Language Independent and Unsupervised Acoustic Models for Speech Recognition and Keyword Spotting

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

Department of Engineering, University of Cambridge
Trumpington Street, Cambridge CB2 1PZ, UK.
kate.knill@eng.cam.ac.uk, mjfg@eng.cam.ac.uk

Abstract

Developing high-performance speech processing systems for low-resource languages is very challenging. 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 train a multi-language bottleneck DNN. Language dependent and/or multi-language (all training languages) Tandem acoustic models are then trained. This work considers a particular scenario where the target language is unseen in multi-language training and has limited language model training data, a limited lexicon, and acoustic training data without transcriptions. A zero acoustic resources case is first described where a multi-language AM is directly applied to an unseen language. Secondly, in an unsupervised training approach a multi-language AM is used to obtain hypotheses for the target language acoustic data transcriptions which are then used in training a language dependent AM. 3 languages from the IARPA Babel project are used for assessment: Vietnamese, Haitian Creole and Bengali. Performance of the zero acoustic resources system is found to be poor, with keyword spotting at best 60% of language dependent performance. Unsupervised language dependent training yields performance gains. For one language (Haitian Creole) the Babel target is achieved on the in-vocabulary data.

Index Terms: speech recognition, low resource, multilingual

1. Introduction

There has been increased interest in recent years in rapidly developing high performance speech processing systems for low resource languages. Although a lot of progress has been made e.g. [1, 2, 3, 4, 5] this is still highly challenging. This paper considers the problem of automatic speech recognition (ASR) and keyword spotting (KWS) under a zero acoustic resource scenario. Here it is assumed that there is a limited lexicon and language model training data available for the new, target, language. Two approaches to tackling this problem are considered: language independent recognition; unsupervised training. These approaches are evaluated on data distributed under the IARPA Babel program [6].

Speech recognition systems built with multi-language "deep" neural networks (DNNs) have been shown to provide consistent improvements over language dependent systems e.g. [7, 3, 4, 5, 8]. The models have primarily been applied to within training set languages or only the feature extraction component has been applied to unseen target languages. In this case, many systems require addition of a new output layer and retuning. However, if a single output layer is used with a common phone set then the multi-language acoustic models can be applied as language independent acoustic models to recognise the target language speech and the recognised lattices used in keyword spotting. In [9] it was seen that the performance is dependent on the coverage of the phone set and acoustic space of the target language by the multi-language training set. [9] used 4 languages for training, here, 7 languages are added to the training set in this paper to produce a broader acoustic model with wider acoustic and phonetic coverage. Testing is performed on 3 languages: Haitian Creole, Bengali and Vietnamese.

If it is assumed that it is possible to obtain audio data for the target language, even if transcriptions are not available, then unsupervised training [10] can be applied. In unsupervised training, transcriptions for untranscribed audio data are automatically generated using a pre-existing recogniser. A subset of the data is selected for use in training through confidence measures [10, 11, 12] or alternatives such as closed captions [13]. Typically the selected data subset is then used to boost the training data set within language. Lööf et al. [14] showed that it could also be applied to the case where no transcribed audio existed for a language. A cross-language mapping was made between a single language (Spanish) system and the target language (Polish). Vu et al. [15, 16, 17] extended this to using a combination of 4-6 language dependent systems. Cross-language mappings are again required. In this paper the language independent acoustic model is used to recognise the audio data of the unseen target language and the resulting, confidence selected, transcriptions used to train a language dependent acoustic model for the target language from scratch.

The language independent acoustic model is described in Section 2, followed by the unsupervised training approach in Section 3. Experimental setup and results are presented in Section 4. Finally conclusions are given in Section 5.

2. Language Independent Acoustic Models

One option to handle languages with no transcribed audio data is to treat the problem as a zero acoustic resources problem. Here it is assumed that a limited lexicon is available, as well as limited language model training data. In this work, a language independent acoustic model approach is applied to this case. To do this a multi-language acoustic model (MLAM) is produced from the set of available training languages such that it can be

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.

applied to unseen languages. For this to be succesful the phones need to be consistent across languages and there should be good phone set coverage of the unseen languages in the MLAM. If the phone attributes are consistently labelled across languages then these attributes can be used to handle missing phones. All languages in the IARPA Babel program are supplied with a X-SAMPA phone set so the first criteria is met. Splitting diphthongs and triphthongs[1] into their constituent phones increases cross-language phone coverage[2]. Since there is no equivalent to X-SAMPA for tones, a new tonal marking scheme is proposed based on 6 tonal levels (top (1), high (2), mid (3), low (4), bottom (5), creaky (6)) and 5 tonal shapes (falling (1), level (2), rising (3), dipping (4), peaking(5)). A 2 digit marker is used to indicate the level and shape of the tone, e.g. mid-falling 31, top-level 12, giving a total of 30 tone labels. It is hoped that this will prove applicable to both contour and register tones. Table 1 shows the tone labels for the two tonal training languages and the tonal unseen (Vietnamese) language. Tone level and shape questions are asked in the decision trees as well as tone label.

Tone			Training		Unseen
Label	Level	Shape	L101	L203	L107
21	high	falling	0	4	—
22	high	level	1	—	—
23	high	rising	2	2	2
32	mid	level	3	1	1
34	mid	dipping	—	—	4
43	low	rising	5	3	—

Table 1: *Tone mapping from IARPA Babel tones for Cantonese (L101), Lao (L203) and Vietnamese (L107).*

A Tandem GMM-HMM approach is taken for the MLAM, pictured in Figure 1. Initially multi-language GMM-HMMs are trained on PLP plus pitch features. These models are built from a flat start using the procedure described in [18]. A multi-language phone set is used, formed from the superset of X-SAMPA phone sets of each training language. Phonetic alignments are generated using language specific lexicons and language models. This avoids an explosion in cross-word contexts and incorrect pronunciations being learned for words that appear in more than one language. To perform GMM state tying [19] state position root phonetic decision trees are constructed using all the training data. Tying at the state position, rather than phone, enables the simple combination of data from multiple languages. It also mitigates rare phones and allows new phones in unseen languages to be supported [9]. The decision tree questions are automatically derived from a table of X-SAMPA phones and their associated attributes (e.g. vowel, front) and the lexicon for each language. Phone, attribute, tone and word boundary questions are asked in these experiments (language questions were not asked here).

A multi-layer perceptron (MLP) with a narrow hidden layer (the bottleneck layer) prior to the output layer is trained on data from multiple languages [20]. Context dependent (CD) output layer targets were adopted as they have been found to yield lower error rates than context independent (CI) targets. To support extension to unseen languages the output layer consists of a set of global CD targets based on the common phone set [9]. A single state-position based decision tree is used as shown in Figure 1, generated with the multi-language GMM-HMMs. This allows the MLP to be used to generate features for an unseen language without any tuning. The MLP features are optimised to discriminate all phones and normalisation is across the whole output layer. By contrast normalisation is on a language specific basis in "top-hat" based multi-language MLPs e.g. [3, 5] where language specific output layers are used. This latter approach is not suited to the zero acoustic resources scenario as (at a minimum) a new output layer is needed to support a new language followed by tuning[3].

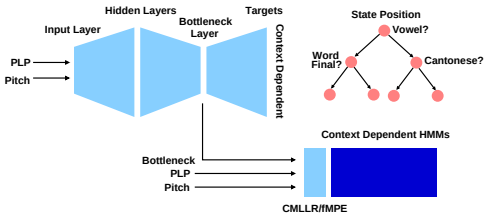

Figure 1: *Multi-language acoustic model.*

All the multi-language training data is presented to the MLP at the same time, with joint optimisation across all the training languages. The order of presentation of data to the MLP is randomised at the frame level across all the languages [21, 5]. The alignment of the context-dependent output states to the training data frames is left fixed during training. Sigmoid and softmax functions are used for the nonlinearities in the hidden and output layers, respectively. The cross-entropy criterion is used as the objective function for optimisation. The parameters of the network are initialised using a discriminative layer-by-layer pre-training algorithm [22]. This is followed by fine tuning of the full network using the error back propagation algorithm.

The bottleneck features are appended to PLP plus pitch features to form the Tandem feature vector for training the Tandem MLAM. Cepstral mean normalisation (CMN) and cepstral variance normalisation (CVN) are applied to conversational sides. Speaker adaptive training (SAT) [23] is applied using global constrained maximum likelihood linear regression (CMLLR) [24] transforms for an entire side, followed by a discriminative transformation of the feature space (fMPE) [25] if desired. The GMM-HMM acoustic models are then trained as described above.

3. Unsupervised training

The previous section described a zero acoustic resources approach to recognising an unseen target language. Transcribing audio data takes time and requires native speakers, however, it is usually not difficult to collect some audio data. Unsupervised training of the new language [10, 14] is then possible. To perform this the language independent acoustic model described in the previous section can be used to produce automatic transcriptions of the audio data. A language dependent system is then trained from scratch on a confidence selected subset of the unsupervised data. The training procedure is shown in Figure 2. Note, the bottleneck MLP is currently left "as is" and no tuning to the target language applied.

If all the data is used for training performance will be poor due to the very low quality of the hypothesised transcriptions.

[1] We add an additional marker to the lexicon to indicate that the phone was derived from a diphthong or triphthong.

[2] In our previous work [9] diphthongs were not split leading to a high number of unseen vowels.

[3] Cross-language mapping of the phone sets from different languages may be possible but would not be straightforward.

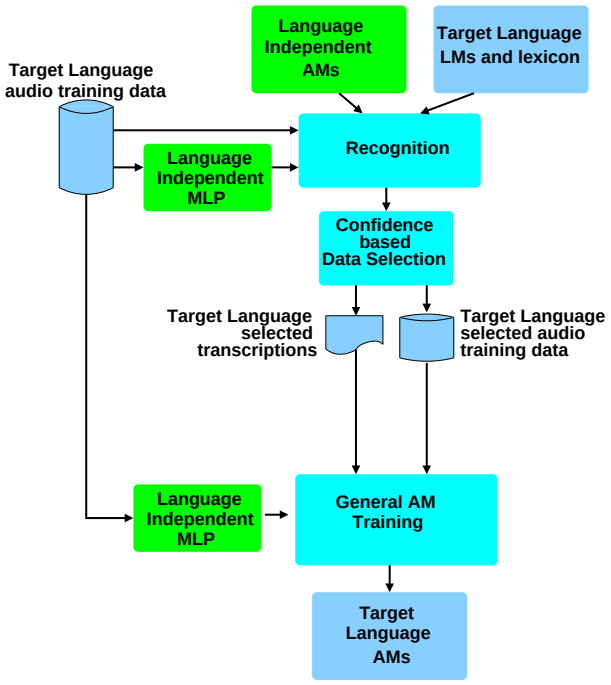

Figure 2: *Boostrapping of language dependent system with no audio transcriptions using a language independent acoustic model.*

Audio segments are selected to form a smaller training set based on frame-weighted word-level confidence score [26]. Mapped word (or token) based confidence scores are obtained from the confusion networks. These are then weighted by the average number of frames to yield an average frame confidence score for each segment. A threshold is applied to select the segments for unsupervised training. Silence frames are excluded from the confidence score computation. MAP adaptation to a smaller, higher confidence, subset of automatically transcribed data may be performed. Further iterations of training could also be added, such as generating new automatic transcriptions using the language dependent model. The latter is not investigated here.

4. Experiments

4.1. Setup

All the experiments are based on language releases from the IARPA Babel program as listed in Table 2. The Limited Language Packs (LLPs) are used for training the LIAM and testing. Each LLP consists of approximately 13 hours of transcribed audio training data and an equivalent development test set. A X-SAMPA phone set and lexicon covering the training vocabulary is provided with each LLP. No changes are 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. 7 languages are used to train the multilanguage acoustic model (MLAM): Assamese, Cantonese, Lao, Pashto, Tagalog, Turkish and Zulu. Bengali, Haitian Creole and Vietnamese are used as the unseen target languages. They have 12, 2 and 7 phones not covered by the MLAM phone set, respectively. Language dependent models using the supplied transcriptions are also trained to provide a baseline. Unsupervised training is performed on a confidence selected subset of the Full Language Pack (FLP) for each of the test languages. About 65 hours of data is automatically transcribed per language. From this $\sim$25 hours are selected for training the unsupervised models. A further stage of MAP adaptation is performed on a reduced set of $\sim$2.5 hours.

Language	Release
Cantonese	IARPA-babel101-v0.4c
Pashto	IARPA-babel104b-v0.4aY
Turkish	IARPA-babel105b-v0.4
Tagalog	IARPA-babel106-v0.2f
Vietnamese	IARPA-babel107b-v0.7
Assamese	IARPA-babel102b-v0.5a
Bengali	IARPA-babel103b-v0.4b
Haitian Creole	IARPA-babel201b-v0.2b
Lao	IARPA-babel203b-v3.1a
Zulu	IARPA-babel206b-v0.1d

Table 2: *IARPA Babel language releases.*

The ASR systems are trained and decoded using HTK [27] and MLPs on an extended version of ICSI's QuickNet [28] software. Speaker adaptive training (SAT) using CMLLR [24] is applied in training and test, with MLLR also used for decoding. Minimum Phone Error (MPE) [29] is used for discriminative training and fMPE for feature-space projection where applied.

The MLAM uses $\sim$7000 states for the MLP output targets and GMM-HMMs. Language dependent (LD) models use $\sim$1000 GMM-HMM states, and for the LD MLP in the supervised training case. Each state has an average of 16 Gaussian components with 32 components for silence. The base GMM-HMMs are trained with PLP plus pitch features. 52-dimensional PLP+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features are projected down to 39 by HLDA. Pitch+Δ+$\Delta\Delta$ features are appended. For the Tandem systems 26 bottleneck (BN) features are also appended. A 504 dimensional input feature vector is used for the MLP, produced by splicing [4] the 52-dimensional PLP+pitch+Δ+$\Delta\Delta$+$\Delta\Delta\Delta$ features. 3 hidden layers plus the BN layer are used in the LD MLPs in configuration 504-1000-500^2-26-1000. The MLAM MLP has 4 hidden layers plus the BN layer in configuration 504-1000^4-26-7000.

Word based (syllable for Vietnamese) bigram language models are used in decoding, with trigram models used for lattice rescoring and confusion network (CN) generation. They are trained on the LLP transcripts with modified Kneser-Ney smoothing using the SRI LM toolkit [30]. At decoding time the language is assumed known and the language specific training lexicon and LM applied. The decoding parameters are kept fixed across all systems. Token error rates are given for trigram CN. Keyword search uses the IBM KWS system without the system combination component [31, 32]. Cascade search is applied with a full phone-to-phone confusion matrix to the bigram decoded lattices. The language model is ignored in the OOV and cascade search (i.e. LM weight set to 0). Keyword search is scored in terms of mean term weighted value (MTWV).

4.2. Results

Table 3 shows the performance of the Haitian Creole baseline language dependent (LD) and language independent (LI) systems. The best LIAM system uses SAT, MPE and fMPE. Even

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

in this case there is an absolute drop in TER of 15.5% and the MTWV is more than halved despite the phone set being largely covered by the LIAM. Bengali exhibits less of a drop in TER (12.6%) and MTWV (66%) as seen in Table 4 whereas Vietnamese has a large drop of 18.3% in TER and the MTWV drops to close to zero.

AM		TER	MTWV		
Data	Type	(%)	IV	OOV	Tot
LD	fMPE	61.7	0.4673	0.2347	0.4317
LI	ML	78.8	0.2126	0.0756	0.1916
	MPE	78.4	0.2067	0.0884	0.1885
	fMPE	77.2	0.2250	0.0966	0.2058
UN	ML	70.4	0.3118	0.1560	0.2880
	MPE	71.7	0.3021	0.1682	0.2815
	fMPE	71.3	0.2956	0.1524	0.2736
	ML-MAP	70.6	0.3123	0.1723	0.2911

Table 3: *Release B Haitian-Creole (L201) LLP performance using Language Dependent (LD), Language Independent (LI), and Unsupervised (UN) models.*

AM		TER	MTWV		
Data	Type	(%)	IV	OOV	Tot
LD	fMPE	68.5	0.3173	0.0987	0.2504
LI	fMPE	81.1	0.1929	0.0775	0.1573
UN	ML	74.9	0.2226	0.1059	0.1872
	ML-MAP	75.1	0.2310	0.1034	0.1920

Table 4: *Release B Bengali (L103) LLP performance.*

AM		TER	MTWV		
Data	Type	(%)	IV	OOV	Tot
LD†	fMPE	69.3	0.1962	0.1081	0.1851
LI	fMPE	87.6	0.0255	0.0268	0.0257
UN	ML	84.7	0.0141	0.0109	0.0137
	ML-MAP	84.8	0.0138	-0.0277	0.0080

Table 5: *Release B Vietnamese (L107) LLP performance.* † *PLP input to MLP.*

Automatic transcription of the FLP audio data for each of the 3 test languages is performed. Figure 3 shows how the percentage of data selected varies with confidence score. The highest confidence is found with Haitian Creole, closely followed by Bengali. Zulu has a very low confidence score, unsurprisingly given the 88% TER.

As seen in Tables 3 and 4, the Unsupervised systems are 25-35% better than the Language Independent system for both Haitian Creole and Bengali. The Haitian Creole Unsupervised system achieves the Babel target of 0.3 MTWV for in-vocabulary terms with both the ML and ML-MAP models, and is < 0.01 off for the overall MTWV. Table 3 shows that discriminative training currently degrades performance of the Unsupervised systems. The TER for Vietnamese is slightly reduced (3%) with the Unsupervised models but the MTWV is degraded even further. Vietnamese's poor performance is partly due to limitations in the multi-language decision tree to discriminate well for Vietnamese phones. This is shown in Figure 4 where red and green indicate the unseen and tonal training languages, respectively.

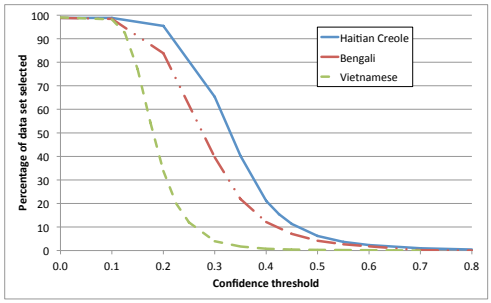

Figure 3: *Percentage of data selected against confidence score.*

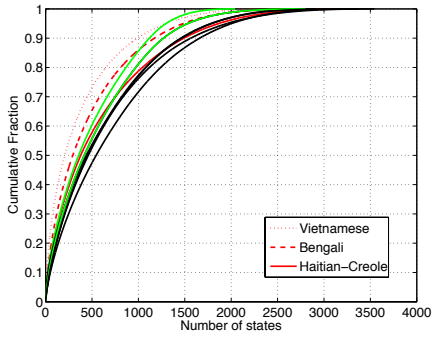

Figure 4: *Cumulative PDF of state coverage of multi-language decision trees in language dependent AMs.*

5. Conclusions

This paper has discussed the problem of automatic speech recognition (ASR) and keyword spotting (KWS) under a zero acoustic resource scenario. Here it is assumed that there is a limited lexicon available, as well as target language model training data available. Two modes of operation are described. First general, language independent, acoustic models are trained and used for recognition. Second, these language systems are used to generate unsupervised transcriptions for the target language. This mode assumes that it is possible to obtain audio data, even if transcriptions are not available. These approaches were evaluated on data distributed under the Babel program. Though the performance of the systems is significantly worse than when there is transcribed audio data available, the results demonstrate that the approaches described do enable ASR and KWS systems to be implemented in this highly challenging scenario. For "simpler" languages, where the phonetic structure is well covered by the training languages, the targets of the Babel project can be achieved for in-vocabulary KWS. However when there is a poor match with the training languages, the performance for both ASR and KWS is poor.

Future work will examine the impact of adding more training languages, as they become available, as well as investigating approaches that allow better use to be made of the phonetic contexts observed in the training languages.

6. Acknowledgements

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

7. References

[1] T. Schultz and K. Kirchhoff, *Multilingual Speech Processing*, 1st ed. Academic Press, 2006.

[2] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low-resource LVCSR systems," in *Proc. Interspeech*, 2010.

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

[4] N. T. Vu and T. Schultz, "Multilingual Multilayer Perceptron for Rapid Language Adaptation Between and Across Languages," in *Proc. Interspeech*, 2013.

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

[6] M. Harper, "IARPA Solicitation IARPA-BAA-11-02," 2011, http://www.iarpa.gov/solicitations_babel.html.

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

[8] F. Grezl, M. Karafiat, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU*, 2011.

[9] K. Knill *et al.*, "Investigation of multilingual deep neural networks for spoken term detection," in *Proc. ASRU*, 2013.

[10] G. Zavaliagkos and T. Colthurst, "Utilizing untranscribed training data to improve performance," in *Proc. Broadcast News Transcription and Understanding Workshop*, 1998, pp. 301–305.

[11] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," in *Proc. ISCA Eur. Conf. Speech Communication Technology*, 1999, pp. 2725–2728.

[12] F. Wessel *et al.*, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.

[13] L. Lamel *et al.*, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–129, 2002.

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

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

[16] ——, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in *Proc. ICASSP*, 2011.

[17] ——, "Rapid building of an ASR system for Under-Resourced Languages Based on Multilingual Unsupervised Training." in *Proc. Interspeech*, 2011.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996.

[24] 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.

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

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

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

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

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

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

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

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