MODELING GENDER DEPENDENCY IN THE SUBSPACE GMM FRAMEWORK

Ngoc Thang Vu, Tanja Schultz

Karlsruhe Institute of Technology
thang.vu@kit.edu, tanja.schultz@kit.edu

Daniel Povey

Microsoft Research, USA
dpovey@microsoft.com

ABSTRACT

The Subspace GMM acoustic model has both globally shared parameters and parameters specific to acoustic states, and this makes it possible to do various kinds of tying. In the past we have investigated sharing the global parameters among systems with distinct acoustic states; this can be useful in a multilingual setting. In the current paper we investigate a related idea: to have different global parameters for different acoustic conditions (gender, in this case) while sharing the acoustic-state-specific parameters. We experiment with modeling gender dependency in this way, and show Word Error Rate improvements on a range of tasks and comparable results to the Vocal Tract Length Normalization (VTLN)-like technique Exponential Transform (ET).

Index Terms— Subspace Gaussian Mixture Models, gender depedency modeling

1. INTRODUCTION

The Subspace Gaussian Mixture Model (SGMM) [1] is a way of compactly representing a large collection of mixture-of-Gaussian models. Let us write a conventional Gaussian mixture model as:

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

where j is the state and the parameters of the model are c_{ji}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{\Sigma}_{jm}$. The basic version of the SGMM, without speaker adaptation or "sub-states", is:

$$\begin{aligned} p(\mathbf{x}|j) &= \sum_{i=1}^{I} w_{ji}\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i) && (2) \\ w_{ji} &= \frac{\exp(\mathbf{w}_i^T \mathbf{v}_j)}{\sum_i \exp(\mathbf{w}_i^T \mathbf{v}_j)} && (3) \\ \boldsymbol{\mu}_{ji} &= \mathbf{M}_i \mathbf{v}_j, && (4) \end{aligned}$$

where the vectors $\mathbf{v}_j$ (normally of dimension around S=40) describe in some abstract space how the states differ from each other; I is the number of Gaussians in the shared GMM structure, and is normally several hundred. The parameters of the system are the state-specific parameters $\mathbf{v}_j$, and the globally shared parameters $\mathbf{w}_i$, $\mathbf{M}_i$ and $\boldsymbol{\Sigma}_i$ (these are full covariances). It is described in [1] how to extend this with sub-states (replacing $\mathbf{v}_j$ with mixtures $\mathbf{v}_{jm}$ and sub-states weights c_{jm}), and how to add speaker-dependent mean offsets via "speaker vector" parameters $\mathbf{v}^{(s)}$ and "speaker projections" $\mathbf{N}_i$.

We sometimes speak of a Universal Background Model (UBM). This 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 UBM Gaussians correspond to the indices i, and when we speak of changing the number of UBM Gaussians, this involves changing the number of parameters $\mathbf{N}_i$ and so on.

As described in [2], it is possible to use the SGMM framework to improve speech recognition performance by leveraging out-of-language data. The basic idea is to share all the global parameters between languages. Since, for smaller systems, the globally shared quantities dominate the parameter count, this can lead to more robust parameter estimates.

In this paper, we explore a related idea, which is to have different sets of shared parameters for different genders, while leaving the state-specific parameters $\mathbf{v}_{jm}$ gender-neutral. We expect that this would be more useful when there is a relatively large amount of training data, because in this case the parameter count tends to be less dominated by the global parameters (so we would increase the parameter count less, relatively, by introducing more Gaussians in the UBM). Experimentally, we implemented this technique and test it on a range of tasks; we found that for all the tasks that had a reasonably large amount of training data, this technique gave an improvement versus the standard SGMM. On one setup, as an additional baseline we compared our method with a VTLN-like technique, the Exponential Transform [3] (used in conjunction with SGMMs), and we find that our method gives better results.

This paper is organized as follows. Section 2 explains how we implemented gender dependency through the Gaussian pruning mechanism of the SGMM framework, Section 3 describes our experimental setup and results, and we conclude in Section 4.

2. GENDER-DEPENDENT SYSTEMS VIA GAUSSIAN PRUNING

Our experiments were done with the open-source Kaldi speech recognition toolkit [4]. In the recipes distributed with Kaldi, the Gaussian selection phase tends to be done just once in a particular stage of system building, and the selected Gaussian indices are stored on disk. We decided that the simplest way to implement gender dependency in the SGMM framework would be to make it part of the Gaussian selection phase: that is, pre-allocate certain Gaussian indices (certain values of i) to male, and certain ones to female. Then, when training or decoding a male utterance, we would limit the Gaussian selection phase to only the "male" indices, and likewise for female. This has almost the same effect as doing it in the most natural and obvious way, which would be to have multiple sets of global parameters and adding a new index corresponding to gender on the $\mathbf{w}_i$, $\mathbf{M}_i$ and $\mathbf{\Sigma}_i$ quantities (so we would have $\mathbf{w}_{ki}$ and so on). The only difference when doing it in the Gaussian selection phase is that the model may now be attempting to model the state-specific probability of being in a particular gender, which is not very optimal. That is, ideally in Equation 3 we would like to normalize the weights per gender, rather than globally, but in our simple implementation based on the Gaussian selection mechanism, it is normalized globally. However, we guess that most acoustic states would have seen similarly balanced statistics, so the male/female probabilities would usually be about the same (typically 0.5) and this should have very little effect on the decoded output. We have verified this experimentally.

We now describe how we adapt the UBM training process to the gender-dependent setup. Suppose we want a total of 800 Gaussians in the UBM (including both male and female), and the corpus is reasonably gender-balanced. We cluster the Gaussians in a traditional HMM-GMM system down to 400, as described in [1]. Then we do four iterations of full-covariance GMM re-estimation on the training data; this is done separately for the male and female training data, so we get two separate UBMs, one for male and one for female. At this point we merge them into a single UBM with about 800 Gaussians (a few may have been lost due to low counts), and we record which Gaussian indices correspond to male, and which correspond to female, in the merged UBM. Compared to other VTLN-like technique Exponential Transform (ET) [3] which requires another model and another pass of decoding, our technique is very efficient.

Our training and test data are both annotated with gender information. During both training and test, we provide the program that does the Gaussian selection with lists of allowable Gaussian indices i for each training or test utterance, and it writes out the top-scoring Gaussians in those allowable sets.

We were concerned that it might be considered "cheating" to use gender information during test time. To forestall this objectsion, for the English, Spanish and French Global Phone data (see the experimental section) we classified the test utterances by gender, by comparing the likelihoods obtained during Gaussian selection based on a male versus female assumption. We got 100% classification accuracy for all languages, so we can be confident that this "cheating" does not affect our results.

Since the use of gender-dependent UBMs can be considered a form of speaker adaptation, we felt that it should be evaluated in conjunction with standard speaker adaptation methods used in SGMMs. Therefore we did our gender-dependent experiments in a system that had the speaker vectors $\mathbf{v}^{(s)}$, and we also tested with Constrained MLLR (CMLLR) adaptation and compared the results with another VTLN-like technique (ET in this case).

3. EXPERIMENTAL SETUP

All our experiments are performed with the Kaldi speech-recognition toolkit, introduced in [4]. The scripts for the Resource Management and Wall Street Journal experiments which we report here, are included with the toolkit (see egs/rm/s1 and egs/wsj/s1).

3.1. Wall Street Journal experiments

The Wall Street Journal database [5] consists of clean, read speech recorded with a high quality microphone (we used the Sennheiser version of the recordings). For results reported in this paper we train on all the SI-284 data— about 80 hours. Our test results are with the Nov'92 and Nov'93 evaluation test sets, using the 20k open vocabulary with non-verbalized pronunciations. This is the hardest test condition so the results may seem higher than expected for WSJ. See [4] for comparison with other published results. We test with a highly-pruned version of the trigram language model supplied with the WSJ corpus (pruned from 6.7 million to 1.5 million entries), since the decoders in Kaldi currently do not support very large language models.

All results we report here are based on MFCC plus delta plus acceleration features. We report results with standard mixture-of-diagonal-Gaussian models, and with SGMMs. We used a dictionary in which phones were marked with stress information and beginning and end-of-word information, and built decision trees corresponding to each "base phone", in which questions could be asked about the stress and word-position information. The HMM-GMM system had 3349 context-dependent states and 40 000 Gaussians, and the SGMM systems had 4780 context-dependent states (for SGMM systems, the optimum number of states tends to be higher) and 35 000 sub-states (i.e. 35 000 vectors $\mathbf{v}_{jm}$). The gender-independent UBM had 600 Gaussians ($I = 600$) and the phonetic subspace dimension (S) was 50; the speaker subspace dimension, where applicable, was 39. For gender-dependent models, we used 800 UBM Gaussians (400 per

Model	System	%WER	
/adaptation	Id	Nov'92	Nov'93
HMM-GMM	tri3a	10.7	13.8
+CMLLR	tri3a	9.5	12.1
SGMM+spk-vecs	sgmm3b	7.8	10.4
+CMLLR	sgmm3b	7.7	10.0
SGMM+spk-vecs+GD	sgmm3c	7.5	9.5
+CMLLR	sgmm3c	7.6	9.2
ET+SGMM+spk-vecs	sgmm3c	7.5	9.9
+CMLLR	sgmm3c	7.4	9.8

Table 1. Results on Wall Street Journal: %WERs

gender). We use an acoustic weight of 1/16 for GMM-based systems, 1/11 for speaker-independent SGMM-based systems, and 1/12 for speaker-adapted SGMM-based systems.

As seen in Table 1, gender dependency improves results by 0.1% and 0.8% absolute on the Nov'92 and Nov'93 test sets respectively, comparing the sgmm3b and sgmm3c systems with CMLLR adaptation. We repeated the gender dependent decoding with gender-specific normalization of the weights w_{ji} (actually, w_{jmi} when we consider the substates). In two out of the four gender-dependent decoding experiments in Table 1 the WER was 0.1% worse, in one it was 0.1% better, and in one it was unchanged. This confirms our intition that global verusus gender-specific normalization does not make a big difference. To clarify: by gender-specific normalization of the weights we mean ensuring that within each sub-state j, m, the weights w_{jmi} sum to one for the indices i corresponding to each gender. Furthermore, compared to the results by using ET, we observed worse results on the Nov'92 test set (0.2% absolute) and better results on the Nov'93 test set (0.6% absolute). The combination of ET and gender-dependent UBM degraded result compared to either baseline (results not shown).

3.2. Resource Management experiments

The Resource Management (RM) dataset [6] is a medium-vocabulary dataset recorded under clean conditions. There are 3.9 hours of training data. The language model used in testing is a word-pair grammar supplied with the corpus. We report results averaged over six test sets, as described in [4]; in total, the testing data we used is about 1.3 hours long.

All results are reported on top of MFCC plus delta plus acceleration features. The models are triphone models with context-dependency and tree clustering. The GMM baseline system had 1459 context-dependent states and 9000 Gaussians, and the SGMM systems had 2039 context-dependent states and 7500 sub-states. The gender-independent SGMM systems had 400 UBM Gaussians; the gender-dependent ones had 500 (300 for male and 200 for female). The phonetic subspace dimension S was 40 and the speaker subspace dimension (if using speaker vectors) was 39. We used an acoustic scale of 1/12 for GMM-based systems and 1/10 for SGMM-based systems.

Model	System	%WER
/adaptation	Id	(average)
HMM-GMM	tri2a	4.0
+CMLLR	tri2a	3.6
SGMM	sgmma	3.3
+CMLLR	sgmma	2.9
SGMM+spk-vecs	sgmmb	2.5
+CMLLR	sgmmb	2.4
SGMM+spk-vecs+GD	sgmmc	2.7
+CMLLR	sgmmc	2.5
ET+SGMM+spk-vecs	sgmmc	2.3
+CMLLR	sgmmc	2.3

Table 2. Results on Resource Management: %WERs

In this case we did not see any improvement from gender dependency; in fact, the WER increased by 0.1%-0.2%. In fact, we did not expect to see improvements with so little training data. The issue is that adding gender dependency doubles the number of global parameters (assuming we keep the same number of UBM Gaussians). Of course, after tuning we have fewer UBM Gaussians per gender than we did for the gender independent system, since with so little data we cannot afford to train many UBM Gaussians per gender.

We reran the gender dependent decoding with gender-dependent normalization of the substate-specific weights. This did not affect results to within the rounding error, for these experiments.

3.3. GlobalPhone experiments

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages, including Arabic, Bulgarian, Chinese (Mandarin and Shanghai), Croatian, Czech, English, French, German, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, Tamil, Thai, Turkish, and Vietnamese [7]. The corpus contains more than 400 hours of speech spoken by more than 1900 adult native speakers. GlobalPhone is available from ELRA, the European Language Resources Association. The read articles cover national and international political news as well as economic news from 1995-2009. The speech data is available in 16bit, 16kHz mono quality, recorded with a close-speaking microphone. Most transcriptions are internally validated and supplemented by special markers for spontaneous effects like stuttering, false starts, and non-verbal effects. For this work we selected English, French, and Spanish from the GlobalPhone corpus. Each language has about 20 hours of training data, and we report results on the development sets which are about 2 hours long.

To build the language models we used our Rapid Language Adaptation Toolkit (RLAT) [8] to crawl for each language several websites with link depth 20 in up to twenty days e.g. as in [9]. Since Kaldi currently only supports decoding with relatively small language model, we used the SRI language model toolkit to prune all the language models using an entropy criterion [10]. Table 3 gives a breakdown of the trigram perplexities, OOV rate, and vocabulary size for the three languages.

Languages	PP	OOV	Vocabulary
English (EN)	340	0.5%	60k
French (FR)	423	2.4%	65k
Spanish (SP)	224	0.1%	19k

Table 3. Perplexities (PP), OOV rate and vocabulary size for English, French and Spanish

For acoustic modeling, we used Kaldi to train HMM-GMM and SGMM systems. These systems used MFCC plus delta and acceleration features. The HMM-GMM-systems had 9000 Gaussians and about 1220 context-dependent acoustic states. The SGMM systems had 7500 sub-states, and about 2100 context-dependent states. The gender-independent SGMM systems had 400 UBM Gaussians (i.e. $I{=}400$), and the gender-dependent systems had 500 UBM Gaussians. For English and French there were 250 per gender, but for Spanish, because of unbalanced data (9 female and 7 male) we trained 290 Gaussians for female speakers and 210 for male.

Systems	English	French	Spanish
HMM-GMM	17.8	27.4	23.8
SGMM	13.3	23.1	18.6
SGMM+spk-vector	11.8	22.5	17.3
+CMLLR	11.4	22.1	16.5
SGMM+spk-vector+GD	11.0	22.6	16.8
+CMLLR	10.7	21.9	15.9

Table 4. WERs for English, French, German and Spanish systems (GlobalPhone)

By adding gender dependency , we can see from Table 4 that WER from the final pass (after applying CMLLR) is improved 0.7%, 0.2%, and 0.6% absolute for English, French, and Spanish respectively.

4. CONCLUSIONS

We have described a simple way to model gender variation within the SGMM framework. It consists of allocating certain Gaussians in the UBM to male, and certain ones to female, and enforcing this allocation during the Gaussian selection process. Experiments on five different training data sets show that the technique almost always gives improvements over gender-independent SGMM based systems, with a fairly typical improvement being 0.4% absolute. Furthermore, we got comparable results compared to the VTLN-like technique Exponential Transform which was shown in [3] to perform about the same as the conventional VTLN technique.

The Gaussian selection based implementation that we describe here is not very optimal as we do not properly normalize the likelihoods for the genders (that is, the model is trying to predict the the male versus female likelihoods, which is not what we want). However, when we tried with normalizing the likelihoods per gender in decoding time, we did not see any improvement in WER.

We may in future investigate the application of this technique to other sources of variation, such as accent and acoustic condition, and its combination with multilingual systems.

5. REFERENCES

[1] D. Povey, L. Burget *et al.*, “The Subspace Gaussian Mixture Model–A Structured Model for Speech Recognition,” *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, April 2011.

[2] L. Burget, P. Schwarz *et al.*, “Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models,” in *ICASSP*, 2010, pp. 4334–4337.

[3] D. Povey, G. Zweig and A. Acero, “Speaker Adaptation with an Exponential Transform,” in *ASRU*, 2011.

[4] D. Povey, A. Ghoshal *et al.*, “The Kaldi Speech Recognition Toolkit,” in *ASRU*, 2011.

[5] D. Paul and J. Baker, “The design for the Wall Street Journal-based CSR corpus,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[6] P. Price, W. Fisher, J. Bernstein, and D. Pallett, “The DARPA 1000-word resource management database for continuous speech recognition,” in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 651–654.

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

[8] T. Schultz and A. Black, “Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing,” *Proc. ICASSP Las Vegas, NV*, 2008.

[9] N. Vu, T. Schlippe, F. Kraus, and T. Schultz, “Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[10] A. Stolcke, “SRILM-an extensible language modeling toolkit,” in *Proceedings of the international conference on spoken language processing*, vol. 2, 2002, pp. 901–904.