# MULTILINGUAL ACOUSTIC MODELING FOR SPEECH RECOGNITION BASED ON SUBSPACE GAUSSIAN MIXTURE MODELS

Lukáš Burget[1], Petr Schwarz[1],
Mohit Agarwal[2], Pinar Akyazi[3], Kai Feng[4], Arnab Ghoshal[5], Ondřej Glembek[1], Nagendra Goel[6],
Martin Karafiát[1], Daniel Povey[7], Ariya Rastrow[8], Richard C. Rose[9], Samuel Thomas[8]

[1] Brno University of Technology, Czech Republic, {burget,schwarzp}@fit.vutbr.cz;
[2] IIIT Allahabad, India; [3] Boğaziçi University, Turkey; [4] HKUST, Hong Kong;
[5] Saarland University, Germany; [6] Virginia, USA; [7] Microsoft Research, Redmond, WA;
[8] Johns Hopkins University, MD; [9] McGill University, Canada

## ABSTRACT

Although research has previously been done on multilingual speech recognition, it has been found to be very difficult to improve over separately trained systems. The usual approach has been to use some kind of "universal phone set" that covers multiple languages. We report experiments on a different approach to multilingual speech recognition, in which the phone sets are entirely distinct but the model has parameters not tied to specific states that are shared across languages. We use a model called a "Subspace Gaussian Mixture Model" where states' distributions are Gaussian Mixture Models with a common structure, constrained to lie in a subspace of the total parameter space. The parameters that define this subspace can be shared across languages. We obtain substantial WER improvements with this approach, especially with very small amounts of in-language training data.

***Index Terms***— Large vocabulary speech recognition, Subspace Gaussian mixture model, Multilingual acoustic modeling

## 1. INTRODUCTION

Nowadays speech technology is mature enough to be useful for many practical applications. Its good performance, however, depends on availability of adequate training resources. There are many applications, where resources for the domain or language of interest are very limited. For example, in intelligence applications, it is often impossible to collect necessary speech resources in advance as it is hard to predict which languages become the next ones of interest. Limited resources were also at the center of interest of our team at the Johns Hopkins University 2009 summer workshop, titled "Low Development Cost, High Quality Speech Recognition for New Languages and Domains". Besides the work on learning lexicons, which is described in separate paper [1], this team explored a new approach to acoustic modeling for automatic speech recognition (ASR) based on Subspace Gaussian Mixture Models (SGMM) [2].

This work was conducted at the Johns Hopkins University Summer Workshop which was (partially) supported by National Science Foundation Grant Number IIS-0833652, with supplemental funding from Google Research, DARPA's GALE program and the Johns Hopkins University Human Language Technology Center of Excellence. BUT researchers were partially supported by Czech MPO project No. FR-TI1/034. Thanks to CLSP staff and faculty, to Tomas Kašpárek for system support, to Patrick Nguyen for introducing the participants, to Mark Gales for advice and HTK help, and to Jan Černocký for proofreading and useful comments.

In conventional acoustic models, the distribution of each (possibly tied) HMM state is represented by relatively large number of parameters completely defining a Gaussian Mixture Model (GMM). The SGMM also uses mixtures of Gaussians as the underlying state distribution, but the high-dimensional super-vector of all the GMM parameters is constrained to live in a relatively low dimensional subspace, which is common to all the states. This constraint is justified by high correlation between states' distributions, since the variety of distributions corresponding to the sounds that the human articulatory tract is able to produce is quite limited, and the number of tied states can be quite large. The majority of the parameters in an SGMM are typically shared across the states; these parameters are those defining a subspace of GMM parameters. Distributions of the individual states are then described using relatively low-dimensional vectors representing co-ordinates in the subspace. Therefore, the SGMM allows for a much more compact representation of HMM state distributions, which results in more robust estimation of parameters and improved performance – especially when the amount of available training data is limited.

In this paper, we concentrate on a set of "multilingual" experiments carried out during the JHU 2009 summer workshop, where the aim was to improve recognition performance for one language by also training the shared parameters of the acoustic model on data from other languages. In the past, other attempts to benefit from availability of data from different languages have been made. The usual approach has been to define a common set of universal phone models with appropriate parameter sharing [3] and train it on data from many languages with eventual adaptation on the data from the language of interest. However, mixed results were reported when using this rather complicated procedure. Also, use of the universal phone models often leads to degradation in performance for the resource-rich languages. However, we note the recent work reported in [4] which uses a similar experimental setup to ours and uses cross-lingual training of MLP-based multi-stream posterior features to leverage out-of-language data.

SGMMs can be very naturally trained in a multilingual fashion. The HMM states are defined and the state-specific SGMM parameters are trained as in the case of individual language-specific models. The common SGMM parameters are, however, shared across the HMM states from all the languages used. Our experiments were on Spanish, German and English, and we were able to show substantial improvements over building individual systems.

In Section 2, we describe the Subspace Gaussian Mixture

Model; in Section 3 we show visually the kind of information the subspace seems to be learning; in Section 4 we describe our experimental setup and results and, in Section 5, we give conclusions.

## 2. SUBSPACE GAUSSIAN MIXTURE MODEL

In the Subspace Gaussian Mixture Model (SGMM), the distribution of features in HMM state $j$ is a mixture of Gaussians:

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

where the same number of mixture components $I$ (typically a few hundreds) is used for all the states. The covariances $\boldsymbol{\Sigma}_i$ are shared across states. Unlike in a conventional GMM, the mean vectors $\boldsymbol{\mu}_{ji}$ and mixture weights $w_{ji}$ are not directly estimated as parameters of the model. Instead, a particular state $j$ is associated with a vector $\mathbf{v}_j$ which determines the means and weights. The mean vectors are derived as:

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j, \quad (2)$$

where the parameters $\mathbf{M}_i$ are shared across state distributions and define the subspace in which the GMM mean vectors can live. The matrices $\mathbf{M}_i$ will typically comprise the majority of the parameters in an SGMM model. The mixture component weights are derived from the vector $\mathbf{v}_j$ using a log-linear model

$$w_{ji} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_j}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_j}, \quad (3)$$

with a globally shared parameter $\mathbf{w}_i$ determining the mapping. This model is of the same form as that used in multi-class linear logistic regression.

To find the right balance between the amounts of shared and state-specific parameters, we have adopted the concept of *substates*. Here, the distribution of a state can be represented by more than one vector $\mathbf{v}_{jm}$, where $m$ is the substate index. Each vector $\mathbf{v}_{jm}$ determines a "substate" distribution, which is again a mixture of Gaussians. The state distribution is then a mixture of the substate distributions defined as follows:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{jmi}, \boldsymbol{\Sigma}_i) \quad (4)$$

$$\boldsymbol{\mu}_{jmi} = \mathbf{M}_i \mathbf{v}_{jm} \quad (5)$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}}, \quad (6)$$

where the substate mixture weights $c_{jm}$ are additional state-specific parameters. In our experiments, we allocate the number of mixture components $M_j$ proportional to a small power of the data count for a state $j$ (e.g. 0.2).

For more discussion and results relating to SGMMs as applied to a single language and for results relating to speaker adaptation, see [5] and [6]. Our experiments reported here did not use speaker adaptation. A more detailed treatment, including derivations and complete estimation formulae, can be found in [2].

## 3. INTERPRETING SUBSPACE DIMENSIONS

The hope with this type of modeling technique is that the vectors $\mathbf{v}_{jm}$ will correspond to some kind of meaningful representation of speech sounds. We decided to see if we could verify this by plotting these vectors for different phonetic classes. We did this for a system without substates. We display the vectors $\mathbf{v}_{jm}$ by working out the two most important directions of variation and plotting those (see [2, Appendix K] for the computation). Figure 1 is a plot of the vector $\mathbf{v}_j$ in the central states of the various tied 3-state HMMs for various different vowels, with each state represented by a single dot and the dots corresponding to each vowel in a different color (the colors will not be visible in black and white, but we have placed the corresponding label at the center of each cluster to indicate where they are). The labels correspond to the PRONLEX dictionary format. Figure 2 consists of the same set of vowels extracted from a standard IPA vowel chart [1] and converted to the same PRONLEX notation. The locations in the subspace display almost the same pattern as the IPA vowel chart. This suggests that the subspace model is learning something meaningful.

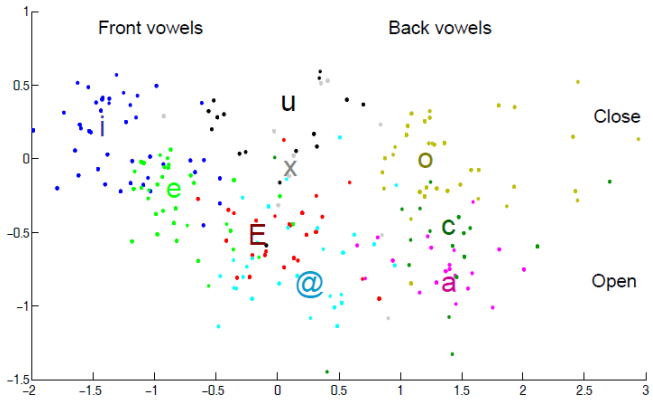


**Fig. 1**. Most significant dimensions in SGMM space

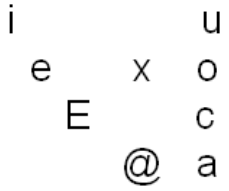


**Fig. 2**. IPA phone chart, converted to PRONLEX format

## 4. EXPERIMENTAL SETUP AND RESULTS

### 4.1. Baseline systems

In the following experiments, systems are trained and results are reported for three languages, namely English, German and Spanish. The corresponding parts of Callhome corpora [7, 8, 9] are used for acoustic model training and to test the recognition performance. The amounts of data used for training and test are sumarized in Table 1.

[1] http://www.langsci.ucl.ac.uk/ipa/vowels.html

**Table 1**. Amounts of data for acoustic model training and testing

| | Training | | Test | |
|---|---|---|---|---|
| | hours | conversations | hours | conversations |
| English | 15.1 | 80 | 1.8 | 20 |
| Spanish | 16.5 | 80 | 2.0 | 20 |
| German | 14.7 | 80 | 3.7 | 20 |

**Table 2**. Word recognition performance for English

| | #Sub-states | #Parameters | | WER [%] |
|---|---|---|---|---|
| | | Shared | State-specific | |
| Baseline | n/a | 0 | 2427k | 52.5 |
| SGMM | 1921 | 952k | 77k | 48.9 |
| SGMM | 12k | 952k | 492k | 47.5 |
| +multilingual | 12k | 952k | 492k | 46.4 |
| +multilingual | 53k | 952k | 2173k | 44.9 |

The language-specific baseline recognition systems are based on conventional 3-state left-to-right HMM triphone models. Decision tree based clustering is used to obtain 1921, 1696 and 1584 tied states for English, German and Spanish respectively. We use 16 Gaussians per state. The tree sizes and numbers of Gaussians per state were tuned to optimize unadapted WER. The features are 13 PLP coefficients [10] including energy, plus $\Delta$ and $\Delta\Delta$ and per-speaker mean and variance normalization.

For English, the language model was a trigram with a word-list of 61k words obtained by interpolating individual models trained on English CallHome, Switchboard [11], GigaWord [12] and some web data. The web data is obtained by crawling the web for sentences containing high frequency bigrams and trigrams occurring in the training text of the Callhome corpus. The Spanish trigram LM used a word-list of 45k words and was trained on the Spanish CallHome transcripts and web data. We did not do word decoding on German because the German CallHome has no development data to estimate LM interpolation weights, and because of lack of time to develop an alternative approach.

The results obtained with language-specific baseline systems can be found in the first line of Table 2 for English and in the first line of Table 3 for Spanish. Although 54.7% word error rate (WER) for English and 68.9% WER for Spanish may seem to be rather high, these results are in agreement with those obtained by other sites [13, 14] with ML-trained unadapted systems on this very challenging task.

We also report recognition performance for all three languages in terms of phone error rate. Phone reference transcripts were obtained using a forced alignment performed with our language-specific baseline systems. For the phone decoding, language-specific phone bigram language models were used. These "phonotactic" models were trained on phone transcripts of acoustic training data, which were again obtained using forced alignment. Phone error rates for all three languages obtained with baseline systems are reported in the first line of Table 4.

**Table 3**. Word recognition performance for Spanish

| | #Sub-states | #Parameters | | WER [%] |
|---|---|---|---|---|
| | | Shared | State-specific | |
| Baseline | n/a | 0 | 2000k | 68.3 |
| SGMM | 1582 | 952k | 63k | 65.9 |
| SGMM | 22k | 952k | 902k | 65.2 |
| +multilingual | 22k | 952k | 902k | 64.6 |
| +multilingual | 40k | 952k | 1640k | 64.4 |

**Table 4**. Phoneme error rate

| | English | Spanish | German |
|---|---|---|---|
| | #phones | | |
| | 42 | 27 | 45 |
| Baseline | 54.9 | 46.2 | 56.3 |
| SGMM | 51.7 | 44.0 | 53.4 |
| +multilingual | 50.2 | 43.5 | 52.4 |

### 4.2. Subspace Gaussian Mixture Model

In the first set of experiments with SGMMs, language-specific models are built with both shared and state-specific parameters trained only using data from the corresponding language. Except for the acoustice model, the systems are the same as the baseline (i.e. we use the same features, state tying, language models, etc.) In these experiments, we use an SGMM configuration with $I = 400$ mixture components and 40 dimensional state-specific vector $\mathbf{v}_j$, which results in 952000 shared parameters. This configuration was found to be close to the optimum. The second line in Table 2 shows results for an English SGMM based system, where each of 1921 tied states is described using only a single state-specific vector $\mathbf{v}_j$. In this case, the overall number of state-specific parameters is 76840, which is only fraction of the number of shared parameters. With this system, we obtain 2.2% absolute improvement in WER compared to the baseline. As can be seen on the next line in the table, an additional 2.8% absolute improvement can be obtained from increasing the number of substates as discussed in section 2. Note that, for English, the overall number of the parameters in this model is still only about half the baseline.

### 4.3. Multilingual experiments

In the previous experiments, each language-specific system was trained using quite a limited amount of acoustic data. Assuming that the simple linear constraints put on state distributions by shared SGMM parameters are not very language-specific (i.e. correspond to the constraints given by human articulatory tract rather than defining subspace of sounds specific to a language), we can attempt to increase the robustness of model estimation by training the shared parameters using larger amounts of data from multiple languages. In the following experiments, state-specific SGMM parameters are still associated with the same tied states specific to each language. However, the shared SGMM parameters are now shared across states from all three languages and are effectively trained using all the training data available for these languages (46.3 hours).

The results obtained with such configuration are those with the tag *multilingual* reported in Table 2 for English and Table 3 for Spanish. Using the same number of substates that was optimal for systems trained on language-specific data (12k and 22k substates for English and Spanish respectively), we observe WER reduction for both English and Spanish when training the shared parameters in the multilingual fashion. Additional performance gains are observed for both languages when increasing the number of substates from 12k to 53k for English and from 22k to 40k for Spanish. Although only

**Table 5**. Results for English system trained using only one hour of English data.

| | Shared parameters trained on | WER [%] |
|---|---|---|
| Baseline | n/a | 70.5 |
| SGMM | 1h English | 67.6 |
| | Spanish + German | 59.8 |
| | +1h English | 59.6 |

the shared SGMM parameters are effectively trained using the increased amount of training data, the more robust estimates of these parameters allow us to benefit from further increasing the number of state-specific parameters. However, we did not see any benefit from increasing the number of the shared parameters by doubling the number of Gaussian components in SGMM model. The advantage of multilingual training can also be seen from the phone recognition results in Table 4, where the phone error rates for German are also reported.

### 4.4. Acoustic modeling for languages with very limited resources

In the last set of experiments (Table 5), we investigate the possibility of building acoustic models for an extreme case, where only 1 hour of conversational speech is available for the language of interest. The first line in the table shows the baseline result obtained with a conventional acoustic model trained only on one hour of English. For this purpose, segments from Callhome English were randomly selected to contain speech from all speakers in the training part of the database. After tuning the model size, the optimal performance of 70.5% WER was obtained with only 500 tied states and 4 Gaussian components per state.

With an SGMM model (second line), the WER decreases to 67.6%. The SGMM configuration used in this case is: 1000 tied states, $I = 400$ mixture components, single 20 dimensional vector $\mathbf{v}_j$ per state (no substates).

In the next experiment (third line), we first train the system on 16.5h of Spanish and 14.7h German in the multilingual fashion as described in the previous section. From this system, only the SGMM shared parameters are retained and are kept fixed while training the state-specific parameters on one hour of English. In this case, the SGMM configuration is: 1500 tied states, $I = 400$ mixture components, single 40 dimensional vector $\mathbf{v}_j$ per state.

The last result (fourth line) was obtained using exactly the same configuration and training procedure with the only exception that, in the first step, the SGMM shared parameters were trained also on the one hour of English. Adding this small amount of English data gives only a small improvement. The total improvement versus the conventional baseline is very large (10.9% absolute).

## 5. CONCLUSIONS

We have reported experiments with the Subspace Gaussian Mixture Model (SGMM), a new kind of acoustic model that uses Gaussian Mixture Models (GMMs) with the parameter space constrained to a subspace of the total parameter space. We have reported experiments on a multilingual setup where we have a limited amount of training data for each language (about 10 hours). We showed that we could get improvements from jointly training the shared parameters of the model on all languages. We also showed that when the amount of training data for the target language is extremely limited (1 hour), we can get an extremely large WER reduction of 10.9% absolute by using data from other languages to train the shared parameters. This suggests that the SGMM shared parameters are to large extent independent of the language as the parameters learned on resourceful languages can be successfully reused to improve performance for a language with a limited resources.

## 6. REFERENCES


[1] N. Goel et al., "Approaches to Automatic Lexicon Learning with Limited Training Examples," 2010, submitted to: ICASSP.

[2] D. Povey, "A Tutorial Introduction to Subspace Gaussian Mixture Models for Speech Recognition," Tech. Rep. MSR-TR-2009-111, Microsoft Research, 2009.

[3] Hui Lin, Li Deng, Dong Yu, Yifan Gong, Alex Acero, and Chin-Hui Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *ICASSP*, 2009, pp. 4333–4336.

[4] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low-Resource LVCSR systems," 2010, submitted to: ICASSP.

[5] D. Povey, Lukas Burget, et al., "Subspace Gaussian Mixture Models for Speech Recognition," 2010, Submitted to: ICASSP.

[6] A. Ghoshal, D. Povey, et al., "A Novel Estimation of Feature-space MLLR for Full Covariance Models," 2010, Submitted to: ICASSP.

[7] A. Canavan, D. Graff, and G. Zipperlen, *CALLHOME American English Speech*, Linguistic Data Consortium, 1997.

[8] A. Canavan, D. Graff, and G. Zipperlen, *CALLHOME German Speech*, Linguistic Data Consortium, 1997.

[9] A. Canavan, , and G. Zipperlen, *CALLHOME Spanish Speech*, Linguistic Data Consortium, 1997.

[10] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

[11] J.J Godfrey et al., "Switchboard: Telephone speech corpus for research and development," in *ICASSP*, 1992.

[12] D. Graff, *English Gigaword*, Linguistic Data Consortium, 2003.

[13] George Zavaliagkos, Manhung Siu, Thomas Colthurst, and Jayadev Billa, "Using Untranscribed Training Data to Improve Performance," in *ICSLP*, 1998.

[14] Thomas Hain, Philip Woodland, Gunnar Evermann, and Dan Povey, "The cu-htk march 2000 hub5e transcription system," in *In: Proceedings Speech Transcription Workshop*, 2000.