SPEECH RECOGNITION AND KEYWORD SPOTTING FOR LOW RESOURCE LANGUAGES: BABEL PROJECT RESEARCH AT CUED

Mark J. F. Gales, Kate M. Knill, Anton Ragni and Shakti P. Rath

Cambridge University Engineering Department
Trumpington Street, Cambridge, CB2 1PZ, UK
{mjfg,kate.knill,ar527,shakti.rath}@eng.cam.ac.uk

ABSTRACT

Recently there has been increased interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) systems for low resource languages. One of the driving forces for this research direction is the IARPA Babel project. This paper describes some of the research funded by this project at Cambridge University, as part of the Lorelei team co-ordinated by IBM. A range of topics are discussed including: deep neural network based acoustic models; data augmentation; and zero acoustic model resource systems. Performance for all approaches is evaluated using the Limited (approximately 10 hours) and/or Full (approximately 80 hours) language packs distributed by IARPA. Both KWS and ASR performance figures are given. Though absolute performance varies from language to language, and keyword list, the approaches described show consistent trends over the languages investigated to date. Using comparable systems over the five Option Period 1 languages indicates a strong correlation between ASR performance and KWS performance.

Index Terms: keyword spotting, deep neural network, low-resource languages, multi-lingual systems.

1. INTRODUCTION

In recent years there has been an increasing interest in Automatic Speech Recognition (ASR) and Key Word Spotting (KWS) for low resource languages. One of the driving forces for this research direction is the IARPA Babel project [1]. The aim of the project is to develop robust KWS and ASR technologies that can be rapidly applied to any human language. To enable both rapid development of systems, and performance on languages for which there has traditionally been little research, to be evaluated the program has focused systems built using *limited* quantities of data. This paper gives an overview of the research undertaken at the Cambridge University Engineering Department (CUED) as part of the Lorelei team led by IBM. Three main research areas will be discussed: deep neural network acoustic models; data augmentation; and zero resource acoustic models. The performance of these approaches is evaluated on data released under the Babel program. Two *language packs* are released for each language: a full language pack (FLP) comprising about 80 hours of transcribed audio data; and a limited language pack (LLP) comprising about 10 hours of transcribed audio data. For the LLPs (and some of the FLPs) additional untranscribed audio data is also available.

Speech recognition systems using neural networks have had a long history [2]. Recently there has been renewed interest in this area with the development of deep neural network (DNN) systems [3, 4]. Currently two configurations of DNN are commonly used. The DNN can be used as feature extractor for a standard GMM-based HMM system [5, 6], this approach is referred to as *Tandem*. The second configuration, known as *Hybrid* uses the network to compute state posteriors, which are then converted into scaled likelihoods by normalising by the state priors. In this work both forms of network are trained for the FLP and LLP releases. System combination for both ASR [7, 8] and KWS [9, 10, 11] are standard approaches for improving final system performance. In addition to the performance of the individual system performance for both ASR and KWS, the impact of combining these two form of DNN system is described.

Data augmentation is a class of approaches where the effective quantity of data used to train the system is increased. In this paper these approaches are split into two distinct groups. The first is where only the data from the target language is considered. In this case it is necessary to use automated approaches to increase the amount of transcribed data. One technique is to artificially create more data with known transcriptions, for example using acoustic data perturbation [12, 13, 14] or speech synthesis [15]. Another scheme assumes that additional, untranscribed, audio data is available. In this scenario it is possible to use semi-supervised training [16, 17, 18]. An alternative class of approaches is to make use of data from other languages to increase the available data. This has become increasingly popular as DNNs are more commonly used as they are well suited to these schemes. Two approaches have been adopted. The first is to build multi-lingual bottleneck features for use in a Tandem system [6, 19, 20, 21]. Alternatively Hybrid systems, with target language specific output layers, have also been investigated [22, 6, 23]. In this work the impact of multi-lingual bottleneck features, using the configuration discussed in [21], will be discussed.

The above approaches have assumed that there is some transcribed audio data available for the target language. For some situations, it may not be possible to transcribe any audio. The final research area will be referred to as *zero acoustic model resources* [21]. Here it is assumed that there is no transcribed audio data, just a limited amount of language model data and a lexicon. The aim here

This work was supported in part 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. The authors would like to thank the Lorelei team for providing the KWS infrastructure and morphological decomposition. The authors would like to thank the following members of the CUED Speech Group who have contributed to the Babel project work and the systems described: X. Chen, X. Liu, P.C. Woodland, T. Yoshioka, and C. Zhang.

is to build a language-independent (LI) acoustic model. This model can then be used directly for ASR, or KWS. Alternatively the LI acoustic model can be used to transcribe audio data in the target language, which can then be used for training. This is effectively an unsupervised acoustic model training process [24, 25].

For all experiments the core ASR toolkit, used for acoustic feature generation, clustering, decoding and GMM-based acoustic model training, was an extended version of the HTK-3.4.1 [26] toolkit. The MLP training used an extended version (to allows deeper network configurations) of ICSI's QuickNet [27], to train both Tandem and Hybrid systems. The results given in this paper were generated at various stages of system development. Thus results are not necessarily consistent across tables, however within a table all results are comparable unless otherwise stated. The focus of this paper is acoustic modelling. The language models for all systems used the vocabulary and training data from the audio transcriptions. For all systems N-gram language models (either bigram or trigram) were used, optionally interpolated with class-based language models.

For all ASR systems in this work, the underlying context-dependent states were specified using state, rather than phone-state, roots of the decision tree. Here questions involving X-SAMPA attributes and position of the phone in the word were used for both left, right and centre context. This was found to yield additional robust to rare phones, for example the X-SAMPA phone /kx/ in Zulu. If phone/state-position decision tree roots are used for these rare phone, there is insufficient data to train any context models. Effectively these rare phones are modelled as monophones. To further improve the ability to model rare phones, diphthongs (and triphthongs) were split into their constituent parts, with additional markers added to indicate that the unit was derived from a diphthong.

2. TASK DESCRIPTION

The work reported in this paper was undertaken as part pf the IARPA Babel [1] program, which aims to foster research on speech recognition and keyword spotting for low resource languages. The Babel speech corpora covers a range of diverse languages and is distributed under two configurations for each language - the "full" language pack (FLP) and the "limited" language pack (LLP). The FLP and LLP packs consist of approximately 80 hours and 10 hours of speech for training, respectively. The data is recorded in "real-life" scenarios, such as conversational telephone speech, over a range of acoustic conditions, such as mobile phone conversation made from car. The FLP and LLP share the same development set of about 10 hours of conversational speech. The phone set and phonetic lexicon are supplied for every language pack and contains only those words occurring in the transcribed audio data for that language pack.

In the Option Period 1 (OP1) phase of the project, five languages were released for development: Assamese; Bengali; Haitian Creole; Lao; and Zulu. The ASR and KWS experiments reported in this paper are primarily conducted on these OP1 languages (both FLP and LLP), and the performance is evaluated on the development data. The official metric to measure the accuracy of the system performance has been defined to be the Maximum Term Weighted Value (MTWV), which is the best term weighted value [28] (TWV) that can be achieved over all choices of detection threshold. The TWV is defined as

$$TWV(\theta) = 1 - [P_{miss}(\theta) + \beta P_{fa}(\theta)] \tag{1}$$

where $P_{miss}(\theta)$ and $P_{fa}(\theta)$ denote the probability of miss and false alarm, respectively and β is 999.9.

Below are listed the releases of the languages that are used in the experiments. The languages marked in bold are the development language from OP1. The languages marked with a † are used as training languages for the multi-language and language-independent system in sections 5.2 and 6 respectively.

Language	**Id**	**Release**
Cantonese†	101	IARPA-babel101-v0.4c
Assamese†	102	IARPA-babel102b-v0.5a
Bengali	103	IARPA-babel103b-v0.4b
Pashto†	104	IARPA-babel104b-v0.4aY
Turkish†	105	IARPA-babel105b-v0.4
Tagalog†	106	IARPA-babel106-v0.2f
Vietnamese	107	IARPA-babel107b-v0.7
Haitian Creole	201	IARPA-babel201b-v0.2b
Lao†	203	IARPA-babel203b-v3.1a
Zulu†	206	IARPA-babel206b-v0.1e

For all the experiments there is approximately 10 hours of audio to recognise, and 2000 KW terms for the KWS task [1]. In this paper Token Error Rate (TER), rather than WER, is used when discussing ASR results. For the broad range of languages investigated under the Babel program, some languages, for example Vietnamese, do not have references at the word level. Thus TER removes the concept of word (though measured in the same fashion). The TER results quoted are based on Confusion Network (CN) decoding [29] applied to the lattices that were used for KWS unless otherwise stated.

It is worth emphasising that given the targets of the project, KWS performance of greater than 0.3, where choices of system configuration have been made they were based on KWS performance, not ASR performance.

3. KWS SYSTEM DESCRIPTION

The focus of the research at CUED is on improving ASR systems for low-resources languages. However, since the Babel program uses KWS to assess performance, this section gives a brief description of the Lorelei team KWS system, and the approaches adopted to handle KWS with low-resource languages.

The KWS system is based on a weighted finite state transducer (WFST) framework [30]. First an ASR system is used to generate word lattices. These lattices are then processed to generate the word indices for the in-vocabulary (IV) search and phonetic indices to accommodate out-of-vocabulary (OOV) search. The timing information is pushed to the output labels of the arcs of the resulting WFSTs. The arcs in the resulting WFST after the push operation can be expressed as a 5-tuple (p, i, o, w, q), where p and q indicate the start and end states, i denotes the input label, which can be a word in case of IV search or a phoneme in case of OOV search, w indicates the posterior probability associated with the input label, and finally o denotes the output label.

The IV queries are searched in the word index, whereas the OOV queries are searched in the phonetic index. More specifically, for the IV search, each query is converted to a word weighted finite state acceptor (WFSA) and a composition operation is carried out with the word index in order to retrieve the hit list for the query. Each hit list is identified by the name of the audio file, the starting time of the query, duration and the score, which is the posterior probability derived from the WSFT. On the other hand, for the OOV search, each

[1] For Vietnamese, the base period surprise language, a more limited set of about 900 keywords was used.

query is first expanded to a reasonable phonetic representation using a grapheme-to-phoneme converter, which may not give accurate pronunciation for all query terms. The resulting pronunciation is then represented as a phonetic WFSA, and a composition with the phonetic WFST is carried out to retrieve the hit lists for the OOV terms. It is possible to vary the number of phone query confusions [31]. In the simplest case no confusions are included, the *identity P2P* case. For the experiments in this work the number of confusions was in the range 100 to 50000. The IV queries that did not return hits were searched again in the phoneme index which is known as the cascaded search. Finally the IV, OOV and cascaded search hit lists are combined and sum-to-one (STO) [30] score normalisation is applied to make sure that sum of all normalised detection scores for each query is 1.0.

For some languages that are morphologically rich the number of OOV terms can become very large impacting performance. For example for the Zulu LLP 61% of the query terms were OOV, compared to 31% for the Bengali LLP. To address this problem a morphological KWS can be used [32]. Here initially IV word terms are found. Then IV morph terms are found, finally OOV morph terms are found.

KWS Process	MTWV		
	IV	OOV	Tot
Word	0.2655	0.0000	0.1033
+phone	0.2596	0.0970	0.1606
+cascade	0.2609	0.0970	0.1611
+lm0	0.2649	0.1338	0.1851
+morph	0.2615	0.2073	0.2287

Table 1: *MTWV scores comparing KWS system stages for Zulu LLP.*

The impact of the various stages for KWS are shown for the Zulu LLP in Table 1. The ASR system is the Tandem system used in section 7. The most basic search just examines the in-vocabulary terms, a word search (*Word*). To handle OOV terms phone confusions can be added (+*phone*). This handles the OOV terms, but the performance on these terms is significantly worse than for the IV terms. To improve IV performance, cascade search can be added (+*cascade*). For Zulu this gave only a small improvement, but for some languages, such as Vietnamese, large gains were observed. For the OOV search, there is not expected to be any benefit from using the language model scores from the IV terms, but these influence the scores associated with the phones. To address this, the lattices generated by the ASR system are mapped to remove the language model component (+*lm0*) for the OOV search. This improves the OOV search. Finally by using morphological decomposition (+*morph*), some of the OOV terms are mapped to be IV in terms of the morphology lattices. This further improves the OOV performance. Note, the slight variation in the IV word performance is due to shifts in the MTWV operating point.

4. DEEP NEURAL NETWORK ACOUSTIC MODELS

In common with most state-of-the-art speech recognition system, significant performance gains can be obtained using DNNs [3, 4] for limited resource systems. In this work both Hybrid and Tandem systems were constructed. The Tandem configuration used a single network with PLP and pitch features at the input. The output of this network was then used in a hybrid system yielding a stacked configuration. This is illustrated in Figure 1. All networks were initialised with layer-by-layer discriminative pre-training [4]. Further details of the two acoustic models are given below. The results for the individual systems, and combination, are given in section 7.

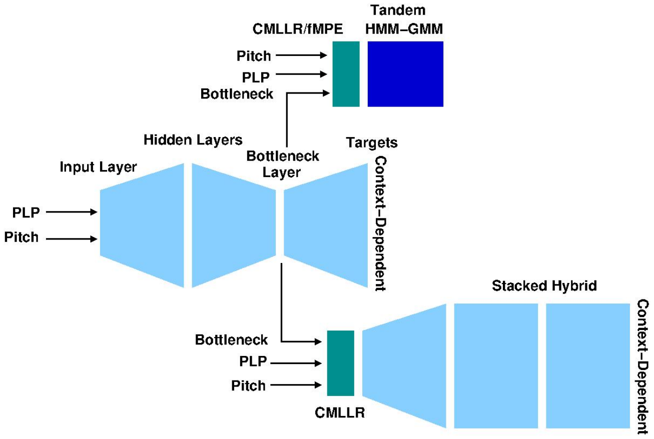

Fig. 1. Tandem and Stacked Hybrid systems

4.1. Tandem System

The development of the Tandem systems was based on [33]. An MLP was trained using cross-entropy, and context dependent targets defined by a phonetic decision tree. The input to the network was 9 frames of PLP with pitch[2] appended, and delta, delta-deltas and triples added. This yields a total input vector size of 504. The network was configured to have a bottleneck layer of 26. The 26 dimensional bottleneck features were transformed using a global semi-tied covariance matrix [34] and then appended to HLDA projected PLP features (39 dimensions) and pitch with delta and delta-delta parameters. This yields a complete feature of 68 dimensions. These are the baseline features for the hybrid system below.

A speaker adaptive training (SAT) system using global constrained maximum likelihood linear regression (CMLLR) at a speaker level [35], was then constructed incorporating both Minimum Phone Error (MPE) [36] training and feature-space MPE (fMPE) [37]. The CMLLR transforms were estimated using maximum likelihood (ML) on the ML estimated acoustic models. These were then fixed and MPE and fMPE estimated using these transforms.

A multi-pass decoding and adaptation process was used for all experiments in this paper:

1. speaker-independent (SI) decoding with a PLP-based MPE system;
2. a global CMLLR transform was estimated for each speaker using the Tandem ML-SAT model;
3. global CMLLR and MLLR transforms were estimated using the Tandem-SAT fMPE+MPE acoustic model;
4. speaker adapted decoding using the Tandem-SAT fMPE+MPE system and a bigram word-based language model;
5. lattice rescore with a class-based language model and confusion network (CN) generation.

[2] Initial experiments showed that using pitch as an input to the MLP significantly improved the performance of tonal languages such as Lao, with smaller improvements for non-tonal languages

The configuration of the Tandem systems for the two language packs was tuned to the quantities of data available.

Full Language Pack: the target number of states was set at about 6000 for both the MLP and HMM system. Five hidden layers, including the bottleneck layer, were used. The network configuration was (including input and target layers): 504x1000^{4}x26x6000.

Limited Language Pack: the target number of states was set at about 1000 for both the MLP and HMM system. Four hidden layers, including the bottleneck layer, were used. The network configuration was (including input and target layers): 504x1000x500^{2}x26x1000.

4.2. Stacked Hybrid System

As shown in Figure 1 the hybrid system was trained in a stacked fashion. First the bottleneck MLP for the Tandem system was constructed. Using the ML Tandem-SAT system, and the ML-estimated CMLLR transforms these features were transformed to be speaker specific. Again 9 vectors, each of 68-dimensions, were then stacked together to yield a total input vector the network of 612 features. Speaker adapted decoding with the Hybrid system, used the transforms generated at stage (2) of the Tandem decoding process to transform the features to be speaker specific. Hybrid decoding with a bigram language model, was then followed by the lattice rescoring and CN generation as in step (5) of the Tandem decoding.

The configuration of the Hybrid systems for the two language packs was tuned to the quantities of data available.

Full Language Pack: the target number of states was set at about 6000 for the MLP. Five hidden layers were used, the network configuration was (including input and target layers): 612x1000^{5}x6000.

Limited Language Pack: the default target number of states was set at about 1000 for the MLP. Four hidden layers were used as the default network configuration, (including input and target layers): 612x1000x500^{3}x1000.

5. DATA AUGMENTATION

When there is very limited training data, approaches that increase the quantity of training data available have been proposed. In this paper the approaches are split into two broad categories. The first is *data and transcription generation*, where audio data is either artificially generated [12, 13, 14], or additional transcriptions generated in a semi-supervised fashion [16, 17, 18]. An alternative approach is to make use of data from other languages [6, 19, 20, 21], *multi-language resources*. The form of these approaches examined at CUED, and preliminary results, are discussed in the next two sections.

5.1. Data and Transcription Generation

Two forms of within language data augmentation were investigated: vocal tract length perturbation; and semi-supervised training. For vocal tract length perturbation (VTLP) [12], 8 warp factors were randomly selected in the range 0.8 to 1.2. The data was then perturbed by the selected warp factor. This increased the quantity of training data to be approximately the same as the FLP. For the semi-supervised training, the LLP system was used to recognise the untranscribed data. Confidence based-selection was then used to select about 50% of the data with no transcriptions. This data was then added to the supervised LLP data and used to train a system. Finally discriminative MAP of the semi-supervised system to the (supervised) LLP data was performed. It is also possible to combine these two approaches to further increase the quantity of data. For further information about the experimental configuration, and additional results, see [38].

Data Augmentation		TER	MTWV
HMM	BN-MLP	(%)	Tot
—	—	78.4	0.1362
—	`vtlp`	77.1	0.1496
—	`semi`	77.7	0.1468
—	`semi+vtlp`	76.7	0.1446
`semi`	`semi`	76.9	0.1490
`semi`	`semi+vtlp`	76.1	0.1441
`semi+vtlp`	`semi+vtlp`	76.1	0.1454

Table 2: *%TER (no CN) and MTWV for Zulu (206) LLP performance using data augmentation approaches - semi-supervised training (*`semi`*) and Vocal Tract Length Perturbation (*`vtlp`*) .*

Table 2 shows the impact of the two data augmentation approaches on the LLP Zulu system. The acoustic model configuration used for these experiments was Tandem-SAT. First considering the ASR performance. Data augmentation for training the BN MLP yielded performance gains. At these high TERs, 78.4%, the gains from semi-supervised training were smaller than those from VTLP. However combining the two approaches yielded additional gains. Applying semi-supervised approaches to also training the acoustic model HMM gave further gains. However for this task combining both semi-supervised training and VTLP to train the HMM did not yield gains over just using semi-supervised training.

The performance on KWS was not as consistent as the ASR performance. Again using any form of data augmentation yielded a performance gain. However the best performing system used only VTLP data augmentation.

5.2. Multi-Language Resources

Rather than artificially generating data or transcriptions, it is also possible to make use of data from other languages. In this work only the Tandem configuration was used and the MLPs used to extract the bottleneck features were trained on multi-lingual data, the LLPs from seven training languages described in section 2 were used for this purpose.

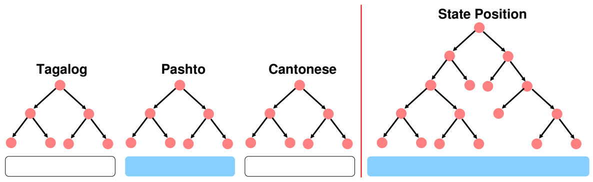

Fig. 2. Context Dependent MLP Targets

When training these multi-language networks there are two forms of targets that can be used, illustrated in Figure 2. The option on the left is where the MLP targets are context-dependent (though context independent targets can be used) and language-specific, for example [20]. Thus the normalisation summation (shown in blue) acts on single language. This approach is useful as there is no requirements for consistency in the phonetic labels from the individual languages. The network will attempt to generate a projection layer that maximises the average within-language discrimination over the training languages.

The second approach, and the one adapted in these experiments, is to have a single decision tree that covers all languages [21]. This requires that there is a consistent phonetic labelling scheme for all languages, which is the case for Babel where X-SAMPA is used. Now the normalisation term is over all context dependent targets. Thus the projection layer is optimised to discriminate between all context-dependent labels. The rationale for this approach is that when the network is to be applied to an unseen language, the phoneset and important phonetic context structure is unknown when the MLP are being trained. By maximising discrimination over all possible context dependent phones, it is hoped that any unseen phonetic contexts will also be easily separated.

Language	Id	BN MLP	TER (%)	MTWV Tot
Assamese†	102	UL	68.0	0.2132
		ML	66.4	0.2382
Zulu†	206	UL	75.8	0.1274
		ML	74.4	0.1396
Bengali*	103	UL	68.6	0.2392
		ML	67.0	0.2551
Haitian Creole	201	UL	62.2	0.4054
		ML	61.1	0.4266
Vietnamese	107	UL	69.3	0.1851
		ML	68.2	0.1908

Table 3: *%TER and MTWV LLP performance using Target Language BN features (UL) or Multi-Language BN (ML). † indicates that the language was seen in the ML BN training data, ⋆ indicates "identity" phone-mapping OOV search.*

Table 5.2 shows the ASR and KWS results for languages seen in the training data (Assamese and Zulu) and languages not seen in the training data (Bengali, Haitian Creole and Vietnamese). The combination of the seven languages yields comparable quantities of data to a single language FLP. Thus for these experiments the FLP BN MLP configuration, $504 \times 1000^4 \times 26 \times 6000$, was used. For all languages, even those that are not represented in the training data, performance gains are obtained for both ASR and KWS.

The systems shown above have only included a limited amount of data from each language. Additional gains have been obtained by including data from the FLPs, and also "fine-tuning" to the target language [39].

6. "ZERO ACOUSTIC MODEL RESOURCE" SYSTEMS

Using phonetic labels from X-SAMPA, for example, it is possible to generate lexicons that have the same set of labels for all languages. However even if the X-SAMPA label is consistent across two languages the realisation of that phone may vary significantly between languages. This limitation impacts the ability to generate language-independent (LI) acoustic models. Despite this, it is still an interesting goal to see what performance can be obtained using state-of-the-art approaches for language-independent modelling as well as investigating whether these approaches can be used to bootstrap acoustic models in an unsupervised fashion.

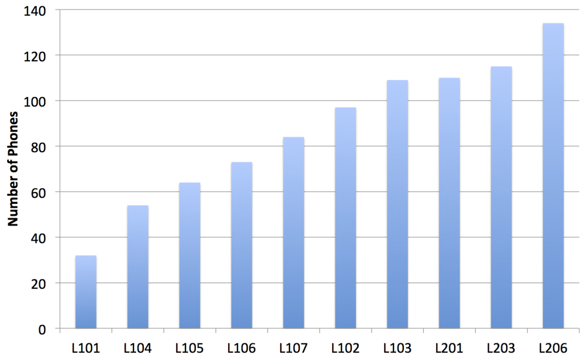

Fig. 3. Cumulative Phone Occurrences against Language Release

One of the first issues to be considered when constructing these LI acoustic models is the phone coverage. Figure 3 shows the cumulative phone coverage over the ten languages considered. The ordering is the base period development languages (101,104,105,106), the base period surprise language (107) and then the five option period 1 development languages (102,103,201,203,206). Note for these plots diphthongs (and triphthongs) are split into their constituent units. It is clear from the plot that the phone coverage has not yet converged. Indeed the overall X-SAMPA attribute file at CUED comprises 215 entries, of which only 62% have currently been seen.

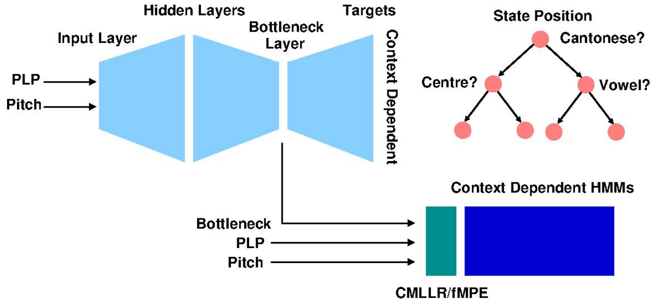

Fig. 4. Language Independent Acoustic Models

The overall structure of the LI acoustic models is shown in Figure 4. The same decision tree is used for the targets of the BN-MLP and the acoustic model, though this is not a requirement. The states for unseen phones for the target language are determined using the phonetic attributes from X-SAMPA. It is possible to use language questions in the decision tree construction process. For the results in this section language questions were not used. For additional details of the experimental set-up see [40].

As discussed in section 2 seven languages were used as training data for the LI acoustic model (these were the same languages as used to train the BN features in the previous section). The LLPs for each of the languages were used. For the three unseen languages, the number of unseen phones were: Vietnamese (107) 7; Bengali (103) 12; and Haitian Creole (201) 2. To handle this issue, the state-position roots to the decision trees were used. Thus unseen phones were mapped to leaf nodes using X-SAMPA phone attributes. The language model training data and lexicon were taken from the LLPs for the training languages. To avoid any bias from using the transcriptions from the LLP to train the language model, the LLP transcribed data was not used. For all experiments a Tandem-SAT system was used.

System		TER	MTWV		
		(%)	IV	OOV	Tot
Haitian Creole (201)					
LD	fMPE	61.7	0.4673	0.2347	0.4317
LI	fMPE	77.2	0.2250	0.0966	0.2058
UN	ML	71.4	0.2907	0.1462	0.2691
Bengali (103)					
LD	fMPE	68.5	0.3173	0.0987	0.2504
LI	fMPE	81.1	0.1929	0.0775	0.1573
UN	ML	75.9	0.2068	0.0913	0.1723
Vietnamese (107)					
LD†	fMPE	69.3	0.1962	0.1081	0.1851
LI	fMPE	87.6	0.0255	0.0268	0.0257
UN	ML	84.9	0.0086	0.0357	0.0174

Table 4: *LLP performance using Language Dependent (LD), Language Independent (LI), and Unsupervised (UN) models. † indicates that plp features were used as the input to the BN MLP.*

Table 4 shows the performance of the LI acoustic models against the language dependent acoustic models (LD) on the three unseen languages. As expected the performance of the LI acoustic models is significantly worse in terms of both ASR and KWS performance for all languages. For Vietnamese (107) the performance is very poor. For additional analysis of these results see [40].

In addition Table 4 shows the performance of using the LI acoustic models to bootstrap a new language in a completely unsupervised fashion. In this preliminary work the transcriptions from the LI acoustic models were used in the standard system build framework. Using these unsupervised transcriptions for discriminative training (either MPE or fMPE) degraded performance [3]. The MLP to obtain the bottleneck features was not retrained, so the multi-lingual BN features from section 5.2 were used. Note none of the languages in Table 4 were in the training data for this network. For Haitian Creole and Bengali unsupervised trained acoustic models (UN) improved performance, both for ASR and KWS, over the LI acoustic models. This indicates the limitations of assuming phone consistency over multiple languages (as used in the LI models).

For Vietnamese, where the ASR performance was significantly worse than Haitian Creole and Bengali, there were slight gains in TER, however no gain in KWS performance. In some way this is not surprising as extrapolating the graph in Figure 5 at ASR performance levels of about 85% the KWS performance is starting to just look like noise.

The above results have been generated using transcriptions from the LI acoustic models. It is also possible to use the unsupervised acoustic models to retranscribe the data. This mode will be investigated in future work.

7. HYBRID AND TANDEM SYSTEM COMBINATION

Given the different forms of classifier being used for the Tandem and Hybrid system, they may be expected to be complementary to one another. To investigate this for ASR , the confusion networks generated by the Tandem and Hybrid systems were combined using CN combination (CNC) [8]. Before combining the two system, the posterior probability associated with the CN of each system, based on the arc posteriors from the lattice, were mapped to remove any biases in the confidence measures. In this work, a simple merging of the posting lists from each of the systems, prior to STO normalisation, was used for combining the KWS systems together, rather than a more complicated approach such as MTWV-weighted CombMNZ method discussed in [30]. In initial experiments, there was a slight degradation in performance by using the merging, rather than CombMNZ, but it simplifies the pipeline.

To moderate the impact of the quality of data in the LLP either Vocal Tract Length Perturbation (VTLP) was used (Bengali, Haitian Creole, and Lao) or semi-supervised approaches (Assamese and Zulu) were used to train the BN MLP for the Tandem system. For both VTLP and semi-supervised training, the approaches described in section 5.1 were used. Due to time constraints, data augmentation was only applied to the Zulu Hybrid system. Here semi-supervised training, in the same fashion as the Tandem system was used, and the number of target states increased to 3000.

Language	Id	LP	TER (%)		
			Tandem	Hybrid	CNC
Assamese	102	FLP	54.2	55.1	52.8
		LLP	65.1	67.8	64.3
Bengali	103	FLP	54.9	56.6	54.3
		LLP	67.0	69.5	66.8
Haitian Creole	201	FLP	48.7	50.3	48.2
		LLP	60.5	63.4	60.4
Lao	203	FLP	48.5	51.9	48.9
		LLP	61.2	65.8	61.3
Zulu	206	FLP	62.1	64.4	61.2
		LLP	71.5	74.1	70.6

Table 5: *%TER with CN decoding for Tandem and Hybrid and CNC for Full (FLP) and Limited (LLP) Language Packs.*

Table 5 shows the STT system performance on each of the languages, and each configuration. There are some general trends. For these DNN systems, the Tandem system consistently outperformed the Hybrid configuration. Part of this difference in performance may be because of the use of cross-entropy, rather than sequence training [41]. The difference in performance was also greater for the LLP than the FLP. This can partly be attributed to the use of data augmentation for the Tandem system, but not the Hybrid system. In general the combination of the Tandem and Hybrid STT results yielded gains. The outlier for this was Lao, where the difference in performance between the Tandem and Hybrid systems was greatest.

Table 6 shows the performance of the KWS system on each of the languages and language packs. For the FLPs the performance of both the Tandem and Hybrid systems was very similar, for Assamese the Hybrid system yielded the best performance. For the LLPs there was still a gap in performance with the Tandem system outperforming the Hybrid. This was also true when comparing the Tandem with no data augmentation to the Hybrid system. In contrast to the ASR combination, merging posting lists improved KWS performance in

[3] The default parameter settings for I-smoothing were used. It is possible to tune the system to ensure than MPE does not degrade performance, but this tuning was not done.

Language	Id	LP	MTWV		
			Tandem	Hybrid	Merge
Assamese	102	FLP	0.4660	0.4730	0.4946
		LLP	0.2569	0.2360	0.2771
Bengali	103	FLP	0.5151	0.5121	0.5388
		LLP	0.2992	0.2615	0.3100
Haitian Creole	201	FLP	0.6387	0.6329	0.6602
		LLP	0.4648	0.4336	0.4867
Lao	203	FLP	0.5951	0.5881	0.6149
		LLP	0.4262	0.3790	0.4439
Zulu	206	FLP	0.3770	0.3654	0.4084
		LLP	0.2287	0.1924	0.2366

Table 6: *MTWV for Tandem and Hybrid and their combination for Full (FLP) and Limited (LLP) Language Packs.*

every configuration, even for Lao LLP where there were large differences in KWS performance. For the combined system, 8 out of the 10 configurations achieved the program goals of 0.3 TWV. Note these numbers are based on the MTWV, not the performance with an automatically determined threshold. However, there is usually only a slight degradation when the threshold is automatically determined rather than using the MTWV.

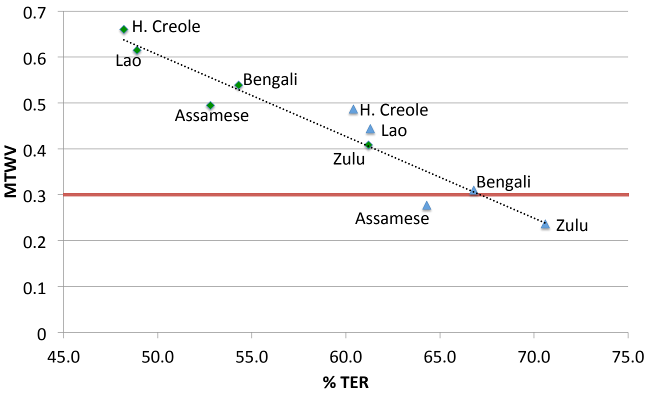

Fig. 5. MTWV against TER, ◇ indicates FLP, △ LLP

Given that configurations for five languages, and two language pack sizes, have been run in a consistent framework, it is interesting to examine the correlation between the ASR performance and the KWS performance. Figure 5 shows the plot of MTWV against TER (%) for all five option period 1 languages in both LLP and FLP configurations. Here the CNC TER% and the Merged MTWV values are given. The correlation between the two is high (Pearson Correlation Coefficient -0.945, R^2 value 0.911). From the plot it is also clear that some languages, Haitian Creole and Lao, are simpler at least for this task. Also the performance of Assamese on the development Keyword List, is lower for both FLP and LLP than expected for the ASR performance.

8. CONCLUSIONS

This paper has described some of the research undertaken at CUED under the Babel program as part of the Lorelei team led by IBM. The aim of the project is to develop robust KWS and ASR technologies that can be rapidly applied to any human language. Data distributed under the Babel project has been used throughout this paper to illustrate both ASR and KWS performance. Primarily the languages from Option Period 1 (OP1) have been used: Assamese, Bengali, Haitian Creole, Lao and Zulu. Two sizes of language pack are distributed for each language: a Full Language Pack (FLP) with approximately 80 hours of transcribed audio data; and a Limited Language Pack (LLP) with about 10 hours of transcribed data. Three main research areas have been described: deep neural networks (DNNs) for acoustic modelling; data augmentation; and zero-acoustic model resource systems. Finally contrasts between Tandem and Hybrid DNN systems, and their combination for both ASR and KWS are described.

It is clear from the level of performance of the systems described in this paper, that though it is possible to construct ASR and KWS systems that can achievable useful levels of performance, there is still a significant amount of work required in the area of low-resource language speech processing systems.

9. REFERENCES

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

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

[3] 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.

[4] 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.

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

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

[7] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc ASRU*, 1997.

[8] G. Evermann and P. C. Woodland, "Large vocabulary decoding and confidence estimation using word posterior probabilities," in *Proc. of ICASSP*, 2000.

[9] D. R. H. Miller, M. Kleber, et al., "Rapid and accurate spoken term detection," in *Proc. Interspeech*, 2007.

[10] D. Vergyri, I. Shafran, et al., "The SRI/OGI 2006 spoken term detection system," in *Proc. Interspeech*, 2007.

[11] I. Szoke, L. Burget, J Cernocky, and M. Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proc. of SLT*, 2008.

[12] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc ICML*, 2013.

[13] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Proc ASRU*, 2013, pp. 309–314.

[14] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Proc ICASSP*, 2014.

[15] M. J. F. Gales, A. Ragni, H. AlDamarki, and C. Gautier, "Support vector machines for noise robust ASR," in *Proc ASRU*, 2009, pp. 205–210.

[16] L. Lamel and J.-L. Gauvain, "Lightly supervised and unsupervised acoustic model training," *Computer speech and language*, vol. 16, pp. 115–129, 2013.

[17] Scott Novotney, Richard M. Schwartz, and Jeff Z. Ma, "Unsupervised acoustic and language model training with small amounts of labelled data," in *ICASSP*, 2009, pp. 4297–4300.

[18] Kai Yu, Mark J. F. Gales, Lan Wang, and Philip C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[19] Ngoc Thang Vu, Florian Metze, and Tanja Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. SLTU*, 2012.

[20] Z. Tüske, J. Pinto, D. Wilett, and R. Schlüter, "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions," in *Proc ICASSP*, 2013, pp. 7349–7353.

[21] K. M. Knill, M. J. F. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc ASRU*, 2013.

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

[23] A. Ghoshal, P. Swietojanski, and S. Renals, *Multilingual training of deep neural networks*, pp. 7319–7323, IEEE, 2013.

[24] Jonas Lööf, Christian Gollan, and Hermann Ney, "Cross-Language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System," in *Proc. Interspeech*, 2009.

[25] N. T. Vu, F. Kraus, and T/ Schultz, "Multilingual A-stabil: A new confidence score for multilingual unsupervised training," in *Proc. SLT*, 2010.

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

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

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

[29] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus among words: Lattice-based word error minimization," in *Proc. Eur. Conf. Speech Commun. Technol.*, 1999.

[30] J. Mamou et al., "System combination and score normalization for spoken term detection," in *Proc. ICASSP*, 2013.

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

[32] M. S. Rasooli, N. Habash, O. Rambow, and T. Lippincott, "Unsupervised morphology-based vocabulary expansion," in *The 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[33] 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.

[34] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transaction of Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, May 1999.

[35] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[36] D. Povey and P.C. Woodland, "Minimum Phone Error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002.

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

[38] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, "Data augmentation for low resource languages," in *Proc Inter-Speech*, 2014, submitted.

[39] Z. Tüske, D. Nolden, R. Schlüter, and H. Ney, "Multilingual mrasta features for low-resource keyword search and speech recognition systems," in *Proc ICASSP*, 2014, pp. 7349–7353.

[40] K. M. Knill, M. J. F. Gales, A. Ragni, and S. P. Rath, "Language independent and unsupervised acoustic models for speech recognition and keyword spotting," in *Proc Inter-Speech*, 2014, submitted.

[41] B. Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *Proc. ICASSP*, 2009.