Cross-Lingual Automatic Speech Recognition Using Tandem Features

Cross-lingual Automatic Speech Recognition using Tandem Features

Partha Lal, Simon King *Senior Member, IEEE*

***Abstract*—Automatic speech recognition depends on large amounts of transcribed speech recordings in order to estimate the parameters of the acoustic model. Recording such large speech corpora is time-consuming and expensive; as a result, sufficient quantities of data exist only for a handful of languages — there are many more languages for which little or no data exist. Given that there are acoustic similarities between speech in different languages, it may be fruitful to use data from a well-resourced source language to estimate the acoustic models for a recogniser in a poorly-resourced target language.**

Previous approaches to this task have often involved making assumptions about shared phonetic inventories between the languages. Unfortunately pairs of languages do not generally share a common phonetic inventory. We propose an indirect way of transferring information from a source language acoustic model to a target language acoustic model without having to make any assumptions about the phonetic inventory overlap. To do this, we employ tandem features, in which class-posteriors from a separate classifier are decorrelated and appended to conventional acoustic features. Tandem features have the advantage that the language of the speech data used to train the classifier need not be the same as the target language to be recognised. This is because the class-posteriors are not used directly, so do not have to be over any particular set of classes.

We demonstrate the use of tandem features in cross-lingual settings, including training on one or several source languages. We also examine factors which may predict a priori how much relative improvement will be brought about by using such tandem features, for a given source and target pair.

In addition to conventional phoneme class-posteriors, we also investigate whether articulatory features (AFs) — a multi-stream, discrete, multi-valued labelling of speech — can be used instead. This is motivated by an assumption that AFs are less language-specific than a phoneme set.

***Index Terms*—Automatic speech recognition, Multilayer perceptrons**

I. Introduction

Training acoustic models for automatic speech recognition (ASR) typically requires hundreds of hours of transcribed speech data (e.g., [1]). Whilst such data exist for English and a handful of other languages, there are thousands of languages for which there is only a little data [2].

In the work presented here, we focus exclusively on acoustic modelling and not other aspects of the recogniser — for instance, we assume that a lexicon and language model exist for the language to be recognised. Our work examines ways in which training speech data in one or more languages can be used to improve the accuracy of a recogniser in a particular target language. This is achieved by encapsulating information learnt from data in the parameters of a *classifier*, which is then applied to the target language. In this paper, we will use the term *acoustic model* to refer specifically to the hidden Markov Models (HMMs) used to perform recognition; this is distinct from the classifier used to encapsulate information from data in one language, in order to transfer it to another language.

This classifier — a neural network — is learned from data in one or more *source* languages and then applied to data in a *target* language. The classes will be sub-word units, such as phonemes. To use the trained classifier *directly* to perform ASR would require that the languages are labelled with a common set of sub-word units or that a mapping is learnt from the sub-word units in the source language(s) to the target language. Neither of these is easy to achieve or entirely satisfactory. We use the classifier output *indirectly* and thus avoid the need for any mapping between label sets.

An established method for cross-lingual acoustic model training is that of Polyphone Decision Tree Specialization (PDTS) [3]. In a standard mono-lingual context-dependent acoustic model, clusters of related contexts are treated as the same, according to a decision tree learnt from source language data. Where PDTS differs is that the tree is then further grown with target language data. We pursue an alternative, tandem feature-based approach in order to avoid difficult decisions about phoneme correspondences across languages; however, it would be possible to use our approach in combination with PDTS.

Phonemes are the most common sub-word unit used in ASR but, since our proposed indirect method does not depend on using any particular set for the *classifier*, we can consider alternatives. The properties we might look for in a sub-word unit set include:

Realized in the same way in different languages
: This means that a classifier trained in one language is more likely to produce useful classifications on data from some other language. Since nominally identical phonemes (e.g. sharing the same IPA symbol) can in fact be realized differently in different languages [4], phonemes may not be the best choice for cross-lingual recognition.

Evenly distributed across languages
: For instance, we if we trained a classifier on a language which has few, or zero, instances of a unit that occurs frequently in the target language, this may result in poor performance.

Easily labelled
: Speech data usually have only word level transcriptions — the lexicon is then used to derive a phone-level

S. King, is with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, EH8 9AB, United Kingdom. TEL: +44-131-650-4434 FAX: +44-131-650-6626 E-mail: Simon.King@ed.ac.uk.

Manuscript received June 17th, 2012.

transcription. Any alternative to phonemes must also be somehow derived from words.

Few in number and acoustically distinct

This makes the classification problem easier.

The alternative to phonemes that we consider here is a set of articulatory features (AFs), described in more detail in Section V-A, which are a discrete multi-stream labelling of speech that encodes some properties of speech production. AFs have most of the desirable properties listed above: the concept applies to all languages; they have similar coverage in different languages ([5, page 98] and Section IV-D1); individual AF classifiers use fewer classes than phonemes. Previous work (e.g. [6]) has shown that AFs can be recognised from acoustic observations alone. Whilst manual labelling of AFs is not easy, we can derive an approximate AF labelling from phoneme labels. Note that the *acoustic models* will still be of conventional phone-like [7] sub-word units.

II. Background

A. Tandem features for automatic speech recognition

1) Phoneme-based Tandem Features: Tandem features were introduced in [8] for a noisy digits task and then in [9] for a noisy, medium-vocabulary spontaneous speech task (both in English). Tandem features are the concatenation of conventional acoustic features (e.g. MFCCs) to posterior-based features; the posterior-based features are typically posterior probabilities provided by a discriminatively-trained classifier, processed by a dimensionality reducing transform — see Section IV. In [9], phone-classifying MLPs were used to generate the posteriors. Substantial improvements were found in conjunction with context-independent acoustic models, and smaller but still significant gains with context-dependent models plus Maximum Likelihood Linear Regression (MLLR). In general, tandem features consistently improve accuracy, compared to conventional acoustic features alone, e.g. [10].

2) Articulatory Feature-based Tandem Features: Since the posterior probabilities from the classifier are processed and then effectively considered as just another form of acoustic feature, there is no requirement for the classes over which the posteriors are estimated to be consistent with any other component of the system (specifically, the sub-word units used in the acoustic models and pronunciation dictionary). They can use a different phoneme inventory or any other set of classes. Articulatory features (AFs) are an obvious alternative to phonemes for tandem features, as first reported in [11] where, on a small/medium vocabulary spontaneous speech task, they performed as well as phonemes.

B. Cross-lingual ASR using Tandem Features

Because the posterior-estimating classifier is effectively acting as a sophisticated form of acoustic feature extraction, it is possible to use a classifier trained on one data set to estimate posterior-based features for other data. The earliest example of this is [12], in which data from a spontaneous speech corpus is used to train tandem features that are then used in a continuous digit recognition task. The effect of adding various amounts of task-specific data to the training set was also investigated.

That work used English language corpora, the English part of the OGI Multilingual Corpus [13] and the Spine corpus [14], with the transfer being between different tasks. The first instance of tandem features used cross-lingually is [15] — MLPs were trained with conversational telephone speech data in English and then used to generate MLP features for use in Mandarin and Levantine Arabic recognisers. Consistent improvements in word error rate were observed in both cases and the authors state that phonetic distinction shared at the level of articulatory features may explain why the MLP features represent the acoustic space well.

Following on from [11], [16] reports the use of AF-based tandem features from a classifier trained on English continuous telephone speech data to generate tandem features for a Mandarin broadcast news task. However, whilst phoneme-based tandem features (from an MLP trained on English data) reduced WER on the Mandarin task from 21.5% to 21.2%, AF-based tandem features (also from MLPs trained on English data) actually increased WER.

In the work we present in this paper, we demonstrate consistent WER reductions using both phoneme- and AF-based tandem features in cross-lingual settings.

Another example of tandem features being used cross-lingually is [17], in which English phoneme MLPs and English AF MLPs[1] are used to generate tandem features for a Hungarian telephone speech recognition task. As well as those two cross-lingual systems and monolingual tandem and non-tandem baselines, a system that used an adapted MLP was produced. The adapted MLP took the English phoneme MLP and retrained some model parameters with Hungarian data.

Some results from that work include

- Both English phoneme and AF MLPs provide an improvement over the non-tandem baseline but do not perform any better than using tandem features from the Hungarian phoneme MLP. Domain and channel differences may have contributed to this result.
- Using the adapted MLP resulted in word error rates statistically significantly better than all other systems.

In [19] tandem features are used but the cross-lingual element comes about through retraining of the MLP. The task addressed is the challenging Callhome corpus of conversational telephone speech. An MLP was trained to classify German and Spanish speech using a pooled phoneme set. It was then applied to English — output activations were observed as English speech was passed forward through the net and the mutual information between English phoneme labels and pooled German-Spanish phonemes was calculated. That information was used to learn a mapping between English and German-Spanish phonemes and the MLP underwent further training with a limited amount of target language data, now relabelled with German-Spanish phonemes.

Recognition accuracy improved with the use of non-target language speech data. The main differences between that work and ours is in their use of MLP training as the tool for cross-lingual transfer as well as their use of novel acoustic features.

[1] Again, the AF MLPs from [18] that were trained on 2000 hours of Fisher corpus data were used.

Our method does not require retraining of the classifiers or relabelling of the MLP outputs with a new phone set.

[20] features the use of AF posteriors in a Kullback-Leibler divergence-based HMM (KL-HMM). A typical KL-HMM takes phoneme posteriors at each frame and computes the KL-divergence between them and a reference multinomial distribution defined for each state. The state sequence that minimizes the total KL-divergence is found by Viterbi decoding.

That paper showed that by using a multi-stage series of AF MLPs to estimate AF posteriors it is possible to perform phoneme recognition as accurately as with phoneme MLPs on the TIMIT corpus. Furthermore, AF posteriors can easily be combined with phoneme posteriors in a KL-HMM system to give an improvement in accuracy relative to a phoneme posterior only system.

C. Novel contributions of our work

In the context of the prior work described in this section, the novel contributions of this work can be summarised as:

- We show that tandem features can be used cross-lingually and can result in a statistically significant improvement over a non-tandem baseline
- We demonstrate that a set of MLPs classifying articulatory features can generate better tandem features than an MLP classifying phonemes, resulting in reduced word error rates
- We investigate the use of cross-lingual tandem features when only limited amounts of target language data are available. In some situations, pooling data from multiple source languages to train a language-independent MLP is shown to be more effective than an MLP trained exclusively on the limited target language corpus.

III. Data

We need a multilingual transcribed speech corpus with which to train our MLPs and GMM-HMMs. The GlobalPhone corpus was chosen because it contains enough data in each language for baseline recognisers to be built and because it covered a wide range of languages. Ten of the available languages were selected such that a wide range of phonetic phenomena are seen and some groups of similar languages exist. So far, experiments have only been performed with six of those languages, due to a lack of language models for the other four.

Our choice of languages covers a range of language families — their relation to each other is described in Figure 1. The phonetic characteristics of each of the language families included, in particular those aspects that differ between families, are briefly given below. Variation in the set of phonetic phenomena exhibited in source and target languages is one of the challenges faced in cross-lingual speech recognition and so choosing a set of languages with a diverse range of properties should provide more widely-applicable results from experiments.

Chinese In Mandarin Chinese, syllables consist of a vowel nucleus, which can be a monophthong, diphthong or triphthong, and optionally have an onset and coda. The tone of the vowel is phonemic. Consonant clusters are rare in the syllable onset. In Mandarin, only /n/ and /ŋ/ are valid codas. [21]

Germanic Swedish features a unique voiceless palatal-velar fricative realization of /ɧ/ [22, pages 171–2, 330; 173–6]. It also possibly has more than one type of lip rounding gesture in vowels [22, page 295]. Both German and Swedish have phonemic vowel length. German and Russian have broadly similar movement patterns for labiodental fricatives [22, page 140].

Romance Spanish has an alveolar trill /r/ that also appears in Russian [22, page 218]. Spanish is unusual amongst the world's languages in having dental fricatives [23]. An uncommon aspect of Portuguese is that, whilst laterals in most languages have some place of articulation, it has completely unoccluded laterals [22, page 193].

- Sino-Tibetan → **Chinese**
- Indo-European
  - Germanic
    - North → East Scandinavian → Danish-Swedish → **Swedish**
    - West → High German → German → Middle German → East Middle German → **German**
  - Balto-Slavic → Slavic → East-Slavic → **Russian**
  - Italic → Romance → Italo-Western → Western → Gallo-Iberian → Ibero-Romance → West Iberian
    - Portuguese-Galician → **Portuguese**
    - Castilian → **Spanish**

Fig. 1. The placement of the languages used in our experiments within the Ethnologue language hierarchy.

Language	Number of speakers									Total
	Training			Development			Evaluation			
	M	F	Σ	M	F	Σ	M	F	Σ	(hours)
Chinese	53	58	111	6	5	11	5	5	10	31
German	62	3	65	4	2	6	4	2	6	18
Portuguese	45	41	86	4	4	8	4	3	7	26
Russian	51	44	95	5	5	10	5	5	10	22
Spanish	34	45	79	5	5	10	4	4	8	22
Swedish	40	39	79	5	4	9	5	5	10	22

TABLE I
THE NUMBER OF SPEAKERS IN GLOBALPHONE IN EACH CORPUS SPLIT, WITH GENDER, AND THE TOTAL SIZE OF THE CORPUS IN HOURS.

GlobalPhone consists of recordings of speakers reading from a newspaper in their native language. Recordings were made under a range of 'quiet' conditions using identical recording equipment although recording location varied. The amount of data available in each language, as well as the standard partitioning into training, cross-validation (development) and test (evaluation), plus the gender split of the speakers is described in Table I. The sizes of the available GlobalPhone lexica in each language are given in Table II. The phoneme inventory for each language is described in Table III.

Language	Pronunciations	Words
Chinese	73388	73387
German	48979	46037
Portuguese	54163	51987
Russian	28818	27062
Spanish	41286	28803
Swedish	25402	25257

TABLE II
GLOBALPHONE LEXICON SIZES FOR EACH LANGUAGE.

Shared by this many languages	Number of phonemes	Polyphonemes	
		Consonants	Vowels
All	10	f, k, l, m, n, p, s, t	i, u
5	7	b, d, g, r	a, e, o
4	5	j, ʃ, v, x, z	
3	4	ŋ, ɴ, ts	y
2	29	ç, dʲ, ɾ́, ʂ, tʲ, zˌw	ɛ, ai, aː, ä, ɐ, au, ɐi, eː, ë, ɘ, eu, iː, ɪ̈, ɔ, ø, øː, oː, ö, yː, uː, ü
Language	Number of monophonemes (total phonemes)	Monophonemes	
Chinese	24(45)	k^h, ɕ, t^h, ts^h, tʂ, $tʂ^h$, tɕ,$tɕ^h$	ɑ,ɑʊ, ɑɪ, ia, iɑʊ, iɛ, iɔ, iou, ou, ɤ, ua, uaɪ, uei, yœ, uɔ
German	1(44)	-	ɐ
Portuguese	15(48)	ʁ	ã, ˈã, ɐ̃, ẽ, ˈẽ, ĩ, ˈĩ, ɪ, õ, ˈõ, ũ, ˈũ, ʊ
Russian	16(49)	bʲ, lʲ, mʲ, pʲ, ʔ, rʲ, sʲ, ɕː, ɕːʲ, zʲ, zˌʲ, ʃʲ, ts, tsʲ, vʲ	ɯ
Spanish	8(43)	ð, ɣ, ɴ, ɾ, θ, tʃ, β	oi
Swedish	14(52)	ɖ, ks, ɭ, ɳ, ʈ	ɑː, ɛː, æ, æː, ɔ, œ, œː, ɵ, ʉː
Σ	78		

TABLE III
PHONEME DISTRIBUTION ACROSS LANGUAGES. THIS TABLE IS IN FACT A VERSION OF [5, TABLE 4.3] LIMITED TO THE SIX LANGUAGES USED HERE. POLYPHONEMES ARE PHONEMES APPEARING IN MORE THAN ONE LANGUAGES, MONOPHONEMES APPEAR IN ONLY ONE.

A. Cross-corpus normalization

Whilst the GlobalPhone corpus benefits from the fact that the same recording equipment was used throughout (and sampling rates, bit depths etc. were also consistent), recording sessions were conducted at different locations around the world, in different sized rooms and occasionally under different noise conditions. This will inevitably result in acoustic differences that are independent of the words being spoken. This section describes efforts to address this issue.

Prior work in this area includes [24], which involved estimating a corpus-normalizing feature transform for each corpus. Training maximised the likelihood of the training data by alternately updating the transforms and the GMM means and variances, until convergence.

Here we focus on the point at which the data from the different corpora meet, that is, when target language acoustic observations are passed through the source language MLP. We apply a linear transformation to those features before feeding them through the net. Note that, whilst the PLPs used by the MLP are transformed, the MFCCs modelled with a GMM are unchanged. The MLP itself remains unchanged throughout this process. We derive the transform as follows:

1) Use two single state HMMs to model the source language training data — one HMM models all speech frames and the other models silence (initial labels are derived from the existing word-level transcriptions). Each HMM state uses a 128 component GMM to model a 39 dimension PLP feature vector
2) Treat the target language training data as if it were adaptation data and compute an MLLR transform that brings it closer to source language speech[2]
3) Apply that transform to all target language data before it is passed through the source language MLP

We are able to apply a model transform as if it were a feature transform here because all speech data is modeled with one HMM. The same transform is applied to all frames even though it would have been preferable to apply the transform learnt for silence to silent frames and the transform learnt for speech to speech frames. Since speech is significantly different to silence, this mistreatment of silent data is assumed to have little effect.

To evaluate the effectiveness of the transform we considered the following methods:

1) measure the reduction in MLP frame error rate when the transform is applied *or*
2) compare the accuracy of tandem systems built from original vs transformed PLPs *or*
3) measure the increase in mutual information (MI) between target language labels and acoustic features before / after applying the transform

The first of these isn't a viable option — it is difficult to draw meaningful conclusions from MLP frame error rates when the MLP is being used cross-lingually. Even if there is some overlap in the source and target phoneme sets, the error rate observed still appears to be high even though the MLP can be used to produce useful tandem features[3]. The second option would be ideal, except that is is computationally expensive. The third option was therefore selected.

The comparisons of MI are shown in Table IV — all features were speaker normalized to zero mean and unit variance, only the development set was used to compute the results in this table. We can see that applying the adaptation method described above generally results in an increase in MI between acoustic features and reference phoneme labels and that this holds true for both the PLP features themselves and the transformed MLP features (the one exception being German).

Note that the use of MI is not contingent on any aspect of the normalization method and so other cross-corpus normalization methods could be compared in the same way.

[2]Only means are adapted in this work, i.e. `MLLRMEAN` in HTK

[3]For example, tandem features for a Portuguese recogniser generated with a Spanish phoneme MLP give a 17% relative drop in WER compared to baseline. However, passing Portuguese data through the Spanish phoneme MLP gives a misleadingly high frame error rate of 67.1%.

Target Lang.	relative increase in mutual information (%) PLP	MLP features
Chinese	9.5	6.8
German	37.9	-6.0
Portuguese	6.1	4.8
Russian	12.1	4.4
Spanish	18.5	12.7

TABLE IV
MUTUAL INFORMATION INCREASES OBTAINED BY USING MLLR FEATURE NORMALISATION FOR BOTH THE PLPS THEMSELVES AND THE RESULTANT MLP FEATURES. FOR THE 2^{nd} AND 3^{rd} COLUMNS, THE MEAN ACROSS ALL SIX SOURCE LANGUAGES IS SHOWN.

Source Language	Word error rate (%) normalized	baseline
Chinese	26.0	27.3
German	25.7	26.3
Portuguese	25.3	25.9
Russian	25.6	25.8
Spanish	22.8	23.2
Swedish	28.5	25.3

TABLE V
WORD ERROR RATES FOR A SPANISH RECOGNISER USING VARIOUS SOURCE LANGUAGE TANDEM FEATURES, WITH AND WITHOUT CROSS-CORPUS NORMALIZATION.

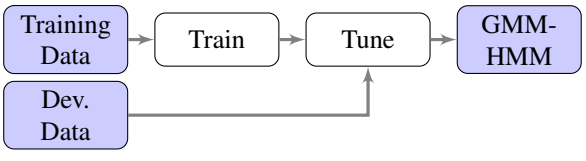
As a final check that the increases in MI due to cross-corpus normalization do result in improvements in WER, we built recognisers for Spanish using tandem features from each of the languages. Results (for the development set) are shown in Table V. Apart from when the Swedish MLP is used, cross-corpus normalization always results in an improvement in word error rate for all source languages when applied to Spanish. This is consistent with the predictions made by the mutual information measure.

Because of the order in which our experiments was conducted, the normalised features generated by this method were not arrived at in time for use in subsequent experiments — no cross-corpus normalization is performed in the work that follows. However, we would predict a small decrease in WER across the board by adding normalization.

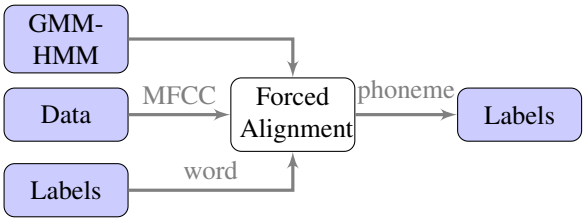
IV. PHONEME TANDEM FEATURES

The steps required to create a recogniser that uses tandem features are now described.

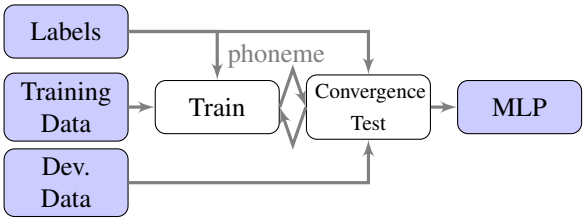
1) **Train the MFCC baseline system**, as described in Section IV-A



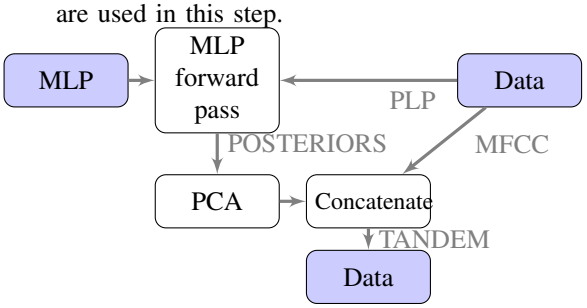
2) **Generate a frame-level phoneme labelling** for the corpus by forced-alignment of the MFCC baseline model. This step also requires a word-level transcription and a lexicon that maps to phonemes.



3) **Train an MLP** using frame-level targets obtained from the previous step. PLPs[4] are extracted from the acoustic data instead of MFCCs.

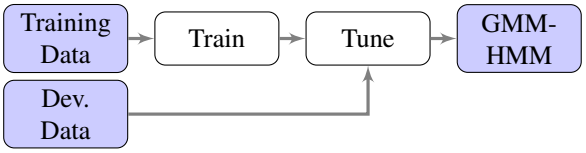


4) **Generate tandem features**:
   a) Apply that trained classifier to the corpus, estimating **phone posteriors** at each frame.
   b) **Take logs** of those posteriors. This is equivalent to omitting the softmax function that is usually applied in the output layer of the net. Taking logs results in features that have a more Gaussian distribution.
   c) Transform them using, for example, **PCA**. PCA is used to decorrelate and reduce the dimensionality of the features that are going to be concatenated; using HLDA (Hetroscedastic Linear Discriminant Analysis, [25]), or some other similar scheme would be an option here. The PCA transform is estimated using the training set[5].
   d) The massaged MLP output vector is now **concatenated** to a vector of acoustic features; these could be the same as those input to the MLPs but a further gain in performance can be attained if complementary acoustic features are used, e.g. the MLPs are trained with PLPs, and then MFCCs are used in this step.



[4] PLP_E_D_A_Z in HTK

5) The new features can be **modelled with a GMM-HMM** again using the training schedule described in Section IV-A



A. GMM-HMM Training

A conventional GMM-HMM training workflow was used — HTK [26] was used for this and the tutorial recipe in HTKBook was used as a starting point. The same training schedule was used for all GMM-HMMs including the baseline MFCC-based model (which is also used for generating training labels for the MLP) and all tandem models. There are a handful of model parameters that are tuned on the development set of the relevant corpus:

- the average number components per Gaussian mixture
- the number of triphone models
- word insertion penalties and grammar scales

The language models were provided by Tanja Schultz (creator of the GlobalPhone corpus). During decoding, a bigram language model was used for first-pass lattice generation. A trigram model was then used to rescore those lattices, with the first-best path through the lattice being used as the recognition result.

B. Multi-layer Perceptrons

Training and classification were both performed with QuickNet [27]. The input to the MLP consists of the PLP coefficients at the frame to be classified and at the adjacent four frames in either direction — a nine frame context window. At the beginning of each utterance the left-hand context consists of some padding frames — the first frame repeated four times – with equivalent padding at the end of each utterance.

1) Training: We use three-layer feed forward MLPs and train them in the conventional way, using back-propagation to minimize errors on the training set. Training consists of two steps, jointly referred to as an epoch, that are iterated until the frame error rate on a cross validation set (identical to the development set mentioned earlier) converges.

Propagation New training patterns are presented to the network. In a batch update setup as used here, multiple patterns are presented before weights are updated (the exact number of patterns is discussed below).

Forward The input is passed forwards through the network and the current weights are used to give output activations at each layer

Backward At the output layer, the activations are compared with the training targets, to give deltas

Update For each weight:

- Compute the gradient at that point using previously computed deltas
- Using the gradient to determine which direction would reduce error, update the weight in that direction by moving an amount proportional to the learning rate

The initial learning rate used when training the MLPs was 0.005. The "newbob" learning rate schedule was used, meaning that after starting with the initial learning rate we repeat epochs until the development set frame accuracy increases by less than 0.5% over the previous epoch. After that, the learning rate is halved, to home-in with increasing precision on the local optimum[6].

The MLP input was normalised to have zero mean and unit variance. The number of units in the hidden layer is set such that the number of free parameters in the net is equal to some percentage of the total number of data frames[7]. The number of free parameters in a net with an $I \times H \times O$ structure is $I + H + O + H(I + O)$ where I is the product of the number of features per frame (39 PLP coefficients) and the size of the context window (9 frames), H is the size of the hidden layer, and O is the size of the output layer (i.e., the number of phonemes in the language being classified). The actual percentage used was around 35–50% but was tuned for each language so as to maximize development set accuracy. A softmax output function is used so that output nodes sum to one and can be treated as the posterior probability of their corresponding class.

All MLPs are gender-independent. The presentation order of the training data was randomized. A batch update of the parameters is applied during training after each 'chunk' of data is processed. The chunk size is determined dynamically by a simple heuristic which refers to the memory available on the executing machine. Chunk size is selected to be as large as can be held in memory but also such that the final chunk is not small (a small final chunk would bias the estimated parameters towards the data appearing in that less representative chunk).

The language-independent MLPs are trained in much the same way as the monolingual MLPs, with the same schedule used in training. For both phoneme and AF MLPs we treated symbols (phonemes or articulator states) as being the same across languages.

2) Classification: If we were using the MLPs as classifiers, the output unit with the highest activation — i.e., the one with the greatest posterior probability — would be selected and the label corresponding to that unit would be taken to be the label for the current frame. Classification accuracy is one way to evaluate the performance of the nets before proceeding to train the full HMM-GMM Tandem system.

Table VI reports the classification performance of the language-specific MLPs described in Table VII (since silence is very easy to classify and, at the same time, not very useful, all frames labelled as silent in the reference labelling are ignored).

[5]The number of dimensions is reduced such that 95% of the variance is still accounted for — this conveniently means that we can fairly compare systems built using different numbers of sub-word units. A script to implement this heuristic was provided by Özgür Çetin.

[6]http://www.icsi.berkeley.edu/speech/faq/nn-train.html

[7]This heuristic was suggested by Joe Frankel

Language	Frame error rate(%) dev	eval
Chinese	31.2(32.2)	34.6(35.6)
German	26.5(30.6)	25.2(29.2)
Portuguese	47.7(49.4)	41.1(42.4)
Russian	38.4(39.5)	37.5(38.6)
Spanish	31.8(32.6)	29.0(29.8)
Swedish	39.6(41.3)	37.4(39.2)

TABLE VI
FRAME ERROR RATES FOR ALL USED LANGUAGES ARE REPORTED HERE — IGNORING THE SILENCE CLASS GIVES THE FIGURES IN PARENTHESES.

Language	units in layer hidden ($\times 10^3$)	output	free params as % of data frames	Training corpus (hh:mm)	Training time (hh:mm)
Chinese	12.1	44	50	25:50	24:48
German	4.85	44	35	14:35	04:48
Portuguese	8.35	48	40	22:23	12:40
Russian	8.07	45	45	18:38	13:53
Spanish	8.00	43	50	16:48	10:07
Swedish	7.11	52	45	17:02	08:07
German, Portuguese & Spanish	19.5	77	-	53:36	> 50
Portuguese, Spanish & Swedish	19.1	91	-	56:06	> 90

TABLE VII
INFORMATION ABOUT THE MLPS USED TO CLASSIFY PHONES IN OUR TANDEM SYSTEM AND THE CORPORA USED TO TRAIN THEM.

C. Results

The results in Table VIII demonstrate a consistent improvement in recognition accuracy is obtained for Tandem features when the source language is identical to the target language, consistent with results in the literature. A matched pairs sentence-segment word error statistical significance test [28] tells us that, at a 95% confidence level, the tandem system performs significantly better than baseline for all languages — statistically significant differences are shown in bold face.

Target language	Word error rate (%) Source language CH	GE	PO	RU	SP	SW	Baseline
Chinese	**17.9**	22.7	22.5	23.4	24.0	23.6	23.3
German	25.3	**23.5**	**24.5**	25.2	24.9	**24.6**	26.1
Portuguese	22.4	**21.0**	**18.4**	**20.4**	**20.2**	**21.3**	23.5
Russian	34.2	33.9	**32.5**	**30.5**	**33.1**	33.2	34.7
Spanish	18.2	17.9	**17.1**	**17.2**	**16.0**	**17.5**	18.3

TABLE VIII
WORD ERROR RATES FOR VARIOUS CROSS-LINGUAL PHONEME TANDEM SYSTEMS.

Table IX shows the word error rates achieved when using MLPs trained on multiple languages to generate tandem features either for a language in the training set, or a different language. The WERs are better than the MFCC baseline, although generally significantly worse than when a matched monolingual MLP is used.

Target Language	Word error rate (%) Multi-language {German, Portuguese, Spanish}	{Portuguese, Spanish, Swedish}	Mono-lingual	Non-tandem baseline
German	23.8	**26.1**	23.5	26.1
Portuguese	**19.8**	**24.9**	18.4	23.5
Spanish	**16.7**	**17.2**	16.0	18.3

TABLE IX
WORD ERROR RATES FOR SYSTEMS USING MLPS TRAINED ON MULTIPLE LANGUAGES – THE OUTPUT LAYER OF THESE NETS IS A SHARED PHONESET – REPORTED ON THE EVALUATION SET. STATISTICALLY SIGNIFICANT DIFFERENCES (EITHER BETTER OR WORSE) RELATIVE TO THE MONOLINGUAL TANDEM SYSTEM ARE SHOWN IN BOLD.

D. Analysis

The improvements in accuracy brought by tandem features can be explained by a number of factors, the main ones being access to a wider time context (i.e. the 9-frame window input to the MLP) and the use of a discriminatively trained classifier. Some general observations that can be made about the results presented here include:

- For both Romance languages, the second most effective source language (after the same language itself) is the other Romance language. Those two languages also have a high degree of lexical similarity[8].
- Looking at the two Germanic languages, a statistically significant improvement over baseline occurs when one of the languages is used to generate tandem features for the other.
- Chinese belongs to a different language family to the others and does not receive any improvement from the use of tandem features generated from nets trained on other languages.

Whilst these observations may be useful, it could be more desirable to have a quantitive measure which can predict the potential gains before a system is actually built. We now introduce some candidate measures and examine how well they correlate with word error rate improvements.

1) Share Factor: In order to quantify the degree of overlap between different phoneme sets, [29, Section 2.3] defines the phoneme share factor sf_N for a set of N languages. The share factor can be interpreted as the average number of languages sharing the phonemes of the global (pooled) phoneset.

Here, only a source and target language are involved ($N = 2$) so the share factor is simply the sum of the sizes of the two monolingual phonesets, divided by the size of the shared phoneset. It will range between 1 and 2, indicating completely distinct or completely overlapping phoneme sets respectively.

2) Triphone Overlap: [30, Table 3.6], somewhat like [29, Table 4.5], shows what proportion of source language triphones are covered by target language triphones.

[8]In places, Ethnologue provides lexical similarity figures. Lexical similarity is defined here as the percentage of overlap in the words appearing in each language. The figures provided by Ethnologue are computed by taking a standardised word list and looking at the similarity of words with a shared meaning. Spanish and Portuguese overlap by 89%, which is deemed by Ethnologue to be a high degree of similarity: comparable to that between dialects.

Variable	Mean correlation coefficient
Share factor	0.91
Triphone overlap	0.90
Mutual information	0.73
MLP FER	0.41

TABLE X
A COMPARISON OF VARIABLES PREDICTING CROSS-LINGUAL PERFORMANCE OF TANDEM FEATURES.

Target lang. data (hh:mm)		Word error rate (%)	
train	dev	Monolingual	Language-independent
14:35	1:58	23.5	23.8
7:11	0:51	24.5	25.4
3:31	0:46	26.2	26.8
1:05	0:21	43.6	39.1

TABLE XI
WORD ERROR RATES FOR GERMAN RECOGNISERS USING PHONEME TANDEM TRAINED WITH VARYING AMOUNTS OF GERMAN DATA.

Unlike [29], the source language model in our work is used indirectly for target language recognition, so source language triphones do not actually make an appearance in the target model. However, triphone overlap may still give some estimate of linguistic similarity if we assume shared labels in different languages refer to the same sound.

3) MLP Accuracy: We can also examine the relationship between the accuracy of the MLP used to generate tandem features, measured in terms of MLP classification frame error rate, and the improvement in WER that those features then provide to the resulting tandem system. Section III-A has already discussed why it is not possible to measure MLP accuracy using target language data, so we use the frame error rate (FER) on source language data as the measure.

4) Mutual Information: The final measure we considered is the mutual information between the tandem features and target language phone labels — we can then correlate that to the tandem system accuracy. The transformed output of the MLP is used here, rather than full tandem features with MFCCs appended (i.e. the output of step 4(c) in Section IV). Cepstral mean and variance normalization is applied on a speaker-level.

5) Comparing Predictors: Correlation coefficients, averaged across each of the target languages, for each measure listed above are shown in Table X. First of all, we can see that relatively simple measures have a high degree of correlation with word error rate. Given a range of options for source language to use in a cross-lingual system, we can make an accurate estimate of the best language to use by examining the monophone share factor or triphone overlap.

However, those measures do not take into account the amount or quality of data available and so can probably only be used when the source corpora are similar to each other in size and type. In fact, these measures only perform so well here because multiple systems from the same language were not included in the comparison. If tandem features generated using less training data, different feature sets or less accurate reference labels were used when computing the correlation coefficient then these measures probably would not perform as well.

Next we look at the mutual information between the tandem features and target language phone labels. Whilst the mean correlation is weaker here than for the simpler measures, this method does have some advantages over them, the primary one being that the acoustic features have some bearing on the measure. It also allows us to draw comparisons between different types of tandem features created from the same MLP, such as features with or without cross-corpus normalization (Section III-A).

Surprisingly, the frame error rate of the source language MLP has the least correlation of all with word error rate reduction. This could be explained by the fact that source language MLP error rates are independent of the choice of target language. An MLP may accurately predict phonemes for the language it was trained for but whether it can be used to produce useful tandem features for some other target language depends on the choice of source and target languages.

6) Language-independent Results: Table IX also showed results where language-independent MLPs performed significantly worse than when a matched monolingual MLP is used. This can perhaps be because there is a mis-match between the MLP training corpus and the data it was applied to. Training with different corpora in the same language might outperform a monolingual MLP. This observation is paralleled by a similar result in [12, Section 3].

The results in this section have shown that monolingual tandem features provide a statistically significant improvement in word error rate and that in many cases cross-lingual tandem features also give a similar improvement. However, the cross-lingual example does raise the question, if you have access to target language data, why use tandem cross-lingually? One circumstance where cross-lingual tandem makes sense is where we don't know the target language in advance, and don't have the resources to train a target language MLP before recognition. Another, more realistic, circumstance is where the amount of source language data is far greater than the amount of source language data available — Table XI shows how a language-independent cross-lingual system outperforms a monolingual system when less than around three hours of training data is available. Experiments with limited data are covered in more detail in [30, p73].

V. ARTICULATORY FEATURE TANDEM

Thus far, we have used a single MLP, classifying speech frames into phonemes. We now compare this to using multiple MLPs, each of which provides posteriors for a different articulatory feature.

A. Articulatory Features

Articulatory features (AFs), as used in this work, are a multi-stream labelling of the speech signal that is loose representation of the actions of the speech articulators. They are an abstraction, rather than a complete representation of the articulators' precise physical positions. The articulatory

Feature	Values	Cardinality
Place	labial, labio-dental, alveolar, post-alveolar, velar, glottal, lateral, none	8
Manner	vowel, approximant, fricative, closure, trill	5
Nasality	+, -	2
Voicing	voiced, voiceless	2
Rounding	+, -	2
Vowel	German: a, ɛ, ɐ, e, ɘ, i, o, ø, u, y	10
	Portuguese: a, ɐ, e, i, ɪ, o, u, ʊ, ɯ	9
	Spanish: a, e, i, o, u	5
Height	very high, high, mid-high, mid, mid-low, low, nil	7
Frontness	back, central, front, mid, nil, reduced-back, reduced-front	7
Stress	+, -	2

TABLE XII
ARTICULATORY FEATURES AND THEIR VALUES.

features used here (based on [11]) and their values are shown in Table XII ("silence" is another valid value for all features).

Our motivation for considering AFs is that they are less language-specific than phonemes: it is easier to devise a language-independent AF set than a language-independent phoneme set. Furthermore, because AFs are a factorial representation, each feature has fewer possible values and therefore will suffer less from data sparsity problems than a phoneme set.

B. Articulatory feature classification

Training the AF MLPs requires frame-level labels for each AF, which were derived from phoneme labels via the following three steps:

1) Use the same forced-alignment as for training phoneme MLPs
2) Split phones that are composed of two parts (including diphthongs and plosives) into two different labels.
3) Map these new labels to their corresponding articulatory feature values. The mapping used is listed in [30, Table A.6]

Details of the AF MLPs are given in Table XIII and the performance of tandem recognisers using those MLPs is provided in Tables XIV and XV. We can see that AF tandem recognisers perform at least as well as phoneme tandem recognisers for each of the languages examined and in one case significantly better. Word error rates achieved with the language-independent MLP are not as low as with the monolingual ones but a similar pattern of improvement over phoneme tandem is observed.

VI. CONCLUSIONS

We have shown that tandem features, using either phoneme MLPs or AF MLPs, can be used successfully in a cross-lingual scenario — recognition accuracy is significantly better than in an MFCC-based baseline model. Articulatory feature MLPs work at least as well as phoneme MLPs for this purpose, and are especially effective where data from multiple source languages is used — AF tandem can sometimes be significantly better than phoneme tandem. An extended report of this work can be found in [30].

Language	units in layer		free params as % of data frames	Training data (hh:mm)	Training time (hh:mm)
	hidden	output			
German	5360± 51	6.44± 3.21	35	14:35	2:19–3:51
Portuguese	9312± 78	6.67± 3.08	40	22:23	6:08–9:41
Spanish	8841± 66	5.78± 2.77	50	16:48	3:38–7:01
Swedish	8007± 100	7.11± 4.65	45	17:01	3:00–5:43
German, Portuguese & Spanish	22.5± 0.27	7.67± 4.30	-	53:36	21.23–52:56

TABLE XIII
INFORMATION ABOUT THE MLPS USED TO CLASSIFY ARTICULATORY FEATURES IN OUR TANDEM SYSTEM, AND THE CORPORA USED TO TRAIN THEM. THE NUMBER OF HIDDEN AND OUTPUT UNITS VARIES BETWEEN AFS: ONLY MEANS AND STANDARD DEVIATIONS ARE GIVEN HERE. LIKEWISE, FOR TRAINING TIMES, THE RANGE FROM ONE STANDARD DEVIATION ABOVE AND BELOW THE MEAN TIME IS STATED.

Target Language	Word error rate (%)		
	Articulatory Feature	Phoneme	Non-tandem baseline
German	23.1(22.5)	23.5(22.1)	26.1(26.9)
Portuguese	**17.2**(21.4)	18.4(21.8)	23.5(26.1)
Spanish	15.6(22.2)	16.0(23.2)	18.3(27.3)

TABLE XIV
WORD ERROR RATES WHEN AF MLPS ARE USED TO GENERATE TANDEM FEATURES, WITH PHONEME TANDEM AND NON-TANDEM SYSTEMS DISPLAYED FOR COMPARISON. RESULTS ARE REPORTED ON THE EVALUATION SET WITH DEVELOPMENT SET FIGURES IN BRACKETS. AF TANDEM SYSTEMS THAT ARE STATISTICALLY SIGNIFICANTLY DIFFERENT TO THEIR CORRESPONDING PHONEME TANDEM SYSTEM ARE SHOWN IN BOLD.

REFERENCES

[1] A. Janin, A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, J. Frankel, and J. Zheng, "The ICSI-SRI Spring 2006 Meeting Recognition System," in *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006*, ser. Lecture Notes in Computer Science, S. Renals, S. Bengio, and J. Fiscus, Eds. Washington, D.C.: Springer, 2007, pp. 444–456.

[2] R. G. Gordon, Jr and B. F. Grimes, Eds., *Ethnologue: Languages of the World*, 15th ed. Dallas, ,TX, USA: SIL International, 2005, www.ethnologue.com.

[3] T. Schultz and A. Waibel, "Language adaptive LVCSR through Polyphone Decision Tree Specialization," in *Workshop on Multi-lingual Interoperability in Speech Technology (MIST)*, Leusden, The Netherlands, Sep. 1999, pp. 85–90.

[4] D. Imseng, H. Bourlard, M. Magimai.-Doss, and J. Dines, "Language Dependent Universal Phoneme Posterior Estimation for Mixed Language

Target Language	Word error rate (%)		
	Articulatory Feature	Phoneme	Non-tandem baseline
German	24.0	23.8	26.1
Portuguese	**18.7**	19.8	23.5
Spanish	**16.1**	16.7	18.3

TABLE XV
WORD ERROR RATES WHEN LANGUAGE-INDEPENDENT AF MLPS ARE USED TO GENERATE TANDEM FEATURES, WITH LANGUAGE-INDEPENDENT PHONEME TANDEM AND NON-TANDEM SYSTEMS DISPLAYED FOR COMPARISON. GERMAN, PORTUGUESE AND SPANISH DATA WERE USED TO TRAIN ALL MLPS. RESULTS ARE REPORTED ON THE EVALUATION SET.

Speech Recognition," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2011, pp. 5012–5015.
[5] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*. Burlington, MA, USA: Academic Press, 2006.
[6] J. Frankel and S. King, "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition," in *Proceedings of the* 9^{th} *European Conference on Speech Communication and Technology*, Lisbon, Portugual, Sep. 2005.
[7] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.
[8] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1635–1638, 2000.
[9] D. P. W. Ellis, R. Singh, and S. Sivadas, "Tandem Acoustic Modeling in Large-Vocabulary Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Salt Lake City, USA, May 2001, pp. 517–520.
[10] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRI's Conversational Speech Recognition System," in *Proceedings of the* 9^{th} *European Conference on Speech Communication and Technology*, Lisbon, Portugual, Sep. 2005, pp. 2141–2144.
[11] Ö. Çetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An Articulatory Feature-based Tandem Approach and Factored Observation Modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, USA, Apr. 2007.
[12] S. Sivadas and H. Hermansky, "On Use of Task Independent Training Data in Tandem Feature Extraction," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Montreal, Quebec, Canada, 2004, pp. I–541–4.
[13] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI Multi-Language Telephone Speech Corpus," in *Proceedings of the* 2^{nd} *International Conference of Spoken Language Processing*, Banff, Alberta, Canada, 1992, pp. 895–898.
[14] "Speech in noisy environments database." [Online]. Available: http://www.speech.sri.com/projects/spine/
[15] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, May 2006, pp. 321–324.
[16] Ö. Çetin, M. Magimai-Doss, A. Kantor, S. King, C. Bartels, J. Frankel, and K. Livescu, "Monolingual and Crosslingual Comparison of Tandem features derived from Articulatory and Phone MLPs," in *Proceedings of the* 10^{th} *biannual IEEE Workshop on Automatic Speech Recognition and Understanding*. Kyoto, Japan: IEEE, Dec. 2007.
[17] L. Toth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual Portability of MLP-Based Tandem Features — A Case Study for English and Hungarian," in *Proceedings of the* 9^{th} *International Conference of the International Speech Communciation Association*, Brisbane, Australia, Sep. 2008, pp. 2695–2698.
[18] J. Frankel, M. Magimai-Doss, S. King, K. Livescu, and Özgür Çetin, "Articulatory Feature Classifiers Trained on 2000 hours of Telephone Speech," in *Proceedings of the* 8^{th} *International Conference of Spoken Language Processing*, Antwerp, Belgium, Aug. 2007.
[19] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low-resource LVCSR Systems," in *Proceedings of the* 11^{th} *Annual Conference of the International Speech Communication Association*, Makuhari, Japan, Sep. 2010.
[20] R. Rasipuram and M. Magimai-Doss, "Integrating Articulatory Features using Kullback-Leibler Divergence based Acoustic Model for Phoneme Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011, pp. 5192–5195.
[21] Y.-R. Chao, *A Grammar of Spoken Chinese*. Berkeley: University of California Press, 1968.
[22] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Blackwell, 1996.
[23] J. W. Harris, *Spanish Phonology*. Cambridge, Massachusetts: MIT Press, 1969.
[24] S. Tsakalidis and W. Byrne, "Acoustic Training from Heterogeneous Data Sources: Experiments in Mandarin Conversational Telephone Speech Transcription," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, Mar. 2005.
[25] N. Kumar, "Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis forImproved Speech Recognition," Ph.D. dissertation, Johns Hopkins University, 1997.
[26] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.
[27] D. Johnson, D. Ellis, C. Oei, C. Wooters, and P. Faerber, "Quicknet," 2011, `www.icsi.berkeley.edu/Speech/qn.html`.
[28] L. Gillick and S. Cox, "Some Statistical Issue in the Comparison of Speech Recognition Algorithms," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Glasgow, UK, May 1989, pp. 532–535.
[29] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, pp. 31–51, Aug. 2001.
[30] P. Lal, "Cross-lingual Automatic Speech Recognition using Tandem Features," Ph.D. dissertation, University of Edinburgh, 2011.

Partha Lal holds an M.Phil. degree from the University of Cambridge and a Ph.D. from the University of Edinburgh.

Simon King (M'95–SM'08) holds M.A.(Cantab) and M.Phil. degrees from Cambridge and a Ph.D. from the University of Edinburgh. He has been with the Centre for Speech Technology Research at the University of Edinburgh since 1993, where he is now Professor of Speech Processing and the director of the centre. His interests include speech synthesis, recognition and signal processing and he has around 120 publications in these areas. He has served on the ISCA SynSIG board and currently co-organises the Blizzard Challenge. He serves on the IEEE SLTC, has served as an associate editor of IEEE Transactions on Audio, Speech and Language Processing and is a current associate editor of Computer Speech and Language.