INVESTIGATION OF MULTILINGUAL DEEP NEURAL NETWORKS FOR SPOKEN TERM DETECTION

K.M. Knill, M.J.F.Gales, S.P. Rath, P.C. Woodland, C. Zhang, S.-X. Zhang

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

The development of high-performance speech processing systems for low-resource languages is a challenging area. One approach to address the lack of resources is to make use of data from multiple languages. A popular direction in recent years is to use bottleneck features, or hybrid systems, trained on multilingual data for speech-to-text (STT) systems. This paper presents an investigation into the application of these multilingual approaches to spoken term detection. Experiments were run using the IARPA Babel limited language pack corpora ($\sim$10 hours/language) with 4 languages for initial multilingual system development and an additional held-out target language. STT gains achieved through using multilingual bottleneck features in a Tandem configuration are shown to also apply to keyword search (KWS). Further improvements in both STT and KWS were observed by incorporating language questions into the Tandem GMM-HMM decision trees for the training set languages. Adapted hybrid systems performed slightly worse on average than the adapted Tandem systems. A language independent acoustic model test on the target language showed that retraining or adapting of the acoustic models to the target language is currently minimally needed to achieve reasonable performance.

Index Terms— Multilingual, speech recognition, spoken term detection, keyword search, neural networks

1. INTRODUCTION

In recent years there has been significant interest in the area of multilingual speech technologies in particular for low resource languages. This trend is set to continue with project funding, such as the IARPA Babel Program the stated aim of which is *developing agile and robust speech recognition technology that can be rapidly applied to any human language in order to provide effective search capability for analysts to efficiently process massive amounts of real-world recorded speech* [1]. In addition the development of techniques such as Tandem systems [2] and deep neural networks (DNN) [3] offer alternative research approaches for developing multilingual speech-to-text (STT) systems. From a commercial perspective, multilingual systems would allow faster and cheaper deployment of speech systems. Improvements in multilingual speech systems may also help us better understand the commonalities and differences across languages.

The majority of work to date has focused on multilingual STT systems (see [4] for a recent review). However for many scenarios STT is simply part of a pipeline for a final speech processing application. This paper investigates these systems from an application performance perspective. In particular this paper will examine keyword search (KWS) performance for spoken term detection (STD). Improvements in STT performance do not necessarily help in KWS. Most of the systems investigated in this paper are multilingual i.e. speech data from multiple (4) languages were used to train a single system and then this is applied to recognition in one of the training languages or adapted to a target language not used in the initial training. In addition the performance of these systems on a held-out language will be investigated. This form of language-independent system is incredibly challenging to develop, but would potentially allow, for example, STD systems to be rapidly brought up in any human language, even when no acoustic training data is available.

Recent STT work in this area has been dominated by systems which use an MLP as a feature extractor, trained to produce phoneme posterior probabilities – both probabilistic features [2] and bottleneck configurations [5] have been investigated and reported in literature – using either the traditional GMM-HMM [6, 7, 8, 9] or the "hybrid" HMM-MLP [10, 11, 12, 13, 14] as the back-end classifier. The use of MLP features for cross-lingual/multilingual speech recognition can be traced back to [15], which showed that concatenating cepstral (MFCC) features with the probabilistic features [2] produced by an MLP trained on a different language (English) helped to improve the recognition performance of the target languages (namely, Mandarin and Arabic).

A common issue that arises is the question of the appropriate phone set to use. Traditionally researchers have either used universal phone sets [16, 17, 7, 6] or mapped phones between languages [18, 19]. Universal phone sets attempt to exploit the commonalities between the speech sounds across languages. They enable data with the same characteristics from multiple language training sets to be pooled to, hopefully, deliver more robust models. Mapping phone sets can be very difficult when the two languages are far apart e.g. Cantonese and English. Recently it has become common to use language dependent phone sets with associated language dependent MLP output layers e.g. [20, 13, 9, 14, 12]. That has the benefit of simply requiring the output layer to be trained for a new language. Keeping the phone sets and output layers separate, however, does not allow full exploitation of the similarities across languages. It also does not allow recognition with no training data. For this work, similarly to [17, 7, 6, 21], a universal phone set was therefore adopted.

For MLP based systems there are a number of open questions about the best MLP/HMM configuration and MLP training proce-

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012.The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

dure to be used in multilingual modelling including: how to train the MLP; which parts of the system should be unilingual/multilingual; how can the information from other languages best be exploited; is Tandem GMM-HMM or hybrid DNN-HMM better; do multilingual models help over simply multilingual Tandem features. This paper presents an investigation into a number of these for the Babel STD task. This consists of keyword and key-phrase searching on conversational telephone data recorded over a range of different conditions including mobile phones in cars. It is therefore a far more challenging data set than other multilingual corpora such as GlobalPhone [22] but very interesting for STD. In these experiments initial multilingual systems were trained on four diverse languages (Cantonese, Pashto, Turkish and Tagalog) with Vietnamese used as the target and held-out language. Only 10 hours of training data were made available for each language.

Section 2 presents the spoken term detection task. The speech-to-text and keyword search systems used are described in Sections 3 and 4 respectively, with specifics relating to the multilingual STT systems in Section 5. Experimental setup and results are given in Sections 6 and 7, followed by conclusions.

2. TASK DESCRIPTION

The present work addresses the STD task defined by NIST for the 2006 STD Evaluation with some modifications introduced by IARPA's Babel program [1]. The task consists of finding all the exact matches of a specific query in a given corpus of speech data. A query is a textual phrase containing one or several terms. In this work the system components and word indices are frozen before the queries are provided. KWS performance is measured in terms of the maximum term weighted value (MTWV), which is the best term-weighted value [23] achievable given a *post-hoc* choice of detection threshold.

The Babel corpora consists of transcribed telephone conversations in a range of languages. There are two database configurations per language: full language pack (FLP) with about 100-200 hours of transcribed audio training data ($\sim$60-80 hours speech); limited language pack (LLP) with $\sim$10 hours of transcribed audio data. The training data is a mix of conversational and scripted speech. The FLP and LLP share the same development set of 10 hours of conversational speech data[1]. A phone set and phonetic lexicon covering the training data are also supplied. Initially four development languages were supplied - Cantonese, Pashto, Turkish and Tagalog - and a fifth language - Vietnamese - for a surprise language evaluation.

The aim of this work is to build STT systems to optimise KWS on the low resource Babel LLP data sets. The development languages are used for training the initial multilingual systems and the surprise language as the target and held-out language. Following the Babel primary condition, the acoustic and language models were trained solely on the LLP audio data and associated transcripts. The development language LLP data sets were combined for the multilingual systems (violating the primary condition).

3. SPEECH-TO-TEXT SYSTEMS

3.1. Phone set and lexicons

The default Babel phone sets are based on X-SAMPA but some inconsistencies were observed between languages. The relevant phones[2] were mapped to 'standard' X-SAMPA. Table 1 shows the overlap across the five Babel languages. 11 phones are common to all languages (which were mostly plosives and fricatives).

Language	Id	Unique	101	104	105	106	107
Cantonese	101	14	**37**	13	15	17	13
Pashto	104	12		**44**	20	28	20
Turkish	105	10			**42**	26	19
Tagalog	106	11				**48**	22
Vietnamese	107	15					**41**

Table 1. *X-SAMPA phone set overlap across Babel languages. Vietnamese diphthongs and triphthongs split into constituent phones.*

3.2. GMM-HMMs training

The GMM-HMM acoustic models (AMs) were trained using the procedure described in [24]. Unilingual and multilingual AMs were each built from a flat start. Cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN) were applied to conversational sides. Speaker adaptive training (SAT) was applied using global constrained maximum likelihood linear regression (CMLLR) transforms for an entire side, followed by a discriminative transformation of the feature space (fMPE) [25] if desired.

The standard GMM-HMMs used PLP plus pitch features. Pitch was adopted for all languages in these experiments as initial experiments showed that tonal languages benefitted from the use of pitch of features in GMM-HMMs and there was no loss in performance for non-tonal languages. For the Tandem GMM-HMMs, bottleneck features were appended to the PLP plus pitch features to form the Tandem feature vector.

For the state tying of GMMs [26], the standard approach is to define the roots of the decision trees to be separate for each state of each phone. With highly limited training data this can result in some phones having insufficient data to generate many, or even any, context-dependent models. To mitigate this issue decision trees with state position roots (thus non-phone specific) were constructed and the set of questions expanded to include center context questions. Initial experiments were run on the Babel corpora with decision trees with state position roots. For large quantities of training data minimal difference in performance was observed between the two tree types. However, when less data was present, as in the 10 hour Babel LLP training sets, the state position root tree outperformed the phone root tree. In addition tying at the state position root allows the simple combination of data from multiple languages. State position root trees were therefore used for all experiments reported in this paper.

The decision tree questions were automatically derived from a SAMPA-based phone attribute file and the lexicon. The former listed all the attributes, such as vowel or nasalised, associated with each phone in the training set. Further attributes were derived from the lexicon, such as word boundaries. Phone, attribute, tone and word boundary questions were asked in these experiments.

3.3. Bottleneck features and hybrid MLP training

For this paper bottleneck features in a Tandem configuration and hybrid systems have been investigated. The MLPs in each case were deep neural network (DNN) multi-layer perceptrons (MLP) [27]. The bottleneck MLPs used a narrow hidden layer (the bottleneck

[1] There are also evaluation data sets but these were not used for the experiments reported here.

[2] In Cantonese, Pashto and Turkish.

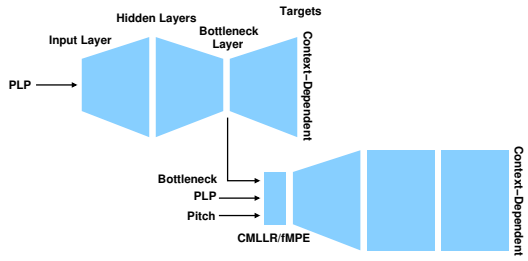

Fig. 1. *Stacked hybrid MLP architecture [27].*

layer) prior to the output layer, as shown in the top network in Figure 1. Following MLP training the bottleneck features were appended to PLP and pitch features to form the Tandem feature vector. Prior to recognition, Tandem GMM-HMMs must be trained based on the new Tandem features as described in Section 3.2. In the hybrid MLPs the hidden layers were all the same size, as shown in the bottom network in Figure 1. For recognition, the hybrid system directly uses the DNN to produce likelihoods for an HMM-based recognition system, replacing the GMMs. No further training is therefore required. For this work, the hybrid DNN was trained in a "stacked" configuration as shown in Figure 1. Following data normalistion the Tandem features were used as training inputs for the hybrid (or 2nd Tandem) MLP. The data was normalised through feature space projections such as heteroscedastic LDA (HLDA), CMLLR and fMPE.

The alignment of the context-dependent output states to the training data frames (required for the supervised MLP training) was derived from the PLP+pitch GMM-HMM systems. This alignment was left fixed during training. Sigmoid and softmax functions were used for the nonlinearities in the hidden and output layers, respectively. All the training data was presented to the network and randomised at the frame level. The objective function used for optimization was the cross-entropy criterion. The parameters of the network were initialised using a discriminative layer-by-layer pre-training algorithm [28]. This was followed by fine tuning of the full network using the error back propagation algorithm.

4. KEYWORD SEARCH SYSTEM

The KWS system was based on using weighted finite state transducers (WFSTs) to represent both the recognition lattice in an pre-processing indexing phase and also the query key-words/key-phrases. The search was performed at the word level for in-vocabulary search terms. For out-of-vocabulary (OOV) search items the recognition lattice was converted to phonetic form and the phonetic form of the query was expanded with a transducer that models phone-to-phone confusions. The KWS search returned approximate posterior probabilities of each search term occurring at a particular point in time. Before MTWV scoring these values were further normalised using a sum-to-one approach which ensures that the sum over the test set of the the scores for each keyword sum to unity. More details of the approach are given in [29].

The arc costs included both the language model and scaled acoustic model log likelihoods. For OOV terms, experiments showed excluding the word level language model scores from the arc costs could improve KWS performance.

5. MULTILINGUAL SYSTEMS

5.1. Speech-to-text

A common X-SAMPA phone set was adopted for the multilingual systems here. The phone set covered the 4 multilingual training languages (Cantonese, Pashto, Turkish and Tagalog) (3.1).

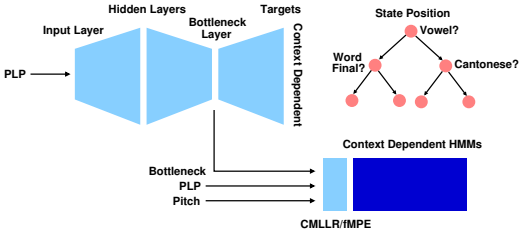

Fig. 2. *Tandem multilingual structure.*

Figure 2 shows the Tandem multilingual structure. The context dependent output targets were created with the multilingual phone set. This meant the output layer (and all other layers) was common to all the languages. All the multilingual training data was presented to the DNN at the same time and joint optimisation took place across all the multilingual training languages. The order of presentation of data to the MLP was randomised at the frame level across all the languages [13, 9]. Unlike the language dependent output layer systems, fine tuning did not require any modifications to the standard back propagation algorithm as all data samples contributed to optimising all network parameters.

The state position root decision trees from multilingual GMM-HMMs with PLP plus pitch input features were used to provide the MLP output targets and tie the HMM states. As in Section 3.2, phone, attribute, tone and word boundary questions were asked, derived from an attribute file across all the training languages and all the training lexicons. All the multilingual training data was jointly used to optimise the decision trees. The attribute questions performed a similar role to the articulatory features in e.g. [30] but without the explicit attribute detectors of these systems. The same phone symbol could correspond to different realisations of the associated phone across languages, especially when co-articulations are taken into account. With the joint training of the decision trees, questions could also be asked about the language the state came from. This is illustrated in the decision tree in Figure 2.

Unlike the phone sets, the lexicons were not merged across languages, both in training and at recognition. In the Babel scenario the speaker's language is assumed to be known and any code switching[3] is already encapsulated in that language's lexicon. Phonetic alignments for all the multilingual AM trainings were generated using language specific lexicons. This avoided an explosion in cross-word contexts and incorrect pronunciations being learned for words that appear in more than one language. A multilingual language model (LM) built from the transcriptions from all 4 training languages was used for alignments. At recognition time a language specific LM, trained on the transcriptions from the language under test was used. The same LM was used for unilingual experiments.

[3] Any code switching takes the form of imported words rather than long phrases in another language.

5.2. Keyword search

Keyword search was carried out on each language separately. No changes were required to the unilingual set-up.

6. EXPERIMENTAL SETUP

Release B of the Babel LLP corpora described in Section 2 was used for these experiments. Multilingual MLPs and AMs were trained on Cantonese (babel101b-v0.4c), Pashto (babel104b-v0.4bY), Turkish (babel105b-v0.4) and Tagalog (babel107b-v0.7) data. Vietnamese (babel107b-v0.7) was used as a target or held-out language. In each case, the STT system configurations used for these experiments were designed to optimise KWS performance. They are therefore not optimal in terms of raw STT performance, for example currently the best KWS MTWV is achieved with a bigram LM rather than a trigram LM which yields a lower recognition error rate.

6.1. STT systems

The STT systems were trained and decoded using HTK [31]. SAT, Minimum Phone Error (MPE) discriminative training and fMPE features were applied in training and CMLLR and maximum likelihood linear regression (MLLR) were applied at decoding. For the GMM-HMM systems, 1000 tied state AMs were trained for unilingual systems and 3000 for multilingual systems. Each state had an average of 16 Gaussian components with 32 components for silence.

All decision tree roots were state position based. Phonetic, attribute and tonal questions were asked in each tree. Questions relating to the language of the training/development data were also asked for a subset of the multilingual AMs. For the hybrid multilingual system 3 silence targets were used corresponding to a 3 emitting state HMM. No changes were made to the supplied pronunciation lexicons except for mapping of a small subset of Cantonese, Pashto and Turkish phones to a 'standard' X-SAMPA phone set.

The base GMM-HMMs were trained with PLP plus pitch features. 52-dimensional PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features were projected down to 39 by HLDA. Pitch+Δ+$\Delta\Delta$ features were appended. For the Tandem systems 26 bottleneck (BN) features were also appended.

The DNN MLPs were trained on an extended version of ICSI's QuickNet [32] software. The key changes made were to support multiple hidden layers and layer-by-layer pre-training. For the non-stacked BN features, the input feature vector had 468 dimensions. This was produced by splicing[4] the 52-dimensional PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features. The MLP targets were context dependent states, tied with decision trees generated in the base GMM-HMM ML training. The unilingual BN MLPs had 3 hidden layers plus the BN layer in configuration 468-1000-2x500-26-403 and the multilingual MLPs 4 hidden layers plus the BN layer 468-4x1000-26-3000. For the stacked hybrid and BN MLPs (Figure 1) the 26-dimensional BN features are de-correlated using a global semi-tied covariance transform.The 26 decorrelated BN features are appended to 39-dimensional HLDA normalised PLP features and 3-dimensional pitch (pitch+Δ+$\Delta\Delta$) features to form a 68-dimensional feature vector per frame. As before, 9 frames are spliced together to form a 612-dimensional input vector. The multilingual stacked BN MLP had the same configuration as before. In the hybrid MLP the BN layer was replaced by another hidden layer to give a configuration 612-5x1000-3000. Note, no unilingual stacked MLP systems were built.

Word based bigram language models were used in decoding. They were trained on the LLP transcripts with modified Kneser-Ney smoothing using the SRI LM toolkit [33]. The language specific training set lexicons were used. At decoding time the language was assumed known and the language specific lexicon and LM applied. The decoding parameters were kept fixed across all systems.

[4] i.e., concatenating the current frame with a certain number of frames in the left and right contexts, for example, ± 4.

6.2. Keyword search

The IBM KWS system was run [29, 34] without the system combination component. As for decoding, each language was treated separately with KWS parameters fixed across languages and AMs. The OOV lattices were pruned to reduce the search space. Unless otherwise noted, the language model was ignored in the OOV search (i.e. LM weight set to 0). This was essential for the Hybrid systems to eliminate dynamic range mismatches but also beneficial for the Tandem systems.

7. EXPERIMENTAL RESULTS

All STT results are Viterbi bigram results. Percentage token error rate (%TER) is presented where a token is a character for Cantonese, a word for Pashto, Turkish and Tagalog, and a Vietnamese syllable or foreign word for Vietnamese. In the tables † indicates this language was not used for the multilingual MLP, AM and/or decision tree training.

7.1. Multilingual MLP features

Unilingual systems were built with Tandem features comprised of PLP, pitch and matched language or multilingual MLP bottleneck features. Table 2 shows the STT and KWS performance for discriminative speaker adaptive trained systems with fMPE. As can be seen the multilingual MLP features reduce the error rate for all the languages, including Vietnamese which was not included in the multilingual BN features training. This concurs with previously reported results. For all the languages in the multilingual training set, the improvements in STT are carried through to KWS with gains in the MTWV score being observed. However, for Vietnamese the KWS performance degrades with the multilingual MLP feature system.

Language	Id	%TER		MTWV	
		uni	multi	uni	multi
Cantonese	L101	65.7	63.9	0.3295	0.3445
Pashto	L104	68.8	67.7	0.1494	0.1621
Turkish	L105	68.5	67.7	0.3479	0.3630
Tagalog	L106	66.6	65.4	0.2607	0.2737
Vietnamese†	L107	71.3	71.1	0.1634	0.1576

Table 2. *Comparison of matched language (uni) and multilingual (multi) MLP features.*

7.2. Multilingual features and acoustic models

The ability to further exploit commonalities between the sounds of different languages to share data to train more robust systems was investigated through combining multilingual MLP features with multilingual acoustic models. Experiments considered whether it is beneficial to ask language questions (LQ) in the MLP and AM decision trees or to have language independent (LI) trees.

Language	Id	multi MLP/ uni AM	multilingual: MLP/AM		
			LI/LI	LI/LQ	LQ/LQ
Cantonese	L101	63.9	63.8	63.6	63.9
Pashto	L104	67.7	68.3	67.8	68.0
Turkish	L105	67.7	68.2	67.7	68.0
Tagalog	L106	65.4	67.6	66.4	66.4
Vietnamese†	L107	71.1	91.4	91.8	91.9

Table 3. *% TER STT performance of unilingual and multilingual AMs with multilingual MLP features.*

Language	Id	multi MLP/ uni AM	multilingual MLP/AM		
			LI/LI	LI/LQ	LQ/LQ
Cantonese	L101	0.3309	0.3395	0.3449	0.3275
Pashto	L104	0.1606	0.1471	0.1528	0.1459
Turkish	L105	0.3595	0.3555	0.3685	0.3676
Tagalog	L106	0.2612	0.2560	0.2722	0.2589
Vietnamese†	L107	0.1440	0.0250	0.0011	0.0004

Table 4. *MTWV KWS performance of unilingual and multilingual AMs with multilingual MLP features. KWS OOV LM was non-zero.*

The tied state outputs for the multilingual MLP in Table 2 did not ask any language specific questions (i.e. language independent (LI)). As seen in Tables 3 and 4 training multilingual AMs using LI decision trees (LI/LI) results in a degradation in both STT and KWS performance across all languages except for Cantonese KWS. For the multilingual training languages, introducing language questions (LQ) into the AMs achieves equivalent STT performance (LI/LQ) to the multilingual MLP features-only case (except for Tagalog where it is 1% worse). Gains are also seen for all languages but Pashto in KWS. Additionally asking language questions in the MLP features trees (LQ/LQ) produces slightly worse STT and lower KWS performance than the LI/LQ system but error rates are lower than for LI/LI. However, the KWS performance compared to LI/LI is mixed.

Stacked LQ/LQ systems were trained to compare Tandem GMM-HMM and Hybrid DNN-HMM modelling. The input features in each case were taken from the bottleneck output of the LQ MLP above. Table 5 presents experiments for the multilingual training set languages. The Tandem system out-performed the Hybrid except for Turkish. Further optimisation of both systems is required e.g. due to time constraints phone level re-alignment of the training data was not performed.

Language	Id	%TER		MTWV	
		Tandem	Hybrid	Tandem	Hybrid
Cantonese	L101	64.3	65.6	0.3301	0.3174
Pashto	L104	68.4	69.3	0.1543	0.1297
Turkish	L105	69.3	67.7	0.3493	0.3535
Tagalog	L106	67.9	68.5	0.2946	0.2624

Table 5. *%TER STT performance of multilingual Tandem and Hybrid LQ/LQ MLP/AM systems.*

7.3. Analysis of language independent models

Tables 3 and 4 show that for the non-training set language, Vietnamese, the performance of the multilingual AM systems was very poor. There are two possible sources for this poor performance. The first is that the acoustic data associated with Vietnamese, even if related to the same X-SAMPA symbol, may be mismatched to that of the multi-lingual training languages. Second, the decision tree used for tying the acoustic model may not be appropriate for Vietnamese. For example there are 15 phones in Vietnamese that will not be seen in decision tree construction, so the leaves will be determined by the broad class questions.

To investigate these effects unilingual systems were constructed using the multilingual decision tree. To ensure robust parameter estimation, CMLLR, MLLR and MAP adaptation approaches were used and for simplicity a basic PLP+pitch, ML trained system was used. From Table 6 the TER performance using the multilingual tree (multi) is comparable to the unilingual tree. It is interesting that Cantonese, with more unique phones than the other training languages, has the least non-zero multilingual tree leaves. However the TER performance of the Vietnamese system is poor, approximately mid way between the Vietnamese single language system and the multilingual Vietnamese performance. This degradation in performance must be due to the tree, as the acoustic training data is the same. As can be seen in Figure 3 the Vietnamese data is concentrated in far fewer leaf nodes, resulting in poor discrimination.

Language	Id	# states		%TER	
		uni	multi	uni	multi
Cantonese	L101	1038	1119	75.7	75.4
Pashto	L104	1038	1623	76.7	76.1
Turkish	L105	1048	1698	77.1	76.2
Tagalog	L106	1033	1698	74.8	74.5
Vietnamese†	L107	1028	1169	78.5	84.9

Table 6. *Use of unilingual (uni) and multilingual (multi) decision trees in a PLP+pitch ML system. # states indicates the number of stated (uni) and used (non-zero count) states of the 2985 multi-state system. No language questions were asked in the multilingual trees.*

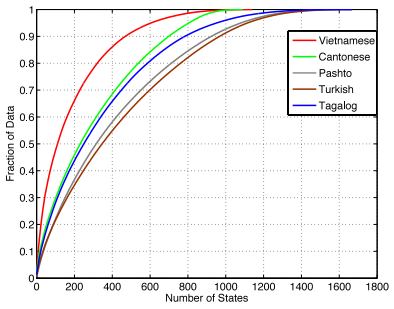

Fig. 3. *Cumulative PDF of state coverage.*

8. CONCLUSIONS AND DISCUSSION

The development of high performance speech processing systems for low resource languages is challenging. This paper has considered spoken term detection (STD) of conversational telephone speech data. Experiments were run using the IARPA Babel limited language pack corpora (∼10 hours/language) with 4 languages for initial multilingual system development and an additional held-out or target language, Vietnamese. Multilingual bottleneck features in a Tandem configuration yielded gains over unilingual systems for both speech-to-text (STT) and STD. Further improvements were observed in both STT and STD by training multilingual Tandem GMM-HMM acoustic models with language questions incorporated into the GMM decision trees. Adapted Hybrid systems performed slightly worse on average than the adapted Tandem systems.

A language independent acoustic model test on the target language showed that retraining or adapting of the acoustic models to the target language is currently essential. It was seen that approximately half the degradation of the multilingual AM systems applied to Vietnamese is due to the decision trees. The ability to handle both decision tree and X-SAMPA label acoustic data mismatches is an important aspect of language independent work and will be considered further in the future.

The best performing Cantonese system achieved 63.6% CER and 0.3449 MTWV. By contrast an equivalent Cantonese unilingual system trained on the full language pack with approximately 8x as much data achieved 46.4% CER and 0.5469 MTWV. As the Babel program goes forward the planned increase in the quantity of resources and languages should facilitate the development of multilingual systems in some configuration.

9. ACKNOWLEDGEMENTS

The authors are grateful to IBM Research's Lorelei Babel team for the KWS system.

10. REFERENCES

[1] Mary Harper, "IARPA Babel Program," http://www.iarpa.gov/Programs/ia/Babel/babel.html.

[2] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," in *Proc. ICASSP*, 2000.

[3] Y Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, pp. 1–127, 2009.

[4] H. Bourlard et al., "Current trends in multilingual speech processing," *Sadhana*, vol. 36, pp. 885–915, 2011.

[5] Frantisek Grezl et al., "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. ICASSP*, 2007.

[6] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Cross-lingual and multi-stream posterior features for low-resource LVCSR systems," in *Proc. Interspeech*, 2010.

[7] David Imseng, Hervé Bourlard, and Mathew Magimai.-Doss, "Towards mixed language speech recognition systems," in *Proc. Interspeech*, 2010.

[8] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP*, 2012.

[9] Zoltán Tüske et al., "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Proc. ICASSP*, 2013.

[10] H. A. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[11] Stefano Scanzio et al., "On the use of a multilingual neural network front-end," in *Proc. Interspeech*, 2008.

[12] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals, "Multilingual training of deep neural networks," in *Proc. ICASSP*, 2013.

[13] Jui-Ting Huang et al., "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proc. ICASSP*, 2013.

[14] Georg Heigold et al., "Multilingual acoustic models using distributed deep neural networks," in *Proc. ICASSP*, 2013.

[15] Andreas Stolcke et al., "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006.

[16] Tanja Schultz and Alex Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech*, 1997.

[17] Stephane Dupont et al., "Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents," in *Proc. ASRU*, 2005.

[18] W. Byrne et al., "Towards language independent acoustic modeling," in *Proc. ICASSP*, 2000.

[19] Khe Chai Sim and Haizhou Li, "Robust phone set mapping using decision tree clustering for cross-lingual phone recognition," in *Proc. ICASSP*, 2008.

[20] Karel Veselý et al., "The language-independent bottleneck features," in *Proc. SLT*, 2012.

[21] Frantisek Grezl, Martin Karafiat, and Milos Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011.

[22] T. Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe Univ.," in *Proc. ICSLP*, 2002.

[23] J. G. Fiscus et al., "Results of the 2006 Spoken Term Detection Evaluation," in *Proc. ACM SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007.

[24] J. Park et al., "The Efficient Incorporation of MLP Features into Automatic Speech Recognition Systems," *Computer Speech and Language*, vol. 25, pp. 519–534, 2010.

[25] Daniel Povey et al., "fMPE: Discriminatively trained features for speech recognition," in *Proc. ICASSP*, 2005.

[26] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings ARPA Workshop on Human Language Technology*, 1994, pp. 307–312.

[27] G. Hinton, L. Deng, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[28] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, Dec 2011.

[29] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. ICASSP*, 2013.

[30] S.M. Siniscalchi et al., "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887, 2012.

[31] S. J. Young et al., *The HTK Book (for HTK version 3.4)*, Cambridge University, 2006.

[32] David Johnson et al., "QuickNet," http://www1.icsi.berkeley.edu/Speech/qn.html.

[33] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, 2002.

[34] B. Kingsbury et al., "A high-performance Cantonese keyword search system," in *Proc. ICASSP*, 2013.