DEALING WITH ACOUSTIC MISMATCH FOR TRAINING MULTILINGUAL SUBSPACE GAUSSIAN MIXTURE MODELS FOR SPEECH RECOGNITION

Aanchan Mohan, Sina Hamidi Ghalehjegh, Richard C Rose

Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

ABSTRACT

The subspace Gaussian mixture model (SGMM) has been recently proposed as an acoustic modeling technique suitable for configuring multilingual speech recognition systems. It is attractive for this purpose since its parametrization allows its "shared" model parameters to be trained with data from multiple languages [1]. In this work, we report on the results of an experimental study carried out with the goal of improving native Spanish language speech recognition performance using an existing telephone speech corpus of English spoken by speakers of Spanish origin. Compensation for sources of acoustic variability between Spanish and English language data sets was found to be important in obtaining good multilingual ASR performance. We conclude with a discussion about the notion of *acoustic similarity* between the state dependent parameters of the SGMM, and its possible use in effectively modelling pronunciation variation.

Index Terms— Multilingual Speech Recognition, Acoustic Modelling, Subspace methods

1. INTRODUCTION

There has been considerable recent interest in configuring ASR systems for a target language using speech data acquired from multiple languages [2, 3]. This work has generally been motivated by scenarios where very limited speech and language resources exist for the given language [4, 5]. In practice, multilingual acoustic model training in ASR has the potential to address a range of problems. For example, in porting a commercial ASR based service to a new language, it may be advantageous to bootstrap the system using an initial model that incorporates data collected from other languages [3, 6]. A more aggressive example is to improve the performance of conversational telephone speech (CTS) ASR system in a given language by incorporating data from existing multilingual speech corpora which is not necessarily from the same task domain. This is the problem which is addressed in this paper.

This problem is particularly interesting if the performance obtained for a particular target language is low due to insufficient language specific training data or other issues that might be inherent to that language. Using existing multilingual data in this manner is attractive since, if it already exists, incorporating the new data can have effectively zero cost. Of course, there are many issues that must be dealt with for this scenario to be practical. These include the issues of defining common phone sets across languages and the extent to which similar phonetic contexts across languages can benefit training for a given language pair [6, 7]. However, just as important, one must account for systematic acoustic differences that might exist between data sets associated with different languages.

Many approaches to multilingual acoustic model training in ASR have focused on sharing data between units by using phoneme inventories which are common across languages [6, 7]. Other approaches have focused on sharing data amongst phones in continuous density hidden Markov models (CDHMMs) which have similar acoustic contexts [8, 6].The performance of all of these approaches are limited by the difficulty associated with specifying phonemic units whose acoustic realizations are consistent across languages. They are also potentially limited by the difficulties associated with sharing data across HMM states in CDHMM models.

Recently, the subspace Gaussian mixture model (SGMM) was shown to be an attractive acoustic modelling technique for modelling acoustic units for multilingual speech recognition [9]. The effectiveness of the SGMM for multilingual ASR in [9] was shown by training the so-called "shared parameters" of the SGMM using combined multilingual data collected under similar acoustic and channel conditions. It could be a cause for concern if the sources of multilingual data are from differing acoustic and channel conditions. Here we describe an experimental study for training an SGMM system in a multilingual fashion as in [9], but using data collected in two significantly different acoustic and channel conditions. We show that the acoustic and channel mismatch between the two sets of data does negatively affect performance of the multilingual SGMM system. We then propose a variant of the Speaker Adaptive Training (SAT) procedure to normalize the data across multiple languages.

The paper is organized as follows. Section 2 briefly describes the SGMM in the context of multilingual ASR. Section 3 describes the task domain and the baseline system. Next, in section 4 we describe our multilingual system, highlighting the effects of the acoustic and channel mismatch between our data sets. We then describe our procedure for compensating for the observed mismatch in Section 5 and present our results. Finally, we provide a discussion on the ability of the SGMM parametrization to model pronunciation variation effectively.

2. THE SUBSPACE GAUSSIAN MIXTURE MODEL

This section provides a brief description of our implementation of the subspace Gaussian mixture model (SGMM) recently proposed by Povey et al. [1]. The description here follows the work of Rose et al. in[10].

For an LVCSR system configured with J states, the observation density for a given D dimensional feature vector, $\mathbf{x}$ for a state $j \in 1 \ldots J$ can be written as,

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji} N(\mathbf{x}|\mu_{ji}, \mathbf{\Sigma_i}), \quad (1)$$

where I full-covariance Gaussians are shared between the J states. The state dependent mean vector, μ_{ji}, for state j is a projection into the ith subspace defined by a linear subspace projection matrix $\mathbf{M}_i$,

$$\mu_{ji} = \mathbf{m}_i + \mathbf{M}_i \mathbf{v}_j \quad (2)$$

This work was partially supported by the FQRNT

In Eq(2) $\mathbf{v}_j$ is the state projection vector for state j. The subspace projection matrix $\mathbf{M}_i$ is of dimension $D \times S$ where S is the dimension of the state projection vector $\mathbf{v}_j$ for state j. In this work, $S = D$. The state specific weights in Eq.(2), are obtained from the state projection vector $\mathbf{v}_j$ using a log-linear model,

$$w_{ji} = \frac{\exp \mathbf{w_i^T} \mathbf{v}_j}{\sum_{i'=1}^{I} \exp \mathbf{w_{i'}^T} \mathbf{v}_j}. \quad (3)$$

In addition, to add more flexibility to the SGMM parametrization at the state level, the concept of substates is adopted where the distribution of a state can be represented by more than one vector $\mathbf{v}_{jm}$, where m is the substate index. This "substate" distribution is again a mixture of Gaussians. The state distribution is then a mixture of substate distributions which are defined as follows:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} N(\mathbf{x}|\mu_{\mathbf{jmi}}, \mathbf{\Sigma_i}) \quad (4)$$

where c_{jm} is the relative weight of substate m in state j and the means and mixture weights are obtained from substate projection vectors, $\mathbf{v}_{jm}$

$$\mu_{jmi} = \mathbf{m}_i + \mathbf{M}_i \mathbf{v}_{jm} \quad (5)$$

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

It is apparent that this acoustic modelling formalism has a large number of "shared" parameters and small number of state specific parameters. For multilingual acoustic modelling the shared parameters, namely $\mathbf{M}_i$, $\mathbf{w_i}$ and $\mathbf{\Sigma_i}$ are trained by pooling data from multiple languages. In addition, while maintaining separate phone sets for each language, the state specific parameters $\mathbf{v}_j$ are only trained from data specific to each language. In [9] the speech recognition performance was shown to be superior when the "shared" parameters were trained with data from multiple languages rather than being trained with data from a single language.

3. TASK DOMAIN AND BASELINE SYSTEMS

The experimental study addressed in this paper deals with a Spanish language (CTS) task as characterized by the Call Home Spanish speech corpus [11]. Multilingual training is performed using a separate English language CTS corpus collected from a population of Hispanic English speakers. This section introduces the task domain and describes how the baseline CallHome and Hispanic English CDHMM and SGMM acoustic models are trained.

3.1. CallHome Spanish and Hispanic English Corpora

The CallHome corpora are known to be a unique challenge for speech recognition [12]. Apart from the small size of the corpora, the data consists of speech between familiar parties with inherent dysfluencies. Burget et al. in [9] use the English, German and Spanish sections of the CallHome speech data for building their multilingual SGMM system. Amongst these individual mono-lingual systems, Spanish was reported as the worst performing system. For this reason we decided to focus our effort on improving the performance on the extremely challenging Spanish language task.

For our multilingual study, we used Spanish inflected English telephone speech data from the Hispanic English corpus [13]. This data has significantly different acoustic and channel conditions in comparison with data from the CallHome database. This corpus itself consists of 20 hours transcribed data of spontaneous telephone conversations between non-native speakers of English whose native language is Spanish. We used only a 6 hour subset of this data, which we call the "clean" subset which consists of utterances whose transcriptions did not have mispronounced words, flagged background events, false starts and any words that were labelled as unrecognizable by the transcribers.

Table 1. Word recognition performance for Spanish. All SGMM systems here are trained with $I = 400$ shared Gaussians. In this table JPI - Joint Posterior Initialization; FSInit - Flat Start Initialization

System	Initialization	# substates	WER [%]
Baseline CDHMM	n/a	n/a	68.61
Monolingual SGMM	FSInit	1604	67.43
Monolingual SGMM	JPI	1604	67.14
Monolingual SGMM	JPI	6000	66.62

3.2. Spanish Language CDHMM and the SGMM Training

The baseline system was based on conventional three state left-to-right HMM triphone models. Decision tree clustering was used to obtain a system with 1604 states. We used 16 Gaussians per state. The features are 13 PLP coefficients, with Δ and $\Delta\Delta$ and speaker-wise mean and variance normalization. We used 16.5 hours of conversational speech data for training, and our test data consisted 2.0 hours conversational speech data. A trigram LM was used with a vocabulary of 45k words. This trigram LM was trained on the Spanish CallHome transcripts and data obtained by crawling the web for sentences containing high frequency bigrams and trigrams occurring in the training text of the Callhome corpus.

Next, we describe SGMM system for the same Spanish Language task. The training data for this system is the same as that used for the baseline CallHome CDHMM system. In addition we used the same language model for recognition as the baseline CallHome CDHMM system. As mentioned in section 2, an SGMM system without any substates is a single substate per state system. The Spanish language single substate per state system consists of $J = 1604$ states, with $I = 400$ full-covariance Gaussians shared between the states. This system was initialized with Gaussians obtained from a Universal Background Model (UBM) obtained from speech-only segments of all the speakers in the training corpus. The SGMM training was carried out as mentioned in [10].

The word error rate performance for the systems described in this section appear in Table 1. The first line indicates the performance of the baseline continuous density hidden Markov model (CDHMM) system, with a word error rate (WER) of 68.61%. While thus baseline WER is high, it is consistent with performance previously reported under similar scenarios for this task[9] and [12]. Next we discuss the WER obtained for the mono-lingual Spanish SGMM system. We experimented with two initialization schemes for the SGMM: flat start initialization (**FSInit** in Table 1) and joint posterior initialization (**JPI** in Table 1) [10]. The advantage of using the SGMM acoustic modelling formalism for this task compared to the CDHMM is apparent from WER performance results in line 2 and line 3 of Table 1. While the joint posterior initialization for the SGMM resulted in the best performing mono-lingual system, multilingual models are initialized using a flat start initialization due to limited computational resources. Also, the concept of substates was briefly introduced in Section 2. To see the effect that the introduction of substates has on the performance, we refer the reader to

line 4 of Table 1. We see that introducing substates does decrease the WER performance even further to 66.62%. It is clear from the table that the monolingual SGMM system provides a 2% absolute decrease in WER with respect to the baseline CDHMM system.

3.3. Hispanic English CDHMM System

The training data that we used for this CDHMM system consists of the"clean" subset of the Hispanic English corpus described in section 3.1. We trained this system with similar specifications for the CDHMM Callhome system mentioned in Section 3.2. We obtained a system with 773 left-to-right tri-phone states after decision tree clustering.

4. MULTILINGUAL SGMM SYSTEM

As mentioned in Section 1, we are interested in studying the impact of configuring a multilingual system with data not necessarily from the same task domain. Having described the procedure for training a Spanish language SGMM in the previous section, here we describe the procedure similar to [9] for training our multilingual SGMM system.

SGMM training was carried out in a multilingual fashion, by using the Callhome and the "clean" Hispanic English data to train the "shared" parameters, while maintaining distinct phone sets for the two languages. Maintaining separate phone sets allows for training the state specific parameter training only with data from each of the two languages individually. Our multilingual SGMM has a total of $J = 2377$ states, with 1604 states coming from Spanish and 773 coming form the Hispanic English system. The system was initialized with a UBM with $I = 400$ Gaussians trained on speech-only segments of speakers from both corpora. The system was initialized only with a flat-start. This was because the joint-posterior initialization procedure for calculating the initial state and mixture dependent posterior probabilities is rather time consuming with only a marginal gain in performance as is reported on our mono-lingual Spanish SGMM results.

We report results for our multilingual system in Table 2 only with respect to its performance on the CallHome Spanish test set. It is apparent that there is a slight degradation in the performance of the system with a WER of 67.94%, when compared with the performance of the single substate per state monolingual SGMM system initialized using the flat start initialization procedure which had a WER 67.43%. Instead of observing a gain in performance due to the ability of the SGMM to capture shared acoustic phonetic structures across languages, we observe the contrary. We believe degradation is due to a mismatch in the acoustic environment and channel conditions between the two sets of data. The cepstrum mean removal and variance normalization used in the feature extraction process appears to be inadequate to deal with this mismatch. If we could remove inter-speaker variation and variation due to environmental acoustic mismatch, then we conjecture that the subspace matrices $\mathbf{M}_i$ would model phonetic variability more accurately.

5. SPEAKER AND ENVIRONMENT MISMATCH

The multilingual SGMM model described in Section 4 was trained using data taken from the CallHome and Hispanic English corpora with no explicit mechanism for accounting for the acoustic mismatch between the two corpora. As a result, there was no improvement in WER obtained using the multilingual model. This section presents an acoustic normalization procedure for dealing with this cross-corpus mismatch. The procedure is applied as part of multilingual SGMM training and corresponds to a straight-forward variant of speaker adaptive training (SAT) [14] and constrained maximum likelihood linear regression (CMLLR). The procedure is explained in two-steps as follows.

Given our training feature sets $\mathbf{X} = [X_r^{CH}, X_r^{HE}]$ and the SI CDHMM models $\mathbf{\Lambda} = [\Lambda^{CH}, \Lambda^{HE}]$ we first estimate speaker dependent CMLLR matrices A_r^{CH} and A_r^{HE}, as is done in standard SAT. Here we use the index r to denote a speaker in our combined training data set. In addition we obtain a transformed feature set $\hat{\mathbf{X}}$ by transforming our original feature set $\mathbf{X}$ using the speaker dependent transformations A_r^{CH} and A_r^{HE}. Interpreting the CMLLR transformation as a feature space transformation [14] we use the transformed features $\hat{\mathbf{X}}$ in the next step of SAT, to train the speaker-normalized models $\hat{\mathbf{\Lambda}} = [\hat{\Lambda}^{CH}, \hat{\Lambda}^{HE}]$. Similarly, we transform our CallHome test features $\mathbf{Y}^{CH} = [Y_r^{CH}]$ using speaker dependent CMLLR transformations to obtain a new set of test features $\hat{\mathbf{Y}}^{CH}$.

Since the first step normalizes only inter-speaker variation and not inter-corpus acoustic mismatch we perform a modified version of the SAT procedure in a second step. Using the model $\hat{\mathbf{\Lambda}}$ and the feature set $\hat{\mathbf{X}}$ we obtain a single CMLLR transformation matrix $\mathbf{B}$. We then use $\mathbf{B}$ to obtain an updated set of features $\hat{\hat{\mathbf{X}}}$ by transforming the features $\hat{\mathbf{X}}$ and a new set of test features $\hat{\hat{\mathbf{Y}}}^{CH}$ is obtained by transforming $\hat{\mathbf{Y}}^{CH}$. We use $\hat{\hat{\mathbf{X}}}$ to obtain the updated model $\hat{\hat{\mathbf{\Lambda}}}$. We then use these new set of training features $\hat{\hat{\mathbf{X}}}$ to train our new multilingual SGMM system trained in the exact same manner as mentioned in section 4. In addition we report results using the transformed CallHome test feature set $\hat{\hat{\mathbf{Y}}}^{CH}$.

Clearly there are many ways of compensating for variability across speakers and across corpora. The advantage of transforming features from both data sets as described in this section is that it serves to reduce mismatch across all stages of SGMM training.

Table 2. Word recognition performance for Multilingual SGMM Systems with $I = 400$ shared Gaussians. All systems were initialized with a Flat Start. NC indicates that speaker and environment compensated features were used in multilingual training

System	# substates	WER [%]
Baseline CDHMM	n/a	68.61
CDHMM+SAT	n/a	65.71
Multilingual SGMM	2377	67.94
Multilingual-NC SGMM	2377	64.82
Multilingual-NC SGMM	6000	64.7

6. RESULTS AND DISCUSSION

We report the ASR results for this speaker and environmental compensated SGMM model which appears as Multilingual-NC SGMM in Table 2. We see a significant decrease in the WER (approximately a 4% absolute with respect to the baseline) after compensating for both speaker and environment variation. In addition a slight decrease in WER is seen by increasing the number of substates in the SGMM model. This result verifies our hypothesis that inter-corpus acoustic mismatch and inter-speaker variability could significantly affect the performance of multilingual systems configured with the SGMM.

Summarizing the results in Table 2, firstly we see that with well known techniques such as Speaker Adaptive Training followed by CMLLR Adaptation of test utterances gives us approximately a 3% absolute decrease in WER compared to the baseline system. Secondly, we see that with multilingual SGMM training without any

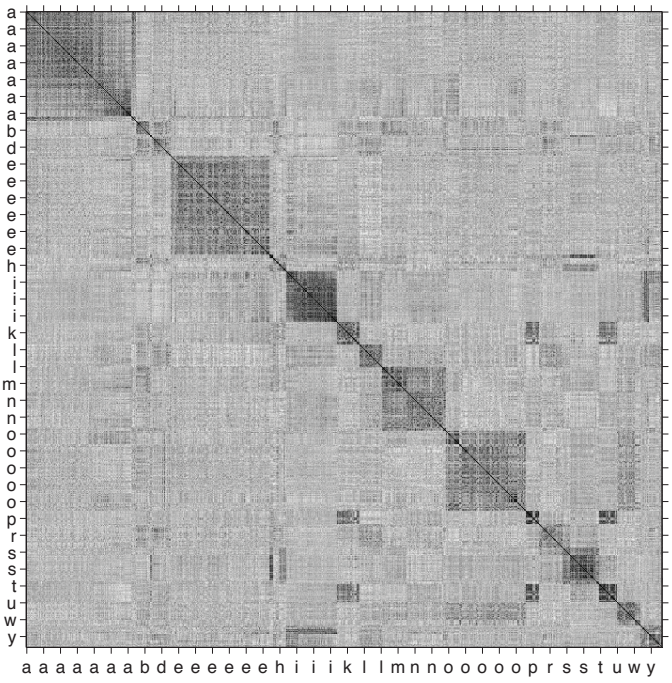

Fig. 1. Dissimilarity matrix between the state projection vectors $\mathbf{v}_j$ of the center states of the phones of the mono-lingual CallHome SGMM system computed as a cosine distance

compensation at all we get an increase in WER compared to the performance of the mono-lingual SGMM system. The WER displayed in the last row of the table shows that the multilingual SGMM model estimated from noise compensated features represents a 4% absolute decrease in WER with respect to the baseline CDHMM and a 1% reduction in WER with respect to the SAT adapted model.

The plot in Figure 1 shows an anecdotal investigation of the ability of the SGMM parametrization to characterize pronunciation variability at the level of the phone and the SGMM state. More specifically, Figure 1 shows a dissimilarity matrix computed between the state projection vectors $\mathbf{v}_j$ of the center states of context-dependent phones of the Spanish language SGMM system. For the purposes of viewing, not all of the center context labels are displayed along the axes of the plot. Dark regions indicate regions of high similarity. From this block diagonal structure, the high-self similarity within a group of phones (e.g., context specific phones with center context "a") is immediately apparent. It is evident from some of the significant off-diagonal distances in the plot that they reflect common consonant confusable pairs differing only in their place of articulation. It is clear that the state projection vectors have the potential to provide a good characterization of pronunciation variability. These state dependent parameters appear to be a compact representation that characterizes pronunciation variability effectively.

7. CONCLUSIONS

A scenario was presented where the two sets of data used for SGMM multilingual training are not necessarily from the same task domain. Instead of observing SGMM's ability to leverage shared acoustic phonetic information across multiple languages for increased ASR performance, we observe contrary results. A simple variant of Speaker Adaptive Training was proposed to obtain features for training and testing that are compensated for speaker and environmental variation. Using these compensated features, a decrease in the word error rate performance is seen for the multilingual system. In addition we found that multilingual SGMM training with noise compensated features proved to give better ASR performance compared to performance with traditional CDHMM Speaker Adaptive Training. We can conclude from this experimental study that multilingual training of SGMMs can be effective despite the existence of acoustic mismatch across the multilingual corpora.

8. REFERENCES

[1] Daniel Povey and et al., "The subspace Gaussian mixture model- a structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, no. 2, April 2011.

[2] A. Mandal, D. Vergyri, M. Akbacak, C. Richey, and A. Kathol, "Acoustic data sharing for Afghan and Persian languages," in *ICASSP '11*, may 2011, pp. 4996 –4999.

[3] Ngoc Thang Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *ICASSP '11*, may 2011, pp. 5000 –5003.

[4] Ngoc Thang Vu, F. Kraus, and T. Schultz, "Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training," in *InterSpeech '11*, August 2011.

[5] Scott Novotney, Rich Schwartz, and Sanjeev Khudanpur, "Unsupervised Arabic dialect adaptation with self-training," in *InterSpeech '11*, August 2011.

[6] Tanja Schultz and Alex Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *SPEECH COMMUNICATION*, vol. 35, pp. 31–51, 2001.

[7] Hui Lin et al., "A study on multilingual acoustic modeling for large vocabulary ASR," in *ICASSP '09*, april 2009.

[8] Thomas Niesler, "Language-dependent state clustering for multilingual acoustic modelling," *Speech Communication*, vol. 49, no. 6, pp. 453 – 463, 2007.

[9] Lukas Burget et al., "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models.," in *ICASSP'10*.

[10] R. Rose, Shou-Chun Yin, and Yun Tang, "An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition," in *ICASSP'11*.

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

[12] George Zavaliagkos, Manhung Siu, Thomas Colthurst, and Jayadev Billa, "Using untranscribed training data to improve performance," in *ICSLP*, 1998.

[13] William Byrne, Eva Knodt, Sanjeev Khudanpur, and Jared Bernstein, "Is automatic speech recognition ready for non-native speech? a data collection effort and initial experiments in modeling conversational Hispanic English," in *In Proceedings of Speech Technology in Language Learning, Marholmen, Sweden. European Speech Communication Association*, 1998, pp. 37–40.

[14] M. J. F. Gales, "Multiple-cluster adaptive training schemes," in *ICASSP '01*, 2001.

[15] T. Schultz and K. Kirchhoff, *Multilingual speech processing*, Elsevier Academic Press, 2006.