PERSONALIZED ACOUSTIC MODELING BY WEAKLY SUPERVISED MULTI-TASK DEEP LEARNING USING ACOUSTIC TOKENS DISCOVERED FROM UNLABELED DATA

Cheng-Kuan Wei[1], *Cheng-Tao Chung*[1], *Hung-Yi Lee*[2] *and Lin-Shan Lee*[2]

[1]Graduate Institute of Electrical Engineering, National Taiwan University
[2]Graduate Institute of Communication Engineering, National Taiwan University
r02921036@ntu.edu.tw, f01921031@ntu.edu.tw, tlkagkb93901106@gmail.com, lslee@gate.sinica.edu.tw

ABSTRACT

It is well known that recognizers personalized to each user are much more effective than user-independent recognizers. With the popularity of smartphones today, although it is not difficult to collect a large set of audio data for each user, it is difficult to transcribe it. However, it is now possible to automatically discover acoustic tokens from unlabeled personal data in an unsupervised way. We therefore propose a multi-task deep learning framework called a phoneme-token deep neural network (PTDNN), jointly trained from unsupervised acoustic tokens discovered from unlabeled data and very limited transcribed data for personalized acoustic modeling. We term this scenario "weakly supervised". The underlying intuition is that the high degree of similarity between the HMM states of acoustic token models and phoneme models may help them learn from each other in this multi-task learning framework. Initial experiments performed over a personalized audio data set recorded from Facebook posts demonstrated that very good improvements can be achieved in both frame accuracy and word accuracy over popularly-considered baselines such as fDLR, speaker code and lightly supervised adaptation. This approach complements existing speaker adaptation approaches and can be used jointly with such techniques to yield improved results.

Index Terms: speech adaptation, unsupervised learning, deep neural network, multitask learning, transfer learning

1. INTRODUCTION

Today most commercially available speech recognizers are user-independent, although it is well known that recognizers personalized to each individual user offer superior performance, because the speaker characteristics and language patterns of each individual user are captured in the recognition models. With the popularity of smartphones today, collecting personal audio data for each individual user is not difficult; annotating this data, however, is difficult. If for each individual user we could use a commercially-available user-independent recognizer to obtain a small quantity of his or her personal audio data (e.g., 10 to 50 utterances) and then make the necessary corrections to this data, this small bit of annotated data could be used together with his or her other much larger set of unlabeled personal audio data to train personalized acoustic models. This "weakly supervised"scenario is the focus of this paper.

Speaker adaptation has been investigated thoroughly not only in the past, but also recently, in particular within the years after the deep learning paradigm became mainstream in the global speech technology community. Improved regularization approaches have been developed for the adaptation training criterion [1][2][3]. Supplementary features are appended to the input to compensate for different acoustic conditions, as with i-vectors [4][5], underlying factors in joint factor analysis (JFA) [6], or the sequence summarizing neural network (SSNN) [7]. DNN's are also trained with a set of automatically-obtained speaker-specific features referred to as speaker codes [8][9]. Meanwhile, many groups use transformation-based schemes that treat speaker-independent (SI) neural networks as canonical models while adding additional linear hidden layers as speaker-dependent (SD) transformations either prior to the input layer – sometimes referred to as feature-discriminative linear regression (fDLR) [10] – or prior to the hidden layer [11][12] or to the output layer [13]. Alternatively, instead of modeling such additional transformations, the Hermitian-based activation function is used in adaptation while keeping the DNN weights fixed [14]. However, these methods all rely primarily on annotated adaptation data and do not take into account the abundant quantities of unlabeled personalized data.

Conventional approaches to use unlabeled data include unsupervised or "lightly supervised" adaptation, very often considered in low-resource speech recognition [15][16][17][18]. The basic idea is to use a speaker-independent (SI) model or a background model to transcribe the unlabeled raw audio data and generate approximate transcriptions used for training, sometimes as an iterative process. Further improvements are possible by for instance removing utterances with less reliable transcriptions or by selecting utterances with more reliable transcriptions [19][20], either based on confidence scores and other useful features, or by using models such as conditional random fields (CRF) [21][22]. In these approaches, all knowledge which can be extracted from the unlabeled data is based on either the SI or background model, or the limited annotated data. This raises the question: is there any knowledge that can be extracted directly from the unlabeled data?

In recent years it has become clear that acoustic tokens can be automatically discovered from unlabeled corpora in an unsupervised fashion [23][24][25][26]; these tokens have been shown successful for both spoken term detection and spoken document retrieval tasks [27][28][29]. This implies that these automatically discovered acoustic tokens correlate well to underlying linguistic units such as phonemes. The acoustic models for these tokens can be trained from unlabeled data, which in turn can be decoded into sequences of these tokens. Thus these tokens are extracted directly from the unlabeled data without using any other knowledge. It is therefore reasonable to consider if such tokens can be used jointly with the limited annotated data in the weakly supervised scenario considered here.

One major problem is that the direct relationship between these automatically discovered acoustic tokens and the phoneme labels is unknown and likely to be noisy. This problem may be solved by multi-task learning, in which multiple related tasks are trained simultaneously, for example with shared hidden layers in DNN, and thus benefit from each another. Multi-task learning has been shown to offer significant improvements in multilingual acoustic models because of cross-language knowledge transfer [30][31]. Such knowledge transfer across tasks is helpful for many reasons: local optima supported by more tasks may be better, more data can be used to learn the same set of parameters, some knowledge may be easier to learn in one task than in others, and so on [32][33].

In this paper, we propose a multi-task deep learning framework for the weakly supervised personalized acoustic modeling scenario mentioned above. In this framework, unlabeled data are transcribed into sequences of automatically discovered tokens, and this knowledge obtained straight from the unlabeled data is used jointly with the phonemes in the annotated data for multi-task learning, or by a DNN with shared hidden layers. The underlying basic idea is that the

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

high degree of similarity between the HMM states of the unsupervised token models and the supervised phoneme models may be mutually beneficial during multi-task learning. If we consider the unlabeled data with its corresponding token sequences to be a different language, then we can view the task as cross-language knowledge transfer for multi-lingual acoustic modeling. In Section 2 we present the proposed approach, and in Sections 3 and 4 we go through the experiments and results.

2. PROPOSED APPROACH

2.1. Automatic Discovery of Acoustic Tokens from Unlabeled Data

Acoustic tokens, which refer to short segments of sounds frequently occurring in a corpus, are discovered automatically by machine and are very similar to human-defined phonemes. Here we seek to discover automatically such acoustic tokens [23][24] from the personal audio data of each individual user. This can be achieved in the following way. Let $\overline{O}$ represent the acoustic feature vector sequences for the entire corpus (the audio data of a user). We begin with an initial set of tokens and the initial token label sequence W_0, including boundaries for each token for the observation $\overline{O}$ as in (1) below. We first use segmentation algorithms to divide the utterances in $\overline{O}$ into small signal segments based on discontinuities in the contours of energy- and MFCC-related features. We then compute the mean of the MFCC vectors for each of these small signal segments, and perform K-means clustering for the mean vectors over the whole corpus $\overline{O}$. We then assign to each cluster a token ID (this is the initial token set), with which we define W_0, the initial token label segments, including the boundaries over $\overline{O}$. We then fine-tune this initial label W_0 using the following iterative optimization approach: In each iteration i, given the label W_{i-1} over $\overline{O}$, we train the HMM model for each token using the short signal segments with the given token ID using the Baum-Welch algorithm. This yields a set of HMMs for all the tokens with parameters θ_i as in (2), which we then use to decode the whole corpus $\overline{O}$ to obtain a new label W_i as in (3). We then repeat (2)(3) iteratively until the convergence of the generated labels W_i, including the label boundaries.

$$W_0 = initialization(\overline{O}) \tag{1}$$

$$\theta_i = \arg\max_{\theta} P(\overline{O}|\theta, W_{i-1}) \tag{2}$$

$$W_i = \arg\max_{W} P(\overline{O}|\theta_i, W) \tag{3}$$

There are two key parameters for the token HMMs. The number of states in each token HMM, m, controls the token lengths, or the temporal granularity of the tokens. The initial total number of the clusters mentioned above or the distinct number of tokens, n, concerns the segmentation of the phonetic space, or the phonetic granularity of the tokens. Hence each parameter set $\psi = (m, n)$ defines a set of tokens learned from $\overline{O}$.

It is difficult to know the ideal parameter set for a given corpus, although when examining our data we noted the token set for $\psi = (3, 100)$ approximated gender-dependent phonemes and the token set for $\psi = (13, 300)$ roughly approximated syllables. Thus we more or less capture the characteristics and behavior of the underlying language described by the corpus by using multi-granular acoustic tokens, that is, by combining acoustic tokens with a variety of model granularities.

2.2. Phoneme-Token DNN (PTDNN)

The phoneme-token DNN (PTDNN) proposed in this paper is depicted in Fig. 1(a). At the top of this figure, the states for each phoneme HMM (in the left) and each token HMM (on the right) are the two tasks to be learned in parallel with a set of shared hidden layers and a shared feature discriminative linear regression (fDLR) transformation. The phoneme state probabilities are learned from the transcribed data (bottom left), while the token state probabilities are learned from the unlabeled data (bottom right) along with the transcribed data (bottom left). If, as suggested above, the acoustic tokens are considered a different language, this corresponds to the structure for multilingual acoustic modeling. Furthermore, acoustic token sets of multiple granularities $\psi = (m, n)$, or other features or labels, can also be learned jointly in this architecture, simply by adding more targets and the corresponding output networks in Fig. 1(a).

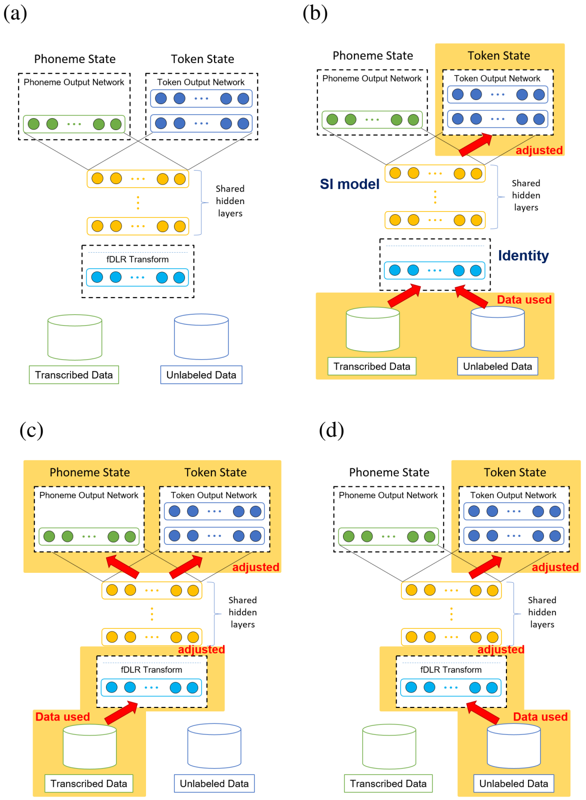

Fig. 1: (a) The phoneme-token DNN (PTDNN); (b) Training step 1: initialization; (c)(d) Training step 2: joint optimization.

2.3. Speaker Adaptation

Speaker adaptation involves the three steps summarized below and is shown in Fig. 1(b)(c)(d).

1. Initialization

   As shown in Fig. 1(b), we first take a speaker-independent (SI) DNN-HMM acoustic model trained on the large set of speaker-independent data, and use its hidden layers for the shared hidden layer needed for initialization. We also initialize the fDLR transformation with an identity matrix. With these parameters in the shared hidden layers and the fDLR transformation all fixed in this step, we train the token output network network only (upper right corner) on both the transcribed and unlabeled data of the personalized audio data set, where the transcribed data are also decoded into tokens to be used as the token state target.

2. Joint optimization

After initialization, we then jointly optimize the phoneme state and the token state output network iteratively. As in Fig. 1(c), we first use the limited set of transcribed data (bottom left) to train both the phoneme and the token targets with the objective function of (4) being the weighted sum of that for the two targets, and then use the large set of unlabeled data to train the acoustic token state only with an objective function of (5), as depicted in Fig. 1(d). In this way, we attempt to optimize the output networks of both targets synchronously and jointly. Thus we shuffle the transcribed and unlabeled data at a mini-batch level to facilitate the mutual learning of the included knowledge.

$$f = W_{phoneme} \cdot f_{phoneme} + W_{token} \cdot f_{token} \tag{4}$$

$$f = W_{token} \cdot f_{token} \tag{5}$$

3. Transferring back

   Finally, in the last step (not shown in the figure), in order to emphasize the desired phoneme state target, we further optimize the phoneme state output network only to transfer all the knowledge learned back, for phoneme recognition and to fine-tune the model.

In the whole training procedure, to prevent overfitting on the very small training set, we update only the parameters of the fDLR transformation and the output networks for phoneme and token states. As the size of the training set increases, we can then attempt to adjust more parameters, or even omit the above initialization step and just begin with step two to better fit the data.

3. EXPERIMENTAL SETUP

To better simulate the personalized recognizer scenario mentioned in the introduction, the experiments were performed on a Facebook post corpus we collected. Each of five male and five female speakers was asked to produce 1000 utterances, all extracted from his or her own Facebook posts in a spontaneous speech style; this yielded a 6.6-hour dataset. These utterances were primarily in Chinese but about 4.1% of the words were in English. We divided the 1000 utterances for each speaker into three sets: 500 utterances as the adaptation set, 250 as the development set, and 250 for testing. Also, we randomly selected as the transcribed data 50 utterances out of the 500 in the adaptation set.

The initial speaker-independent (SI) model was trained using a mixed corpus of the ASTMIC corpus (read speech in Mandarin, 31.8 hours) [34] and the EATMIC corpus (read speech in English produced by Taiwanese speakers, 29.7 hours) [34] with 4 hidden layers, each with 2048 units. The acoustic features used were the 13-dimensional MFCCs plus their first and second order delta features. The features were normalized to zero mean and unit variance, and a context window of 9 frames (4 frames on each side) was used. A trigram language model was used in the decoding, which was trained on data crawled from the PTT bulletin board system (BBS), a popular system in Taiwan with more than 1.5 million registered users and over 20000 new posts daily. Before training the model, we generated the personal acoustic token sets for each speaker using all 500 utterances of the adaptation data for the speaker. For the parameter set $\psi = (m, n)$, where m is the number of states in each token HMM and n is the number of distinct tokens in the set, we set $m = 3, 5, 9, 13$ and $n = 50, 100, 200, 300$, aiming for 16 sets of acoustic tokens, each with a different granularity.

As in Section 2.3 and Fig. 1(b)(c)(d), we initialized the token output network using the state-aligned labels for the given acoustic token set. We pretrained the token output network with a stacked RBM, and then fine-tuned the output network for 100 epochs with a relatively large learning rate of 0.01, after which we jointly optimized the output networks for phoneme and token states for 50 epochs iteratively with the learning rate set to 10^{-3}, $W_{phoneme} = 4$, and $W_{token} = 1$. Finally, we transferred the knowledge learned back to the phoneme output network for 50 epochs with a learning rate of 10^{-4}.

4. EXPERIMENTAL RESULTS

4.1. Speaker Adaptation Experiment

Models		Frame accuracy	Word accuracy
(A)	SI (DNN-HMM)	31.91%	57.45%
(B-1)	fDLR	41.66%	65.70%
(B-2)	Speaker code	42.11%	65.92%
(C)	Lightly supervised adaptation	45.04%	62.10%
(D-1)	PTDNN, $\psi = (5, 200)$	43.49%	66.94%
(D-2)	[PTDNN, $\psi = (5, 200)$] + fDLR	**48.74%**	**69.83%**

Table 1: Basic speaker adaptation results for the proposed PTDNN and the baselines

As mentioned above, we used all the 500 utterances of adaptation data for each speaker to generate the personal token sets. We also randomly selected a subset of 50 utterances out of the 500 as the transcribed data, and used the other 450 as the unlabeled data. The test was performed on the test set of 250 utterances, disjoint from the adaptation set for each speaker.

The results are listed in Table 1. As baselines we also include the SI model with DNN-HMM (row A), fDLR speaker adaptation (row (B-1)), speaker code adaptation (row (B-2)), and lightly supervised adaptation [15][16] (row (C)). In fDLR and speaker code in rows (B-1) and (B-2), only the 50 utterances of transcribed data were used in adaptation, while the remaining 450 utterances of unlabeled data were not used. In the lightly supervised adaptation in row (C), the 450 unlabeled utterances were first recognized by the SI models to produce machine-generated transcriptions. These data were then combined with the 50 utterances of transcribed data with human-generated transcriptions to form a complete set of 500 utterances for adaptation. We see that both fDLR and the speaker code improved significantly both the frame and word accuracies over the SI model (row (B-1)(B-2) vs (A)), while speaker code was slightly better than fDLR in the scenario considered here (rows (B-2) vs (B-1)). We also see that lightly supervised adaptation yielded similar improvements over SI models (rows (C) vs (A)), although it slightly underperformed fDLR and speaker code in word accuracy (rows (C) vs (B-1)(B-2)) but was better in frame accuracy (rows (C) vs (B-1)(B-2)), in the scenario considered, probably due to the added 450 utterances of adaptation data (which yielded better frame accuracy) and the errors in the machine-generated transcriptions (incorrect phoneme-to-word mappings can degrade word accuracy).

In row (D-1) is the proposed approach (PTDNN) with the parameter set $\psi = (5, 200)$ using 50 utterances of transcribed data and 450 utterances of unlabeled data. This parameter set was chosen from all the sets tested and yielded the best results. It is likely that this set of tokens $(5, 200)$ better modeled the linguistic units in the corpus tested (phonemes or high-frequency syllables, for example). We see that PTDNN outperformed fDLR (rows (D-1) vs (B-1)) by 1.83% in frame accuracy and 1.24% in word accuracy absolutely. It was also much better than speaker code in a similar way (rows (D-1) vs (B-2)), except by a slightly smaller range. Note that the proposed approach in row (D-1) did not suffer from the coarse tokens discovered from the unlabeled data, because the fine phoneme states and the coarse token states were jointly learned and converged in the iterative training step. PTDNN also outperformed lightly supervised adaptation (rows (D-1) vs (C)) in word accuracy (by 4.84% absolutely), although it was slightly worse in frame accuracy, probably because the 450 utterances of unlabeled data were transcribed by a supervised SI model in row (C) but used to produce the token set only in an unsupervised way in row (D-1). Thus the frame-level phoneme information that was used was more precise.

We further integrated PTDNN with $\psi = (5, 200)$ in row (D-1) with fDLR in row (B-1) via a weighted summing of the output state posteriors (weights selected with the development set) with the results in row (D-2). This resulted in the best performance in this series of experiments: it improved the frame accuracy by 7.08% and word accuracy by 4.13% over taking fDLR alone as the baseline (rows (D-2) vs (B-1)). This verified that PTDNN is complementary to other adaptation approaches and capable of yielding additive improvements.

4.2. Granularity parameter sets $\psi = (m, n)$

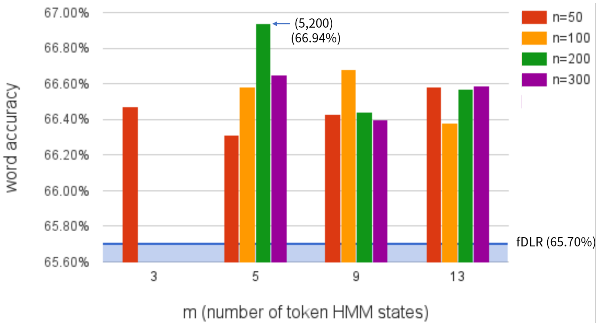

Fig. 2: Word accuracies for the proposed PTDNN with tokens sets for different parameters $\psi = (m, n)$

As mentioned above, there are many choices for the granularity parameter set $\psi = (m, n)$. In the experiments we chose $m = 3, 5, 9, 13$ and $n = 50, 100, 200, 300$. Out of the resulting 16 combinations of $\psi = (m, n)$, the token sets with $\psi = (3, 100), (3, 200)$, and $(3, 300)$ would not converge when generating the tokens in Equations (2) and (3) of Section 2.1, probably because such short models with $m = 3$ was too short for phonemes, but for a single speaker $n = 100, 200$, or 300 exceeded the number of phonemes by too much. As a result, many different redundant tokens were chosen, and therefore the sets did not converge.

The remaining 13 sets of parameters were successfully used in the PTDNN training. The word accuracies obtained with 50 utterances of transcribed data and 450 unlabeled data, corresponding to the last column of Table 1, are shown in Fig. 2. The blue horizontal line on the bottom is for the fDLR baseline (65.70%, row (B-1) of the table), while the highest bar is for $\psi = (5, 200)$ (66.94%, row (D-1)).

We observe that all the token sets offered reasonable improvements over fDLR regardless of the choice of (m, n); this suggests that the proposed approach is robust to the choice of these parameters. However different token sets yielded slightly different performance. The average performance of the 13 sets of tokens was 66.54%, 0.84% higher than fDLR absolutely. These results verify that these multi-granular tokens did indeed capture differing speaker-specific phonetic characteristics or information from the unlabeled data, and therefore that the approach was helpful here.

Also, we note that the better token sets may have to do with the phonetic structures of the underlying language for the data. For example, for the best token set $(5, 200)$, $m = 5$ could be a good number to model tokens close to phonemes, while $n = 200$ was not too far from the order of Chinese phonemes plus English phonemes. On the other hand, $m = 13$ could be a good number to model syllables. The majority of the data were in Mandarin, a syllable-based language. Of the roughly 400 Mandarin syllables, about 200 are frequently used. This may be why in Fig. 2 with $m = 13$, the results were consistent and reasonably good for most cases.

4.3. More or Less Transcribed Data and More Token Sets

We are also curious to know what happens when given greater or fewer numbers of transcribed utterances. In addition to the 50 utterances of transcribed data from Table 1, in Table 2 we list more word accuracy results when given 10 and 100 utterances of transcribed data (and 490 and 400 of unlabeled data respectively, always making 500 utterances of personalized data in total). The middle 50-utterance column results are copied over from Table 1, and two columns of results for 10 and 100 utterances are added on either side. Here rows (A)(B-1)(B-2)(C)(D) are copied from rows (A)(B-1)(B-2)(C)(D-1) in Table 1 for different sets of models. We observe reasonably degraded performance for 10 and reasonably improved performance for 100 utterances, as expected for the proposed approach (row (D)); the trend for other approaches is similar (row (D) vs rows (A)(B-1)(B-2)(C) for 10, 50, or 100 utterances). This demonstrates the suitability of the proposed approach for personalized acoustic modeling.

Word accuracy		Number of transcribed utterances		
Models		10	50	100
(A)	SI (DNN-HMM)	57.45%		
(B-1)	fDLR	60.08%	65.70%	67.73%
(B-2)	Speaker code	60.34%	65.92%	67.89%
(C)	Lightly supervised adaptation	61.45%	62.10%	63.18%
(D)	PTDNN, $\psi = (5, 200)$	63.05%	66.94%	68.65%
(E)	PTDNN with 4 token sets	**64.89%**	**67.15%**	**68.75%**

Table 2: Word accuracy for more or less transcribed data and more token sets

We also attempted integrating the knowledge learned from different token sets with different granularity parameters. This can be easily done by adding more sets of training targets and output networks to the architecture in Fig. 1. In the joint optimization step during training, we simply use the limited transcribed data to train all the targets, and then use the unlabeled data to train the token state targets. In row (E) of Table 2 we chose to integate token sets $(3, 50), (5, 200), (13, 50), (13, 300)$ to capture adequate diversity: slightly better results were obtained than with a single token set (rows (E) vs (D)). We observe that the knowledge learned from different token sets was slightly complementary, but yielded only limited extra improvement, perhaps because they were all extracted in a completely unsupervised way and therefore were less precise. In addition, in comparison to fDLR we see the proposed approach offered more improvements when given fewer transcribed utterances (rows (D) vs (B-1), 2.97%, 1.24% and 0.95% absolute improvements for 10, 50, and 100 utterances respectively). A similar situation is observe with respect to speaker code.

5. CONCLUSIONS

In this paper we propose for personalized acoustic modeling a weakly supervised multi-task deep learning framework based on acoustic tokens discovered from unlabeled data. Output networks for both phoneme states and acoustic token states are jointly learned iteratively during training, such that only very limited amounts of transcribed data need be used with a large set of unlabeled data in the proposed personalized recognizer scenario. Very encouraging initial experimental results were obtained.

6. REFERENCES

[1] J. Stadermann and G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models." in *ICASSP (1)*, 2005, pp. 977–980.

[2] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[3] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "Kl-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal*

Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 7893–7897.

[4] A. Senior and I. Lopez-Moreno, "Improving dnn speaker independence with i-vector inputs," in *Proc. ICASSP*, 2014.

[5] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[6] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5537–5541.

[7] K. Vesel, S. Watanabe, M. Karafi, J. Honza *et al.*, "Sequence summarizing neural network for speaker adaptation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5315–5319.

[8] O. Abdel-Hamid and H. Jiang, "Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7942–7946.

[9] S. Xue, O. Abdel-Hamid, H. Jiang, and L. Dai, "Direct adaptation of hybrid dnn/hmm model for fast speaker adaptation in lvcsr based on speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6339–6343.

[10] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[11] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ann/hmm models," *Speech Communication*, vol. 49, no. 10, pp. 827–835, 2007.

[12] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 366–369.

[13] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid nn/hmm systems," 2010.

[14] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian based hidden activation functions for adaptation of hybrid hmm/ann models." in *INTERSPEECH*, 2012.

[15] L. Lamel, J.-L. Gauvain, and G. Adda, "Unsupervised acoustic model training," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–877.

[16] ——, "Lightly supervised and unsupervised acoustic model training," *Computer Speech & Language*, vol. 16, no. 1, pp. 115–129, 2002.

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

[18] J. Z. Ma and R. M. Schwartz, "Unsupervised versus supervised training of acoustic models." in *Interspeech*, 2008, pp. 2374–2377.

[19] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[20] M. Doulaty Bashkand, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA (International Speech Communication Association), 2015, pp. 2897–2901.

[21] L. Sheng, A. Yuya, K. Tatsuya *et al.*, "Discriminative data selection from multiple asr systems' hypotheses for unsupervised acoustic model training," *SLP*, vol. 2015, no. 8, pp. 1–6, 2015.

[22] S. Li, Y. Akita, and T. Kawahara, "Data selection from multiple asr systems' hypotheses for unsupervised acoustic model training," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5875–5879.

[23] C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8081–8085.

[24] ——, "Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7814–7818.

[25] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[26] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models." in *INTERSPEECH*, 2011, pp. 1693–1692.

[27] Y.-C. Li, H.-y. Lee, C.-T. Chung, C.-a. Chan, and L.-s. Lee, "Towards unsupervised semantic retrieval of spoken content with query expansion based on automatically discovered acoustic patterns," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 198–203.

[28] H.-y. Lee, Y.-C. Li, C.-T. Chung, and L.-s. Lee, "Enhancing query expansion for semantic retrieval of spoken content with automatically discovered acoustic patterns," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8297–8301.

[29] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources." in *INTERSPEECH*, 2010, pp. 1676–1679.

[30] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[31] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[32] S. Thrun, "Is learning the n-th thing any easier than learning the first?" *Advances in neural information processing systems*, pp. 640–646, 1996.

[33] K. Kovac, "Multitask learning for bayesian neural networks," Ph.D. dissertation, University of Toronto, 2005.

[34] H.-y. Lee, S.-R. Shiang, C.-f. Yeh, Y.-N. Chen, Y. Huang, S.-Y. Kong, and L.-s. Lee, "Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 883–898, 2014.