Multilingual Exemplar-Based Acoustic Model for the NIST Open KWS 2015 Evaluation

Van Hai Do*†, Xiong Xiao†, Haihua Xu†, Eng Siong Chng*† and Haizhou Li*†‡
* School of Computer Engineering, Nanyang Technological University, Singapore
† Temasek Laboratories@NTU, Nanyang Technological University, Singapore
‡ Institute for Infocomm Research, A*STAR, Singapore
E-mail: {dovanhai, xiaoxiong, haihuaxu, aseschng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

***Abstract*—In this paper, we investigate the use of the proposed non-parametric exemplar-based acoustic modeling for the NIST Open Keyword Search 2015 Evaluation. Specifically, kernel-density model is used to replace GMM in HMM/GMM (Hidden Markov Model / Gaussian Mixture Model) or DNN in HMM/DNN (Hidden Markov Model / Deep Neural Network) acoustic model to predict the emission probability of HMM states. To get further improvement, likelihood score generated by the kernel-density model is discriminatively tuned by the score tuning module realized by a neural network. Various configurations for score tuning module have been examined to show that simple neural network with 1 hidden layer is sufficient to fine tune the likelihood score generated by the kernel-density model. With this architecture, our exemplar-based model outperforms the 9-layer-DNN acoustic model significantly for both the speech recognition and keyword search tasks. In addition, our proposed exemplar-based system provides complementary information to other systems and we can further benefit from system combination.**

I. Introduction

Among the several thousands of spoken languages used today, few of them are studied by the speech recognition community [1]. One of the major hurdles of ASR (Automatic Speech Recognition) system deployment in new languages is that ASR system relies on a large amount of training data for acoustic modeling. This makes a full fledged acoustic modeling process impractical for under-resourced languages. Popular approaches are to transfer well-trained acoustic models to under-resourced languages such as universal phone set [2, 3], tandem approach [4–6], subspace GMMs (SGMMs) [7, 8], Kullback-Leibler divergence HMM (KL-HMM) [9, 10], cross-lingual phone mapping [11–13] and its extension, context-dependent phone mapping [14–16, 19].

Note that all above methods use a parametric way such as GMM or SGMM to model input feature distribution. Exemplar-based methods are non-parametric techniques that use the training samples directly. Unlike parametric methods, exemplar-based methods, such as k-nearest neighbors (k-NN) [20] for classification and kernel density (or Parzen window) [20] for density estimation, do not assume a parametric form for the discriminant or density functions. This makes them attractive when the distribution of the parameters or their decision boundary is unknown or difficult to estimate.

Recently, several studies apply exemplar-based methods for acoustic modelling [22–25]. In our recent study [33], we successfully applied the exemplar-based method for cross-lingual speech recognition even with few minutes of target language training data. This promising result motivates us to apply the proposed exemplar-based system in [33] to the NIST Open Keyword Search 2015 Evaluation (OpenKWS15)[1] where only 3 hours of speech data can be used to build the speech recognition system. In this paper, various configurations for exemplar-based acoustic model are investigated and experimental results show that our exemplar-based system outperforms DNN and other acoustic models significantly for both the speech recognition and keyword search tasks.

The rest of this paper is organized as follows: Section II describes the exemplar-based acoustic model. Section III presents the experimental procedures and results. Finally, we conclude in Section IV.

II. Multilingual Exemplar-based system

A. System overview

Fig. 1 illustrates the proposed multilingual exemplar-based system for speech recognition. There are six steps to build the system.

1) Generate the multilingual bottleneck features (MBNF) $\boldsymbol{x}_t$. The MBNF extracting network (BN-DNN) is well trained from resource rich source languages.
2) Build a target language triphone-based HMM/GMM acoustic model with MBNF. Generate frame level state label for training data using forced alignment.
3) Generate fMLLR (feature space Maximum Likelihood Linear Regression) feature $\boldsymbol{o}_t$ from MBNF $\boldsymbol{o}_t$ to reduce the speaker effect.
4) Use kernel density estimation and fMLLR feature to estimate HMM state emission probability $\widehat{p}(\boldsymbol{o}_t|s_j)$.
5) Apply discriminative score tuning to refine the likelihood scores in step 4.
6) Plug in the state emission probability to a standard decoder for decoding.

There are three key components in the multilingual exemplar-based acoustic model, i.e. the kernel density estimation, multilingual bottleneck network, and discriminative score tuning. In the following subsections, these components will be presented in detail.

[1] http://www.nist.gov/itl/iad/mig/openkws15.cfm

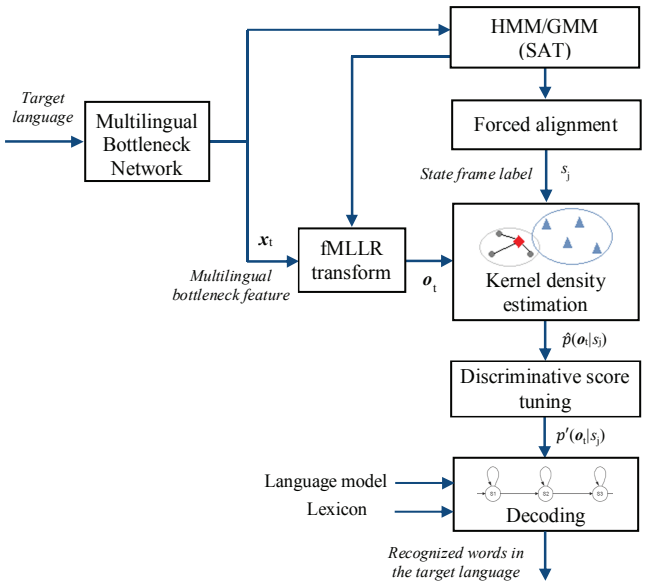

Fig. 1. *Multilingual exemplar-based system.*

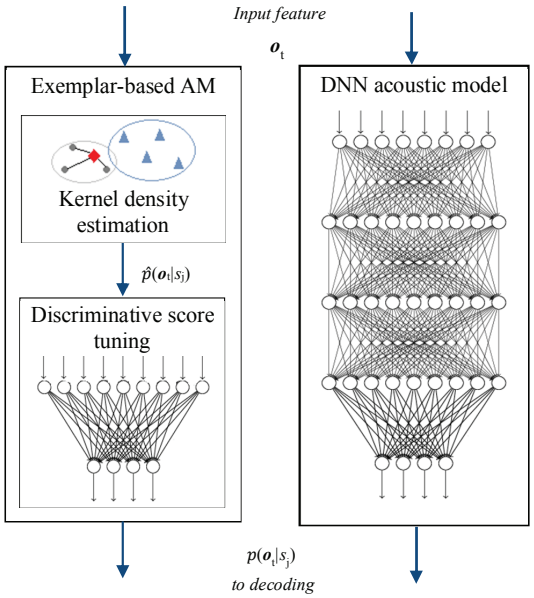

Fig. 2. *Comparation of the proposed exemplar-based and DNN acoustic models.*

B. Kernel density estimation for speech classes

We use kernel density estimation similar to the one used in [22, 33] to model the feature distribution of a triphone tied states. Specifically, the likelihood of a feature vector for a speech class (i.e. a tied state) is estimated as follows:

$$\widehat{p}(\boldsymbol{o}_t|s_j) = \frac{1}{ZN_j}\sum_{i=1}^{N_j}\exp\left(-\frac{||\boldsymbol{o}_t-\boldsymbol{e}_{ij}||^2}{\sigma}\right) \quad (1)$$

where $\boldsymbol{o}_t$ is the feature vector at frame t, $\boldsymbol{e}_{ij}$ is the i^{th} exemplar of class j, N_j is the number of exemplars in class j, and Z is a normalization term to make (1) a valid distribution. In this study, the Euclidean distance between the test feature vector and the exemplars is used. The Euclidean distance has been proved to work well with bottleneck feature [33, 35].

From (1), the likelihood function is mathematically similar to a GMM with shared scalar variance for all dimensions and Gaussians. As our final target is speech recognition rather than density estimation, the term we are interested in is actually the class posteriors. Hence, the normalization term Z will never need to be computed as it is the same for all classes due to the use of single σ in all classes. The parameter σ is used to control the scale of the Gaussians and hence the smoothness of the resulting distribution. If σ is too big, the resulting distribution will be very smooth and vice versa [29]. In this study, σ is simply set to 1 for all classes.

C. Multilingual bottleneck features

In the OpenKWS15, only 3 hours of target language data are provided. To improve performance, we borrow acoustic information from resource rich languages. Specifically, multilingual bottleneck network [16, 17, 26–28] is well trained from several well-resourced languages then used as the feature extractor of the resource limited language. The detailed our bottleneck network can be found in [17].

D. Discriminative score tuning

In the previous section, the kernel density estimation for acoustic modeling is presented. This acoustic model can be considered as a generative model since the state likelihood $\widehat{p}(\boldsymbol{o}_t|s_j)$ is estimated for each state s_j independently. It is well known that using discriminative models e.g., multilayer perceptron (MLP), deep neural network (DNN) [36, 37] or discriminative training criteria [38] can significantly improve performance of speech recognition. To achieve a further gain from the kernel density estimation, in [33], we proposed a technique called discriminative score tuning method. The basic idea of this approach is to use a neural network to discriminatively tune the likelihood scores generated by the kernel density model.

As shown in Fig. 2, in the conventional DNN acoustic model, DNN is used to directly map from input feature space such as MFCC or bottleneck features to HMM states and hence DNN with many layers is used. In our exemplar-based model, score tuning neural network is just used to tune the scores in a discriminative way and hence a very simple network architecture can be used. In [33], we demonstrated that even with 2-layer-neural network i.e. no hidden layer, score tuning can work well and enable the exemplar system to outperform the DNN acoustic model with many layers.

III. Experiments

A. Experimental procedures

To promote keyword search study, the National Institute of Standards and Technology, USA (NIST) has organized the KWS evaluation since 2013. Participants are given a surprise language that is unknown until the evaluation date. The surprise language is Swahili in 2015. In the OpenKWS15, NIST focuses on the keyword search ability for under-resourced

languages. In this evaluation, NIST has formulated a very limited language pack (VLLP) condition, in which only about 3 hours of transcribed speech can be used to build the ASR system together with 10 hours of development data. The acoustic data are collected from various real noisy scenes and telephony conditions. No manual lexicon is available. Text data for language modeling is provided by the organizer. The data is collected from various public available websites. The text data contains 84M words altogether. It is used to establish the lexicon of 350K size, and to build trigram language models. Since Swahili is an agglutinative language, new words and long words are common, leading to large OOV (Out-Of-Vocabulary) rate for a given vocabulary. For instance, even with the 350K vocabulary lexicon, the OOV rate on the dev data is still 7.4%. Besides, the pronunciation of each word is represented as a grapheme string. This is because no lexicon expertise knowledge is available.

System performance is evaluated in both two metrics i.e. WER (Word Error Rate) for speech recognition and ATWV (Actual Term-Weighted Value) [34] for keyword search performance. NIST also provided two sets of keyword lists (KW list). One is the development set (*Dev KW list*), and the other is the evaluation set (*Eval KW list*). Both these keyword lists will be used to evaluate our system.

For acoustic modeling, DNN with 7 hidden layers with 1024 hidden units in each layer. To train DNN, the whole training set is splitted into two subsets i.e. training set to train the network parameters and cross-validation set to examine the training process. The sequential training criterion is applied to train the DNN.

B. Baseline acoustic models

The first row of Table 1 presents performance of monolingual HMM/DNN system. In this case, PLP+pitch feature is used as the input for a 9-layer-DNN acoustic model. We can see that the monolingual system achieves a poor performance with 67.9% WER. We note that there are only 3 hours of training data for the DNN training.

To improve performance, multilingual approaches are applied [17]. In the first multilingual approach, DNN is first initialized with multilingual DNN [17] which is trained with 6 Babel languages which include Cantonese, Pashto, Tamil, Tagalog, Turkish, Vietnamese and then limited training data of the target language is used to tune the DNN. As shown in the second row of Table 1, performance of the multilingual model is significantly improved over the monolingual system in both speech recognition and keyword search.

Another multilingual approach is to use MBNF (Multilingual BottleNeck Feature) [17]. In this case, a BN-DNN is well-trained with 6 Babel languages and then adapted to the target language. MBNF is then used to train the target language DNN acoustic model. Performance of this approach is listed in the third row of Table 1. Similar to our observation in [17], this approach achieves better performance than multilingual HMM/DNN approach in the second row.

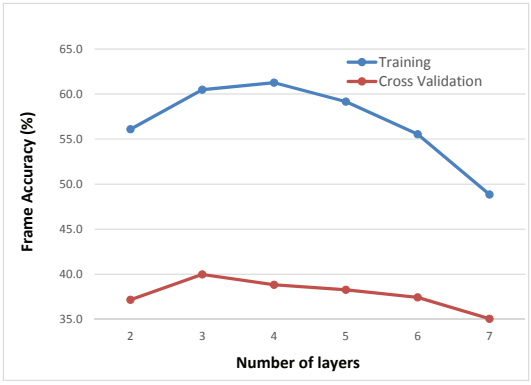

Fig. 3. *Frame accuracy on the training and cross validation sets for different discriminative score tuning architectures.*

C. Exemplar-based acoustic model

In our exemplar-based system, state likelihood generated by the kernel density model is fine tuned using the discriminative score tuning module. This module is realized by a neural network. We now investigate different configurations for score tuning module.

1) Different architectures for discriminative score tuning: In [33], we showed that under very limited training data conditions e.g. several minutes using a 2-layer neural network placed on the top of the kernel density model can tune the likelihood scores. In this paper, we examine whether we can benefit from using more complicated neural network architectures for score tuning.

Fig. 3 shows frame accuracy on the training and cross validation sets for different score tuning architectures. The number of neural network layers increases from 2 to 7. From Fig. 3, we can see that for score tuning module, 3-layer-neural network is sufficient to achieve a good performance since it only slightly tunes the state scores in a discriminative way [33]. This observation is different from the conventional hybrid HMM/DNN model where very deep architectures are normally used to achieve the good performance [36, 37].

2) Context expansion: Recently, using multiple frames as the input for DNN has demonstrated significant improvement for speech recognition. In this experiment, we simply concatenate several likelihood score frames generated by the kernel density model as in input of the score tuning neural network. Surprisingly, as shown in Fig. 4, using more than 1 frame as the input for the score tuning does not achieve any improvement. It can be explained that the multilingual bottleneck feature used for the kernel density model is generated by a multilingual bottleneck DNN which already uses a 31-frame-input. Hence, there is no further benefit to use more frames as the input for the score tuning network. In addition, since input frame of the score tuning network is the high dimensional likelihood score (1,000 dimensions), concatenating multiple frames can lead to over-fitting in the case of limited training data. This phenomenon is observed in Fig. 4, when the number of input frames increases from 1 to 5, the frame accuracy in the

TABLE I
WORD ERROR RATE (WER) AND ACTUAL - TERM WEIGHTED VALUE (ATWV) FOR DIFFERENT ACOUSTIC MODELS

No	System	WER (%)	ATWV	
			Dev KW list	*Eval KW list*
1	Monolingual HMM/DNN acoustic model	67.9	0.2517	0.2917
2	Multilingual HMM/DNN acoustic model	59.7	0.3333	0.3820
3	Multilingual bottleneck feature with HMM/DNN acoustic model	56.5	0.3703	0.4136
4	Multilingual bottleneck feature with examplar-based acoustic model	54.9	0.3800	0.4205

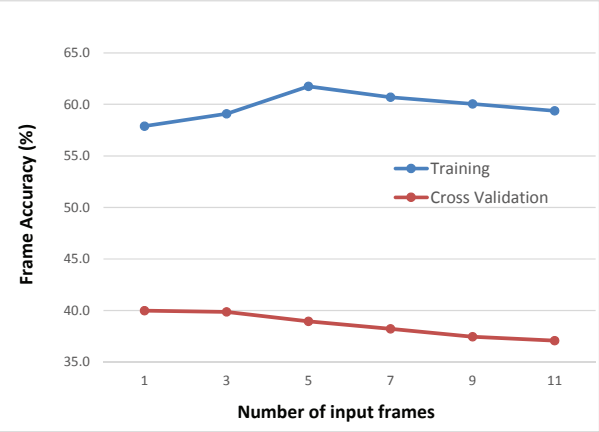

Fig. 4. *Frame accuracy on the training and cross validation sets for different input context sizes of the score tuning network.*

cross-validation reduces as the frame accuracy in the training set increases.

With above analysis, our final score tuning is 3-layer-neural network with only 1 likelihood frame is used to form the input of the neural network. Performance of our final exemplar-based system for speech recognition and key work search is shown in the last row of Table 1. We can see that our exemplar-based system significantly outperforms other systems in term of both WER and ATWV although we use only 3-layer neural network for the score tuning module while the DNN acoustic models in row 1, 2, 3 have the 9-layer architecture. In addition to achieving better performance than other acoustic models, the exemplar-based system provides complementary information which is suitable to our system combination [34].

IV. CONCLUSION

This paper presented our proposed exemplar-based acoustic model for the NIST Open Keyword Search 2015 Evaluation. In our system, kernel-density model is used to estimate the emission probability of HMM states. To improve performance, a score tuning module with different architectures was examined. The experimental results revealed that the exemplar-based model significantly outperforms the 9-layer-DNN acoustic model for both the speech recognition and keyword search tasks.

ACKNOWLEDGEMENT

This work is supported by DSO National Laboratories, Singapore, Project MAISON DSOCL14045.

REFERENCES

[1] H. Li, K. A. Lee, and B. Ma, "Spoken Language Recognition: From Fundamentals to Practice," Proceedings of the IEEE, Vol. 101, No. 5, May 2013, pp. 1136–1159.

[2] T. Schultz and A. Waibel, "Experiments On Cross-Language Acoustic Modeling," in Proc. International Conference on Spoken Language Processing (ICSLP), 2001, pp. 2721-2724.

[3] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5000–5003.

[4] A. Stolcke, F. Grezl, M. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006, pp. 321-324.

[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 877-880.

[6] P. Lal, "Cross-lingual Automatic Speech Recognition using Tandem Features," Ph.D. thesis, The University of Edinburgh, 2011.

[7] L. Burget, et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 4334-4337.

[8] L. Lu, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4877–4880.

[9] D. Imseng, H. Bourlard, and P. N. Garner, "Using KL-divergence and multilingual information to improve ASR for under-resourced languages," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4869–4872.

[10] D. Imseng, P. Motlicek, H. Bourlard, and P. N. Garner, "Using out-of-language data to improve an under-resourced speech recognizer," Speech communication, vol. 56, 2014, 142–151.

[11] K. C. Sim and H. Li, "Context Sensitive Probabilistic Phone Mapping Model for Cross-lingual Speech Recognition," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2008, pp. 2715-2718.

[12] K. C. Sim and H. Li, "Stream-based Context-sensitive Phone Mapping for Cross-lingual Speech Recognition," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2009, pp. 3019-3022.

[13] K. C. Sim, "Discriminative Product-of-expert Acoustic Mapping for Crosslingual Phone Recognition," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2009, pp. 546-551.

[14] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Context dependant phone mapping for cross-lingual acoustic modeling," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP), 2012, pp. 16–20.

[15] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "A Phone Mapping Technique for Acoustic Modeling of Under-resourced Languages," in Proc. International Conference on Asian Language Processing (IALP), 2012, pp. 233–236.

[16] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Context-dependent phone mapping for LVCSR of under-resourced languages," in Proc of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2013, pp. 500–504.

[17] H. Xu, V. H. Do, X. Xiao, E. S. Chng, "A Comparative Study of BNF and DNN Multilingual Training on Cross-lingual Low-resource Speech Recognition," in Proc of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015.
[18] V. T. Pham, H. Xu, V. H. Do, X. Xiao, E. S. Chng, H. Li "On the study of very low-resource language keyword search," submitted to APSIPA 2015.
[19] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages," the IEICE Transactions on Information and Systems, Vol. E97-D, No. 2, Feb. 2014, pp.285–295.
[20] R. O. Duda, P. E. Hart, and D. G. Stork. "Pattern classification". John Wiley & Sons, 2000.
[21] K. Weinberger, and L. Saul, "Distance metric learning for large margin nearest neighbor classification", Journal of Machine Learning Research, vol. 10, 2009, pp. 207–244.
[22] T. Deselaers, G. Heigold, and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in Proc of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2007, pp. 2093-2096.
[23] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," IEEE Signal Processing Magazine, vol. 29, no. 6, 2012, pp. 98-113.
[24] J. Labiak and K. Livescu, "Nearest neighbor classifiers with learned distances for phonetic frame classification," in Proc of the Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011, pp. 2337–2340.
[25] N. Singh-Miller and M. Collins, "Learning label embeddings for nearest-neighbor multi-class classification with an application to speech recognition," In Advances in Neural Information Processing Systems 22, 2009, pp 1678-1686.
[26] F. Grezl, M. Karafiat, S. Kontar, and J. Cernock, "Probabilistic and bottleneck features for LVCSR of meetings," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007, pp. 757–760.
[27] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The Language-Independent Bottleneck Features," in Proc. IEEE Workshop on Spoken Language Technology (SLT), 2012, pp. 336–341.
[28] N. T. Vu, F. Metze, and T. Schultz, "Multilingual Bottle-Neck Features and its Application for Under-resourced Languages," in Proc. International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU), 2012.
[29] B. W. Silverman, "Density Estimation for Statistics and Data Analysis", Chapman and Hall, New York, 1986.
[30] X. Xiao, E. S. Chng, T. P. Tan, and H. Li, "Development of a Malay LVCSR System," in Proc. Oriental COCOSDA, 2010.
[31] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation 18, 2006, pp. 1527–1554
[32] S. Young and others, "The HTK book", Cambridge university engineering department, 2006.
[33] V. H. Do, X. Xiao, E. S. Chng, H. Li, "Kernel Density-based Acoustic Model with Cross-lingual Bottleneck Features for Resource Limited LVCSR," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2014, pp. 6–10.
[34] V. T. Pham, H. Xu, V. H. Do, T. Z. Chong, X. Xiao, E. S. Chng, H. Li, "On the study of very low-resource language keyword search," submitted to Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2015.
[35] V. H. Do, X. Xiao, E. S. Chng, H. Li, "Distance Metric Learning for Kernel Density-Based Acoustic Model Under Limited Training Data Conditions," submitted to Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2015.
[36] V. H. Do, X. Xiao, and E. S. Chng, "Comparison and Combination of Multilayer Perceptrons and Deep Belief Networks in Hybrid Automatic Speech Recognition Systems," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2011.
[37] A-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," in IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No.1, 2012, pp. 14-22.
[38] McDermott, Erik. "Discriminative training for speech recognition." PhD thesis, Waseda University, 1997.