Cross-Lingual Subspace Gaussian Mixture Models for Low-Resource Speech Recognition

Liang Lu, *Student Member, IEEE*, Arnab Ghoshal, *Member, IEEE*, and Steve Renals, *Senior Member, IEEE*

***Abstract*—This paper studies cross-lingual acoustic modeling in the context of subspace Gaussian mixture models (SGMMs). SGMMs factorize the acoustic model parameters into a set that is globally shared between all the states of a hidden Markov model (HMM) and another that is specific to the HMM states. We demonstrate that the SGMM global parameters are transferable between languages, particularly when the parameters are trained multilingually. As a result, acoustic models may be trained using limited amounts of transcribed audio by borrowing the SGMM global parameters from one or more source languages, and only training the state-specific parameters on the target language audio. Model regularization using ℓ_1-norm penalty is shown to be particularly effective at avoiding overtraining and leading to lower word error rates. We investigate maximum *a posteriori* (MAP) adaptation of subspace parameters in order to reduce the mismatch between the SGMM global parameters of the source and target languages. In addition, monolingual and cross-lingual speaker adaptive training is used to reduce the model variance introduced by speakers. We have systematically evaluated these techniques by experiments on the GlobalPhone corpus.**

***Index Terms*—Acoustic modeling, subspace Gaussian mixture model, cross-lingual speech recognition, regularization, adaptation.**

I. Introduction

LARGE vocabulary continuous speech recognition systems rely on the availability of substantial resources including transcribed speech for acoustic model estimation, in-domain 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 gathering these resources. For a new language with limited resources, conventional approaches to acoustic modeling normally result in much lower accuracy. There has been extensive amount of work to improve the accuracy of speech recognizers in low-resource conditions, focusing on estimating models from limited amounts of transcribed audio in the target language [1]–[5] or when a pronunciation dictionary is not available [6]–[8]. This paper studies cross-lingual acoustic modeling with the objective of porting information from one or more source language systems which are built using larger amounts of training data, in order to build a system for a target language for which only limited amounts of transcribed audio are available. However, owing to differences such as different sets of subword units, sharing the knowledge among multiple languages is not a straightforward task. The main approaches to cross-lingual acoustic modeling, discussed below, include the use of *global phone sets*, cross-lingual *phone/acoustic* mapping, cross-lingual *tandem features* and the use of *KL-divergence HMMs*.

Schultz and colleagues [1], [2], [9], [10] 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 adapted to the target language [1], [9]. However, an important problem with this approach is that the number of phone units grows as the number of languages to be covered increases. This may lead to inconsistent parameter estimation and, consequently, degradation in accuracy [11], especially in case of context-dependent modeling. To overcome this problem, instead of using a universal phone set, a set of universal speech attributes may be used which represent similar sounds across language than phone units [12]. The fundamental speech attributes which can be viewed as a clustering of phonetic features, such as voicing, nasality and frication, can be modeled from a particular source language and shared across many different target languages. In practice, a bank of detectors using neural networks [12], for instance, may be employed to extract the universal attributes.

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 [3], [4] and data-driven [13], [14] approaches have been investigated. Given a cross-lingual mapping, either the target acoustic model is derived from the source acoustic model, or the transcription of target speech is performed using the mapped source acoustic model [14].

Tandem features, based on phone posterior probability estimates, were originally proposed to improve monolingual speech recognition [15], but they have also proven effective in the cross-lingual setting. In this approach, multi-layer perceptrons (MLPs), trained using source language acoustic data, are used to generate MLP phone posterior features for the target language [5], [16]–[20]. In addition, the training data of the

Manuscript received December 14, 2012; revised August 23, 2013; accepted August 28, 2013. Date of publication September 16, 2013; date of current version November 13, 2013. This work was supported by the EU FP7 Programme under grant agreement number 213850 (SCALE), and by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. James Glass.

The authors are with the School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, U.K. (e-mail: liang.lu@ed.ac.uk; a.ghoshal@ed.ac.uk; s.renals@ed.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2013.2281575

target language may also be used to adapt the MLPs to fit the target system better [5]. Recent advances in using MLPs with multiple hidden layers (deep neural networks, DNNs) [21] have shown great promise for DNN-based cross-lingual acoustic modeling [22].

KL-divergence HMM based acoustic modeling [23] is a recently proposed approach which has shown good performance in low-resource conditions [24], [25]. In this framework, a global phone set is first obtained by manually mapping the phones in the different languages to a common phone set (for example, IPA or X-SAMPA). A multilingual MLP phoneme classifier is trained using the data from all the source languages. For the target language system, the phoneme posterior features are extracted given the MLP. Each HMM state is parameterized by a multinomial distribution, and the model is estimated by minimizing the KL-divergence between the posterior features and HMM state multinomial coefficients. The benefits of this approach are that the multilingual information can be explored by the MLP classifier and the number of multinomial parameters is much smaller than conventional GMMs which is particularly suitable for low-resource speech recognition.

The recently proposed subspace Gaussian mixture model (SGMM) [26] enjoys a particular advantage in cross-lingual modeling [27], [28]. In an SGMM, the emission densities of a hidden Markov model (HMM) are modeled as mixtures of Gaussians whose parameters are constrained to a globally shared set of subspaces. In other words, the SGMM factorizes the acoustic model parameters into a globally-shared set that does not depend on the HMM states and a state-specific set. Since the global parameters do not directly depend on the phone units, they may be shared between languages without sharing phones. This multilingual model subspace may be used to estimate models for a new language with limited training data [27], and in this case, only state-dependent parameters need to be estimated while the model subspace can be fixed. This reduces the amount of training data required to train the recognizer and is especially suitable for speech recognition in low-resource conditions.

In this paper, we organize our previous findings on cross-lingual SGMMs for low-resource speech recognition [28], [29] and extend them with additional experiments and analysis. In particular, we investigate the speaker subspace for cross-lingual speaker adaptive training and show that:

- while the accuracy of conventional speech recognizers degrades significantly in low-resource conditions, a cross-lingual SGMM acoustic model can achieve a substantial improvement in accuracy, since a large proportion of the model parameters can be estimated using the training data of source languages;
- building systems with limited training data may lead to numerical problems in the estimation and overfitting, as we observed in cross-lingual SGMMs. We demonstrate that ℓ_1-norm regularization is an effective way to improve the robustness of model estimation and to achieve increased recognition accuracy;
- a potential mismatch may exist between the training data from the source and target languages owing to phoneme characteristic, corpus recording conditions and speaking style. This may reduce the improvements in accuracy obtained by sharing the SGMM subspace parameters in cross-lingual SGMMs. To address this issue, maximum a posteriori (MAP) adaptation is investigated to adapt the subspace parameters towards the target system;
- with limited amounts of training data, the number of speakers may be too small to estimate the speaker subspace directly for speaker adaptive training. However, the model structure naturally lends itself to cross-lingual speaker adaptive training, in which the speaker subspace is estimated from the source language and applied to the target language.

II. Subspace Gaussian Mixture Models

In conventional hidden Markov model (HMM) based speech recognizers, the emitting states are modeled by Gaussian mixture models (GMMs) with parameters estimated directly from the training data. However, in a subspace Gaussian mixture model (SGMM), the GMM parameters are inferred using a set of model subspaces that capture the correlations among the triphone states and speaker variability. In the SGMM acoustic model [26], the HMM state is modeled as:

$$p(\mathbf{y}_t \,|\, j, s) = \sum_{k=1}^{K_j} c_{jk} \sum_{i=1}^{I} w_{jki}\, \mathcal{N}\left(\mathbf{y}_t \,|\, \boldsymbol{\mu}_{jki}^{(s)}, \boldsymbol{\Sigma}_i\right) \tag{1}$$

$$\boldsymbol{\mu}_{jki}^{(s)} = \mathbf{M}_i \mathbf{v}_{jk} + \mathbf{N}_i \mathbf{v}^{(s)} \tag{2}$$

$$w_{jki} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jk}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jk}} \tag{3}$$

where $\mathbf{y}_t$ denotes the D-dimensional feature vector at time t, j is the HMM state index, k is a sub-state [26], i is the Gaussian index, and s denotes the speaker. $\mathbf{v}_{jk} \in \mathbb{R}^S$ is the phone vector (also referred to as the sub-state vector), where S denotes the phonetic subspace dimension; $\mathbf{v}^{(s)} \in \mathbb{R}^T$ is referred to as the speaker vector, and T denotes the speaker subspace dimension. The matrices $\mathbf{M}_i$, $\mathbf{N}_i$ and the vectors $\mathbf{w}_i$ span the model subspaces for Gaussian means and weights respectively, and $\boldsymbol{\Sigma}_i$ is the i-th globally shared covariance matrix. Specifically, the columns of $\mathbf{M}_i$ are a set of basis vectors spanning the phonetic subspace and $\mathbf{v}_{jk}$ models the corresponding Gaussian mean as a point in this space, while the columns of $\mathbf{N}_i$ are a set of bases spanning the speaker subspaces and $\mathbf{v}^{(s)}$ models the contribution from speaker s as a point in this space. In other words, the model factorizes the phonetic- and speaker-specific contributions to the Gaussian means.

Fig. 1 shows the structure of an SGMM acoustic model. We can divide the total set of parameters into two sets, the globally shared parameters $\boldsymbol{\Phi}_i = \{\mathbf{M}_i, \mathbf{N}_i, \mathbf{w}_i, \boldsymbol{\Sigma}_i\}$ and the state-dependent parameters $(\mathbf{v}_{jk}, c_{jk})$. The sub-state weights c_{jk} are not shown in the figure to reduce clutter. The speaker vector $\mathbf{v}^{(s)}$ is used to adapt the model to speaker s. For each Gaussian component, the parameters are derived from both the globally shared and state-dependent parameter sets. This model is quite different from the conventional GMM based acoustic model, as a large portion of the parameters are globally shared between states (Table I). The number of state-dependent parameters $(\mathbf{v}_{jk}, c_{jk})$ is relatively small if we use a low dimensional model space. This

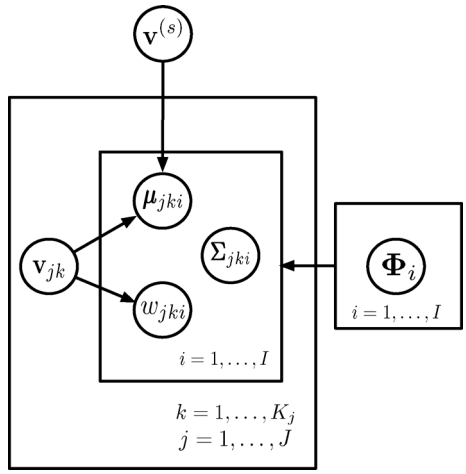


Fig. 1. Model structure of a SGMM acoustic model, with total J HMM states, and each has K_j sub-states. Each sub-state is modeled by a GMM with I components, whose parameters are derived from $\boldsymbol{\Phi}_i = \{\mathbf{M}_i, \mathbf{N}_i, \mathbf{w}_i, \boldsymbol{\Sigma}_i\}$ and $(\mathbf{v}_{jk}, \mathbf{v}^{(s)})$ using (2) and (3), and for covariance $\boldsymbol{\Sigma}_{jki} = \boldsymbol{\Sigma}_i$.

TABLE I
THE NUMBER OF PARAMETERS OF AN SGMM ACOUSTIC MODEL. Q DENOTES THE TOTAL NUMBER OF SUB-SATES. A LARGE PORTION OF THE TOTAL PARAMETERS, E.G. MORE THAN 60% FOR SYSTEMS IN [30], ARE GLOBALLY SHARED

Type	globally shared				state dependent	
	$\mathbf{M}_i$	$\mathbf{N}_i$	$\mathbf{w}_i$	$\boldsymbol{\Sigma}_i$	$\mathbf{v}_{jk}$	c_{jk}
#Parm	IDS	IDT	IS	$ID(D+1)/2$	QD	Q
Total	$I(D(D+1)/2 + DS + DT + S)$				$Q(D+1)$	

allows the model to be trained with less training data, since the number of active parameters in an SGMM acoustic model can be much smaller that its GMM based counterpart [26]. In addition, since the globally shared parameters $\boldsymbol{\Phi}_i$ do not depend on the model topology, they may be estimated by tying across multiple systems or by using out-of-domain data, which inspires its application in multilingual and cross-lingual speech recognition [27], [28], discussed in Section III.

A. Maximum Likelihood Model Estimation

Compared to conventional GMM based acoustic modeling, it is more complex to train an SGMM-based system. The parameters to be estimated for an SGMM may be split into the state-independent parameters $\boldsymbol{\Phi}_i$ and the state-dependent parameters $(\mathbf{v}_{jk}, c_{jk})$, as well as the speaker vector $\mathbf{v}^{(s)}$. Since they depend on each other, no closed form solution is available for the global optimum. However, using the maximum likelihood (ML) criterion, they can be updated iteratively by employing the expectation-maximization (EM) algorithm [26]. For instance, the auxiliary function used in EM for sub-state vector $\mathbf{v}_{jk}$ is

$$\mathcal{Q}(\mathbf{v}_{jk}) = -0.5\mathbf{v}_{jk}^T\mathbf{H}_{jk}\mathbf{v}_{jk} + \mathbf{v}_{jk}^T\mathbf{g}_{jk} + \text{const}, \tag{4}$$

where $\mathbf{H}_{jk}$ and $\mathbf{g}_{jk}$ are an $S \times S$ matrix and an S-dimensional vector capturing the sufficient statistics for the estimation of $\mathbf{v}_{jk}$, and const denotes the independent constant value. If the matrix $\mathbf{H}_{jk}$ is invertible, the update formula is readily available as

$$\mathbf{v}_{jk} = \mathbf{H}_{jk}^{-1}\mathbf{g}_{jk}. \tag{5}$$

A more numerically stable algorithm for this estimation is given in [26] in case $\mathbf{H}_{jk}$ is poorly conditioned. Similarly, the auxiliary function to update the phonetic subspace $\mathbf{M}_i$ is

$$\mathcal{Q}(\mathbf{M}_i) = \text{Tr}\left(\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{Y}_i\right) - 0.5\text{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\mathbf{Q}_i\mathbf{M}_i^T\right) + \text{const} \tag{6}$$

where $\mathbf{Y}_i$ and $\mathbf{Q}_i$ are sufficient statistics defined as

$$\begin{cases} \mathbf{Y}_i = \sum_{jkt}\tilde{\gamma}_{jki}(t)\mathbf{y}_t\mathbf{v}_{jk}^T \\ \mathbf{Q}_i = \sum_{jk}\gamma_{jki}\mathbf{v}_{jk}\mathbf{v}_{jk}^T \end{cases}, \tag{7}$$

where $\tilde{\gamma}_{jki}(t)$ denotes the Gaussian component posterior for acoustic frame $\mathbf{y}_t$, and $\gamma_{jki} = \sum_t \tilde{\gamma}_{jki}(t)$. If $\mathbf{Q}_i$ is invertible, we can obtain

$$\mathbf{M}_i = \mathbf{Y}_i\mathbf{Q}_i^{-1}. \tag{8}$$

Again, a more numerical stable algorithm is given in [26], and also refer it for the estimation of $\mathbf{N}_i$, $\mathbf{w}_i$, $\boldsymbol{\Sigma}_i$, c_{jk} and $\mathbf{v}^{(s)}$.

B. Regularized Model Estimation

Standard maximum-likelihood (ML) estimation of SGMMs can result in overfitting when the amount of training data is small [28]. This problem is most acute for the state-dependent vectors $\mathbf{v}_{jk}$—unlike the globally shared parameters $\boldsymbol{\Phi}_i$, they are only trained on those speech frames which align with the corresponding sub-state. To overcome this problem, we proposed a regularized ML estimate for the state vectors [30] in which penalties based on the ℓ_1-norm and ℓ_2-norm of the state vectors, as well as their linear combination (the elastic net [31]), were investigated. Regularization using the ℓ_1-norm penalty was found to be best suited in cross-lingual settings where the amount of target training data is very limited [28]. With an ℓ_1-norm penalty, the auxiliary function for sub-state vector estimation becomes:

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} \mathcal{Q}(\mathbf{v}) - \lambda\|\mathbf{v}\|_{\ell_1}, \quad \lambda > 0, \tag{9}$$

where λ is the global penalty parameter (we have dropped the subscripts on $\mathbf{v}$ for brevity).

Intuitively, the ℓ_1-norm penalty performs an element-wise shrinkage of $\mathbf{v}$ towards zero in the absence of an opposing data-driven force [31], which enables more robust estimation. The ℓ_1-norm penalty also has the effect of driving some elements to be zero, thus leading to a form of variable selection which has been used in sparse representation of speech features [32], [33], as well as compressed sensing [34]. For the case of cross-lingual SGMMs, the ℓ_1-norm penalty can be used to select the relevant basis in $\mathbf{M}_i$ according to the amount of available data to estimate $\mathbf{v}_{jk}$ while avoiding overtraining. However, the solution of the auxiliary function is not readily available for the ℓ_1-norm penalty, since the derivative of the auxiliary function is not continuous. We have previously applied the gradient projection based optimization approach [35] to obtain the solution [30]. The idea of regularization can also be applied to other types of parameters in SGMMs. In fact, while doing MAP adaptation of $\mathbf{M}_i$ using a Gaussian prior, as described in Section IV, if we set the prior mean to be $\mathbf{0}$ and the row and column covariances to the identity matrix $\mathbf{I}$, then the MAP adaptation is equivalent to ℓ_2-norm regularization of $\mathbf{M}_i$.

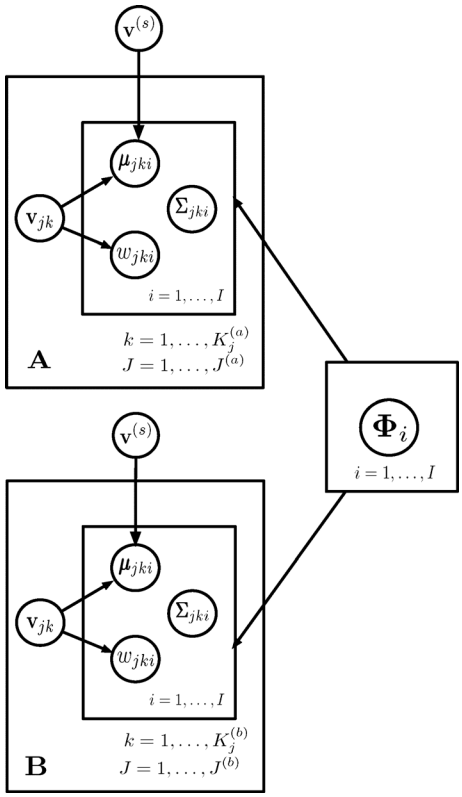


Fig. 2. An example of multilingual estimation of the globally shared parameters $\boldsymbol{\Phi}_i$ where we tie them across two source language system $\mathbf{A}$ and $\mathbf{B}$.

III. Multilingual Model Estimation

One of the main barriers preventing acoustic knowledge being shared across different languages is the mismatch of phone units between languages. Conventional methods tackle this problem by using global phone units or through the use of tandem features. However in an SGMM acoustic model the globally shared parameters $\boldsymbol{\Phi}_i$ do not depend on the HMM topology, and hence are independent of the definition of the phone units. Therefore, when using SGMMs for cross-lingual acoustic modeling, the phoneme unit mismatch problem is to some degree bypassed, since we can estimate the globally shared parameters using multilingual training data by tying the globally shared parameters across the available source language systems.

Fig. 2 demonstrates an example of the multilingual SGMM system in which source language systems $\mathbf{A}$ and $\mathbf{B}$ may have different phone units and HMM topologies, provided that the acoustic feature parameterization and the dimensionality of model subspace are the same. By training a multilingual SGMM system in this way the accuracy for each of the source languages may be improved [27], and the multilingual globally shared parameters can be ported to a new target language system with limited training data [27], [28]. In an SGMM the globally shared parameters typically account for a large proportion of the total number of parameters (Table I). The reuse of the globally shared parameters across languages thus significantly reduces the required amount of acoustic training data—only the state dependent parameters $(\mathbf{v}_{jk}, c_{jk})$ need be estimated from target language data.

Using multiple source language systems to estimate the globally shared parameters $\boldsymbol{\Phi}_i$ involves some modifications in the SGMM training procedure. However, these modifications are minor and relatively simple, since given $\boldsymbol{\Phi}_i$ each source language system is independent—therefore the statistics for each source language system can be accumulated in the standard fashion using either the Viterbi alignment or the Baum-Welch algorithm. In each iteration, the corresponding statistics are then summed across languages to update the globally shared parameters. The state dependent parameters $(\mathbf{v}_{jk}, c_{jk})$ are updated in the standard fashion, for each language separately. Consider $\mathbf{M}_i$: for the system of Fig. 2, after obtaining the statistics for each source language system $(\mathbf{Y}_i^{(a)}, \mathbf{Y}_i^{(b)})$ and $(\mathbf{Q}_i^{(a)}, \mathbf{Q}_i^{(b)})$, the final statistics are obtained simply by

$$\mathbf{Y}_i = \mathbf{Y}_i^{(a)} + \mathbf{Y}_i^{(b)}, \quad \mathbf{Q}_i = \mathbf{Q}_i^{(a)} + \mathbf{Q}_i^{(b)}. \tag{10}$$

Then $\mathbf{M}_i$ can be updated as usual (8). A similar approach can be used to update $\mathbf{N}_i, \mathbf{w}_i$ and $\boldsymbol{\Sigma}_i$ using the multilingual data. To build a cross-lingual SGMM system, these parameters are ported into target language system directly, and only the state dependent parameters $\mathbf{v}_{jk}$ and c_{jk} are estimated using the (limited) in-domain training data. Our previous experimental results [28] indicate that this approach can significantly reduce the word error rate (WER) in low-resource conditions.

IV. MAP Adaptation of Model Subspace

In a cross-lingual SGMM system for a target language with limited acoustic training data, the globally shared parameters are trained using source language data. This may introduce a mismatch with the target language system because of differences in phonetic characteristics, corpus recording conditions, and speaking styles. Since the amount of training data may not be sufficient to allow the global parameters to be updated using ML, the mismatch may be alleviated by an adaptation approach based on the maximum a posteriori (MAP) criterion. In particular, we have studied the adaptation of $\mathbf{M}_i$ using MAP [29].

In ML estimation of the phonetic subspace [26], the auxiliary function for $\mathbf{M}_i$ is given by (6). If a prior term is introduced, then the auxiliary function becomes:

$$\tilde{\mathcal{Q}}(\mathbf{M}_i) = \mathcal{Q}(\mathbf{M}_i) + \tau \log P(\mathbf{M}_i), \tag{11}$$

where $P(\mathbf{M}_i)$ denotes the prior distribution of matrix $\mathbf{M}_i$, and τ is the smoothing parameter which balances the relative contributions of the likelihood and prior. Although any valid form of $P(\mathbf{M}_i)$ may be used, in practical MAP applications a conjugate prior distribution is often preferred for reasons of simplicity. We set $P(\mathbf{M}_i)$ to be a Gaussian distribution which is conjugate to the auxiliary function $\mathcal{Q}(\mathbf{M}_i)$.

A. Matrix Variate Gaussian Prior

The Gaussian distribution of random matrices is well understood [36]. A typical example of its application in speech recognition is maximum a posteriori linear regression (MAPLR) [37] for speaker adaptation, in which a matrix variate prior is used

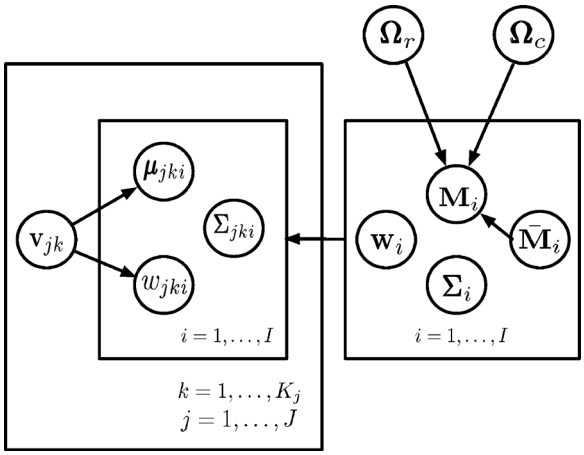


Fig. 3. MAP adaptation of $\mathbf{M}_i$ in SGMM acoustic model. $(\bar{\mathbf{M}}_i, \boldsymbol{\Omega}_r, \boldsymbol{\Omega}_c)$ denote the hyper-parameters of the Gaussian prior $P(\mathbf{M}_i)$, in which the mean $\bar{\mathbf{M}}_i$ is indexed by I while the covariances $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_c$ are global.

for the linear regression transformation matrix. The Gaussian distribution of a $D \times S$ matrix $\mathbf{M}$ is defined as:

$$\log P(\mathbf{M}) = -\frac{1}{2}\left(DS\log(2\pi) + D\log|\boldsymbol{\Omega}_r| + S\log|\boldsymbol{\Omega}_c| + \mathrm{Tr}\left(\boldsymbol{\Omega}_r^{-1}(\mathbf{M}-\bar{\mathbf{M}})\boldsymbol{\Omega}_c^{-1}(\mathbf{M}-\bar{\mathbf{M}})^T\right)\right), \quad (12)$$

where $\bar{\mathbf{M}}$ is a matrix containing the expectation of each element of $\mathbf{M}$, and $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_c$ are $D \times D$ and $S \times S$ positive definite matrices representing the covariance between the rows and columns of $\mathbf{M}$, respectively. $|\cdot|$ and $\mathrm{Tr}(\cdot)$ denote the determinant and trace of a square matrix. This matrix density Gaussian distribution may be written as:

$$\mathrm{Vec}(\mathbf{M}) \sim \mathcal{N}(\mathrm{Vec}(\bar{\mathbf{M}}), \boldsymbol{\Omega}_r \otimes \boldsymbol{\Omega}_c), \quad (13)$$

where $\mathrm{Vec}(\cdot)$ is the vectorization operation which maps a $D \times S$ matrix into a $DS \times 1$ vector, and $\otimes$ denotes the Kronecker product of two matrices. In this formulation, only $\boldsymbol{\Omega}_r \otimes \boldsymbol{\Omega}_c$ is uniquely defined, and not the individual covariances $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_c$, since for any $\alpha > 0$, $(\alpha\boldsymbol{\Omega}_r, \frac{1}{\alpha}\boldsymbol{\Omega}_c)$ would lead to the same distribution. This is not of concern in the current application to MAP adaptation. Fig. 3 illustrates the concept of using the Gaussian prior to adapt the model subspace $\mathbf{M}_i$. In this case, the auxiliary function for MAP adaptation is:

$$\tilde{\mathcal{Q}}(\mathbf{M}_i) \propto \mathrm{Tr}\left(\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{M}_i^T\boldsymbol{\Omega}_r^{-1}\bar{\mathbf{M}}_i\boldsymbol{\Omega}_c^{-1}\right) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\mathbf{Q}_i\mathbf{M}_i^T + \tau\boldsymbol{\Omega}_r^{-1}\mathbf{M}_i\boldsymbol{\Omega}_c^{-1}\mathbf{M}_i^T\right). \quad (14)$$

B. Prior Distribution Estimation

To apply MAP, the prior distribution $P(\mathbf{M}_i)$ for each $\mathbf{M}_i$, should be estimated first. This requires the estimation of the mean matrices $\bar{\mathbf{M}}_i$, and the row and column covariances $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_c$. Given a set of samples generated by $P(\mathbf{M}_i)$, the ML estimation of the mean, and the row and column covariances, is described by Dutilleul [38]. This is used with some heuristic rules for cross-lingual SGMMs [29], in which, the MAP formulation is based on the assumption that the multilingual estimate of the global subspace parameters serves a good starting point, which has been empirically verified earlier [28]. To apply MAP adaptation, we set these multilingual parameters to be the mean of the prior $P(\mathbf{M}_i)$ and update both the state-specific $\mathbf{v}_{jm}$ and the global $\mathbf{M}_i$. With a sufficiently large value of τ in (11), we can shrink the system back to the cross-lingual baseline, whereas $\tau = 0$ corresponds to the ML update.

The covariance matrices for each $P(\mathbf{M}_i)$ are set to be global in order to reduce the number of hyper-parameters in the prior distributions. In [29], we compared different forms of the two covariance matrices $(\boldsymbol{\Omega}_r, \boldsymbol{\Omega}_c)$ and the experimental results indicated that using the identity matrix $\mathbf{I}$ for $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_c$ worked well. Using this configuration, MAP adaptation of $\mathbf{M}_i$ is equivalent to applying ℓ_2-norm regularization by setting the multilingual estimate as the model origin. In this case, the auxiliary function (14) will become

$$\tilde{\mathcal{Q}}(\mathbf{M}_i) \propto \mathrm{Tr}\left(\mathbf{M}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{Y}_i + \tau\mathbf{M}_i^T\bar{\mathbf{M}}_i\right) - \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{M}_i\mathbf{Q}_i\mathbf{M}_i^T + \tau\mathbf{M}_i\mathbf{M}_i^T\right). \quad (15)$$

The solution can be obtained in [29], [39]. In this work, this configuration is adopted in the MAP adaptation experiments.

TABLE II
NUMBERS OF PHONES AND SPEAKERS, AND THE AMOUNT OF TRAINING DATA (HOURS) FOR EACH OF THE 4 LANGUAGES USED IN THIS PAPER

Language	#Phones	#Speakers	Train/hours
German (GE)	44	77	14.8
Spanish (SP)	43	97	17.2
Portuguese (PT)	48	101	22.6
Swedish (SW)	52	98	17.4

V. EXPERIMENTS AND RESULTS

We performed cross-lingual speech recognition experiments using SGMMs on the GlobalPhone corpus [10]. GlobalPhone contains around 20 languages including Arabic, Chinese and a number of European languages, with read newspaper speech from about 100 native speakers per language. Recordings were made under relatively quiet conditions using close-talking microphones. Acoustic conditions may vary within a language and between languages, hence acoustic mismatches may affect 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 II describes the data for each language used in the experiments in terms of the number of phonemes and speakers, and the amount of available audio.

To investigate the effect of limited acoustic training data, we constructed two randomly selected training subsets of the target language German data each containing 1 hour (8 speakers) and 5 hours (40 speakers) of data, with 7–8 minutes of recorded speech for each of the selected speakers. We used these data subsets, in addition to the full 14.8 hours (referred to as 15 hours) of German training data, as the three target language training sets in the following experiments.

A. Baseline Monolingual Systems

We constructed baseline systems using the three training sets (1h/5h/15h) in a monolingual fashion, using conventional GMM and SGMM acoustic modeling. The systems were built using

TABLE III
WERs OF BASELINE GMM AND SGMM SYSTEMS USING 1 HOUR, 5 HOUR AND 15 HOUR TRAINING DATA

System	1 hour		5 hour		15 hour	
	dev	eval	dev	eval	dev	eval
GMM	23.2	34.1	18.5	28.0	15.4	24.8
SGMM	20.4	31.4	14.9	24.9	13.0	22.1
#states	831		1800		2537	

the Kaldi speech recognition toolkit [40]. We used 39-dimensional MFCC feature vectors for the experiments. Each feature vector consisted of 13-dimensional static features with the zeroth cepstral coefficient and their delta and delta-delta components. Cepstral mean and variance normalization (CMN/CVN) was then applied on a per speaker basis. The GMM and SGMM systems shared the same decision tree to determine the tied state clustering used for context-dependent phone modeling; therefore, the differences in recognition accuracies of the GMM and SGMM systems are purely due to the different parameterization of the GMMs. In the SGMM systems, we set the number of UBM Gaussians $I = 400$, and phonetic subspace dimension $S = 40$ for 15 hour training data case, whereas we use $S = 20$ when the training data is limited to 1 hour and 5 hours. Since the estimation of UBM model does not require the labels, we estimated it on the whole training dataset and use it for all German SGMM systems. Table III shows the word error rates (WERs) of baseline systems. As expected, the WERs for both the GMM and SGMM systems increase significantly as the amount of training data is reduced. The monolingual SGMM system has a significantly lower WER than the monolingual GMM system for each of the three training conditions.

There is a large difference between the WERs achieved on the development (dev) and evaluation (eval) sets in Table III. This is due to the language model that we used. In [28] we used a trigram language model obtained with an earlier release of the GlobalPhone corpus, and achieved accuracies on the development dataset that were comparable to these on the evaluation dataset. Here, we interpolated that previously used language model with one estimated on the training corpus, and we obtained a significant reduction in WER on the development dataset (e.g., 24.0% in [28] to 13.0% for SGMM system with 15 hour training data). But the improvements disappear on the evaluation dataset which indicates that the text in the training set matches the text of the development set better than that of the evaluation dataset. In the cross-lingual acoustic modeling presented in this paper we observe similar trends on both the development and evaluation sets (as will be shown in Section V-G), so the linguistic variation between training, development, and evaluation sets is not a confounding factor.

B. Cross-Lingual System Configuration

Each cross-lingual SGMM used the same context dependent tied state clustering as the corresponding monolingual SGMM trained on the same data set. Sharing global parameters between source languages, together with the constraints imposed by the structure of the SGMM, leads to better parameter estimates with limited amounts of training data. This also allows bigger models to be trained, either using more context-dependent tied states [27], or using a model with the same state clustering, but with more substates per state. We do the latter in this paper. In both cases, the combination of improved parameter estimation and bigger models, is predicted to lead to lower WER.

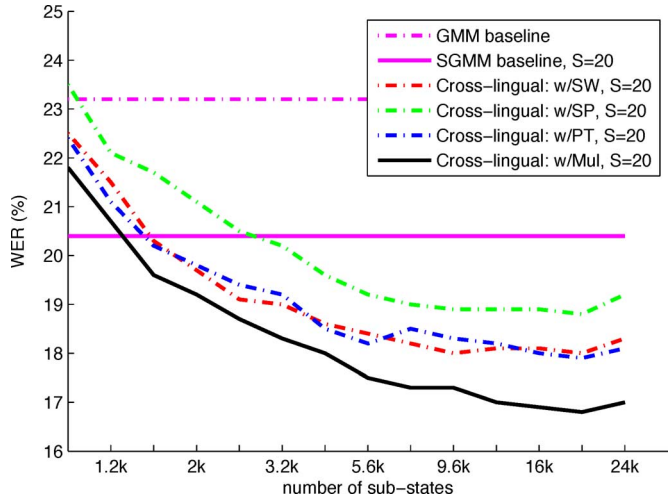


Fig. 4. WER of baseline cross-lingual systems, 1h training data, tested on the development dataset. Only the lowest WER of SGMM baseline system by tuning the number of sub-states is given for clarity.

The UBM was the same as the one that was used to train the globally shared parameters $\mathbf{\Phi}_i$ on the source language(s). This is important, since the globally shared parameters correspond to the segmentation of the acoustic space as determined by the UBM [26]. First, we train $\mathbf{\Phi}_i$ for the source language systems in either a monolingual or a multilingual fashion. We then ported the shared parameters to the corresponding cross-lingual SGMM system. In the baseline SGMM systems, all the parameters in (1)–(3) were updated: the sub-state vectors $\mathbf{v}_{jm}$ and the globally shared parameters $\mathbf{\Phi}_i$. In a cross-lingual system, however, only the sub-state vectors $\mathbf{v}_{jm}$ were re-estimated, with the globally shared parameters fixed unless stated otherwise.

C. Cross-Lingual Experiments: Baseline

The baseline results of the cross-lingual systems are shown for 1h, 5h, and 15h training data (Figs. 4–6). We contrast the shared parameters $\mathbf{\Phi}_i$ obtained from each of the source language systems, as well as the tied multilingual system. In these initial experiments, we do not use the speaker subspace $\mathbf{N}_i$. The dimension of sub-state vectors is set to be $S = 20$. With 1 hour training data, we achieved a relative WER reduction of up to 17% by reusing the globally shared parameters from source language systems trained in either a monolingual or multilingual fashion, demonstrating that out-of-domain knowledge can be used to improve significantly the accuracy of a target language system. In addition, we also observe that the system with multilingually trained subspace parameters "w/Mul" in Fig. 4 results in considerably lower WERs compared with the other cross-lingual systems derived from a single source language. This may be because that there is much larger amount of training data in the multilingual system, and furthermore, the linguistic differences and corpus mismatch may be averaged out by the multilingual estimation which alleviates the mismatch between the multilingual parameters and target language system.

We observed a similar trend in the 5 hour training data case (Fig. 5), although in this case the WER reduction is smaller (up to 10% relative) which is expected as the amount of training

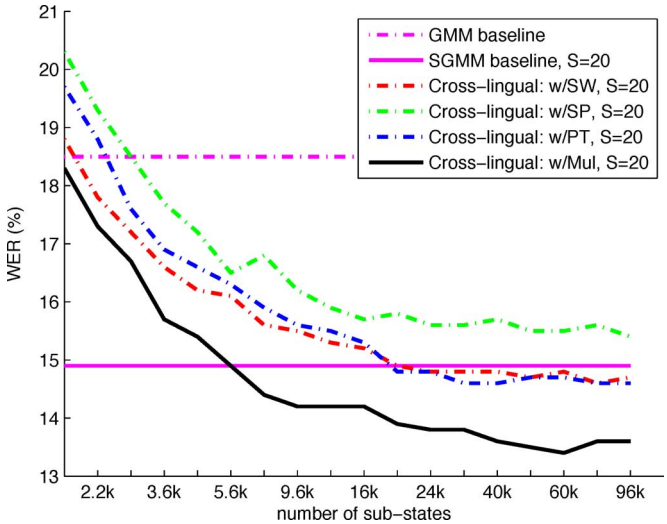


Fig. 5. WER of baseline cross-lingual systems, 5h training data, tested on the development dataset. Only the lowest WER of SGMM baseline system by tuning the number of sub-states is given for clarity.

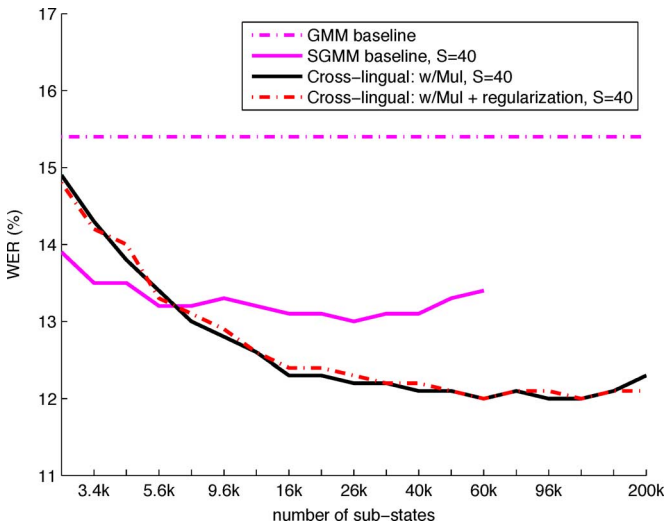


Fig. 6. WER of baseline cross-lingual systems, 15h training data, tested on the development dataset.

data increases. In order to evaluate if the cross-lingual frameworks can achieve improvement when the target training data is more abundant, we carried out the experiments using the entire 15 hour training data. Since we can draw the conclusion from the previous experiments that the multilingual $\mathbf{\Phi}_i$ perform better than their monolingual counterparts, we only use the multilingual parameters for the cross-lingual setups. Results are shown in Fig. 6 where the dimensions of sub-state vectors were set to be $S = 40$. In this case, the cross-lingual SGMM system still reduces the WER by 8% relative (1% absolute).

D. Cross-Lingual Experiments: With Regularization

With limited amounts of training data, it is often necessary to limit the dimensionality of the state vectors $\mathbf{v}_{jk}$, since increasing the phonetic subspace dimension S increases the number of both global and state-specific parameters. When the global parameters $\mathbf{\Phi}_i$ are trained on separate data, state vectors of larger dimensionality may be used. Comparing Figs. 4 and 7, we see that for the cross-lingual system trained on 1 hour of

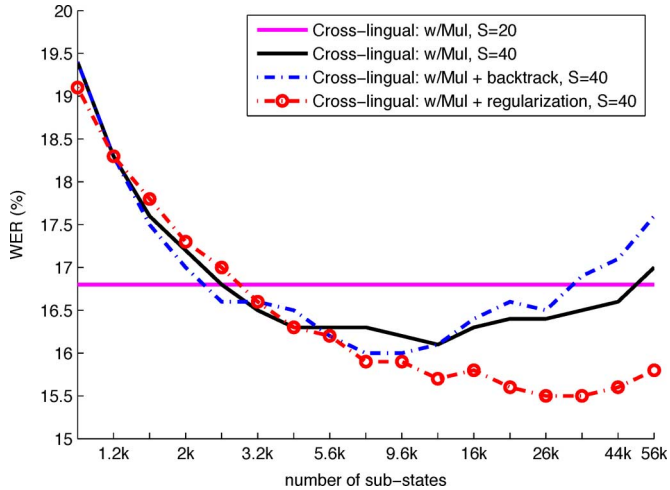


Fig. 7. WER of regularized cross-lingual systems, 1h training data, tested on the development dataset. "Cross-lingual: w/Mul, S $= 20$" corresponds to the best results of the same system by tuning the number of sub-states that is shown in Fig. 4.

speech using a phonetic subspace dimension of $S = 40$ lowers the WER compared to a subspace of dimension $S = 20$[1].

Fig. 7 also compares the standard ML update with a more conservative one that "backtracks" to the previous parameter values if the auxiliary function decreases due to the update. Models trained using both these criteria are found to have larger WER when the number of substates is increased, showing that the models tend to overtrain when using very small amounts of training data. However, when the state vectors are estimated with the ℓ_1-norm regularization, the updates are more stable and allow models with a larger number of substates to be trained leading to lower WER overall. In fact, the WER of 15.5% achieved by the cross-lingual SGMM trained on 1 hour of speech using ℓ_1-norm regularization is comparable to the GMM baseline with the entire 15 hour training data.

Fig. 8 shows the results with 5 hour training data. Not surprisingly, the difference between the regularized model and the one without regularization is smaller than that seen when training on 1 hour of data. However, when the number of sub-states is very large, regularization still helps to avoid model overfitting and results in a small gain in terms of accuracy. Again, the more conservative update with backtracking did not work better than the regularized update. After increasing the amount of training data to be 15 hours, we did not obtain improvement by applying the ℓ_1-norm regularization as shown in Fig. 6. This agrees with our previous experience of using ℓ_1-norm regularization for SGMMs [30] on a different task.

E. Cross-Lingual Experiments: With MAP Adaptation

As discussed above, if $\mathbf{\Phi}_i$ is estimated from out-of-domain data, then there may be a mismatch between the target language system and these parameters. One approach to address this mismatch is via MAP adaptation of $\mathbf{\Phi}_i$. We applied MAP adaptation of $\mathbf{M}_i$ to the systems "w/Mul, S $= 40$" and "w/Mul + regularization, S $= 40$" to the 1h and 5h training data conditions

[1]In [28], we used a preliminary version of Kaldi toolkit that was used in [26] and faced numerical instability when building the baseline system without regularization. We did not have that experience using a more recent version of Kaldi (revision 710).

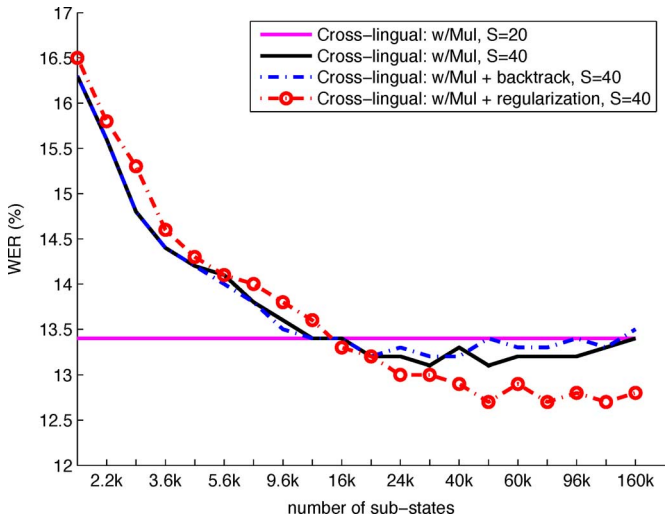


Fig. 8. WER of regularized cross-lingual systems, 5h training data, tested on the development dataset. "Cross-lingual: w/Mul, S = 20" corresponds to the best results of the same system by tuning the number of sub-states that is shown in Fig. 5.

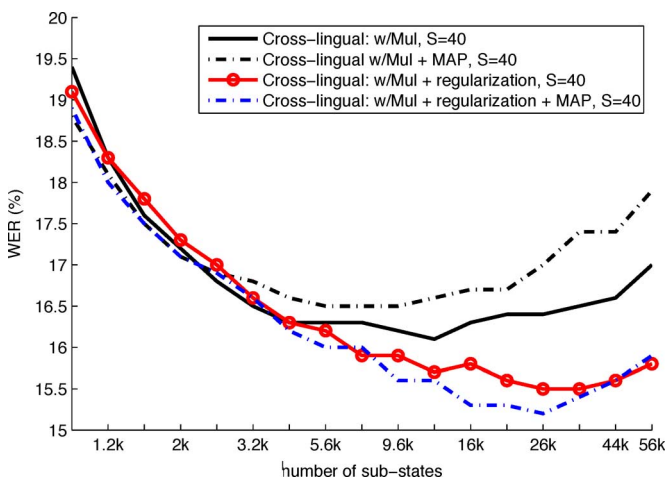


Fig. 9. WER of MAP-adapted cross-lingual systems, 1h training data, tested on the development dataset.

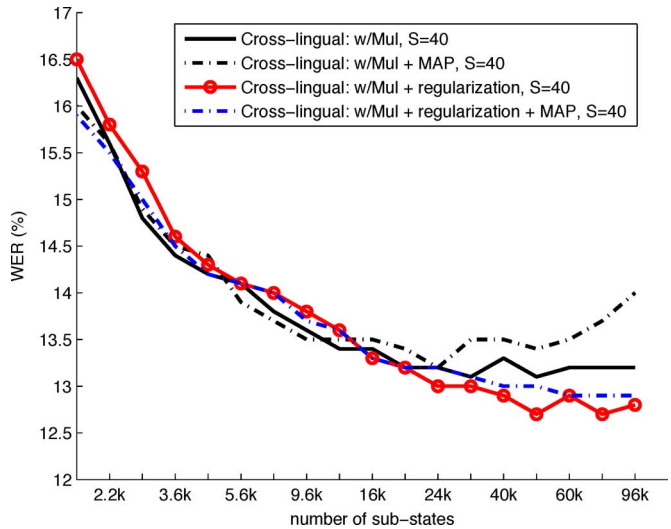


Fig. 10. WER of MAP-adapted cross-lingual systems, 5h training data, tested on the development dataset.

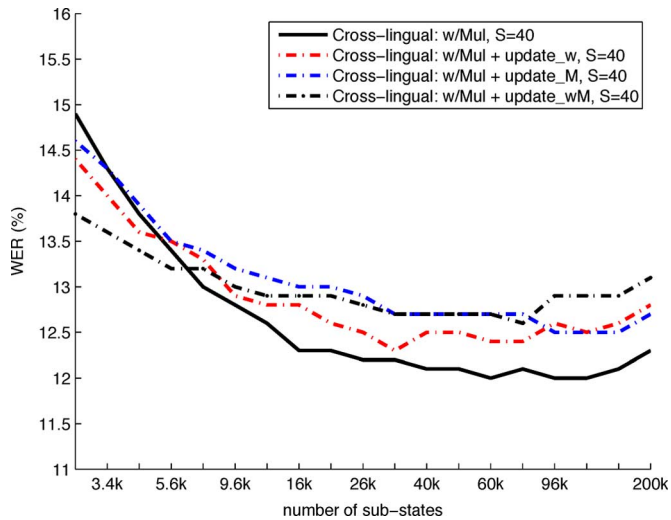


Fig. 11. WER of cross-lingual systems with global parameter update, 15h training data, tested on the development dataset.

(Figs. 9 and 10). As stated in Section IV, the two covariance matrices $\mathbf{\Omega}_r$ and $\mathbf{\Omega}_c$ are set to be the identity matrix $\mathbf{I}$. For the 1h training data case, we set smoothing parameter used in equation (11) $\tau = 500$. By using MAP adaptation, we obtained a small reduction in WER (2% relative) compared to the regularized system. The improvement is not comparable to our previous results [29] since the baseline is much stronger here. When we applied MAP adaptation to the baseline without regularization, we did not observe a reduction in WER when the number of sub-states was large. This may be because the sub-state vectors $\mathbf{v}_{jk}$ are not well estimated due to overfitting and hence we not have sufficient and accurate statistics for equation (7) to adapt $\mathbf{M}_i$.

In the 5h training data case (Fig. 10), we did not observe any reduction in WER using MAP adaptation for both systems with and without regularization, even though the amount of adaptation data was increased. When applying MAP adaptation, the likelihood on the training data increased, but the higher WER suggests that it overfits to the training data. We increased the smoothing term τ but this resulted in moving the adapted system closer to the baseline with no gain being observed. This may further demonstrate that the multilingual parameters are more robust and match the target training data well. We also did not achieve gains by using MAP adaptation of $\mathbf{M}_i$ in the 15h training data case.

For the 15h training data case, we investigated the update of the globally shared parameters $\mathbf{\Phi}_i$. We updated $\mathbf{w}_i$ and $\mathbf{M}_i$ to maximize the likelihood for the target language system. While this resulted in lower WER for models with fewer sub-states, the WER increased for larger models (Fig. 11). This is not unexpected since the multilingual estimation of $\mathbf{w}_i$ and $\mathbf{M}_i$ would be expected to be more accurate and robust than the monolingual estimate. Although updating $\mathbf{M}_i$ and $\mathbf{w}_i$ increases WERs compared with keeping them fixed at the multilingually estimated values, the results are similar to (and in some cases slightly better than) the monolingual system (Fig. 6). This indicates that a better initialization of the iterative ML updates of the subspace parameters (i.e., the multilingually trained parameters) finally does not make a substantial difference. We also carried out the experiments where $\mathbf{\Sigma}_i$ were updated, and similar results were obtained to the ML updates of $\mathbf{M}_i$ and $\mathbf{w}_i$.

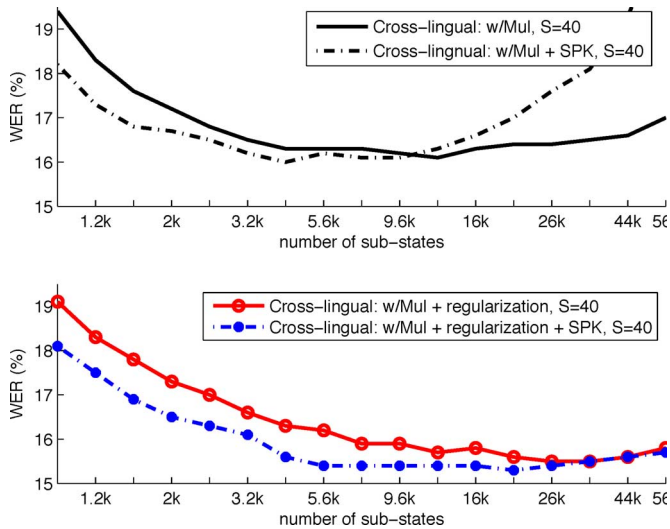


Fig. 12. WER of baseline (above) and regularized (below) cross-lingual systems using speaker subspace, 1h training data, tested on the development dataset.

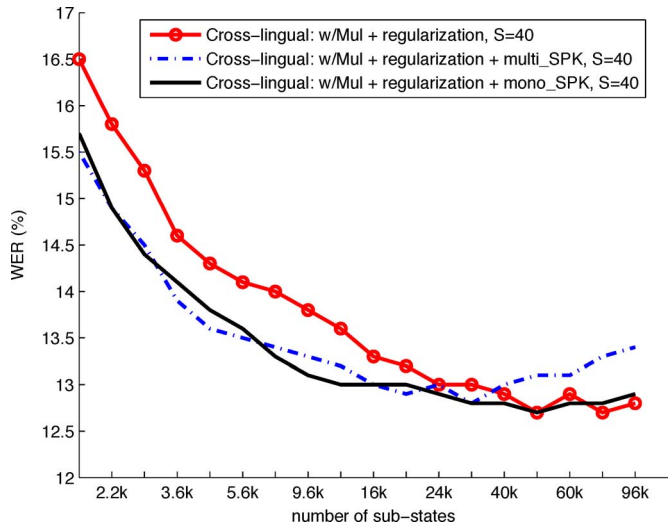


Fig. 13. WER of regularized cross-lingual systems using speaker subspace, 5h training data, tested on the development dataset.

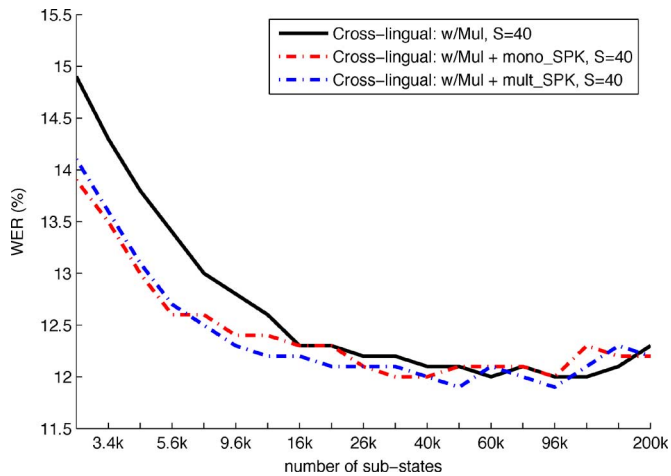


Fig. 14. WER of cross-lingual systems using speaker subspace, 15h training data, tested on the development dataset.

F. Cross-Lingual Experiments: With Speaker Subspace

Our final set of experiments concerned speaker adaptive training using the speaker subspace for cross-lingual SGMM systems for the 1h, 5h, and 15h training data cases (Figs. 12–14). In the 1h training data case, there are only 8 speakers in the training set, which is not sufficient to train the speaker subspace $\mathbf{N}_i$ on a per speaker basis for our baseline SGMM system. We trained $\mathbf{N}_i$ on a per utterance basis for the baseline but did not observe an improvement. However, we can estimate $\mathbf{N}_i$ in multilingual fashion by tying it across the source language system similar to the other globally shared parameters. We then rebuilt the target system "$\mathrm{w/Mul} + \mathrm{regularization}, \mathrm{S} = 40$" using the resultant speaker subspace. Results are given in Fig. 12. Here the dimension of speaker vector was set to be $T = 39$. We can see that for the regularized system, using the multilingual $\mathbf{N}_i$ results in significant gains when the number of sub-states is relatively small. The gains, however, vanish as we further increased the number of sub-states. The system without regularization is more prone to overtraining when using speaker subspace adaptive training.

In the 5h training data case, there are 40 speakers in the training set, enough to estimate $\mathbf{N}_i$ from the in-domain data. This system is referred as "$\mathrm{w/Mul} + \mathrm{regularization} + \mathrm{mono\_SPK}, \mathrm{S} = 40$" in Fig. 13. For the system using the multilingual speaker subspace $\mathbf{N}_i$, we refer it as "$\mathrm{w/Mul} + \mathrm{regularization} + \mathrm{multi\_SPK}, \mathrm{S} = 40$." In both systems, $T = 39$. We can see that both systems achieve large reductions in WER when the number of sub-states is small—again, the gains vanish when using a large number of sub-states. In addition, the multilingual speaker subspace $\mathbf{N}_i$ achieves a similar WER to the monolingual one. This indicates that the speaker information from the out-of-domain data can fit the target system well.

We did not observe notable WER differences between using either a monolingual or a multilingual speaker subspace in the 15h training data case (Fig. 14), as for the 5h training data case. Just as with 1 hour and 5 hours of training data, using the speaker subspace lowers the WER for smaller model sizes, but the difference between the adaptively trained and unadapted models vanishes when using a very large number of substates. Although the speaker adaptive training does not provide an overall reduction in WER, it provides a practical advantage: it is computationally cheaper to use a smaller model with speaker subspace than a larger model without it. In the future, we plan to investigate using feature space (constrained) MLLR for cross-lingual speaker adaptive training as a comparison to the results using the speaker subspace.

G. Cross-Lingual Experiments: Summary

Table IV summarizes the results on the development and evaluation datasets with 1h training data. We observed a similar trend of results on both datasets. The lowest WER on the evaluation set (26.7%) was achieved by using multilingual parameter estimation with regularization, followed by speaker subspace adaptive training. This is significantly better than the GMM and SGMM baseline using the same training data (34.1% and 31.4%) and it is only 2% worse than the GMM baseline using the entire 15h training dataset (24.8%). Hence, by leveraging

TABLE IV
RESULTS OF CROSS-LINGUAL SGMM SYSTEMS WITH 1 HOUR TRAINING DATA ON THE DEVELOPMENT (DEV) AND EVALUATION DATASET (EVAL)

System	Dev	Eval
GMM baseline	23.2	34.1
SGMM baseline	20.4	31.4
Cross-lingual: w/SP, $S = 20$	18.8	32.4
Cross-lingual: w/PO, $S = 20$	17.9	30.9
Cross-lingual: w/SW, $S = 20$	18.0	31.0
Cross-lingual: w/Mul, $S = 20$	16.8	29.3
Cross-lingual: w/Mul + ℓ_1, $S = 40$	15.5	26.9
+speaker subspace	15.3	26.7

TABLE V
RESULTS OF CROSS-LINGUAL SGMM SYSTEMS WITH 5 HOUR TRAINING DATA ON THE DEVELOPMENT (DEV) AND EVALUATION DATASET (EVAL)

System	Dev	Eval
GMM baseline	18.5	28.0
SGMM baseline	14.9	24.9
+speaker subspace	14.6	24.7
Cross-lingual: w/SP, $S = 20$	15.4	26.5
Cross-lingual: w/PO, $S = 20$	14.6	25.2
Cross-lingual: w/SW, $S = 20$	14.6	25.4
Cross-lingual: w/Mul, $S = 20$	13.4	24.5
Cross-lingual: w/Mul + ℓ_1, $S = 40$	12.7	22.1

TABLE VI
RESULTS OF CROSS-LINGUAL SGMM SYSTEMS WITH 15 HOUR TRAINING DATA FOR DEVELOPMENT (DEV) AND EVALUATION DATASET (EVAL)

System	Dev	Eval
GMM baseline	15.4	24.8
SGMM baseline	13.0	22.1
+speaker subspace	12.4	21.5
Cross-lingual: w/Mul + ℓ_1, $S = 40$	12.0	21.6

the out-of-domain data, the cross-lingual SGMM system can mitigate increases in WER arising from limited training data.

Table V summarizes the WERs of systems with 5h training data on both the development and evaluation datasets. Using multilingual parameter estimation and ℓ_1-norm regularization, the cross-lingual system obtains 12.7% on the development dataset and 22.1% on the evaluation dataset, a reduction of about 2% absolute compared to the speaker adaptively trained SGMM baseline using a monolingual subspace.

A summary of the results using the entire 15h training data is given in Table VI. In this condition, the cross-lingual system outperformed the baseline with speaker subspace adaptive training by 0.4% absolute on the development dataset and they achieved around the same accuracy on the evaluation dataset.

VI. CONCLUSION

In this paper, we have studied cross-lingual speech recognition using SGMM acoustic models in low-resource conditions. We first present a systematic review of the techniques used to build the cross-lingual SGMM system. We then carried out a set of experiments using the GlobalPhone corpus with three source languages (Portuguese, Spanish, and Swedish), using German as the target language. Our results indicate that the globally shared parameters in the SGMM acoustic model can be borrowed from the source language system. This leads to large reductions in WER when the amount of target language acoustic training data is limited (e.g. 1 hour). In addition, estimating the globally shared parameters using multilingual training data is particularly beneficial. We observed that the cross-lingual system using the multilingual parameters outperforms other cross-lingual systems using the monolingual parameters.

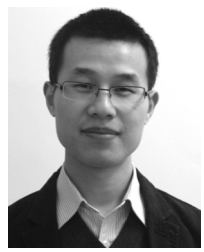
Our results also demonstrate the effectiveness of regularization using an ℓ_1-norm penalty for the state vectors. With a limited amount of training data, regularization is able to improve the numerical stability of the system, enabling the use of a model subspace of higher dimension and with more sub-state vectors. The benefits were demonstrated by experimental results using 1 hour and 5 hour training data in our study, in which substantial reductions in WER were obtained by using a higher dimensional model subspace together with regularization.

We also investigated the MAP adaptation of the model subspace, and cross-lingual speaker adaptive training using a speaker subspace. In both cases, however, they did not achieve further WER reduction on top of the multilingual parameter estimation and regularization in the low-resource settings according to our experimental results. In addition, we compared speaker adaptive training using monolingual and multilingual speaker subspaces and obtained comparable recognition accuracy in 5 hour and 15 hour training data conditions. This may indicate that the speaker subspace may also be portable across languages. Finally, the software and recipe for this work can be found in the Kaldi toolkit—`http://kaldi.sf.net`, released under the Apache License v2.0.

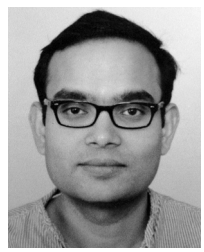
REFERENCES

[1] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech*, 1997, pp. 371–374.
[2] T. Schultz and A. Waibel, "Multilingual and crosslingual speech recognition," in *Proc. DARPA Workshop Broadcast News Transcript. Understand.*, 1998.
[3] W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and W. Wang, "Towards language independent acoustic modeling," in *Proc. ICASSP*, 2000, pp. 1029–1032.
[4] V. B. Le and L. Besacier, "First steps in fast acoustic modeling for a new target language: Application to Vietnamese," in *Proc. ICASSP*, 2005, pp. 821–824.
[5] 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.
[6] T. Slobada and A. Waibel, "Dictionary learning for spontaneous speech recognition," in *Proc. ICSLP*, 1996, pp. 2328–2331.
[7] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of phone sets and lexical transcriptions," in *Proc. ICASSP*, 2000, pp. 1691–1694.
[8] N. Goel, S. Thomas, M. Agarwal, P. Akyazi, L. Burget, K. Feng, A. Ghoshal, O. Glembek, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and P. Schwarz, "Approaches to automatic lexicon learning with limited training examples," in *Proc. ICASSP*, 2010, pp. 5094–5097.
[9] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Commun.*, vol. 35, no. 1, pp. 31–52, 2001.
[10] T. Schultz, "GlobalPhone: A multilingual speech and text database developed at Karlsruhe University," in *Proc. ICLSP*, 2002, pp. 345–348.
[11] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. ICSLP*, 1996, pp. 2195–2198.
[12] S. M. Siniscalchi, D. C. Lyu, T. Svendsen, and C. H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 875–887, Mar. 2012.
[13] K. C. Sim and H. Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proc. ICASSP*, 2008, pp. 4309–4312.

[14] K. C. Sim, "Discriminative product-of-expert acoustic mapping for cross-lingual phone recognition," in *Proc. ASRU*, 2009, pp. 546–551.
[15] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP*, 2000, pp. 1635–1638.
[16] A. Stolcke, F. Grézl, M. Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006, pp. 321–324.
[17] 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*, 2007, pp. 36–41.
[18] Y. Qian, J. Xu, D. Povey, and L. Jia, "Strategies for using MLP based features with limited target-language training data," in *Proc. ASRU*, 2011, pp. 354–358.
[19] C. Plahl, R. Schluter, and H. Ney, "Cross-lingual portability of Chinese and English neural network features for French and German LVCSR," in *Proc. ASRU*, 2011, pp. 371–376.
[20] P. Lal, "Cross-lingual automatic speech recognition using tandem features," Ph.D. thesis, The Univ. of Edinburgh, Edinburgh, U.K., 2011.
[21] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[22] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. IEEE SLT*, 2012, pp. 246–251.
[23] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task," in *Proc. INTERSPEECH*, 2008.
[24] D. Imseng, R. Rasipuram, and M. Magimai-Doss, "Fast and flexible kullback-leibler divergence based acoustic modelling for non-native speech recognition," in *Proc. ASRU*, 2011, pp. 348–353.
[25] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual informaiton to improve ASR for under-resourced languages," in *Proc. ICASSP*, 2012, pp. 4869–4872.
[26] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.
[27] 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. IEEE ICASSP*, 2010, pp. 4334–4337.
[28] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. IEEE ASRU*, 2011, pp. 922–932.
[29] L. Lu, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," in *Proc. ICASSP*, 2012, pp. 4887–4877–4880.
[30] L. Lu, A. Ghoshal, and S. Renals, "Regularized subspace Gaussian mixture models for speech recognition," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 419–422, 2011.
[31] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, NY, USA: Springer, 2005.
[32] G. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. ICASSP*, 2010, pp. 4346–4349.
[33] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. ICASSP*, 2010, pp. 4370–4373.
[34] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
[35] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec.. 2007.
[36] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions*, ser. Monographs and Surveys in Pure and Applied Mathematics. Boca Raton, FL, USA: Chapman & Hall/CRC, 1999, vol. 104.
[37] O. Siohan, C. Chesta, and C. H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 417–428, May 2001.
[38] P. Dutilleul, "The MLE algorithm for the matrix normal distribution," *J. Statist. Comput. Simulat.*, vol. 64, no. 2, pp. 105–123, 1999.
[39] D. Povey, "A tutorial-style introduction to subspace Gaussian mixture models for speech recognition," Microsoft Research, MSR-TR-2009-111, 2009, Tech. Rep..
[40] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlıcek, Y. Qian, P. Schwarz, J. Silovský, G. Semmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.



Liang Lu is a Research Associate at the University of Edinburgh. He received the B.Sc. and M.Sc. degrees in information engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2009, respectively. He then obtained the Ph.D. degree from the University of Edinburgh in 2013. His current research mainly focuses on noise robustness, cross-lingual/multilingual acoustic modeling and pronunciation modeling for speech recognition.



Arnab Ghoshal is a Research Associate at the University of Edinburgh. He received the B.Tech. degree from Indian Institute of Technology, Kharagpur, India in 2002, and the M.S.E. and Ph.D. degrees from the Johns Hopkins University, Baltimore, USA in 2005 and 2009, respectively. He was a Marie Curie Fellow at the Saarland University, Saarbrücken, Germany, from 2009 to 2011. His research interests include acoustic modeling for large-vocabulary automatic speech recognition, multilingual speech recognition, pronunciation modeling and adaptation.

Steve Renals is Professor of Speech Technology at the University of Edinburgh, where he is the director of the Institute for Language, Cognition, and Computation in the School of Informatics. He received the B.Sc. degree from the University of Sheffield, Sheffield, UK, in 1986, and the M.Sc. and Ph.D. degrees from the University of Edinburgh, Edinburgh, UK, in 1987 and 1991, respectively. He held postdoctoral fellowships at the International Computer Science Institute, Berkeley, CA, (1991–1992) and at the University of Cambridge, Cambridge, UK (1992–1994). He was a member of academic staff at the University of Sheffield for nine years as a Lecturer (1994–1999), then Reader (1999–2003), moving to Edinburgh in 2003. He has made significant contributions to speech technology in areas such as neural network-based acoustic modeling, information access from speech, and the automatic recognition and interpretation of multimodal meeting recordings. He is co-editor-in-chief of the ACM Transactions on Speech and Language Processing, an associate editor of the IEEE TRANSACTIONS ON AUDIO SPEECH AND LANGUAGE PROCESSING, and a member of the ISCA Advisory Council.