Regularized subspace Gaussian mixture models for cross-lingual speech recognition

Lu, L., Ghoshal, A., & Renals, S. (2011). Regularized subspace Gaussian mixture models for cross-lingual speech recognition. In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. (pp. 365-370). IEEE. doi: 10.1109/ASRU.2011.6163959

Regularized Subspace Gaussian Mixture Models for Cross-lingual Speech Recognition

Liang Lu[1], Arnab Ghoshal[2], and Steve Renals[1]

[1] *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, UK*
{liang.lu, s.renals}@ed.ac.uk

[2] *Saarland University, 66123 Saarbrücken, Germany*
aghoshal@lsv.uni-saarland.de

***Abstract*—We investigate cross-lingual acoustic modelling for low resource languages using the subspace Gaussian mixture model (SGMM). We assume the presence of acoustic models trained on multiple source languages, and use the global subspace parameters from those models for improved modelling in a target language with limited amounts of transcribed speech. Experiments on the GlobalPhone corpus using Spanish, Portuguese, and Swedish as source languages and German as target language (with 1 hour and 5 hours of transcribed audio) show that multilingually trained SGMM shared parameters result in lower word error rates (WERs) than using those from a single source language. We also show that regularizing the estimation of the SGMM state vectors by penalizing their ℓ_1-norm help to overcome numerical instabilities and lead to lower WER.**

I. Introduction

Large vocabulary continuous speech recognition systems rely on the availability of substantial resources including transcribed speech for acoustic model estimation, text for language model estimation, and a pronunciation dictionary. Building a speech recognition system from scratch for a new language thus requires considerable investment in such resources. Cross-lingual acoustic modelling has the aim of significantly reducing the amount of acoustic training data for a new target language, by leveraging on existing acoustic models for other, source, languages. However, owing to differences such as different sets of subword units, this is not a straightforward task. There have been three main approaches to cross-lingual acoustic modelling: the use of *global phone sets*, cross-lingual *phone/acoustic* mapping, and cross-lingual *tandem features*.

Schultz and co-researchers [1], [2], [3], [4] have investigated the construction of language-independent speech recognition systems by pooling together all the phoneme units, as well as the acoustic training data, from a set of monolingual systems. The resultant multilingual acoustic model may be used to perform transcription directly, or may serve as a seed model to be bootstrapped or adapted to the target language [1], [3]. More recently, this approach has been extended to include confidence scoring for cross-language bootstrapping and unsupervised training [5], [6].

Rather than constructing a global phone set, the mismatch of phone units between source and target languages may be addressed by a direct cross-lingual mapping between phones or between acoustic models. Both knowledge-based [7], [8] and data-driven [9], [10] approaches have been investigated. Given a cross-lingual mapping, either the target language acoustic model is derived from the source language model, or the transcription of target language speech is performed using the mapped source language acoustic model [10].

Tandem features, based on phone posterior probability estimates, were originally proposed to improve monolingual speech recognition [11], but they have also proven effective in the cross-lingual setting. In this approach, multi-layer perceptrons (MLPs) trained using source language acoustic data of source language, are used to generate the MLP phone posterior features for the target language [12], [13], [14], [15]. As tandem acoustic features are not directly dependent on the lexicon, this approach is simple to apply. In addition, the training data of the target language can also be used to adapt the MLPs to fit the target system better [14].

In this paper, we investigate a new approach for cross-lingual acoustic modelling, based on the framework of the subspace Gaussian mixture model (SGMM) [16]. In SGMMs, the hidden Markov model (HMM) parameters are inferred subject to a globally shared model subspace which captures the principal model variations, as opposed to conventional direct estimation. Phonetic and speaker variabilities, which are key factors affecting recognition accuracy, can be modelled using separate model subspaces in SGMMs [16]. As the model subspace is independent of the HMM architecture, it may be shown that the phonetic model subspace can be shared across languages and can be trained using multilingual acoustic data. This model subspace may be used to estimate models for a new language with limited training data [17].

We have further developed these ideas, comparing the performance of model subspaces estimated from the source language in both monolingual and multi-lingual settings, and using a regularized SGMM estimation approach [18] to address numerical instabilities and possible overfitting that arise in the case of highly limited training data. We have performed experiments using the GlobalPhone corpus [4], using German as target language and Spanish, Portuguese and Swedish as source languages. We examined two evaluation conditions with 1 hour and 5 hours of target language training data respectively. We observed considerable reductions in word error rate (WER) when using a multilingual subspace, and achieved further WER reductions using ℓ_1-norm regularization.

II. SGMM Acoustic Models

SGMM acoustic models are similar to conventional Gaussian mixture model (GMM) systems, in that the output pdf of each HMM state is a GMM. The principal difference is that the Gaussian means and the mixture component weights are derived from the phonetic and speaker subspaces, in order to capture the corresponding variability together with the weight projection [16]. In addition, the covariance matrices (which are normally full rather than diagonal) are shared between all the HMM states. The model may be expressed formally as:

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

$$\boldsymbol{\mu}_{jmi}^{(s)} = \mathbf{M}_i \mathbf{v}_{jm} + \mathbf{N}_i \mathbf{v}^{(s)} \quad (2)$$

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

where $\mathbf{o}_t \in \mathbb{R}^F$ denotes the t-th F-dimensional acoustic frame, j is the HMM state index, m is a sub-state [16], I is the number of Gaussians, and $\boldsymbol{\Sigma}_i$ is the i-th covariance matrix. $\mathbf{v}_{jm} \in \mathbb{R}^S$ is referred to as the sub-state vector, and S denotes the subspace dimension. The matrices $\mathbf{M}_i$ and the vectors $\mathbf{w}_i$ span the model subspaces for Gaussian means and weights respectively, and are used to derive the GMM parameters given sub-state vectors (equations (2) and (3)). Similarly, $\mathbf{N}_i$ defines the speaker subspace for Gaussian means, and $\mathbf{v}^{(s)} \in \mathbb{R}^T$ is referred as the speaker vector where T denotes the dimension of the speaker subspace. However, in this paper, we do not perform speaker adaptive training using the speaker subspace.

A. Regularized State Vector Estimation

Maximum likelihood parameter estimation for SGMMs is discussed by Povey et al. [16], in which the sub-state vector $\hat{\mathbf{v}}$ is estimated by maximizing the following objective function:

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} -\frac{1}{2}\mathbf{v}^T\mathbf{H}\mathbf{v} + \mathbf{v}^T\mathbf{y}, \quad (4)$$

where $\mathbf{y}$ is an S-dimensional vector and $\mathbf{H}$ is an $S \times S$ matrix, representing the first- and second-order statistics respectively. While the solution for non-singular $\mathbf{H}$ is $\hat{\mathbf{v}} = \mathbf{H}^{-1}\mathbf{y}$, Povey et al. [16] presented a more practical approach to cover the general case in which $\mathbf{H}$ may be singular.

In the case of limited acoustic training data, the system of equations for estimating the sub-state vectors may be underdetermined or ill-conditioned. Moreover, maximum likelihood estimation can lead to model overfitting. In our previous work, we addressed these by regularizing the objective function [18]:

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} -\frac{1}{2}\mathbf{v}^T\mathbf{H}\mathbf{v} + \mathbf{v}^T\mathbf{y} - J_{\boldsymbol{\lambda}}(\mathbf{v}), \quad (5)$$

where $J_{\boldsymbol{\lambda}}(\mathbf{v})$ denotes a regularization function parametrized by $\boldsymbol{\lambda}$. From a Bayesian perspective, $J_{\boldsymbol{\lambda}}(\mathbf{v})$ may be interpreted as a negative log-prior for the sub-state vector, in which case equation (5) may be viewed as a *maximum a posteriori* (MAP) estimate. Regularization functions investigated include the ℓ_1-norm and the ℓ_2-norm penalties, corresponding to Laplace and Gaussian priors respectively, as well as their combination, referred to as elastic net regularization. Our previous experiments indicated that ℓ_1-norm regularization offers slightly better performance in terms of both recognition accuracy and model robustness [18]. Hence, in this paper, we have concentrated on ℓ_1-norm regularization.

B. Modified Regularization

Regularizing all coefficients in the sub-state vectors (5), forces the sub-state vector to shrink towards zero, corresponding to a prior on the Gaussian means centered at the origin. This is clearly demonstrated in Figure 6, where hundreds of sub-state vectors are shrunk to zero for systems with a larger number of sub-states. Intuitively, we would prefer a prior that fits the general distribution of the data. To achieve this, we modify the regularization such that the Gaussian means shrink towards a universal background model (UBM). This is done by setting the first coefficient of sub-state vector $\mathbf{v}$ to be 1, which also forces the first column of phonetic subspace matrices $\mathbf{M}_i$ to be the UBM means during model update. The regularization penalty is applied to the remaining sub-state coefficients. For ℓ_1-norm regularization, the objective function becomes:

$$\begin{aligned} \hat{\mathbf{v}} = \arg\max_{\mathbf{v}} & -\frac{1}{2}\mathbf{v}^T\mathbf{H}\mathbf{v} + \mathbf{v}^T\mathbf{y} - \lambda\|\mathbf{v}\|_{\ell_1}, \\ & s.t. \quad \lambda \geq 0, \quad \mathbf{v}[1] = 1. \end{aligned} \quad (6)$$

where we are fixing the first coefficient of $\mathbf{v}$ to be exactly one. If we adopt the following expressions:

$$\hat{\mathbf{v}} = \begin{bmatrix} 1 \\ \hat{\mathbf{v}}^* \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 1 \\ \mathbf{v}^* \end{bmatrix}, \mathbf{y} = \begin{bmatrix} a \\ \mathbf{y}^* \end{bmatrix}, \mathbf{H} = \begin{bmatrix} b & \mathbf{h}^T \\ \mathbf{h} & \mathbf{H}^* \end{bmatrix},$$

then equation (6) is equivalent to

$$\begin{aligned} \hat{\mathbf{v}}^* = \arg\max_{\mathbf{v}^*} & -\frac{1}{2}\mathbf{v}^{*T}\mathbf{H}^*\mathbf{v}^* + \mathbf{v}^{*T}(\mathbf{y}^* - \mathbf{h}) - \lambda\|\mathbf{v}^*\|_{\ell_1}, \\ & s.t. \quad \lambda \geq 0. \end{aligned} \quad (7)$$

This modified regularization, does not involve any change in accumulating the statistics of $\mathbf{H}$ and $\mathbf{y}$, and the code for the original regularization [18] can be reused.

III. Experimental Setup

We have performed a set of speech recognition experiments using the GlobalPhone corpus [4], with German as the target language. Our main experiments have investigated low-resource cases, in which we have used one hour and five hours of target language acoustic training data. Before presenting these cross-lingual experiments, we give a brief description of the corpus and system configuration for our experiments, and give results for the baseline monolingual systems, trained on the complete data sets.

A. GlobalPhone Corpus

The GlobalPhone corpus [4] contains up to 20 languages including English, Arabic, Chinese and a number of European languages, and consists of recordings of a range of speakers reading newspapers in their native language. There are about

TABLE I
THE NUMBER OF PHONES AND SPEAKERS, THE AMOUNT OF TRAINING AND DEVELOPMENT DATA (HOURS) FOR THE 4 LANGUAGES USED IN THIS PAPER.

Language	#Phones	#Speakers	Trn(h)	Dev(h)
German (GE)	44	77	14.8	2.0
Spanish (SP)	43	97	17.2	2.0
Portuguese (PT)	48	101	22.6	1.5
Swedish (SW)	52	98	17.4	2.0

TABLE II
WORD ERROR RATES (WER %) OF GMM AND SGMM BASELINE OF TARGET AND SOURCE LANGUAGES ON DEV DATASET.

Language	LM	PPL	OOV	Dict	GMM	SGMM
GE	Trigram	422	5.2%	17k	25.7	24.0
SP	Bigram	306	4.8%	17k	33.7	30.4
PT	Bigram	393	4.3%	52k	29.3	25.9
SW	Trigram	940	0%	23k	47.2	40.8

100 speakers for each language, and recordings were made under a range of 'quiet' conditions, resulting in about 15–20 hours of high quality speech for each language. However, since the recording locations vary, acoustic conditions also vary both within and between each language. Hence, corpus mismatch may degrade the performance of cross-lingual systems.

In these experiments, German (GE) was used as the target language, and Spanish (SP), Portuguese (PT), and Swedish (SW) as the source languages. Table I describes the data for each language used in the experiments in terms of the number of phonemes and speakers, the size of lexicon, and the amount of training and development data.

Our baseline monolingual systems, described below, used the complete training sets for each language. In the cross-lingual experiments (section IV), we used the full training sets for the source languages, but limited training sets (1 hour and 5 hours) for the target language.

B. Baseline Monolingual Systems

We constructed GMM-based baseline systems for each of the four languages, based on the system of Lal [15]. We used 12th order mel-frequency cepstral coefficients, plus energy, with first and second derivatives, to give a 39-dimension acoustic feature vector. We applied cepstral mean and variance normalization, and used HTK[1] to build the acoustic models. For the German GMM baseline, we used 3125 triphone states, and 16 mixture components.

The baseline monolingual SGMM systems used the same acoustic feature vectors and the same context dependent phone clustering as the corresponding baseline GMM system. We set the number of Gaussians $I = 400$, and the sub-state vector dimension $S = 40$. The open source Kaldi software[2] was used for the SGMM systems in this paper. Table II gives the baseline monolingual WERs for each of the four languages. We used trigram language models for all experiments, except the baseline monolingual systems for Spanish and Portuguese.

[1]http://htk.eng.cam.ac.uk
[2]http://kaldi.sourceforge.net/

TABLE III
TOTAL TRACE OF COVARIANCE AND SUBSPACE MATRICES GIVEN BY THE SOURCE SGMM SYSTEMS, $S = 40$.

	SP	PT	SW	Multilingual
# of states	2298	3140	3153	-
# of sub-states	20k	20k	20k	-
$\sum_i tr(\mathbf{\Sigma}_i)/10^3$	8.02	8.07	8.14	8.15
$\sum_i tr(\mathbf{M}_i\mathbf{M}_i^T)/10^3$	16.1	12.9	11.9	11.2

In our cross-lingual experiments, described below, the source language (SP, PT, SW) baseline monolingual SGMM systems were used to provide the globally shared subspace parameters $\mathbf{\Sigma}_i$, $\mathbf{M}_i$ and $\mathbf{w}_i$. Following Burget et al. [17], a multilingual SGMM was constructed by tying $\mathbf{\Sigma}_i$, $\mathbf{M}_i$ and $\mathbf{w}_i$ across the source language systems. We renormalized the phonetic subspace [19] (Appendix K) to concentrate the most important variation in the lower-numbered dimensions. This also allowed us to use a lower dimension subspace for cross-lingual systems without retraining the subspace parameters $\mathbf{M}_i$ and $\mathbf{w}_i$. Table III shows the size of covariance matrices $\mathbf{\Sigma}_i$ and phonetic subspace matrices $\mathbf{M}_i$ estimated in both monolingual and multilingual fashion.

IV. CROSS-LINGUAL EXPERIMENTS

In this section we present results of the cross-lingual acoustic modelling experiments, in cases where we have limited target language resources. Our experiments use German as the target language and we have investigated two levels of limited acoustic training data, 1 hour and 5 hours, in order to provide an estimate of the amount of transcribed speech needed for an acceptable recognition accuracy. These training data subsets were randomly selected from the complete German training set, from which we selected 7–8 minutes of recorded speech from each of 8 and 40 speakers for the 1 hour and 5 hour systems respectively. The globally-shared SGMM parameters for these cross-lingual systems were obtained from monolingual source language systems, or from the multilingually-trained SGMM using all the source languages.

A. Cross-lingual Experiments: 1 Hour Training Data

We trained baseline monolingual systems and cross-lingual systems using 1 hour of target language (GE) data. The GMM baseline system had 620 triphone states, each of which was modelled using 4-component GMMs. The baseline and the cross-lingual SGMM systems used the same context-dependent phonetic clustering as the GMM system, and the dimension of sub-state vector is set to be $S = 20$ for all the SGMM systems. In the baseline SGMM systems, all the parameters in equations (1–3) were updated: the sub-state vectors $\mathbf{v}_{jm}$ and the globally shared parameters $\mathbf{M}_i, \mathbf{w}_i$ and $\mathbf{\Sigma}_i$. In the cross-lingual systems, only the sub-state vectors $\mathbf{v}_{jm}$ were re-estimated, with the globally shared parameters taken from the source language or multilingual systems. As discussed earlier, the subspace parameters were renormalized and we only used a 20-dimension subspace for the cross-lingual system without retraining these parameters.

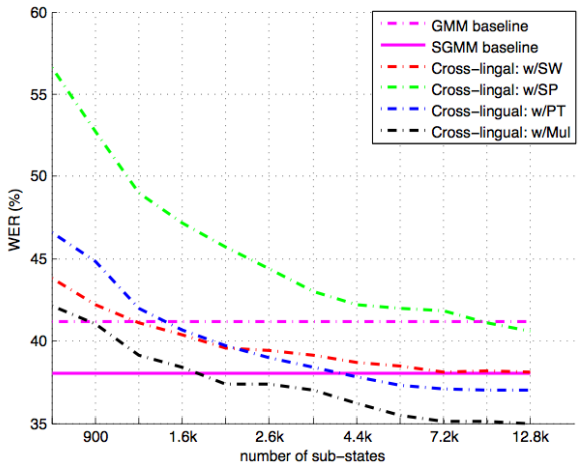

Fig. 1. 1 Hour Training Data: WER of baseline GMM and SGMM system (41.2% vs. 38.0%) as well as cross-lingual systems. For all SGMM systems, the dimensionality of the sub-state vectors is set to be $S = 20$. The lowest WER (35%) is obtained from the multilingual subspace system (w/Mul).

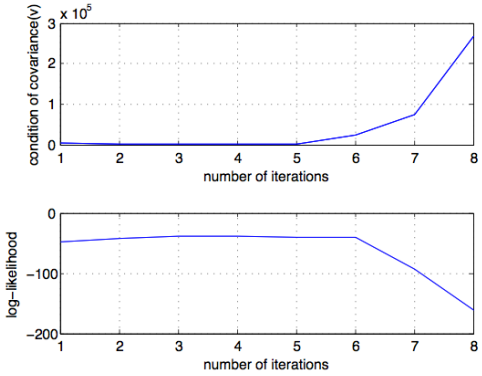

Fig. 2. 1 Hour Training Data: Training of 40-dimensional sub-state vectors showed numerical instability after a few iterations. Condition number of the covariance matrix of the sub-state vectors start increasing rapidly and the estimation failed, leading to a decrease in the log-likelihood. This was fixed using a regularized sub-state vector update.

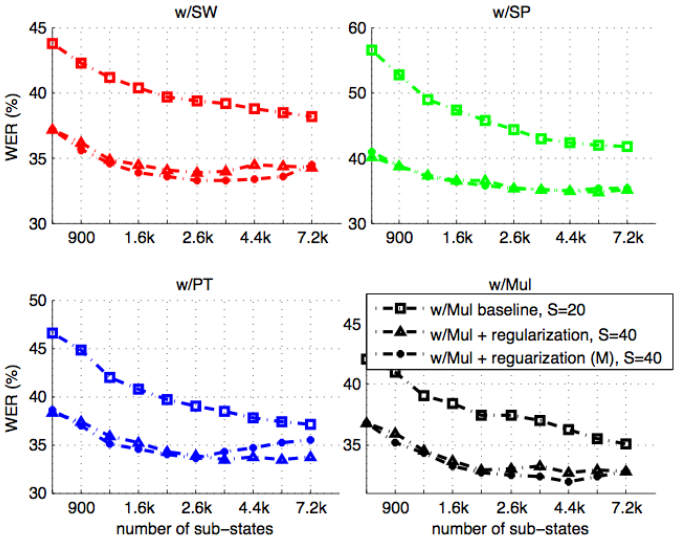

Fig. 3. 1 Hour Training Data: WER of cross-lingual systems with and without ℓ_1-norm regularization. Regularization does not bring performance gains but slight degradation for systems with 20-dimension state vectors, and there are omitted in the figure for clarity. For the multilingual subspace system (w/Mul) with $S = 40$, the best performance is 32.7% by original regularization and 31.9% by modified regularization (M).

In contrast to Burget et al. [17], the number of tied states was fixed rather than being increased (from 500 (GMM) to 1000 (SGMM) to 1500 (multilingual SGMM), in their experiments on CallHome). This clearly demonstrates that the improvements are due to better estimation of the parameters, and not because the SGMM systems allows for estimation of a larger number of context-dependent models for the same amount of data. We expect that the results reported here will improve further by increasing the number of tied states.

The WERs for the monolingual GMM and SGMM systems trained on 1 hour of data were 41.2% and 38.0% respectively, a significant increase in WER compared with the case when they were trained with the complete 14.8 hour training data set. The SGMM system again has a considerably lower WER than the GMM system, as was observed by Burget et al. [17]. The performance of the cross-lingual systems is shown in Figure 1 with the globally shared parameters obtained from each of the source language systems, as well as the tied multilingual system. The results indicate that the system with multilingually trained subspace parameters results in considerably lower WERs compared with the other cross-lingual systems derived from a single source language, as well as compared with the SGMM baseline. We may also observe that the cross-lingual system with Spanish subspace (denoted as "w/SP") results in a higher WER compared with the other cross-lingual systems. In addition to factors such as linguistic differences and corpus mismatch, this difference may also be due to the larger model subspace in the Spanish system (Table III) which may make it harder for the data to saturate the model.

While training a model with 40-dimensional sub-state vectors (i.e. $S = 40$) we faced numerical instabilities. This is shown in Figure 2, where the condition number of the covariance matrix of the sub-state vectors start increasing rapidly and the estimation failed, leading to a decrease in the log-likelihood. Other measures like the determinant and trace of the covariance of $\mathbf{v}_{jm}$ as well as the maximum and minimum elements of the vectors show similar trends. The precise reason for this instability is currently under investigation.

However, using a regurlarized estimation of the 40-dimensional sub-state vectors, with a relatively strong penalty on their ℓ_1-norm, we could not only train the systems but also observe significantly lower WERs for all cross-lingual systems (Figure 3). Applying ℓ_1-norm regularization to the 20-dimension cross-lingual SGMM systems results in slightly higher WERs, even with a relatively weak regularization penalty. This is probably due to the relative simplicity of the

TABLE IV
MEAN AND VARIANCE OF THE 1^{st} COEFFICIENT IN THE STATE VECTORS OF THE CROSS-LINGUAL SYSTEMS WITHOUT REGULARIZATION.

System	w/SW	w/SP	w/PT	w/Mul
Mean	1.006	0.889	0.946	1.019
Varance ($\times 10^{-3}$)	9.75	16.83	17.53	9.33

TABLE V
WER OF GMM AND SGMM BASELINE SYSTEMS WITH 5 HOUR TRAINING DATA.

System	WER(%)
GMM baseline	34.3
SGMM baseline, $S = 20$	31.1
SGMM baseline, $S = 40$	32.0

20-dimensional subspace model.

Results for the modified regularization (equation 7) are also shown in Figure 3. The modification is based on the assumption that the first column of $\mathbf{M}_i$ corresponds to global means, which should be learned by fixing the first coefficient of $\mathbf{v}_{jm}$ to 1. However, the source language systems were not trained with this constraint, leading to a potential mismatch. Yet we found that modified regularization led to modest improvements in WER for subspaces trained using Swedish and multilingual data, but not for Spanish and Portuguese. Table IV shows that this was due to serendipity, as the first element of the state vectors had a mean value of nearly 1 with low variance for the Swedish and multilingual systems, but that was not the case for the other two systems in which modified regularization was ineffective.

B. Cross-lingual Experiments: 5 Hour Training Data

We increased the amount of target language acoustic training data to 5 hours and trained the baseline monolingual GMM and SGMM systems. The GMM system had 1561 tied triphone states and each state is modelled by an 8-component GMM. As before, all the following SGMM systems share a context-dependent phonetic clustering with GMM system. WERs of these baseline monolingual systems are shown in Table V.

The results of the cross-lingual systems are shown in Figure 4 where the dimension of sub-state vectors are still $S = 20$. Again, the cross-lingual system with multilingual subspace parameters denoted as ("w/Mul") results in the lowest WER. The results of the regularized cross-lingual SGMM systems are shown (for the multilingual subspace) in Figure 5, and the results are broadly consistent with the 1 hour case. We did not observe any reductions in WER when regularizing the baseline system ($S = 20$), but we were again able to observe significant reductions in WER when regularizing a cross-lingual SGMM with dimension $S = 40$.[3] A small improvement in WER was observed using the modified regularization approach, from 26.8% (original) to 26.6% (modified).

C. Sparsity Analysis

The ℓ_1-norm regularization used in this paper is able to penalize the model complexity of cross-lingual systems in order to achieve model robustness. In addition, it also has the effect of driving some coefficients to zero, thus leading to a kind of variable selection, where the most relevant bases from $\mathbf{M}_i$ and $\mathbf{w}_i$ get used. In Figure 6 we can see the proportion of parameters set to zero by the ℓ_1-regularization. Not surprisingly, with sub-state splitting, the sub-state vectors are driven to be increasingly sparse as the amount of acoustic frames aligned to each sub-state decrease accordingly.

[3]Once again, it was not possible to train a system with $S = 40$ without regularization.

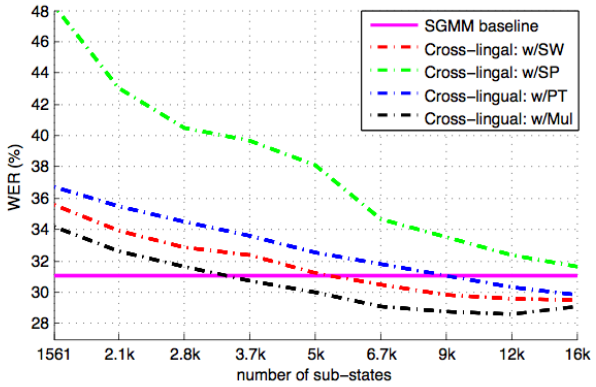

Fig. 4. 5 Hour Training Data Case: WER of cross-lingual systems. In these experiments, the dimension of state vectors is 20. The best performance is achieved by multilingual subspace system denoted as "w/Mul" and the WER is 28.6% which is considerably better than 31.1% by SGMM baseline.

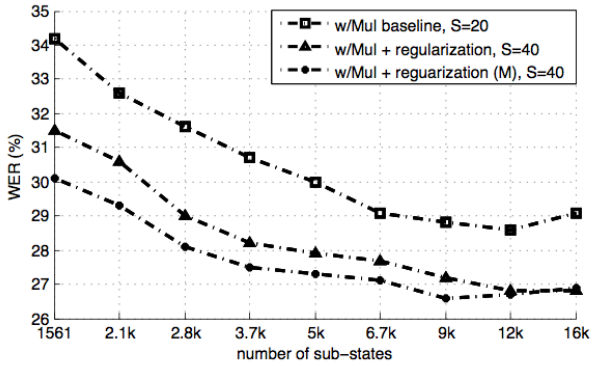

Fig. 5. 5 Hour Training Data Case: WER of cross-lingual systems with and without the original ℓ_1-norm regularization and the modified regularization (M).

V. CONCLUSIONS AND FUTURE WORK

In this paper we have investigated the use of SGMMs for cross-lingual speech recognition when the target language has limited acoustic training data. Following experiments using one and five hours of target language acoustic training data, we are able to draw four principal conclusions:

1) The SGMM-based systems are consistently better than the GMM-based systems, in terms of WER.
2) Cross-lingual SGMM systems with global subspace parameters estimated from tied multilingual systems result in lower WERs compared to systems with global

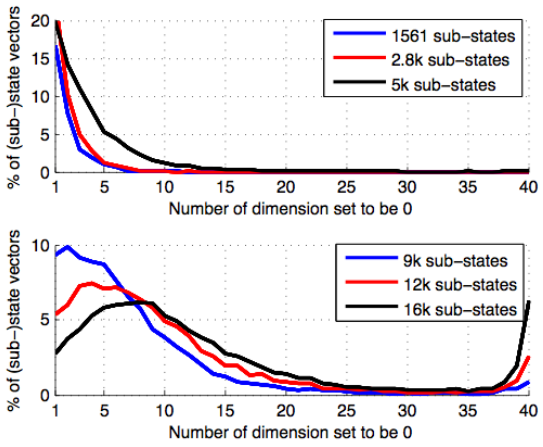

Fig. 6. Sparsity achieved by ℓ_1-norm regularization for the "w/Mul + regularization" system in Figure 5. With larger number of sub-states, hundreds of sub-state vectors are set to the zero-vector.

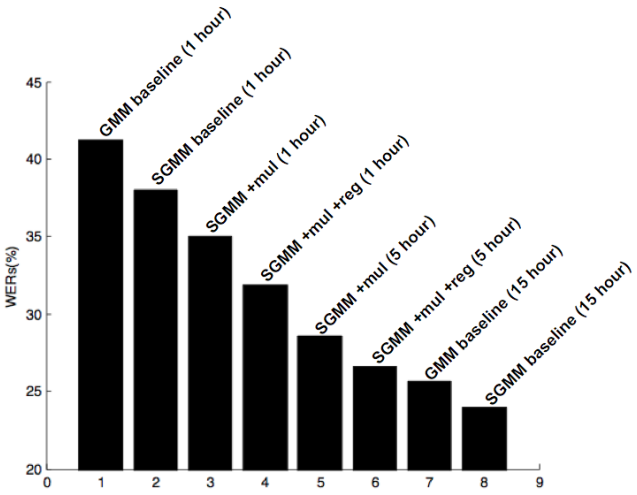

Fig. 7. A summary of results: the WER of the 5-hour SGMM system with multilingual subspaces and 40-dimensional sub-state vectors estimated with ℓ_1-norm regularization is within 1% of the GMM system and within 3% of the SGMM system trained on the full training set.

subspace parameters estimated from invididual source language systems.

3) Cross-lingual SGMM systems with a multilingual subspace result in lower WERs than monolingual SGMM systems built from scratch with the same data.
4) Regularising cross-lingual SGMM systems (using the ℓ_1-norm) enables a higher dimension sub-state vector to be used, resulting in a reduced WER.

In this work, the out-of-domain subspace parameters are fixed in all the cross-lingual systems. In future work, we would like to investigate the adaptation of those parameters using the target language acoustic training data. In particular, we are interested in the MAP adaptation of the phonetic subspace matrix $\mathbf{M}_i$ where the cross-lingual and multilingual subspace will serve as priors in the model estimation.

ACKNOWLEDGMENT

The research leading to these results was supported by the European Community's Seventh Framework Programme under grant agreement number 213850 (SCALE) and by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). We thank Partha Lal for baseline systems on GlobalPhone corpus. We also thank the reviewers of [18] whose comments led us to investigate modified regularization in this paper.

REFERENCES

[1] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech*. Citeseer, 1997, pp. 371–374.
[2] ——, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop on Broadcast News Transcription and Understanding*. Citeseer, 1998.
[3] ——, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–52, 2001.
[4] T. Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University," in *Proc. ICLSP*, 2002, pp. 345–348.
[5] N. Vu, F. Kraus, and T. Schultz, "Multilingual A-stabil: A new confidence score for multilingual unsupervised training," in *IEEE Workshop on Spoken Language Technology, SLT*, 2010.
[6] ——, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *Proc. ICASSP*. IEEE, 2011, pp. 5000–5003.
[7] P. Beyerlein, W. Byrne, J. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, and W. Wang, "Towards language independent acoustic modeling," in *Proc. ICASSP*. IEEE, 2000, pp. 1029–1032.
[8] V. Le and L. Besacier, "First steps in fast acoustic modeling for a new target language: application to Vietnamese," in *Proc. ICASSP*. IEEE, 2005, pp. 821–824.
[9] K. Sim and H. Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proc. ICASSP*. IEEE, 2008, pp. 4309–4312.
[10] K. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proc. ASRU*. IEEE, 2009, pp. 546–551.
[11] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*. IEEE, 2000, pp. 1635–1638.
[12] A. Stolcke, F. Grézl, M. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*. IEEE, 2006, pp. 321–324.
[13] O. Çetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in *Proc. ASRU*. IEEE, 2007, pp. 36–41.
[14] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource LVCSR systems," in *Proc. INTERSPEECH*, 2010, pp. 877–880.
[15] P. Lal, "Cross-lingual Automatic Speech Recognition using Tandem Features," Ph.D. dissertation, The University of Edinburgh, 2011.
[16] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.
[17] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Proc. ICASSP*. IEEE, 2010, pp. 4334–4337.
[18] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for speech recognition," *IEEE Signal Processing Letters*, vol. 18, no. 7, pp. 419–422, 2011.
[19] D. Povey, "A tutorial-style introduction to subspace Gaussian mixture models for speech recognition," MSR-TR-2009-111, Microsoft Research, Tech. Rep., 2009.