Improving Large Vocabulary Accented Mandarin Speech Recognition with Attribute-based I-vectors

Hao Zheng[1], *Shanshan Zhang*[1], *Liwei Qiao*[2], *Jianping Li*[2], *Wenju Liu*[1]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, P.R.China
[2]Electric Power Research Institute of ShanXi Electric Power Company, China State Grid Corp
hzheng@nlpr.ia.ac.cn

Abstract

It has been well-recognized that the accent has a great impact on the ASR of Chinese Mandarin, therefore, how to improve the performance on the accented speech has become a critical issue in this field. The attribute feature has been proven effective on modelling accented speech, resulting in a significantly improved performance in accent recognition. In this paper, we propose an attribute-based i-vector to improve the performance of speech recognition system on large vocabulary accented Mandarine speech task. The system with proposed attribute features works well especially with sufficient training data. To further promote the performance on conditions such as resource limited condition or training data mismatched condition, we also develop Multi-Task Learning-Deep Neural Networks (MTL-DNNs) with attribute classification as the secondary task to improve the discriminative ability on Mandarin speech. Experiments on the 450-hour Intel accented Mandarin speech corpus demonstrate that the system with attribute-based i-vectors achieves a significant performance improvement on sufficient training data compared with the baseline DNN-HMM system. The MTL-DNNs complement the shortage of attribute-based i-vectors on data limited and mismatched conditions and obtain obvious CER reductions.

Index Terms: accented speech recognition, large vocabulary continuous speech recognition, attribute

1. Introduction

As one of the most important factors of speaker variability [1], accent plays an important role in automatic speech recognition (ASR) tasks. Chinese is a language with a number of accents, which is spoken extremely differently by speakers living in different dialectal regions of China [2]. As a result, compared to standard Mandarin recognition, performance degrades when dealing with accented speech. Due to the fact that speakers with strong accents can replace an unfamiliar phoneme in the language [3], which is absent in the standard pronunciation, the accented speech recognition is not properly handled by traditional acoustic model. Some speaker adaptation approaches may have effect on weakening the impact of strong accents. However, experiments show that it is preferable to have models only for a small number of large speaker populations than for many small groups [4], which indicates that the adaptation on accent-level is more useful than that on speaker-level when the speech is strongly accented.

Research has been carried out on dialectal or foreign accented speech recognition [5]. General speaker adaptation techniques, such as fMLLR [6] and MAP [7], are applied to fit the characteristics of foreign accents [8]. Besides the feature transformation, some multilingual approaches, such as multilingual HMM [9] and recognizer combination method and multilingual acoustical models [10, 11], are also applied for accented speech recognition [12].

The speech spoken by speakers with different accents differ in the phonetic pronunciation. Meanwhile, the knowledge-based modelling, such as attribute feature [13], which is known to be linguistically and phonetically relevant [14], has an excellent representation ability on speaker pronunciation. Experiments in [15] demonstrate that the attribute feature has a strong capacity on modelling accented speech. However, traditional acoustic features for speech recognition, such as mel frequency cepstrum coefficient (MFCC) and mel filter bank (FBANK), along with the frame-level attribute feature [16], are all short-term features. The frame-level attributes that extracted from traditional acoustic features contains no more information than the traditional features. The utterance-level attribute features, which are the statistics of frame-level attributes, are more competent in providing accent information for DNNs. The i-vector approach [17] has been proven to be successful in speaker [18], language [19] and accent [20] recognition and speaker adaptation for LVCSR [21]. By extracting statistic information, the i-vector feature contains information of speakers, accents, environments, channels, etc.

Besides the attribute features, multi-task learning provides a model-based discrimination for knowledge integration. In the conventional deep neural network - hidden Markov model (DNN-HMM) hybrid systems, neural networks are typically trained for one task of senones classification. Recent works try to take advantage of the multi-task learning approach to boost the generalization performance of senones classification by jointly learning related tasks, such as grapheme classification [22] and phoneme/state context learning [23]. These studies show that when the classifier uses the same architecture of network to perform more than one related task, it learns the shared structure of tasks.

In this article, we propose an attribute-based feature, which is extracted with the i-vector methodology and represents a strong discrimination on accented Mandarin speech, to improve the performance of conventional DNN acoustic models on accented Mandarin speech recognition task. The system with proposed attribute features works well especially on sufficient training data. To further improve the system on resource-limited and data-mismatched conditions, MTL-DNNs with shared input and hidden layers and individual output layer for each task are jointly trained on a uniform loss criterion. We evaluate and analyze the performance of attribute knowledge integration with different dimensional attribute features, different amounts of training data and data-mismatched conditions. Experimental results demonstrate that the proposed framework achieves a significant performance improvement compared with baseline DNN systems.

Table 1: *Phonological features (attributes) and their associated phones used in this study.*

	Attribute	Phonemes
initial	Voiced	m n l r y w
	Voiced nasal	m n
	Lateral	l
	Stop	b p d t g h
	Fricative	z c zh ch j q
	Retroflex	zh ch sh r
	Alveolar	z c s
	Affricate	f s sh x h r
Final	Simple vowel	ia a e o i u v er
	Head-dominant	ai ei ao ou
	Centre-dominant	iao iou uai uei
	Tail-dominant	ia ua uo ve
	Front nasal	an ian van uan in en uen ven vn
	Back nasal	ang iang uang eng ong ing iong
Silence	Silence	sil

2. The attribute of Mandarin

The speeches spoken by speakers with different accents differ in the phonetic pronunciation. For Mandarin speech, one example is that the Shanghainese tend to replace the standard retroflex fricatives and affricates /zh/ch/sh/ with their alveolar equivalents /z/c/s/, in which case the attributes retroflex and alveolar have obvious discriminations between standard Mandarin and Shanghai-accented Mandarin. Another example is that some people from south China, such as Wuhanese and Chengdunese, tend to replace voiced nasal /n/ with lateral /l/ and back nasals with their corresponding front nasals. In consideration of these characteristics of Chinese accented pronunciation, we select 15 phonological features (attributes) as shown in Table 1. The usual set of 61 Mandarin phone classes (including silence, which is presented as sil in this paper) are then mapped to 15 pairs of 2-dim features, which is 30-dim in total. The value of each pair of phonological features indicates the presence/absence of this attribute in the pronunciation of this phone, with $(1, 0)$ meaning the presence and $(0, 1)$ meaning the absence.

3. Attribute-based i-vector

The DNNs for acoustic modelling in ASR are designed to reveal the text content and be invariant to other information including gender, accent, channel, etc. However, over-fitting problem occurs in practice. The attribute-based i-vector with a strong discrimination on accents is applied to DNNs as accent adaptation in a sense. By this approach, DNNs tend to normalise the signal with respect of the accent information and as a result to be more relevant to the target text.

3.1. The extraction of attribute

A three-layer DNN with 1024 sigmoid units for each layer is utilized as the attribute extractor. The input to the DNN can be any speech features, and we use conventional 120-dim FBANK+$\triangle$+$\triangle\triangle$ along with their left and right contexts in this paper. The number of output units is 30, with presence and absence for each attribute, as described in section 2. To further improve the performance of the attribute extractor, we have tried the MTL-DNN with context-dependent state classification as the secondary task is realized to promote the discrimination on attributes. However, the MTL-DNN attribute extractor shows comparable performance with the STL-DNN extractor. So all the extractors in this article are traditional STL-DNNs.

Given an input vector $\mathbf{x}$, each pair of attribute exactor outputs denotes the presence probability $p(s_i^p|\mathbf{x})$ and absence probability $p(s_i^a|\mathbf{x})$ of the target class i. The two outputs are transformed by a softmax function, which normalizes the two values into the range of $(0, 1)$ and to have a summation of 1. Then the 15 pairs of outputs are concatenated to form a 30-dim vector to be the input of i-vector modelling. Different from experiments in [15], which prove the effectiveness of i-vector modelling with long-term attributes in foreign accent recognition, in this paper, short-term attributes in frame-level are used for i-vector modelling, since short-term features contain more detailed information which is more helpful for speech recognition.

3.2. I-vector modelling

The classical extraction of i-vector is based on the total variability model (TVM) presented by N. Dehak in [17]. An arbitrary duration utterance is represented by a several hundred-dimensional vector in this model. Each utterance is first represented by its zero- and first-order Baum-Welch (BW) statistics extracted from the universal background model (UBM), which is a Gaussian mixture model (GMM) trained with a large amount of speech to represent the distribution of features. Then the super-vector $\boldsymbol{m}$ which is composed by stacking the first-order BW statistics of the utterance, is projected onto the total variability space according to the generative equation:

$$\boldsymbol{m} = \boldsymbol{m}_0 + T\boldsymbol{w} \tag{1}$$

where $\boldsymbol{m}_0$ is the mean vector which is generally taken to be the UBM supervector, T is a rectangular matrix of low rank and the i-vector $\boldsymbol{w}$ is a random vector with standard normally distributed prior. In this modelling, $\boldsymbol{m}$ is normally distributed with mean vector $\boldsymbol{m}_0$ and covariance matrix TT^{T}. The estimations of total variability matrix T and latent variable $\boldsymbol{w}$ are realized by Expectation-maximization (EM) algorithm. For each utterance, the i-vector is the maximum a posteriori (MAP) point estimate of the latent variable $\boldsymbol{w}$.

3.3. Improving DNNs with i-vectors

The utterance-level i-vectors are then appended to the original frame-level acoustic features, as described in [21]. Given a context window with c frames of d dimensional acoustic features and v dimensional i-vector, the augmented feature of $cd + v$ dimension is provided as the input to DNNs. All frames from a given utterance are augmented with the same v dimensional utterance i-vector. Frames from training data are randomly selected while batch-based training to prevent the model from over-fitting the continues appearance of the same i-vector.

4. Multi-task learning

Multi-task learning [24] is a machine learning technique that improves single-task learning by training the model with several related tasks using a shared representation. The effectiveness of multi-task learning depends on the relations between each task and the shared learned structure across the tasks [24]. These secondary tasks are used for the training stage and are dropped while testing the unseen data.

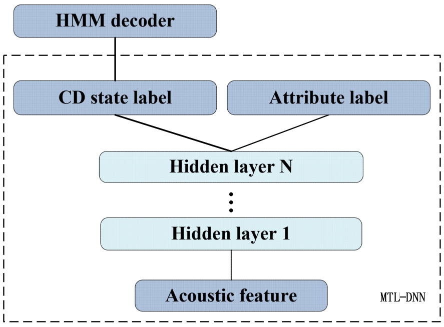

Figure 1: An MTL-DNN system for the joint training of context-dependent state (CD state) and attribute label.

4.1. Understanding multi-task learning

One aspect of the effectiveness of secondary task learning, which is similar to dropout strategy and sparse penalty in a sense, can be explained as a regularization to avoid over-fitting. As a result, the multi-task learning is effective especially when the training data is limited or mismatched with the test data, in which case the over-fitting problem is more likely to occur. By adding extra knowledge-based targets, secondary tasks weaken the excessive dependence between the model and the primary task.

Multi-task learning can also improve the performance of model by taking advantage of extra information, including accent, speaker, lingual and so on. Taking the MTL-DNN for example, the secondary task learning increases the discrimination of the hidden layer outputs on these extra areas, which leads to a more discriminative hidden layer for primary task classification.

4.2. Applying attribute classification as the secondary task

In order to improve the performance of the speech recognition task, we apply attribute classification as the secondary task to integrate articulatory knowledge to conventional STL-DNN. As shown in Figure 1, the MTL-DNN uses shared input layer and hidden layers but individual output layers for each task. Both of the labels are used to update the weights, including shared weights and individual weights, in the training stage. And in the test stage, only the output layer of context-dependent state classification task is used to calculate the posteriors for the H-MM decoder.

The MTL-DNN is trained to minimize the objective function as follows:

$$E = (1 - \alpha)E_p + \alpha E_s, \tag{2}$$

where E_p and E_s denote loss functions of primary and secondary tasks separately, and α is the weight that controls the proportion of gradients which are calculated from the secondary task. We use the factor $(1 - \alpha)$ to scale E_p so that the summation of the two factors is 1, in which case there's no need for scaling the learning rate.

5. Experiments

5.1. Corpus and setup

The experiments are carried out on the 450-hour (gender balanced) Intel accented Mandarin speech corpus. Six typical accents, consisting of Beijing (BJ), Chengdu (CD), Guangzhou (GZ), Haerbin (HB), Shanghai (SH) and Wuhan (WH), are considered. The details of speakers (SPKs) and utterances (UTTs) of training (TRN), development (DEV) and test (TST) sets are shown in Table 2.

Table 2: *Summary of intel corpus used in this article.*

	BJ	CD	GZ	HB	SH	WH
TRN_SPK	213	105	231	101	207	105
DEV_SPK	4	4	4	4	4	4
TST_SPK	20	20	20	20	20	20
TRN_UTT	83490	41446	89737	39897	80506	31636
DEV_UTT	1527	1550	1414	1573	1396	1572
TST_UTT	1000	1000	1000	1000	1000	1000

Table 3: *CERs (%) of DNNs with different dimensions of attribute-based and MFCC-based i-vector input.*

dimension	0	50	100	150	200	300
MFCC	14.57	14.27	13.84	13.85	**13.76**	14.05
Attribute	14.57	11.35	11.08	11.89	**11.02**	11.02

Conventional 40-dim FBANK feature, along with their first and second order derivatives, are used as feature of each frame. The features of 15 consecutive frames are combined as input features of the baseline DNN and feature extractor described in section 3. Each dimension of the input features (including attribute features) is normalized to have zero mean and unit variance over the whole training set. Mini-batches of size 256 are used for all batch-based training procedure. The training labels of DNNs are generated by a well-trained GMM-HMM systems with 8152 tied context dependent HMM states. A 3-gram language model (LM) is used during the recognition.

All the networks, which are set with 6 hidden layers and 2048 sigmoid nodes for each layer, are initialized with greedy layer-wised pre-training [25]. The cross-entropy discriminative training is done by back-propagation algorithm to fine-tune the parameters with exponentially decaying learning rates, which are set to 0.008 at start. We choose 0.2 as the secondary task weight (α in Eq. (2)) after having tested a list of values. As the momentum is very tricky which may affect the final results significantly [26], and we just intend to make a fair comparison between different approaches in the experiment, the momentum is not used as a part of our experiments.

For i-vector extraction, a full-covariance UBM with 1024 Gaussian components and TVMs of corresponding dimensions are trained with all the speeches in the training set. Systems operate on two kinds of features, the MFCC and the attribute, both appended with their first and second order derivatives computed over a 25ms window every 10ms.

5.2. Analysis of attribute feature

Firstly, we experiment on different dimensions of i-vector features. DNNs with attribute-based and MFCC-based i-vector inputs are carried out. As shown in Table 3, DNNs with attribute-based i-vectors (attribute features) outperform DNNs with MFCC-based i-vectors in all dimensions. Both the two approaches achieve their best performance with 200-dimension. Then, the dimension of attribute-based i-vector applied in the following experiments is set to 200.

Table 4: *Comparison of different systems using STL-DNN and MTL-DNN with original FBANK feature and attribute+FBANK feature. The data is selected randomly. The performance are given in CERs (%) for all test conditions.*

Amount of data	50h	100h	200h	400h
STL-DNN	18.40	16.68	15.69	14.57
STL-DNN+Attribute	17.80	15.18	13.22	11.02
MTL-DNN	**16.65**	16.01	15.23	14.45
MTL-DNN+Attribute	16.68	**14.11**	**12.85**	**10.52**

5.3. The influence of training data amount

The experiments are repeated with DNNs trained on different hours of speech to evaluate the relationship between system performance and training data size. The results are shown in Table 4. We observe that all systems achieve CER reductions with the increase of training data. However, more relative reductions are obtained by the MTL-DNNs when the training data size is limited and, oppositely, more relative reductions are attained by DNNs with attribute features (STL-DNN+Attribute) with more training data. The reason is that the multi-task learning improves the DNN training by providing extra information which helps the parameters estimation and prevents from over-fitting, which occurs especially with limited training data. DNNs with additional attribute features learn more complex structures than that with original acoustic features, which leads to a requirement for more training data to estimate the parameters. The hybrid structure (MTL-DNN+Attribute) of the two approaches integrates both of their advantages and overcomes the shortcomings, and results in a further improvement (27.8% relative reduction on CER over DNN baseline) compared with the baseline DNN system and each individual approach with the 400-hour training data.

5.4. The mismatched condition

To systematically evaluate the performance of the proposed approach and other comparison methods on data mismatched condition, we remove Shanghai-accented and Guangzhou-accented speeches out of the training and development sets, after which the amount of training data reduces from 400 to about 210 hours. The attribute extractor is trained with the 210-hour data, however, the UBM for modelling i-vector is trained with the 400-hour training data, this is because the UBM training is done with the unlabeled data, which is easy to get in practice. The results in Table 5 demonstrate that the STL-DNN with attribute feature and the MTL-DNN with original FBANK feature outperform the baseline DNN on both in-domain and out-domain test sets. It indicates that the additional attribute feature leads to a generalization by providing the pronunciation information to the DNN. By supplying extra knowledge and avoiding overfitting, the MTL-DNN achieves more obvious improvement on out-domain test sets. The hybrid architecture with both of their advantages achieves a further improvement on in-domain set compared with the MTL-DNN and outperforms the baseline STL-DNN on all test sets.

6. Discussions

By providing the additional utterance-level attribute feature as input, DNNs achieve valuable improvement (24.4% relative CER reduction with 400-hour training data) compared with baseline DNNs. However, with limited training data, the additional attribute features lead to a limited improvement due to the over-fitting problem, which indicates that more training data is required for DNNs with attribute features. With the increase of training data, the effect of attribute feature enhances and becomes significant. In mismatched condition, DNN with attribute feature achieves obvious improvements on out-domain test sets compared with the baseline DNN.

Table 5: *Comparison of different systems using STL-DNN and MTL-DNN with original FBANK feature and attribute+FBANK feature on mismatched condition. The performance are given in CERs (%) for 2 mismatch test sets (including SH and GZ), average of matched sets and average of all.*

Accent	SH	GZ	others	all
STL-DNN	18.67	25.06	13.23	16.28
STL-DNN+Attribute	15.71	22.41	10.76	13.61
MTL-DNN	17.38	23.69	12.75	15.42
MTL-DNN+Attribute	**15.31**	**21.76**	**10.13**	**13.04**

Multi-task learning boosts the system performance by providing additional knowledge and avoiding over-fitting. Opposite to DNNs with additional attribute features, MTL-DNNs produce the desired improvement especially when training data is limited. A 9.51% relative CER reduction with 50-hour training data is achieved, compared with the DNN baseline. Experimental results also show the robustness of MTL-DNNs in the mismatched condition. Significant improvements are achieved on both Guangdong (5.5% relative) and Shanghai (6.9% relative) test sets by the MTL-DNN.

The combination of the modifications in both feature and model domains keeps up both of their superiorities and recovers the weaknesses. With limited training data, which is insufficient to estimate the parameters of DNNs with additional attribute features, the MTL-DNN boosts the capability by supplying extra information, which makes parameters trained more sufficiently. With training data increasing, although MTL-DNNs loss some relative CER reduction, the supplement of attribute features achieves a further improvement on the MTL-DNN.

7. Conclusions

To handle with accented Mandarin speech, we propose an attribute-based i-vector feature, which represents a strong discrimination on accented Mandarine speech, to improve the performance of conventional DNN acoustic modelling on accented Mandarin speech recognition tasks. The system with proposed features works well with sufficient training data but resource limited and data mismatched conditions. To solve this problem, the multi-task learning with attribute classification as the secondary task is provided. Experimental results show that the DNNs with attribute-based i-vectors obtain significant reductions with sufficient training while DNNs with multi-task learning attain more CER reductions with limited or mismatched training data. The proposed approaches on feature and model domains complement each other's shortcomings, therefore the combination of them is reasonable to achieve a further improvement over individual approach.

Additionally, the proposed architecture is applied on utterances without any speaker or accent information provided, thus no extra adaptation or model storage is required, which have a potential to realize for online speech recognition. As future works, we plan to investigate our architecture to other languages and online speech recognition task.

8. References

[1] C. Huang, T. Chen, S. Z. Li, E. Chang, and J.-L. Zhou, "Analysis of speaker variability." in *INTERSPEECH*, 2001, pp. 1377–1380.

[2] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin." in *Interspeech*. Citeseer, 2005, pp. 217–220.

[3] K. Bartkova and D. Jouvet, "On using units trained on foreign data for improved multiple accent speech recognition," *Speech Communication*, vol. 49, no. 10, pp. 836–846, 2007.

[4] V. Beattie, S. Edmondson, D. Miller, Y. Patel, and G. Talvola, "An integrated multi-dialect speech recognition system with optional speaker adaptation," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[5] Y. Huang, D. Yu, C. Liu, and Y. Gong, "Multi-accent deep neural network acoustic model with accent-specific top layer using the kld-regularized model adaptation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[7] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[8] G. Zavaliagkos, R. Schwartz, and J. Makhoul, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 676–679.

[9] L. Tomokiyo, "Recognizing non-native speech: Characterizing and adapting to non-native usage in speech recognition," 2001.

[10] V. Fischer, E. Janke, and S. Kunzmann, "Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual hmms," in *Seventh International Conference on Spoken Language Processing*, 2002.

[11] 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.

[12] Z. Wang, T. Schultz, and A. Waibel, "Comparison of acoustic model adaptation techniques on non-native speech," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1. IEEE, 2003, pp. I–540.

[13] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4169–4172.

[14] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[15] H. Behravan, V. Hautamauki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "Introducing attribute features to foreign accent recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5332–5336.

[16] L. Q. J. L. Hao Zheng, Zhanlei Yang and W. Liu, "Attribute knowledge integration for speech recognition based on multi-task learning neural networks." in *INTERSPEECH*, 2015, pp. 1377–1380.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[18] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[19] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861–864, 2011.

[20] A. DeMarco and S. J. Cox, "Iterative classification of regional british accents in i-vector space." in *MLSLP*, 2012, pp. 1–4.

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

[22] D. Chen, B. Mak, and S. Sivadas, "Joint sequence training of phone and grapheme acoustic model based on multi-task learning deep neural networks," in *INTERSPEECH*, 2014.

[23] M. L. Seltzer and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6965–6969.

[24] R. Caruna, "Multitask learning: A knowledge-based source of inductive bias," in *Machine Learning: Proceedings of the Tenth International Conference*, 1993, pp. 41–48.

[25] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," *Advances in neural information processing systems*, vol. 19, p. 153, 2007.

[26] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1139–1147.