Boosting Universal Speech Attributes Classification with Deep Neural Network for Foreign Accent Characterization

Ville Hautamäki[1], *Sabato Marco Siniscalchi*[2,3], *Hamid Behravan*[1], *Valerio Mario Salerno*[2] *and Ivan Kukanov*[1]

[1]School of Computing, University of Eastern Finland, Finland
[2]Faculty of Architecture and Engineering, University of Enna "Kore", Italy
[3]School of ECE, Georgia Institute of Technology, USA

villeh@cs.uef.fi

Abstract

We have recently proposed a universal acoustic characterisation to foreign accent recognition, in which any spoken foreign accent was described in terms of a common set of fundamental speech attributes. Although experimental evidence demonstrated the feasibility of our approach, we belive that speech attributes, namely manner and place of articulation, can be better modelled by a deep neural network. In this work, we propose the use of deep neural network trained on telephone bandwidth material from different languages to improve the proposed universal acoustic characterisation. We demonstrate that deeper neural architectures enhance the attribute classification accuracy. Furthermore, we show that improvements in attribute classification carry over to foreign accent recognition by producing a 21% relative improvement over previous baseline on spoken Finnish, and a 5.8% relative improvement on spoken English.

Index Terms: Deep neural networks, data-driven speech attributes, manner of articulation, place of articulation, i-Vector, foreign accent recognition

1. Introduction

In automatic foreign accent recognition the mother tongue (L1) of non-native speakers has to be recognised given a spoken segment in a second language (L2) [1]. We may think of L1 recognition as a language recognition task [2], where L1 is the target language to be recognised. However, language recognition techniques based on n-gram phoneme statistics are not directly usable, as the collected phoneme statistics would match the L2 language. In our previous work [3], we advocated the use of speech attributes, namely manner and place of articulation, to universally characterise all language and accents, and experimental evidence proved their effectiveness in foreign accent recognition. Foreign accent variation is a nuisance factor that negatively affects automatic speech, speaker and language recognition systems [4, 5]. Most of the speech technology systems have been tailored to native speech, but those systems rarely work well on non-native or accented speech, such as the *automatic speech recognition* (ASR) [6, 7].

The most direct way to overcome the problem of non-native speech is to train separate statistical models for each L1-L2 pair. But by using the accent universal units, we would be able to *compensate* against the L1 nuisance effects. Similarly, such units can be used in foreign accent conversion [8] with the idea of reducing the perceptual effect of accentedness. In [8], the accent universal units were articulatory gestures, namely manner and place of articulation recorded using the *electromagnetic articulography* (EMA). Accent conversion is achieved by obtaining parallel audio and EMA recordings from the L1 and L2 targets. Being limited to EMA recordings to obtain articulatory gesture scores is by its vary nature practically very restricted. The *automatic speech attribute transcription* (ASAT) framework [9], is bottom-up detection-based framework, where speech attributes are extracted using data-driven machine. We were able to successfully use these detector scores in foreign accent recognition [3], and regional dialect recognition [10] by modeling the stream of detector scores using the i-Vector methodology [11]. In contrast to phonotactic language recognition systems, the i-Vector based method defers all decisions until the final accent recognition is made. Experimental results demonstrated the effectiveness of our i-Vector modelling of attributes, and a significant system performance improvement over conventional spectrum-based techniques was demonstrated on the Finnish national foreign language certificate corpus. Nonetheless, we also observed that some speech attributes were not properly modelled by the shallow neural networks (SNN), employing a single-hidden non-linear layer. In fact, the baseline speech attribute front-end exhibit a large error rate variance [12].

We believe that accent recognition accuracy can be greatly enhanced if more powerful data-driven learning systems replace shallow networks for speech attribute modelling. Deep neural networks (DNNs), e.g., [13], have been successfully applied across a range of different speech processing tasks in recent years, such as conversational-style speech recognition, e.g., [14], noise robust applications [15], multi- and cross-lingual learning techniques, e.g., [16]. Inspired by the success of those applications, we want to explore the use of DNNs to extract manner and place of articulation attributes to be used in automatic accent recognition systems. DNNs are chosen because they (i) can be easily trained on high dimensional features, (ii) have the potential to learn more efficient and effectively non-linear feature mappings, and (iii) may better capture the complex relationships among speech attributes. Two speech attribute classifiers for manner and place of articulation, respectively, are built using DNNs trained on telephone bandwidth speech material from the six different languages in the OGI Multi-language Telephone Speech corpus [17]. We show that improved attribute modeling positively affects foreign accent recognition performance, and we observed an relative performance improvement over our previous results of 21% and 5.8% on the Finnish and English as second language task, respectively.

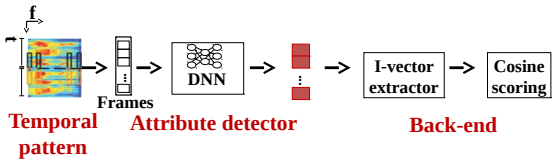

Figure 1: Block diagram of the proposed system. The DNN based attribute detector outputs one feature vector per input frame. We use i-Vector approach with cosine scoring to classify target accents.

2. Foreign Accent Recognition System

Figure 1 shows the proposed foreign accent recognition system. The front-end consists of DNN based classifier that generates either manner or place of articulation posterior probabilities. Finally, the sequence of features of a one utterance is compressed into one i-Vector as representative of that utterance.

2.1. Speech attribute extraction

Manner of articulation classes, namely, **glide, fricative, nasal, stop, and vowel**, and *place of articulation* classes, namely **coronal, dental, glottal, high, labial, low, mid, palatal, and velar**, are the speech attributes used in this work. Speech attributes can be obtained for a particular language and shared across many different languages, and they can thereby be used to derive a universal set of speech units. Furthermore, data-sharing across languages at the attribute level is naturally facilitated by the nature of these classes as shown in [18]. In [19], the authors have demonstrated that manner and place of articulation attributes can compactly characterise any spoken language along the same lines as in the ASAT paradigm for ASR [9]. Furthermore, it was shown that off-the-shelf data-driven attribute detectors built to address automatic language identification tasks [18] can be employed without either acoustic adaptation or re-training for characterising speaker accents never observed before [3]. In [3], attribute detectors were built using shallow neural networks, namely single-hidden layer, feed-forward neural networks. Here we want to test deeper architectures.

In DNNs, hidden layers are usually constructed by sigmoid units, and the output layer is a softmax layer. The values of the nodes can therefore be expressed as:

$$\mathbf{x}^i = \begin{cases} \mathbf{W}_1\mathbf{o}_t + \mathbf{b}_1, & i = 1 \\ \mathbf{W}_i\mathbf{y}^{i-1} + \mathbf{b}_i, & i > 1 \end{cases}, \tag{1}$$

$$\mathbf{y}^i = \begin{cases} \mathrm{sigmoid}(\mathbf{x}^i), & i < L \\ \mathrm{softmax}(\mathbf{x}^i), & i = L \end{cases}, \tag{2}$$

where W_1, and W_i are the weight matrices, $\mathbf{b_1}$, and $\mathbf{b_i}$ are the bias vectors, $\mathbf{o}_t$ is the input frame at time t, L is the total number of the hidden layers, and both sigmoid and softmax functions are element-wise operations. The vector $\mathbf{x}^i$ corresponds to pre-nonlinearity activations, and $\mathbf{y}^i$ and $\mathbf{y}^L$ are the vectors of neuron outputs at the i^{th} hidden layer and the output layer, respectively. The softmax outputs were considered as an estimate of either manner, or place posterior probability according to the set of attributes that we want to model:

$$p(C_j|\mathbf{o}_t) = \mathbf{y}_t^L(j) = \frac{\exp(\mathbf{x}_t^L(j))}{\sum_i \exp(\mathbf{x}_t^L(i))}, \tag{3}$$

where C_j represents the j^{th} manner (or place) and $\mathbf{y}^L(j)$ is the j^{th} element of $\mathbf{y}^L$. The DNN is trained by maximizing the log posterior probability over the training frames. This is equivalent to minimizing the cross-entropy objective function. Let $\mathcal{X}$ be the whole training set, which contains T frames, *i.e.* $\mathbf{o}_{1:T} \in \mathcal{X}$, then the loss with respect to $\mathcal{X}$ is given by

$$\mathcal{L}_{1:T} = -\sum_{t=1}^{T}\sum_{j=1}^{J} \tilde{p}_t(j) \log p(C_j|\mathbf{o}_t), \tag{4}$$

where $p(C_j|\mathbf{o}_t)$ is defined in Eq. (3); $\tilde{p}_t$ is the target probability of frame t. In real practices of DNN systems, the target probability $\tilde{p}_t$ is often obtained by a forced alignment with an existing system resulting in only the target entry that is equal to 1. Mini-batch stochastic gradient descent (SGD) [20] was used to update all neural parameters during training. Pre-training was performed to initialise DNN parameters [21].

2.2. i-Vector Modeling

The idea behind i-Vector model is that the feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, where N is the number of speech attribute feature vectors, can be compressed into a fixed length vector. All variability, such as accent speaker and channel, are retained in that representation of an utterance. For that reason, i-Vector model is also called *total variability modeling* [11]. It stems from the idea that feature stream can be modeled by *Gaussian mixture model* (GMM) that is adapted by relevance *maximum a posteriori* (MAP) from the *universal background model* (UBM). Then stacking the adapted GMM mean vectors creates a fixed length representation of the utterance. But the dimensionality of the GMM supervector space is very high, easily more than 100000. In the i-Vector model, the utterance dependent supervector $\mathbf{M}$ is defined to be [11]:

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} + \epsilon, \tag{5}$$

where $\mathbf{m}$ is the utterance independent mean vector, copied from the UBM by stacking the mean vectors, $\mathbf{T}$ is a rectangular low rank matrix and the latent vector $\mathbf{w}$ is distributed according to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the $\mathbf{T}$ represents the captured variabilities in the supervector space and ϵ captures the residual variability. The residual is distributed $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is copied directly from the GMM. The $\mathbf{T}$-matrix is estimated from the held-out corpus, typically same as the where UBM is estimated from, via an *expectatiom maximization* (EM) algorithm [22]. The idea of the algorithm is that we infer $\mathbf{w}$, which is the posterior mean, for each training utterance given an estimate of $\mathbf{T}$-matrix and then estimate new $\mathbf{T}$-matrix and so on. The estimation is very CPU intensive, so typically only few, for example five, iterations is used in practice.

2.3. Inter-session Variability Compensation

As the extracted i-Vector contain both within- and between accents variation, we use dimensionality reduction technique to project the i-Vectors onto a space to minimize the within-accent and maximize the between-accent variation. To perform dimensionality reduction, we used *heteroscedastic* linear discriminant

Table 1: Manner of articulation accuracies on DNN with one hidden layer and six hidden layers on OGI-TS corpus.

Attribute	1 layer	6 layers
fricative	69.2	**72.0**
glide	27.6	**30.0**
nasal	75.3	**76.8**
silence	**92.5**	92.3
stop	72.3	**75.4**
vowel	91.4	**91.6**
Total	79.2	**80.1**

Table 2: Place of articulation accuracies on DNN with one hidden layer and six hidden layers on OGI-TS corpus.

Attribute	1 layer	6 layers
coronal	55.0	**57.7**
dental	27.7	**32.5**
glottal	39.1	**43.3**
high	54.0	**56.5**
labial	53.3	**56.4**
low	66.0	**68.5**
mid	61.6	**62.3**
palatal	42.3	**45.6**
silence	**93.8**	93.4
velar	49.4	**56.2**
Total	61.8	**63.7**

analysis (HLDA) [23], as it allows to use higher output dimensionality than then number of classes. HLDA is considered as an extension of *linear discriminant analysis* (LDA). In this technique, i-Vector of dimension n is projected into a p-dimensional feature space with $p < n$, using HLDA transformation matrix denoted by $\mathbf{A}$. The matrix $\mathbf{A}$ is estimated by an efficient row-by-row iteration with EM algorithm as represented in [24].

Followed by HLDA, *within-class covariance normalization* (WCCN) is then used to further compensate for unwanted intra-class variations in the total variability space [25]. The WCCN transformation matrix, $\mathbf{B}$, is trained using the HLDA-projected i-Vectors obtained by Cholesky decomposition of $\mathbf{B}\mathbf{B}^\top = \mathbf{\Lambda}^{-1}$, where a within-class covariance matrix, $\mathbf{\Lambda}$, is computed using,

$$\mathbf{\Lambda} = \frac{1}{L}\sum_{a=1}^{L}\frac{1}{N}\sum_{i=1}^{N}(\mathbf{w}_i^a - \overline{\mathbf{w}}_a)(\mathbf{w}_i^a - \overline{\mathbf{w}}_a)^\top, \qquad (6)$$

where $\overline{\mathbf{w}}_a$ is the mean i-Vector for each target accent a, L is the number of target accents and N is the number of training utterances in target accent a. The HLDA-WCCN inter-session variability compensated i-Vector, $\hat{\mathbf{w}}$, is calculated as,

$$\hat{\mathbf{w}} = \mathbf{B}^\top\mathbf{A}^\top\mathbf{w}. \qquad (7)$$

2.4. Scoring Against Accent Models

We use *cosine scoring* to measure similarity of two i-Vectors [11]. The cosine score, t, between the inter-session variability compensated test i-Vector, $\hat{\mathbf{w}}_{\text{test}}$, and target i-Vector, $\hat{\mathbf{w}}_{\text{target}}$, is computed as the dot product between them,

$$t = \frac{\hat{\mathbf{w}}_{\text{test}}^\top \ \hat{\mathbf{w}}_{\text{target}}}{\|\hat{\mathbf{w}}_{\text{test}}\| \ \|\hat{\mathbf{w}}_{\text{target}}\|}, \qquad (8)$$

where $\hat{\mathbf{w}}_{\text{target}}$ is the average i-Vector over all the training utterances of the target accent,

$$\hat{\mathbf{w}}_{\text{target}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{w}}_i, \qquad (9)$$

Table 3: Summary of corpora statistics.

Corpus	# accents	# speakers	# utterances
Finnish	8	415	1146
English	7	511	1262

Table 4: Baseline and attribute results in terms of EER_{avg} and C_{avg} in the FSD corpus.

Feature (dimension)	Classifier	$\text{EER}_{\text{avg}}(\%)$	$\text{C}_{\text{avg}} \times 100$
SDC+MFCC (56)	GMM-UBM	19.03	10.5
SDC+MFCC (56)	i-Vector	12.60	6.85
SNN Place (27)	i-Vector	10.37	6.00
DNN Place (11)	i-Vector	9.33	5.88
SNN Manner (18)	i-Vector	9.21	5.80
DNN Manner (7)	i-Vector	**7.26**	**5.52**

where $\hat{\mathbf{w}}_i$ is the inter-session variability compensated i-Vector of training utterance i in the target accent.

Obtaining scores $\{t_a, a = 1, \ldots, L\}$ for a particular test utterance of accent a, compared against all the L target accent models, scores are further post-processed as,

$$t'_a = \log\frac{\exp(t_a)}{\frac{1}{L-1}\sum_{k\neq a}\exp(t_k)}, \qquad (10)$$

where t'_a is the detection log-likelihood ratio, for a particular test utterance of accent a, scored against all the L target accent models.

3. Speech attribute detection

The front-end shown in Figure 1 is built using two independent DNNs having six hidden layers and 1024 hidden nodes. The input feature vector is a 45-dimension mean-normalized log-filter bank feature with up to second-order derivatives and a context window of 11 frames, forming a vector of 495-dimension (45×11) input. The number of output classes is equal to 6 for manner, and 10 for place. In addition, a further output class is added to both DNNs to handle possible unlabelled frames. The DNN was trained with an initial learning rate of 0.008 using the cross-entropy objective function. It was initialised with the stacked *restricted Boltzmann machines* (RBM) by using layer by layer generative pre-training. An initial learning rate of 0.01 was then used to train the Gaussian-Bernoulli RBM and a learning rate of 0.4 was applied to the Bernoulli-Bernoulli RBMs. This DNN architecture follows conventional configurations used in the speech community, and it was not optimised for the corpora and task at hand. The "stories" part of the OGI Multi-language telephone speech corpus [17] was used to train the attribute detectors. This corpus has phonetic transcriptions for six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Data from each language were pooled together to obtain 5.57 hours of training and 0.52 hours of validation data.

In Tables 1 and 2, we report manner and place of articulation accuracies for each specific attribute using either one or six hidden layers. Classification accuracies increased consistently for all attributes except silence when moving from one to six hidden layers, as we expected. It should be noted that although silence classification accuracy does not increase, it is already above 90%. Glide and dental are instead still very hard to detect, and even with 6 hidden layers an accuracy of only 30% can be attained.

Table 5: Comparison between per language results in the manner and DNN manner systems. Results are reported in terms of EER (%) on the NIST 2008 corpus.

Accent	SNN Manner	DNN Manner
Cantonese	17.68	13.50
Hindi	15.75	13.15
Vietnamese	15.44	12.22
Russian	13.16	10.00
Korean	12.54	11.97
Japanese	11.75	10.76
Thai	11.70	9.31
Total (average)	14.00	11.55

Table 6: English results in terms of $\mathrm{EER}_{\mathrm{avg}}(\%)$ and $\mathrm{C}_{\mathrm{avg}}$ on the NIST 2008 SRE task.

Feature (dimension)	Classifier	$\mathrm{EER}_{\mathrm{avg}}(\%)$	$\mathrm{C}_{\mathrm{avg}} \times 100$
SDC+MFCC (56)	GMM-UBM	16.94	9.00
SDC+MFCC (56)	i-Vector	13.82	7.87
SNN Place (27)	i-Vector	12.00	7.27
DNN Place (11)	i-Vector	11.11	7.00
SNN Manner (18)	i-Vector	11.09	6.70
DNN Manner (7)	i-Vector	**10.45**	**6.50**

4. Foreign accent recognition

To better appreciate experimental results reported in this paper, we compared our attribute-based systems against two spectral-based accent recognition systems based on SDC and MFCC feature vectors, respectively, which have proven to give best performance in foreign accent recognition tasks [26]. Accent classifiers in these two systems were built using either GMM-UBM [27] or i-Vector approach. According to [26], the UBM size was set to 512, i-Vector dimension to 1000 and HLDA output dimension to 180. The UBM and T-matrix were estimated from the same held-out set, not used in either training or testing the foreign accent models. In the following experiments, Finnish and English play the role of being the second language, respectively. Finnish experiments are carried out using the *Finnish language proficiency test* (FSD) (see [26] for details). English experiments are performed using a subset of NIST SRE 2008 corpus. In our experiments, we selected the test utterances from the original 10sec NIST SER 2008 cuts in order to keep the test setup inline with the standard language and accent recognition test. Statistics for the two corpora are given in Table 3.

Table 4 shows foreign accent recognition results on the Finnish corpus. The first two rows indicate that the i-Vector system outperforms the baseline GMM-UBM system when the same input features are used, which is inline with findings in [28, 29]. Replacing the spectral SDC+MFCC features with the attribute features, results indicate that manner and place systems outperform the SDC+MFCC-based i-Vector system by 27% and 18% relative improvements, in terms of $\mathrm{EER}_{\mathrm{avg}}$, respectively. In particular, the best performance is achieved using DNN manner within the i-Vector approach, yielding a $\mathrm{EER}_{\mathrm{avg}}$ of 7.26%, which represents relative improvements of 21% over the shallow neural network (SNN) manner based accent recognition system. Similarly, in Table 6, we see same experimental setup but with the English corpus. We can again observe an improvement in English foreign accent recognition task using DNNs.

Above results indicate the effectiveness of the DNN attribute features over spectral SDC+MFCC and attribute features. Next, we compare the language-wise results achieved by shallow and deep architecture for the manner case for the both the Finnish and the English task. We compensate against the lack of data, by performing a leave-one-speaker-out (LOSO) evaluation. More details of the experimental setup can be found in [28]. Table 7 shows per accent results in the Finnish corpus. We notice that the DNN modeling systematically improves per accent detection error rate. We also note that the difficulty of detection is now more clearly revealed to consist of three groups easiest being {Turkish, Spanish}, medium difficulty being {Albanian, Arabic} and the most difficult being {English, Russian, Estonian, Kurdish}. Table 5 shows per-accent recognition accuracy on the English task. In both Manner and DNN manner systems, Cantonese attains the lowest recognition accuracy with EER of 17.68% and 13.50%, respectively; and the easiest accent is Thai with EER of 11.70% and 9.31%, respectively, in both systems. We note that Russian is the only common accent in both Finnish and English experiments, residing in a medium difficulty group in both experiments.

Table 7: Comparison between per language results in the manner and DNN manner systems. Results are reported in terms of EER (%) on the FSD corpus.

Accent	SNN Manner	DNN Manner
English	16.03	12.58
Estonian	15.44	13.18
Russian	14.21	13.03
Kurdish	14.00	13.67
Arabic	13.15	10.00
Albanian	12.32	10.11
Spanish	11.74	8.82
Turkish	10.41	9.00
Total (average)	13.37	11.29

5. Summary

In this paper, we have investigated into the use of deep architectures to improve accent recognition performances. In particular, we have designed two deep neural networks having six hidden layers with 1024 nodes per each for modelling speech attributes, namely manner and place of articulation. Experimental results have demonstrated that not only the effectiveness of DNNs for attribute classification, but also that deep neural modelling is useful in foreign accent recognition tasks. Specifically, an accent recognition performance improvement of 21% and 5.8% has been observed by moving from shallow to deep architecture for the Finnish and English task, respectively. We intend to expand further this line of research by exploiting multi-task learning approach at the front-end level in Figure 1, and evaluating other neural architectures, such as convolutional and recurrent deep neural networks.

6. Acknowledgements

This project was partially supported by the Academy of Finland projects 253120, 253000 and 283256 and Finnish Scientific Advisory Board for Defence (MATINE) project nr. 2500M-0036. Dr. Hautamäki and Dr. Siniscalchi were supported by the Nokia Visiting Professor Grants 201500062 and 201600008.

7. References

[1] J. H. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. of ICASSP*, 1995, pp. 836–839.

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

[3] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "Introducing attribute features to foreign accent recognition," in *Acoustics, Speech and Signal Processing, 2014. ICASSP 2014. IEEE International Conference on*. IEEE, 2014.

[4] L. M. Arslan and J. H. Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[5] P. Angkititraku and J. H. Hansen, "Advances in phone-based modeling for automatic accent classification," in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, 2006, pp. 634–646.

[6] V. Gupta and P. Mermelstein, "Effect of speaker accent on the performance of a speaker-independent, isolated word recognizer," *J. Acoust. Soc. Amer.*, vol. 71, no. 1, pp. 1581–1587, 1982.

[7] R. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, vol. 42, no. 1, pp. 109–123, 2004.

[8] S. Aryal and R. Gutierrez-Osuna, "Accent conversion through cross-speaker articulatory synthesis," *JASA*, vol. 137, no. 1, pp. 433–446, January 2015.

[9] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

[10] H. Behravan, V. Hautamäki, S. M. Siniscalchi, E. el Khoury, T. Kurki, T. Kinnunen, and C. Lee, "Dialect levelling in finnish: a universal speech attribute approach," in *INTERSPEECH 2014*, 2014, pp. 2165–2169.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[12] V. H. Do, X. Xiao, V. Hautamäki, and E. S. Chng, "Speech attribute recognition using context-dependent modeling," in *APSIPA ASC*, Xi'an, China, October 2011.

[13] G. Hinton, L. Deng, D. 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. 82–97, 2012.

[14] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks." in *Proc. Interspeech*, 2011, pp. 437–440.

[15] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *ACM/IEEE Trans. Audio Speech and Lang. Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[16] Y. Miao and F. Metze, "Improving low-resource CD-DNN-HMM using dropout and multilingual DNN training." in *Proc. Interspeech*, 2013, pp. 2237–2241.

[17] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The ogi multi-language telephone speech corpus," in *Proc. of ICSLP'92*, 1992.

[18] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target specific training data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887, 2012.

[19] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Universal attribute characterization of spoken languages for automatic spoken language recognition," *Computer Speech & Language*, vol. 27, no. 1, pp. 209–227, 2013.

[20] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao, "Optimal distributed online prediction," in *Proc. ICML*, 2011, pp. 713–720.

[21] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[22] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–359, May 2005.

[23] M. Loog and R. P. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 732–739, 2004.

[24] M. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272–281, 1999.

[25] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. of INTERSPEECH*, 2006, pp. 1471–1474.

[26] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign accent detection from spoken Finnish using i-vectors," in *Proc. of INTERSPEECH*, 2013, pp. 79–83.

[27] P. Torres-Carrasquillo, T. Gleason, and D. Reynolds, "Dialect identification using Gaussian mixture models," in *Proc. of Odyssey*, 2004, pp. 757–760.

[28] H. Behravan, V. Hautamäki, and T. Kinnunen, "Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish," *Speech Communication*, vol. 66, pp. 118–129, 2015.

[29] A. DeMarco and S. J. Cox, "Native accent classification via i-vectors and speaker compensation fusion," in *Proc. of INTERSPEECH*, 2013, pp. 1472–1476.