Investigating the Learning Effect of Multilingual Bottle-Neck Features for ASR

Ngoc Thang Vu, Jochen Weiner, Tanja Schultz

Karlsruhe Institute of Technology, Germany

{thang.vu, jochen.weiner, tanja.schultz}@kit.edu

Abstract

Deep neural networks (DNNs) have become state-of-the-art techniques of automatic speech recognition in the last few years. They can be used at the preprocessing level (Tandem or Bottle-Neck features) or at the acoustic model level (hybrid Hidden Markov Model/DNN). Moreover, they allow exploiting multilingual data to improve monolingual systems. This paper presents our investigation of the learning effect of neural networks in the context of multilingual Bottle-Neck features. For this, we perform a visual analysis of the output of the Bottle-Neck layer of a neural network using t-Distributed Stochastic Neighbor Embedding. Our results show that multilingual Bottle-Neck features seem to learn phoneme characteristics, such as the *F1* and *F2* formants which characterize different vowels, and other articulatory features, such as fricatives and nasals which characterize consonants. Furthermore, they seem to normalize language dependent variations and transfer the learned representation to unseen languages.

Index Terms: multilingual Bottle-Neck features, visualization

1. Introduction

The performance of speech and language processing technologies has improved dramatically over the past decade with an increasing number of systems being deployed in a large variety of languages and applications. However, most efforts to date are still focused on a small number of languages. With more than 6,900 languages in the world, one of the important challenges today is to rapidly port speech processing systems to new languages with little manual effort. Therefore, many studies have been conducted in multilingual and crosslingual speech processing. One of the central ideas is the exploration of the sharing factor between languages to improve the performance of a speech processing system and also to benefit by rapidly building a system for a new language [1].

Neural networks (NNs) have become one of the most important techniques to improve ASR performance in the last few years. Two ways to incorporate the NNs techniques into the ASR framework are commonly used: The first way is using the Tandem approach [2] or Bottle-Neck features [3] to integrate DNNs into ASR systems at the preprocessing level. The output of a neural network or a small hidden layer (Bottle-Neck layer) is used as speech features for recognition task. Another way is to use a HMM/DNN hybrid system in which DNNs estimate the emission probabilities of the Hidden Markov Model (HMM) states [4, 5, 6, 7, 8]. Both approaches were successfully applied to large vocabulary continuous speech recognition (LVCSR) where they lead to a significant improvement in various tasks with different data sets.

In the context of rapid language adaptation, the use of NNs allows exploiting multilingual data to improve the monolingual ASR performance. At the preprocessing level, previous studies [10, 11, 12, 13, 14, 15, 16, 17] showed that a multilayer perceptron (MLP) trained with data of one or several languages can be used to extract features for a new target language. Their results revealed that using additional data of other languages to train the MLP improved the ASR performance. Moreover, recent studies [18, 19, 20, 21, 22] utilized multilingual data during DNN training for acoustic modeling in different unsupervised and supervised ways to improve the monolingual ASR performance. All the results indicate that the shared hidden layer is language independent and can be used to bootstrap the DNN for new languages.

However, an analysis to figure out what exactly was learned and why multilingual data can be used to improve the ASR performance for new languages is still missing. In one of the very few works on visualizing deep neural networks the authors concentrated on analyzing the input features which should be used for DNN based acoustic modeling [23]. In contrast, the target of this paper is to achieve a better understanding of the learning effect of a neural network. For this, we perform a visual analysis of the output of the multilingual Bottle-Neck features supplemented by measuring the clustering quality using the Davies-Bouldin index [24]. We aim at finding potential answers to what the multilingual MLP learns and whether the BN representation transfers to new languages.

2. Data Resource and Baseline System

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [25]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work, we select French, German, Spanish, Bulgarian, Polish, Croatian, Russian, Portuguese, Mandarin, Korean, Thai, Japanese and Vietnamese data from the GlobalPhone corpus. In addition, we use the English speech data from WSJ0. In our experiments, we simulate the case of low-resource languages in which we select only 2 hours of Vietnamese training data as target language and the remaining ones as source languages.

The baseline recognizer for the target languages can be described as follows: the language model is built with a large amount of text data which was crawled using the Rapid Language Adaptation Toolkit [26]. For acoustic modeling, we apply the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory trained from seven GlobalPhone languages [1]. To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds. The standard front-end is used by applying a Hamming window of 16ms length every 10ms. Each feature vector has 143 dimensions resulting from stacking 11 adjacent frames of 13 MFCC coefficients each. A Linear Discriminant Analysis transformation reduces the feature vector size to 42 di-

mensions. For Vietnamese ASR, we merge monosyllable words to bi-syllable words to enlarge the context in acoustic modeling and the history of the language model [27]. The trigram perplexity, out-of-vocabulary rates and vocabulary size of the LM are 176, 0% and 30k, respectively. The syllable error rate (SyllER) is 26.0% on the test set.

3. Multilingual MLP training

In this work, audio data of 12 different languages, namely English, French, German, Spanish, Bulgarian, Polish, Croatian, Russian, Mandarin, Korean, Thai, and Japanese is used to train the multilingual multilayer perceptron. We use the knowledge-driven approach to create a universal phone set, i.e. the phone sets of all languages are pooled together and then merged based on their IPA symbols. Afterwards, some training iterations are applied to create the multilingual model and, thereafter, the alignment for the complete data set. As input for the MLP network, we stack 11 adjacent MFCC feature vectors and use phonemes as target classes. A five layer MLP is trained with a 143-1500-42-1500-152 feed-forward architecture using ICSI QuickNet3 software [28]. We use a learning rate of 0.008 and a scale factor of successive learning rates of 0.5. The initial values of the network were chosen randomly. On the cross-validation data (10% of the training data) the frame-wise classification accuracy of the multilingual MLP is 60.15%. We directly use the multilingual MLP to extract the Bottle-Neck features for the Vietnamese ASR. The SyllER is improved to 21.2% on the test set which is more than 20% relative improvement compared to the baseline system (Section 2).

4. Visualization and Evaluation

The multilingual BN features are visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE) [29], an extension of SNE [30]. This technique allows visualizing high-dimensional data by assigning each data point a location in two or three-dimensional space. SNE starts by converting high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. The similarity of data point x_j to data point x_i is the conditional probability $p_{j|i}$ that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j a similar conditional probability $q_{j|i}$ is computed. If the mapped points y_i and y_j correctly model the similarity between the high-dimensional data points x_i and x_j, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. Based on this observation, SNE aims at finding a low-dimensional data representation that minimizes the mismatch between $p_{j|i}$ and $q_{j|i}$. A natural measure for this mismatch is the Kullback-Leibler divergence. SNE minimizes the sum of Kullback-Leibler divergences over all the data points using a gradient descent method. The cost function C used by SNE is given by

$$C = \sum_i KL(P_i|Q_i) = \sum_i \sum_j p_{j|i} log \frac{p_{j|i}}{q_{j|i}} \qquad (1)$$

in which P_i represents the conditional probability distribution over all other data points given data point x_i, and Q_i represents the conditional probability distribution over all other mapped points given mapped point y_i. Although SNE constructs reasonably good visualizations, the cost function is difficult to optimize. Therefore t-SNE uses a cost function that differs from the one used by SNE in two ways: (1) it uses a symmetrized version of the SNE cost function with simpler gradients and (2) it uses a Student-t distribution rather than a Gaussian to compute the similarity between two points in the low-dimensional space. The t-SNE software [31] is used in our further experiments.

To support the visualization, we apply the Davies-Bouldin index (DB) [24] for cluster evaluation:

$$DB = \frac{1}{n} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \qquad (2)$$

where n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of elements in x from c_x and $d(\cdot, \cdot)$ is the distance between two elements. This index takes into account the variance within each cluster and the distance of the cluster means: a lower value means that the elements in the same cluster are clustered closer together and that the clusters are further away from each other, hence a lower value signifies better clusters.

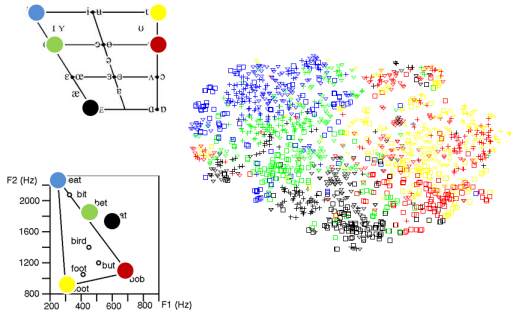


Figure 1: Multilingual BN features of five vowels from French (+), German (□) and Spanish (▽): /a/ (black), /i/ (blue), /e/ (green), /o/ (red), and /u/ (yellow)

5. Analysis

We apply t-SNE to visualize the BN features to find potential answers to the following questions:

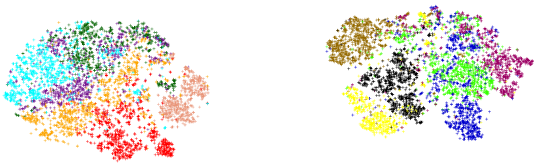
- What does the multilingual MLP learn?
- Does the BN representation transfer to new languages?

The following paragraphs discuss the visualization of the BN features and possible implications.

The multilingual MLP trained on 12 languages (Section 3) is used to extract the BN features. Since the number of languages is quite large for the visualization, only the data of German, French, Spanish, and Vietnamese are plotted. Note that Vietnamese is not related to German, French, and Spanish. To explore the crosslingual effect, only the phonemes are selected that occur in all four languages according to their IPA representation. These are the five vowels /a/, /i/, /e/, /o/, /u/ and twelve consonants /f/, /j/, /k/, /l/, /m/, /n/, /ng/, /p/, /s/, /t/, /v/, /z/. The consonants are further divided into voiced (/j/, /l/, /m/, /n/, /ng/, /z/) and unvoiced (/f/, /k/, /p/, /s/, /t/, /v/) consonants.

5.1. What does the multilingual MLP learn?

We use t-SNE to plot multilingual BN features of the five vowels on the right in Figure 1. The data points are collected by using French (+), German (□) and Spanish (▽) speech data. On the left of Figure 1, we show the IPA vowel chart and the



(a) (b)

Figure 2: Multilingual BN features from French, German and Spanish for (a) voiced consonants /j/ (red), /l/ (orange), /m/ (turquoise), /n/ (violet), /ng/ (green), /z/ (salmon), (b) unvoiced consonants /f/ (black), /k/ (blue), /p/ (purple), /s/ (brown), /t/ (green), /v/ (yellow)

vowel-triangle with the five vowels annotated with corresponding colors. The vowel-triangle expresses which vowels have which formants on average. Interestingly, an analogy of the visualization with the other two pictures can be observed: The data points of the five vowels from the three different languages resemble the relations of the vowels in the vowel chart and the vowel-triangle.

In Figure 2 we plot the multilingual BN features of voiced and unvoiced consonants. We observe that phonemes sharing articulatory features are clustered together while phonemes with different features are clearly separated. In the plot of the voiced consonants (Figure 2a) the features for the nasals /m/, /n/, /ng/ are close together and clearly separated from the palatal /j/ and the alveolars /l/ and /z/. In the plot of the unvoiced consonants (Figure 2b) the fricatives /f/, /s/, /v/ are clearly separated form the plosives /k/, /p/, /t/. We observe that within the clusters for articulatory features the BN features of the consonants from the different languages are also clustered together by phoneme. However, the separation between phonemes sharing articulatory features is not as clear as the separation between phonemes that do not share these features. In particular, the palatal, alveolar and fricative phonemes form distinct individual clusters while the clusters of the different nasal and plosive phonemes overlap.

The observations on vowels and consonants data suggest the following implications:

- BN features seem to discriminate articulatory features: 1) The vowels resemble the pattern of the IPA vowel chart and the vowel-triangle; the MLP has learned spectral characteristics of different vowels, namely the first two formants *F1* and *F2*. 2) The consonants that share articulatory features form distinct clusters.

- BN features seem to normalize the language dependent variations of phonemes. Although the data points are from different languages, the phonemes representing the same IPA symbol are clustered together.

5.2. Does the BN representation transfer to new languages?

As described in Section 3, we obtain significant improvements in terms of SyllER by using the multilingual MLP directly without re-training to extract the BN features for Vietnamese ASR. This indicates that some language independent information has been learned by the multilingual MLP. However, it is not clear how exactly the language independent information is represented in this context. In the previous paragraph, we observe that the multilingual MLP captures articulatory features such as *F1* and *F2* for vowels and normalizes language variations.

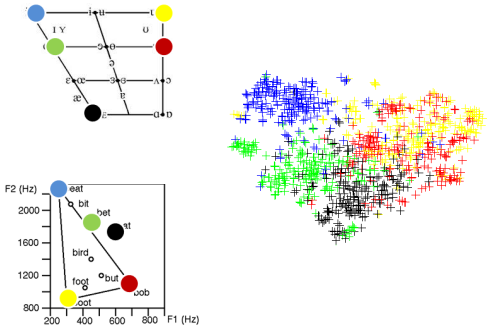


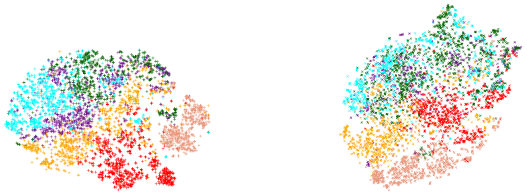
Figure 3: Multilingual BN features for Vietnamese vowels: /a/ (black), /i/ (blue), /e/ (green), /o/ (red), and /u/ (yellow)

5.2.1. *Multilingual BN representation for an unseen language*

In this paragraph, we visualize the BN features of Vietnamese data using this multilingual MLP to obtain a better understanding of the crosslingual transfer effect. Moreover, we look at two further effects: The language independence of the BN features and the discriminability of the multilingual BN features for unseen languages. Note that German, French and Spanish were among the languages used to train the multilingual MLP while Vietnamese is the unseen language in our example. We plot multilingual BN features of the five Vietnamese vowels on the right hand side of Figure 3. On the left hand side of Figure 3, we show the vowel chart and the vowel-triangle again. We observe the same effect as by visualizing the multilingual phones in Figure 1. The data points of the five Vietnamese vowels again represent the relations in the vowel chart and the vowel-triangle. This indicates that the learning effect, in this case the *F1* and *F2* information, can be transferred to the new language.

The multilingual BN features for voiced and unvoiced consonants are plotted in Figures 4 and 5, respectively. The left hand side shows the multilingual features for French, German and Spanish, the right hand side shows the multilingual features for Vietnamese. The effect for the Vietnamese phonemes is the same as for the multilingual phonemes: the features are clustered by the articulatory features of the phonemes. Within each articulatory feature cluster, the features are grouped by phonemes. As with the multilingual features the phoneme clusters are better separated for some phonemes than for others.

We observe for both consonants and vowels that the learning effect has been transferred to the unseen language. This means the multilingual BN features are language independent and can be used for feature extraction for an unseen language.



(a) (b)

Figure 4: Multilingual BN features for voiced consonants from (a) French, German and Spanish, and (b) Vietnamese: /j/ (red), /l/ (orange), /m/ (turquoise), /n/ (violet), /ng/ (green), /z/ (salmon)

(a) (b)

Figure 5: Multilingual BN features for unvoiced consonants from (a) French, German and Spanish, and (b) Vietnamese: /f/ (black), /k/ (blue), /p/ (purple), /s/ (brown), /t/ (green), /v/ (yellow)

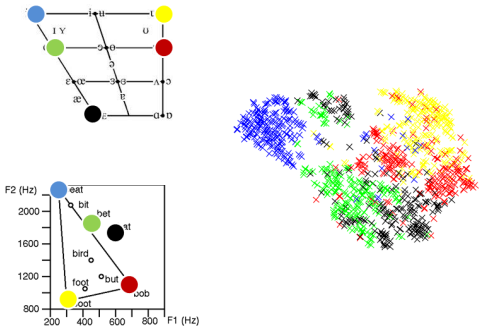


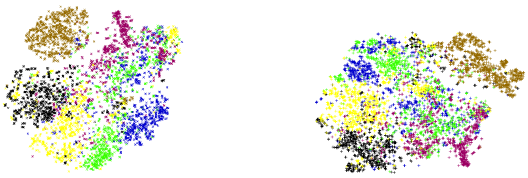
Figure 6: BN features of Vietnamese vowels using a French MLP: /a/ (black), /i/ (blue), /e/ (green), /o/ (red), and /u/ (yellow)

5.2.2. *Monolingual BN representation for an unseen language*

Since the BN representation transfers to unseen languages, we investigate whether is is important to use a multilingual MLP or if the same result can be achieved with a monolingual MLP. For this investigation we plot again the BN features of the same Vietnamese vowels and consonants as before. However, in this case only a monolingual MLP – a French MLP trained with random initialization – was used to extract the features. Again, Vietnamese data is not used for the MLP training. Figure 6 illustrates the IPA vowel chart and the vowel-triangle on the left and on the right the Vietnamese data points. Again, the same effect as in Figures 1 and 3 is observed: The data points of the five Vietnamese vowels illustrate the relations in the vowel chart and the vowel-triangle. Figures 7 and 8 show the data points of voiced and unvoiced consonants. As with the multilingual features the consonants are clustered by articulatory features and by phoneme. This indicates that the MLP learned the articulatory features such as spectral characteristics, namely *F1* and *F2* of different vowels. The MLP transfers this knowledge to an unseen language independent of whether monolingual or multilingual data are used to train the MLP. However, the analogy between the pattern in the plotted data points and the vowel charts in Figure 3 is clearer than in Figure 6, and the separation of the consonant clusters is clearer in Figures 4 and 5 than in Figures 7 and 8. It can be observed in Figure 6 that some data points of phonemes /a/ and /e/ are spread and form a pattern close to phoneme /i/. In Figure 5 we observe that plosive features divide the phoneme /s/ from the other fricatives. One possible explanation for this effect is that the more languages and more data are used to train the MLP, the stronger is the normalization process between languages at the phoneme level.

(a) (b)

Figure 7: BN features for voiced Vietnamese consonants from (a) a multilingual MLP , (b) a French MLP: /j/ (red), /l/ (orange), /m/ (turquoise), /n/ (violet), /ng/ (green), /z/ (salmon)



(a) (b)

Figure 8: BN features for unvoiced Vietnamese consonants from (a) a multilingual MLP , (b) a French MLP: /f/ (black), /k/ (blue), /p/ (purple), /s/ (brown), /t/ (green), /v/ (yellow)

5.2.3. *Comparing Cluster Quality*

In addition to the visual analysis we use the Davies-Bouldin index (DB) to measure the quality of the clusters formed by the BN features for vowels, voice and unvoiced consonants. For each category, we use the original 42-dimensional BN features of multilingual data (FR, GE, and SP), and Vietnamese data using multilingual MLP and French MLP and compute the DB indexes. Note that they are exact the same data which were used for the visualization. The results in Table 1 support our observation based on the visualization results especially for the case of consonants. For vowels, it seems that using the French MLP is more accurate to extract features for recognition task than using the multilingual MLP.

MLP	**Language(s)**	**DB Index**		
		v	vc	uc
Multilingual	French, German, Spanish	4.2	3.1	3.8
Multilingual	Vietnamese	4.0	3.3	3.7
French	Vietnamese	3.3	3.8	4.1

Table 1: Davies-Bouldin index (DB) for the analyzed BN features for vowels (v), voiced (vc) and unvoiced (uc) consonants.

6. Conclusions

This paper presents our investigation of the learning effect of the neural networks in the context of multilingual Bottle-Neck features. We show that their visualization using t-SNE provides useful information to better understand the multilingual BN features. Our results reveal that multilingual BN features seem to learn articulatory characteristics of the phonemes. For vowels these are the *F1* and *F2* formants, while for consonants these are features, such as fricatives, nasals and plosives. Furthermore, the BN features seem to normalize language dependent variations of the phonemes. Their representation is transferred to unseen languages which further indicates their language independence. In the future, we plan to perform an analysis of all the layers of a multilingual DNN.

7. References

[1] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication, Volume 35, Issue 1-2, pp 31-51, 2001.

[2] H. Hermansky, D. Wellis, and S. Sharma. Tandem connectionist feature extraction for conventional HMM systems. In Proc. ICASSP, Turkey, 2000.

[3] F. Grezl et al.. Probabilistic and bottle-neck features for LVCSR of meetings. In Proc. ICASSP, USA, 2007.

[4] N. Morgan and H. Bourlard. Continuous speech recognition: an introduction to the hybrid HMM/connectionist approach. IEEE Signal Processing Magazine, vol. 12, no. 3, pp. 2442, 1995.

[5] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In Proc. of Interspeech, 2011.

[6] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. Audio, Speech, and Language Processing, IEEE Transactions, vol. 20, no. 1, pp. 3042, 2012.

[7] A. Mohamed, G. Dahl, and G. Hinton. Acoustic modeling using deep belief networks. Audio, Speech, and Language Processing, IEEE Transactions, vol. 20, no. 1, pp. 1422, 2012.

[8] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 8297, 2012.

[9] A. Stolcke, F. Grezl, M-Y Hwang, X. Lei, N. Morgan, D. Vergyri. Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons. In Proc. ICASSP 2006.

[10] L. Toth, J. Frankel, G. Gosztolya, S. King. Cross-lingual portability of MLP-based tandem features - a case study for English and Hungarian. In Proc. Interspeech, 2008.

[11] O. Cetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel. Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs. In Proc. ASRU, 2007.

[12] D. Imseng, H. Bourlard, M. Magimai.-Doss. Towards mixed language speech recognition systems. In Proc. Interspeech, Japan, 2010.

[13] C. Plahl, R. Schlueter and H. Ney. Cross-lingual Portability of Chinese and English Neural Network Features for French and German LVCSR. In Proc. ASRU, USA 2011.

[14] S. Thomas, S. Ganapathy, A. Jansen and H. Hermansky. Data-driven Posterior Features for Low Resource Speech Recognition Applications. In Proc. Interspeech, USA, 2012.

[15] K. Vesely, M. Karafiat, F. Grezl, M. Janda, E. Egorova. The language-independent bottleneck features. In Proc. SLT, USA, 2012.

[16] N.T. Vu, F. Metze, T. Schultz. Multilingual bottle-neck feature for under resourced languages. In Proc. SLTU, South Africa, 2012.

[17] N.T. Vu, T. Schultz. Multilingual Multilayer Perceptron For Rapid Language Adaptation Between and Across Language Families. In Proc. Interspeech, France, 2013.

[18] P. Swietojanski, A. Ghoshal, and S. Renals. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In Proc. SLT, USA, 2012.

[19] J.T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. Crosslanguage knowledge transfer using multilingual deep neural network with shared hidden layers. In Proc. ICASSP, 2013.

[20] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. Multilingual acoustic models using distributed deep neural networks. In Proc. ICASSP, 2013.

[21] A. Ghoshal, P. Swietojanski, and S. Renals. Multilingual training of deep neural networks. In Proc. ICASSP, 2013.

[22] N.T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, H. Bourlard. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In Proc. ICASSP, 2014.

[23] A. Mohamed, G. Hinton, G. Penn. Understanding how Deep Belief Networks perform acoustic modelling. In Proc. ICASSP, 2012.

[24] D. Davies, D. Bouldin. A Cluster Separation Measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979.

[25] T. Schultz, N.T. Vu, T. Schlippe. GlobalPhone: A Multilingual Text & Speech Database in 20 Languages. In Proc. ICASSP, Canada, 2013.

[26] N.T. Vu, Tim Schlippe, Franziska Kraus, Tanja Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Proc. Interspeech, Japan, 2010.

[27] N.T. Vu, T. Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In Proc. ASRU, Italy, 2009.

[28] http://www.icsi.berkeley.edu/Speech/qn.

[29] Van der Maaten, L. and Hinton, G.E. Visualizing data using t-SNE, Journal of Machine Learning Research, Volume 9, pp 2579-2605, 2008.

[30] G. Hinton, and R. Sam. Stochastic neighbor embedding. NIPS. Vol. 2. 2002.

[31] http://homepage.tudelft.nl/19j49/t-SNE.html