multilingual ASR task, use of conventional HMM/GMMs is rather complex, rendering comparison of techniques similarly difficult. Accent adaptation represents a task lying conceptually between speaker adaptation and multilingual ASR. Acoustically, it represents a different surface realization of essentially the same phonemic sequence. It is a difficult adaptation task but, crucially, does not involve difficulties with different phone sets, lexicons and language models. The use of a homogeneous phone set in turn allows direct comparison with conventional adaptation techniques such as Bayesian-based (MAP) or linear transformation techniques used on HMM/GMMs. In addition, more complex acoustic model training procedures (e.g., efficient speaker adaptive training) can also be exploited on top of adapted HMM/GMMs and thus directly compared to the novel SGMM acoustic modeling approach. Ultimately, we also investigate the performance dependency of HMM/GMMs and SGMMs on the amount of adaptation data and the number of free parameters in both systems.

Experiments were performed using the well-known Wall Street Journal (WSJ) data (US English) [4] and the accent mismatch was simulated using UK version WSJCAM [5]. The rationale behind this is that US and UK English are mutually intelligible. Further, the newspaper English of the WSJ derived databases is rather formal, and representative of the overlap of the two accents rather than the differences. This suggests use of the same phonetic lexicon throughout the experiments.

The paper is organized as follows: in Sect. 2, we describe acoustic model adaptation techniques in conventional

HMM/GMMs as well as in SGMMs. Sect. 3 presents adaptation experiments with results followed by Sect. 4 which concludes the paper.

2. ACOUSTIC MODEL ADAPTATION

2.1. Background

Much research has been carried out on dialectical and foreign accented speech recognition during the past few years. In [6], German-accented English speakers in a conversational meeting task were investigated. Similar experiments were carried out on Japanese-accented English [7]. Both tasks [6, 7] show that training on non-native speech data yields the largest gains in performance on accented speech. The adaptation provided using Maximum Likelihood Linear Regression (MLLR) [8] (e.g., applied individually to each test speaker), using MAP re-estimation [9] (also known as Bayesian adaptation), or by combining both, revealed them to be effective approaches for accented speech.

2.2. SGMM adaptation

The Subspace Gaussian Mixture Model (SGMM) [1] is a way of compactly representing a large collection of mixture-of-Gaussian models. Unlike conventional HMM/GMMs in which state model parameters are directly estimated from the data, SGMM model parameters are derived from a set of state-specific parameters, and from a set of globally-shared parameters which can capture phonetic and speaker variation [1]. In the case of a conventional Gaussian Mixture Model (GMM), the likelihood is given as

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{M_j} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{ji}), \quad (1)$$

where j is the state and the parameters of the model are w_{ji}, $\boldsymbol{\mu}_{ji}$ and $\boldsymbol{\Sigma}_{ji}$. The SGMM in the basic case is given as

$$p(\mathbf{x} \mid j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_{i}) \quad (2)$$

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j \quad (3)$$

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{l=1}^{I} \exp \mathbf{w}_l^T \mathbf{v}_j}, \quad (4)$$

where v_j are state-specific vectors (with dimension similar to that of the speech features), and $\mathbf{w}_i$, $\mathbf{M}_i$, and $\boldsymbol{\Sigma}_i$ are globally-shared parameters. I is the number of Gaussians in the shared GMM structure. In fact, we employ a Universal Background Model (UBM), which is a mixture of full-covariance Gaussians of size I that is used to initialize the system and to prune the Gaussian indices during training and decoding. The basic concept of SGMMs can be extended towards large-scale acoustic models by adding sub-states (i.e., each state j is assigned with sub-states - each with its own mixture weight and sub-state specific parameters) and speaker-dependent mean offsets via speaker vector parameters $\mathbf{v}^{(s)}$ and "speaker projections" $\mathbf{N}_i$ [2].

Usually in a multilingual SGMM framework, the globally-shared model parameters $\mathbf{w}_i$, $\mathbf{M}_i$, and $\boldsymbol{\Sigma}_i$ embody most of free parameters in the system and are initially trained using out-of-domain data (i.e., data from well-resourced corpora) in a Maximum-Likelihood (ML) fashion. Then, the state-specific parameters $\mathbf{v}_j$ are ML re-trained using in-domain (adaptation) data [1]. If the amount of training data is not sufficient to allow the global parameters to be trained using ML, it has been shown that MAP adaptation of the phonetic subspace parameters with a matrix variate Gaussian prior distribution can be employed in a multilingual scenario [10]. Since a homogeneous phone set is used in our adaptation scenario, the globally-shared SGMM parameters initially estimated using out-of-domain data can be directly re-estimated using in-domain data in a ML fashion.

In this paper, we aim to evaluate SGMMs in AM adaptation task towards accented speech. Since conventional HMM/GMMs can fully adopt Bayesian and transform-based adaptation techniques in this task, a good baseline system can be built capitalizing on combining these and other state-of-the-art algorithms (e.g., MAP, fMLLR, SAT).

3. ADAPTATION EXPERIMENTS

Based on the above discussion, we hypothesize that SGMMs are capable of outperforming an HMM/GMM baseline system for diverse training and adaptation conditions.

All the experiments were done with the open-source Kaldi speech recognition toolkit [11]. As noted before, the unique phone set and lexicon is used throughout all the experiments.

3.1. Core Corpus

The Wall Street Journal (WSJ) database consists of clean, read speech recorded with high quality microphones. In our experiments, recordings made with the Sennheiser (close talking) microphones were used. WSJ was used as out-of-domain data only for training of the acoustic models. In particular, SI-84 (WSJ0) training data (about 15 hours of speech with 84 training speakers) was used, which allowed fast turnaround of our experiments.

3.2. Dialect corpus

The UK English equivalent WSJCAM0 recorded at University of Cambridge was used as in-domain (i.e., adaptation and evaluation) data. WSJCAM0 was derived from the WSJ0 text corpus and primarily designed for the construction and

System *Amount of adaptation data*	(a) WSJ0 -	(b) WSJCAM0 *1 hour*	(c) WSJCAM0 *18 hours*	(d) Adapted: #1 *1 hour*	(e) Adapted: #2 *1 hour*
HMM/GMMs	37.0 (+59)	23.2 (baseline)	14.6	20.6	19.4 (-16)
+fmllr	32.9 (+60)	20.6 (baseline)	13.0	17.7	17.0 (-17)
+SAT(fmllr) + MLLT	32.3 (+70)	18.9 (baseline)	11.2	16.8	15.5 (-18)
#STATES/#GAUSSIANS	2K/10K	1K/2.5K	2K/10K	2K/10K	2.5K/10K
SGMMs	33.4 (+50)	22.3 (baseline)	11.3	17.4	15.9 (-29)
+spkvcs	32.7 (+55)	21.1 (baseline)	10.7	16.6	14.8 (-30)
+spkvcs+fmllr	32.1 (+60)	20.1 (baseline)	10.4	15.9	14.4 (-28)
#STATES/#SUB-STATES	2K/8K	1K/3K	2K/8K	2K/8K	2.5K/8K

Table 1. WERs [%]: Experimental results on the WSJCAM0 evaluation set. WERs given in brackets are relative with respect to the WSJCAM0 baseline trained on 1 hour. Table also shows number of parameters used for building the acoustic models.

evaluation of speaker-independent speech recognition systems [12]. As for WSJ0, we used the high-quality Sennheiser head-mounted microphone recordings. As the training and test sentences in WSJCAM0 were taken from WSJ0 corpus (non-verbalized pronunciation texts) [5], the lexicon provided with WSJ0 was used during all the experiments.

The full training set contains about 18 hours of speech with 92 training speakers. In order to simulate lack of adaptation data, most of the experiments employ 1 hour of training data randomly selected from the WSJCAM0 training corpus (92 speakers are still kept as for the full training set). In addition to 1 hour, we also created subsets with $\frac{1}{2}$, 1, 2, 4 and 8 hours of train data to investigate the dependency of the adaptation on amount of data. The first evaluation set in WSJCAM0 with 14 speakers with a total of 2.5 hours was used for testing of our adapted acoustic models.

3.3. Experimental setup

All reported results are based on mean and variance (per-speaker) normalized 39-dimensional MFCC plus delta plus acceleration features. The WSJCAM0 test set was decoded with the 20 K open vocabulary (with UNK) with non-verbalized pronunciations, which is included with WSJ0 corpus. As the Language-Model (LM), we used a highly-pruned version of the trigram LM ($\sim$0.6 M instead of $\sim$3 M trigrams) included also with the WSJ0 corpus. The acoustic scale factor was always tuned for the best Word Error Rates (WERs) during our experiments.

The conventional context-dependent HMM/GMM triphone system uses standard mixture-of-diagonal-Gaussian models. Both systems (i.e., HMM/GMMs and SGMMs) use the same decision-tree clustered tri-phones trained on the respective corpora (i.e., data used for GMM training). In fact, an extended phone set with position and stress dependent phones, where decision-trees correspond to "real" phones, was used.

3.4. Experimental results

As stated before, Bayesian and transform-based adaptation techniques have shown their effectiveness when applied on accented speech. According to past experimental results on dialectical or accented speech adaptation (i.e., [13] or [14]), MAP usually outperformed MLLR adaptation (applied as a set of phone-based transforms). Although MLLR offers fast adaptation rates, our recent multilingual studies indicate that MLLR was dominant only in cases of very small amounts of adaptation data (i.e., around 5 minutes) [15]. We therefore decided to build a baseline HMM/GMM system around MAP (by exploiting adaptation data for an acoustic model trained using out-of-domain data). We presume that additional significant gain will rather be achieved by implementing a speaker-dependent ASR system which in fact will be provided by transform-based adaptation.

An overview of WER performance of the complete ASR system exploiting differently trained or adapted Acoustic Models (AMs) evaluated on the WSJCAM0 evaluation part is given in Tab. 1. It also describes AMs in terms of number of parameters[1]. For SGMMs, the phonetic subspace dimension S was 40 and the speaker subspace dimension (if applying speaker vectors) was 39. The UBM was trained on corresponding data and had 400 Gaussians (100 Gaussians for the 1 hour WSJCAM0 training). More particularly for Tab. 1:

- **(a)** AM trained on WSJ0 data only.
- **(b)** AM trained on 1 hour of WSJCAM0.
- **(c)** AM trained on the full WSJCAM0 training set.
- **(d)** Adapted: #1 - AM initially trained using WSJ0 data and then adapted on 1 hour of WSJCAM0 (using MAP ($\tau = 10$) in case of HMM/GMMs and using ML re-estimating state-specific parameters in case of SGMMs).
- **(e)** Adapted: #2 - AM initially trained using WSJ0 together

[1]Note: although we aim to minimize WER, SGMMs are built to have approximately the same number of state-specific vectors as the total number of GMMs in the HMM/GMM system.

with 1 hour of WSJCAM0 data in ML fashion and then adapted on 1 hour of WSJCAM0 (using MAP ($\tau = 10$) in case of HMM/GMMs and using ML re-estimating state-specific parameters in case of SGMMs). Unlike (d), here, we perform an experiment where the adaptation data is first used in initial ML training together with out-of-domain data. Then, the same data is used to adapt previously developed AM.

Further, the adaptation scenario has been extended towards speaker-dependent AMs. In the case of HMM/GMMs, first, speaker adaptation using feature-space Maximum Likelihood Linear Regression (fMLLR), also known as constrained MLLR [16], was applied during the decoding. Then, fMLLR was also used in Speaker Adaptive Training (SAT) [17], together with Maximum Likelihood Linear Transform (MLLT) [18] aiming to decorrelate the feature vectors. In the case of SGMMs, the speaker subspace for Gaussian means together with speaker vectors (denoted to as "spkvcs") was applied as a linear transform towards speaker adaptation. Similar to the HMM/GMM case, speaker-based fMLLR was then applied during the decoding.

In addition to Tab. 1, Fig. 1 graphically visualizes recognition results for different amounts of in-domain data used to adapt HMM/GMMs and SGMMs (i.e., the case #1). More specifically, we show results for speaker-independent AMs. Baseline HMM/GMMs and SGMMs were trained only using a corresponding amount of in-domain data. The number of parameters in the HMM/GMM as well as the SGMM system were adapted accordingly. In the adapted systems, the models were initially trained using WSJ0 and then adapted (HMM/GMMs using MAP ($\tau = 10$), state-specific parameters re-trained in SGMMs) using a corresponding amount of in-domain data. Overall, Fig. 1 shows that adapted SGMMs outperform HMM/GMMs for all the chosen sizes of adaptation data. Interestingly, this is also the case for very small amounts of data (0.5 hours) where non-adapted SGMMs yield very poor performance. Further, we also performed an experiment exploiting all in-domain (full corpus) WSJCAM0 training data to adapt AMs initially trained on WSJ0. In this case, adapted HMM/GMMs (WER about 13.8%) as well as adapted SGMMs (WER about 10.9%) outperformed the baseline AMs trained only using the full 18 hours WSJCAM0 corpus (see Tab. 1).

4. CONCLUSIONS

Overall, SGMMs outperformed the HMM/GMM baseline. For the 1 hour adaptation scenario, relative WER improvements of the adapted SGMMs are about 20% and 8% for speaker independent and speaker-dependent acoustic models, respectively, over the adapted HMM/GMMs. When compared to AMs trained uniquely on 1 hour of in-domain data, SGMMs benefit better from out-of-domain data (29% relative improvement) than HMM/GMMs (16% relative improvement). Both types of models are able to profit from in-domain data available during initial ML training. Whilst for very small amounts of in-domain data non-adapted SGMMs fail, the adapted SGMM system significantly outperforms the HMM/GMM baseline. Finally, experimental results indicate that SGMMs adapted using about 1.5 hours of in-domain data achieve similar performance as HMM/GMMs trained on the full 18 hours WSJCAM0 corpus.

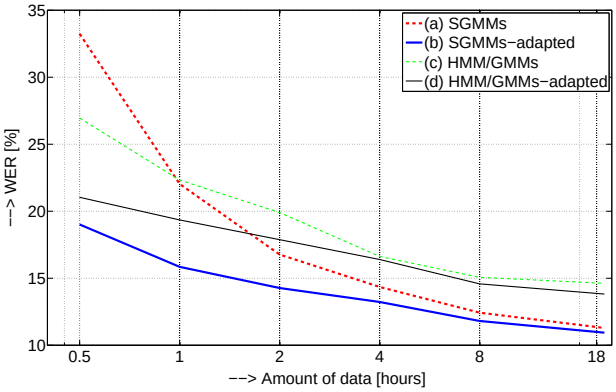

Fig. 1. WERs [%]: Experimental results on the WSJCAM0 evaluation set with respect to amount of in-domain data used during training or adaptation. Non-adapted AMs were directly trained using respective amount of in-domain data. Adapted AMs were first trained using WSJ0 data and then adapted using respective amount of in-domain data.

5. REFERENCES

[1] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. karafiat, A. Rastrow, R. C. Rose, P. Schwarz and S. Thomas, "The Subspace Gaussian mixture model - A structured model for speech recognition," In Computer Speech & Language, vol. 25, no. 2, pp. 404-439, 2011.

[2] D. Povey, M. Karafiat, A. Ghoshal and P. Schwarz, "A Symmetrization of the Subspace Gaussian Mixture Model", *in Proc. of ICASSP*, pp. 4504-4507, Prague, Czech R., 2011.

[3] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. karafiat, D. Povey, A. Rastrow, R. C. Rose and S. Thomas, "Multilingual Acoustic Modeling For Speech Recognition Based On Subspace Gaussian Mixture Models", *in Proc. of ICASSP*, pp. 4334-4337, Dallas, USA, 2010.

[4] D. Paul and J. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *in Proc. of the DARPA SLS Workshop*, USA, February 1992.

[5] J. Fransen, D. Pye, T. Robinson, P. Woodland and S. Young, "WSJCAM0 Corpus and Recordings Description," Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department, September 1994.

[6] Z. Wang, T. Schultz and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," *in Proc. of ICASSP*, pp. 540-543, Hong Kong, 2003.

[7] L. M. Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," *in Proc. of Multilinguality in Spoken Language Processing*, Aalborg, 2001.

[8] 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, pp. 171-185, 1995.

[9] C. H. Lee and J. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters,", *in Proc. of ICASSP*, p. II-558, Minneapolis, USA, 1993.

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

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely, "The Kaldi Speech Recognition Toolkit," *in Proc. of ASRU*, Hawai, USA, December 2011.

[12] T. Robinson, J. Fransen, D. Pye, J. Foote and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," *in Proc. of ICASSP*, pp. 81-84, Detroit, USA, 1995.

[13] M. Elmahdy, R. Gruhn, W. Minker and S. Abdennadher, "Cross-lingual acoustic modeling for dialectal Arabic speech recognition," *in Proc. of Interspeech*, pp. 873-876, Japan, 2010.

[14] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr and S. Yoon, "Accent detection and speech recognition for Shanghai-accented Mandarin," *in Proc. of Interspeech*, pp. 217-220, Portugal, 2005.

[15] D. Imseng, J. Dines, P. Motlicek, P. Garner, H. Bourlard, "Comparing different acoustic modeling techniques for mutlilingual boosting," *in Proc. of Interspeech*, Portland, USA, 2012.

[16] M. J. F. Gales, "Maximum likelihood linear transformations for HMM based speech recognition," Computer Speech and Language, vol. 12, no. 2, pp. 75-98, April 1998.

[17] T. Anastasakos, J. McDonough, R. Schwartz and J. Makhoul, "A compact model for speaker adaptive training," *in Proc. of ICSLP*, pp. 1137-1140, Philadelphia, 1996.

[18] R. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *in Proc. of ICASSP*, vol. 2, pp. 661-664, Seattle, USA, 1998.