To appear in: Proc. Interspeech 2015, Dresden

Vowel mispronunciation detection using DNN acoustic models with cross-lingual training

Shrikant Joshi, Nachiket Deo, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India

{shrikant,nachiket.deo,prao}@ee.iitb.ac.in

Abstract

We address the automatic detection of phone-level mispronunciation for feedback in a computer-aided language learning task where the target language data (Indian English) is limited. Based on the recent success of DNN acoustic models on limited resource recognition tasks, we compare different methods of utilizing the limited target language data in the training of acoustic models that are initialized with multilingual data. Frame-level DNN posteriors obtained by the different training methods are compared in a phone classification task with a baseline GMM/HMM system. A judicious use of domain knowledge in terms of L2 phonology and L1 interference, that includes influence on phone quality and duration, are applied to the design of confidence scores for mispronunciation detection of vowels of Indian English as spoken by Gujarati L1 learners. We also show that the pronunciation error detection system benefits from a more precise signal-based segmentation of the test speech vowels, as would be expected due to the now more reliable frame-based confidence scores.

Index Terms: Computer-assisted learning; pronunciation assessment; Deep Neural Networks; General Indian English; Bilingual models.

1. Introduction

Computer aided language learning tools have recently been the focus of active research due to their potential to complement classroom teaching of spoken language, including the facilitation of automatic assessment of language skills. The two dimensions used to rate a language learner's speech are correctness of pronunciation in terms of the segmentals, and correctness of prosody (suprasegmentals) including fluency and speech rate. To achieve acceptable pronunciation free of a perceptible non-native accent, a learner must acquire familiarity with the phonology and phonotactics of the target language. Mismatches in the phonologies of the learner's L1 and the target language (L2) lead to the typical error of substitution of L2 phones with similar, but not identical, L1 phones. Such subtle differences in phone realization are particularly true for vowels [1]. Further, phone insertion and deletion errors can arise due to differing phonotactic constraints. While an overall rating of pronunciation quality, as obtained from human raters, is useful, focused feedback on specific phone segment-level errors can be invaluable in terms of improving a learner's attention to specific phonetic contrasts.

Automatic speech recognition (ASR) systems, designed essentially to decode speech utterances to text in a given language, can be adapted to the task of mispronunciation detection via a measure of the match of the non-native utterance to acoustic models trained on native speech. Whereas most applications of automatic speech recognition strive towards achieving robustness to variations due to speakers and accents, the task of spoken language assessment on the contrary, requires the system to be *sensitive* to deviations from native speech. Language models, which typically compensate for the weaknesses of acoustic models in ASR systems, must instead be replaced by models that can effectively discriminate non-native from native speech. Well-known approaches to segmental error detection have involved a "goodness of pronunciation" measure that captures the degree of match with respect to the target native speech acoustic models in a standard GMM/HMM ASR system framework [2]. Further, adding scores that explicitly incorporate the degree of match with expected substitutions (e.g. other phones in the same broad class) have been found to lead to performance improvements [3,4]. The requirement of sufficient labeled native and non-native training speech may not always be met however, thus motivating the search for alternate methods. Recent advances related to inherently discriminative acoustic models such as deep neural networks (DNN) have demonstrated significant gains in ASR performance, especially in the context of limited labeled training data [5,6]. The superiority of the DNN hybrid system for ASR has also been demonstrated for low-resourced languages, given unlabeled training data from other languages [7]. Further, very recent work on phone-level mispronunciation detection has shown the effectiveness of DNN acoustic models, followed by a Neural Network based classifier [8].

We address the assessment of spoken English by learners of an Indian L1, specifically, Gujarati. A desirable goal for an Indian English learner is the acquisition of the standard form of spoken English in India known as General Indian English (GIE), which is devoid of regional influences and intelligible across the country and outside it [9,10]. Focusing on a specific L1 allows us to potentially exploit non-native error patterns to improve system reliability. In our task, the L1 is Gujarati and the target language is GIE. In this case, we do not have non-native training speech (i.e. Gujarati L1 speakers' IE speech) and, more significantly, we do not have a large enough database of the target speech (i.e. GIE). Thus our target language is resource limited. Although the IE phone set is a subset of American English (AE), our previous work on mispronunciation detection of IE vowels on isolated words uttered by Gujarati L1 speakers showed that better performance was obtained when target acoustic models were trained on the same IPA phone segments from a Hindi database with unavailable phones drawn from the TIMIT AE database over models trained purely on AE [10]. Further, the

MAP adaptation of the GMM/HMM models with a limited amount of target (GIE) data provided significant improvement. We may thus view the present scenario as a case for the cross-lingual training of acoustic models. Since, unlike the case for GMMs, adaptation methods for DNN models are still a topic of research, we consider different methods for the language based adaptation of a DNN hybrid system.

The present work extends our previous work on vowel mispronunciation detection towards a more realistic evaluation on read text and further includes newer acoustic modeling approaches. The overall system follows the framework of van Dormelan et al. [4,1] where they used error modeling in a phone-based logistic regression classifier for correct/wrong pronunciation based on GMM/HMM posterior features. Our new contributions are (i) using DNN/HMM posteriors with an investigation of how to effectively use the limited GIE (i.e. target language) data for the language adaptation of the network; alternate approaches are compared via a GIE vowel classification experiment. (ii) applying the knowledge of L1 interference errors in determining the best feature set for the binary classifier for correct/wrong pronunciation; (iii) incorporating a robust vowel segmentation method for more accurate state-level alignment of non-native speech for feature computation.

2. Training and testing datasets

As seen from our task description above, our target language is GIE. However we have only limited labeled data of GIE (spoken by "natives" i.e. proficient Indian speakers). We therefore consider training data from other related language datasets. There are 11 pure vowels in GIE [9]. It has been observed that Indian English vowels are more similar to AE vowels than to British R.P. [11]. Table 1 compares the vowel phonologies of all the languages involved. The closely spaced AE central vowels are collapsed into one in GIE. Previous work [10] results indicated that IE vowels are phonetically closer to the corresponding Hindi vowels, when available, rather than to AE realizations. Gujarati (the L1 of our "non-native" GIE speakers) has 6 pure vowels. As seen in Table 1, these correspond to the collapsed forms of each of 4 sets of GIE vowels, and the two remaining GIE vowels [10]. The collapsed phonemes are of an "intermediate" quality and when used for English pronunciation give rise to the corresponding ambiguities, i.e. confusions between long and short vowels: /i:, I/, /u, u:/ , /e:, ɛ, æ /, and /ɒ, o:/ [12,10]. Some of these distinctions exist in written script but are in disuse in the spoken form, similar to the case of the Marathi language [13].

2.1. Training and adaptation datasets

Based on the observations above, we use available AE and Hindi datasets to train native (GIE) vowel models. The TIMIT AE database is used for the AE vowel models unavailable in Hindi [14]. TIMIT train set comprises of 462 speakers across 8 dialect regions uttering 8 unique phonetically balanced sentences each. We also use a standard Hindi speech database [15] to train the 8 acoustic models selected from 10 vowels listed in Table 1 that are common to GIE and Hindi (the 8 Hindi vowels are from among the 10 listed in Table 1, by omitting /ɒ, o:/ (bat, caught) since they have too few instances in the training data). The Hindi database is patterned on TIMIT but much smaller with 100 native Hindi speakers uttering 8 unique sentences each.

Also available is a relatively small dataset of GIE by 20 proficient (i.e. "native") speakers of Indian English, each with 18 unique sentences from TIMIT prompts. These speakers of various L1 were students of IIT Bombay identified by their absence of any recognizable L1 accent. Thus we have the following transcribed datasets: 3 hours of AE, 1 hour of Hindi and 25 min of IE, all at 16 kHz sampling, 16-bit word length.

2.2. Test datasets

Read out text comprised the 3 standard paragraphs of 'Stella', 'The North Wind and the Sun', 'The Rainbow Passage' [16, 17, and 18]. Two speaker groups (10 "native" IE speakers and 21 Gujarati-L1 speakers of low to moderate English proficiency) recorded the 3 paragraphs each. The former were Mumbai college students schooled in English with no discernable accent. Gujarati L1 speakers were drawn from college students in Gujarat state schooled in Gujarati, learning English as a second language. Further, the Gujarati-L1 data is manually annotated at the level of the vowel segments of interest in each paragraph. 3 annotators (students of speech processing) labeled each vowel by its GIE surface transcription. Any segment found to be rated different from its text transcript by 2 or more raters was treated as a pronunciation error. Table 2 shows the resulting confusion matrix after omitting 2 GIE vowels that are never confused. The major confusions observed are within the short-long (lax-tense) high vowel pairs: / iː , ɪ / and / uː, ʊ /. Equally important are confusions within the clusters /eː, ɛ, æ / and / oː, ɒ/. These observations conform to previous literature [12].

Table 1. *Mapping of AE, GIE, Hindi and Gujarati vowels with example words [10]*

Sr. No	AE word	AE IPA	GIE word	GIE IPA	Hindi IPA	Gujarati IPA
1	beat	iː	beat	iː	iː	I
2	bit	ɪ	bit	ɪ	ɪ	
3	bait	eɪ	gate	eː	eː	E
4	bet	ɛ	get	ɛ	---	
5	bat	Æ	bat	Æ	æ	
6	about, bird, butter	ʌ, ə, ɜː	cut	ə	ə	ə
7	father	ɑ	past	aː	aː	aː
8	boat	ɔː	coat	oː	oː	ɔ
9	or, golf	ɔ	caught	ɒ	ɒ	
10	boot	uː	fool	uː	uː	U
11	book	ʊ	put	ʊ	ʊ	

Table 2. *Confusion matrix of Gujarati test data: 'P' indicates perceived and 'I' indicates intended vowels*

P / I	iː	ɪ	eː	ɛ	æ	ɒ	oː	uː	ʊ	Total count
iː	**286**	139		1						426
ɪ	65	**554**								621
eː		1	**299**	23	10					333
ɛ			14	**138**	40					292
æ			13	99	**152**					265
ɒ						**471**	136			633
oː						52	**143**			196
uː						1	1	**193**	23	218
ʊ								28	**194**	222

3. Acoustic model training and evaluation

3.1. GMM/HMM baseline system

As we are mainly concerned with pronounced vowel quality, the non-vowel speech is absorbed into 5 broad class models. Thus we have a total of 22 context independent models (11 vowels + 6 diphthongs + 5 broad classes). Each is a 3-state HMM. A 12 component GMM with diagonal covariances models each state. 13 dimensional MFCCs and short time energy with their delta and acceleration coefficients extracted every 10 ms comprise the feature vectors. The 11 vowels consist of 3 AE models and 8 Hindi models. The AE vowel models are trained using the TIMIT dataset while the Hindi vowel models and the broad class models are trained using the Hindi dataset. The HMMs are trained based on the ML criterion using the CU-HTK[19]. Finally the "native" GIE data is used for adapting the trained GMMs using MAP estimation. We use context-independent models throughout our work due to training data limitations.

3.2. DNN/HMM hybrid system

In a DNN/HMM hybrid model [20] the outputs of a trained DNN are estimates of the posterior probabilities $P(s_i|O_t)$ for each HMM state s_i for a given observation vector O_t at time t. These can be divided by the prior probabilities $p(s_i)$ over the HMM states to give scaled likelihoods which can be used in place of GMM likelihoods in the HMM framework. We make the simplifying assumption of equal prior probabilities across HMM states as in [5].

For DNN training in the current "under resourced" setting (i.e. the small native GIE dataset), we employ cross-lingual training where available data from different languages is used for initializing the DNN weights for an under resourced target language with the same or similar phone set [21,22,7]. Here, we use the TIMIT and Hindi datasets for unsupervised Deep Belief Network (DBN) training followed by supervised Error Back Propagation training of the DNN on the same datasets. The so initialized DNN weights are next "adapted" to the target dataset by one of 2 distinct approaches.

(1) Retraining where the GIE data is used for supervised error back propagation training of the DNN allowing all model weights to be modified.

(2) Model adaptation where a limited set of model parameters is tuned, conceptually similar to GMM adaptation. We apply some recent work where affine transforms applied to the input vector and to the top hidden layer of the DNN are shown to reduce error rates in a speaker-adaptation context over unadapted models. Feature Discriminative Linear Regression (fDLR) [23] applies a block diagonal affine transform to the input vector before it is given to the DNN. The shared weights and the bias term of the affine transform are learned using error back propagation with the rest of the DNN weights kept fixed. Similarly if an affine transform is applied to the outputs of the final hidden layer of the DNN before being given to the softmax layer, learning the affine transform weights amounts to re-training of the final layer of DNN weights [24]. Model adaptation thus involves updating only the final layer DNN weights and the weights and bias terms of the fDLR transform. Thus it requires update of much fewer parameters as compared to retraining the DNN using the GIE data. Specifically in our case, retraining the DNN involves 25 times the number of parameters compared to fDLR and final layer adaptation.

3.2.1. DNN training specifics

All the DNNs used in this work [25] have a fixed topology of 5 hidden layers of 1000 sigmoidal units each and a softmax output layer having 66 units corresponding to the number of HMM states for the 22 phone classes. The input layer uses 9 consecutive frames of input features, here 13 dimensional MFCCs and short time energy with their delta and acceleration coefficients.

For pre-training the first layer of the DBN is trained as a Gauss-Bernoulli RBM while the subsequent layers are trained as Bernoulli RBMs. Supervised fine tuning of the network is carried out using stochastic gradient descent for minimizing the cross entropy error. Mini batches of size 100 are used for weight updates. An exponential learning rate update and a momentum term are used for speeding up training. The HMM state labels required for training are obtained using forced alignment with the baseline GMM/HMM system.

3.3. Acoustic model evaluation

It is expected that mispronunciation detection is highly correlated with the ASR performance of the native acoustic models [8].We compare our DNN hybrid systems by the 11-way classification of vowel segments from the paragraphs read by the native IE speakers. The vowel segments are extracted by forced alignment using the baseline system. (We found that both GMM and DNN hybrid systems provide similar alignment performance.) Next the state-level segmentation is achieved using re-estimation with the corresponding posteriors from the system under consideration. Table 3 shows the vowel classification accuracies. We observe that the DNN/HMM acoustic models clearly outperform the GMM/HMM. We also observe that the DNNs which used the GIE data for re-training of all the model weights outperform the DNN which used the GIE data for fDLR and final layer adaptation. This indicates that the amount of GIE data available was sufficient for updating all the parameters of the DNN rather than just the output layer and the fDLR weights (probably unlike the speaker adaptation context the methods were proposed for). For all our further experiments we retain only the best performing DNN models i.e. the one initialized using cross lingual data via pre training and supervised back propagation, and then retrained with GIE native data.

Table 3.*Vowel classification accuracy in % (over 3947 instances of 11 GIE vowels across 10 native speakers)*

GMM/HMM (GIE adaptation)	DNN/HMM (GIE retraining)	DNN/HMM (GIE adaptation)
61.1	**69.8**	**64.6**

4. Vowel mispronunciation detection

After forced aligning with the text transcript using a language model that allowed for hesitation pauses and commonly observed deletion errors such as those of the articles 'a', 'the', the vowel segments in the non-native test data are extracted for pronunciation classifier training and evaluation. Due to the expected mismatch between non-native speech and native acoustic models, the forced alignment of the transcript may not be accurate [1, 8] in terms of phone as well

as state (senone) boundaries. Hu et al. [8] deal with this by ignoring senone boundaries and using the average of all state-level posteriors across the phone for the confidence score. We instead address the problem at its root by using a phone boundary refinement stage in the processing before feature extraction. The vowel boundaries obtained by the GMM/HMM system forced alignment are subjected to a fine-search (2.5 ms resolution) in a neighborhood of 100 ms based on acoustic cues to vowel onset and offset. The acoustic cues depend on the vowel context in terms of the broad class of the adjacent phone and include formant band energy changes and periodicity strength, and have been shown to significantly reduce the error between automatically detected and manually marked vowel onsets over using forced alignment alone [26].

Phone mispronunciation detection is usually carried out with a discriminative binary classifier using one or more confidence measures computed for the phone segments [1,8, 27]. The scores typically involve estimates of posterior probabilities obtained from an ASR system. The posterior probability of the target phone given the vowel segment is a measure of the match of the acoustics with the underlying model. The Average Posterior probability Estimate (APE) is a representative measure [1,3]. It is an estimate of the posterior probability $P(p_{target}|O)$ of the target phone p_{target} given the observation vectors O_t in the phone segment. It is computed by averaging the frame posterior probabilities for a given phone segment. The expression for APE for a target phone is:

$$\mathrm{APE}(\mathrm{p_{target}}) = \frac{1}{t_e - t_b} \sum_{t=t_b}^{t_e} \log(\mathrm{P}(\mathrm{p_{target}}|\mathrm{O_t}))$$

Where t_b, t_e are the start and end frames of the extracted vowel segment. For the DNN/HMM system, the posteriors $P(p_{target}|O_t)$ are directly obtained using the DNN outputs after obtaining the state wise forced alignment for phone p_{target} for that segment from the DNN/HMM system. For the GMM/HMM system the posteriors are computed using the likelihoods $P(O_t|p_i)$ under the assumption of equal priors.

Non-native vowel realizations can often lie on a continuum between native vowels rather than involve outright substitution of one vowel with another. We thus expect the APE scores of vowels other than the target vowel to contain complementary information for determining whether the vowel was mispronounced. Therefore the APE is extended to "weighted" APE (wAPE) using an approach similar to [1]. This measure takes a weighted combination of APE scores with respect to all the 11 vowels for a given vowel segment. The weights are learnt using a logistic regression classifier. A separate logistic regression classifier is trained for each vowel.

Finally, we consider using our knowledge of important phone confusions as depicted in Table 2 to select only the *relevant* competing vowel scores for the weighted APE. For example, for the target vowel /e:/, we use the following APEs: APE(e:), APE(ɛ), APE(æ). We term this the "subset wAPE". Once again, a logistic regression classifier is used with this input scores vector.

We also consider augmenting the scores vector with log(duration) of vowel segments since duration is a discriminating feature in the vowel set. We expect duration to provide complementary information to the posterior based measures which essentially capture spectral differences. The logistic regression models for each of the vowels are trained and tested in a cross validation setting leaving one speaker out (all 3 paragraphs) on the 21 non-native speakers' test data using the correct/wrong pronunciation labels obtained as described in Sec. 2.2. The logistic regression models are evaluated in terms of equal error rate (EER), i.e. at the point where false acceptance rate for mispronunciation detection equals false rejection rate.

5. Results and Discussion

Table 4 shows the EER results for the different configurations of scores presented in Sec. 4 across GMM/HMM and DNN/HMM acoustic models. For compactness, EERs are averaged across the confusable groups of vowels as in Table 2. We note that the DNN/HMM acoustic models provide lower EER over the baseline GMM/HMM in every configuration given the same training and adaption datasets. In particular, the best system using DNN/HMM acoustic models shows an EER of 25.2% as compared to 28.9% for the best GMM/HMM based system, a relative improvement of 12.8%.

We also observe by comparing columns 2 and 3 that adding the log duration score improves the average EERs. In particular, significant improvement is seen in case of the vowel groups u: U and i: I which are long-short vowel pairs, confirming our hypothesis that vowel segment duration conveys useful cues for certain vowel pairs with the quantity distinction. Comparing columns 3 and 4, we observe that using subset wAPE improves EER for every vowel group as compared to the (full) wAPE in case of the DNN/HMM acoustic models, with a relative improvement of 6.3% in terms of average EER. This improvement is ambiguous in case of the GMM/HMM acoustic models.

Finally we note the contribution of vowel boundary refinement by comparing configurations with and without it (col. 4 and 5 in Table 4) to find that it clearly benefits the DNN/HMM system. The best system in terms of average EER uses the DNN/HMM acoustic models, the subset wAPE restricted to the confused vowel set scores with log duration and vowel boundary refinement. Future work involves incorporating context-dependent acoustic models.

Table 4. *Average EER of confusing vowel groups and overall average EER in %*

Confusable groups	wAPE	wAPE + duration	subset wAPE + duration	subset wAPE + duration (w/o refinement)
GMM/HMM				
i:, ɪ	30.1	27.6	23.9	25.6
e:, ɛ, æ	23.6	22.4	23.1	24.0
ɒ, o:	42.3	41.0	42.1	38.7
u:, ʊ	28.0	22.7	31.8	29.5
Average	**31.2**	**28.9**	**29.5**	**29.1**
DNN/HMM				
i:, ɪ	28.1	25.7	23.6	24.9
e:, ɛ, æ	22.9	23.0	21.5	22.1
ɒ, o:	35.7	35.3	34.6	36.8
u:, ʊ	24.4	22.2	18.8	18.9
Average	**28.1**	**26.9**	**25.2**	**26.4**

Acknowledgement: This work was supported in part by Bharti Centre for Communication at IIT Bombay

6. References

[1] van Doremalen, J., Cucchiarini, C. and Strik, H., "Automatic pronunciation error detection in non-native speech: The case of vowel errors in Dutch", J. Acoust. Soc. Am. 134(2), pp. 1336-1347, Aug. 2013

[2] Witt, S. M. and Young, S. J., "Phone-level pronunciation scoring and accessment for interactive language learning", Speech Communication, vol. 30, no. 2-3, pp. 95-108, Feb. 2000.

[3] Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O., "Combination of machine scores for automatic grading of pronunciation quality", Speech Communication, 30(2), pp. 121-130, Feb. 2000.

[4] van Doremalen, J., Cucchiarini, C. and Strik, H., "Using non-native error patterns to improve pronunciation verification", In: Proc. INTERSPEECH 2010, Makuhari, Japan,pp. 590-593, Sept. 2010.

[5] Mohamed, A. R., Dahl, G. E. and Hinton, G., "Acoustic modeling using deep belief networks",IEEE Transactions on Audio, Speech and Language Processing, 20(1), pp. 14-22, 2012.

[6] Dahl, G. E., Yu, D., Deng, L. and Acero, A., "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition",IEEE Transactions on Audio, Speech and Language Processing, 20(1), pp. 30-42, Dec. 2012.

[7] Swietojanski, P., Ghoshal A. and Renals, S., "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR", In: Proc. SLT-2012, Miami, USA, pp. 246-251, Dec. 2012.

[8] Hu, W., Qian, Y., Soong, F. K., and Wang, Y., "Improved Mispronunciation Detection with Deep Neural Network Trained Acoustic Models and Transfer Learning based Logistic Regression Classifiers", Speech Communication. 67, pp. 154-166, Mar. 2015.

[9] Bansal, R. K. and Harrison, J. B., *Spoken English*. Mumbai: Orient Blackswan Private Limited, 2009.

[10] Joshi, S. and Rao, P., "Acoustic models for pronunciation assessment of vowels of Indian English", In: Proc. OCOCOSDA- 2013,Gurgaon, India, pp. 1-6, Nov. 2013.

[11] Wiltshire, C. R. and Harnsberger, J. D., "The influence of Gujarati and Tamil L1s on Indian English: a preliminary study", World Englishes, 25(1), pp. 91-104, 2006.

[12] Vyas, K. B., "A Comparative study of English and Gujarati phonological systems", Ph.D. thesis, Saurashtra Uni. English Dept, Rajkot, 2010.

[13] Yardi, V., "Teaching English pure vowels to Marathi learners: some suggestions", ELT Journal. 4, pp. 303-307, 1978.

[14] Garofolo, J. S. et al., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", Linguistic Data Consortium, Philadelphia, 1993.

[15] Samudravijaya, K., Rao, P.V. S., Agrawal, S. S., "Hindi speech database", In: Proc. ICSLP – 2000, Beijing, China, pp. 456-459, Oct. 2000.

[16] Weinberger, S., Speech accent archive, 2010. George Mason University: *http://accent. gmu. edu/index. php*.

[17] Visceglia, T., Tseng, C. Y., Kondo, M., Meng H. and Sagisaka, Y., "Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)", In: Proc. OCOCOSDA-2009,pp. 60-65, Aug. 2009.

[18] Markham, D. and Hazan, V., UCL Speaker database, 2002.

[19] Young S. et al., *The HTK Book v3.4*. Cambridge University, 2006.

[20] Bourlard, H. A., and Morgan, N., "Connectionist speech recognition: a hybrid approach", Springer Science & Business Media, 247, 1994.

[21] Thomas, S., Ganapathy, S. and Hermansky, H., "Multilingual MLP features for low-resource LVCSR systems", In: Proc. ICASSP-2012, Kyoto, Japan, pp. 4269-4272, Mar. 2012.

[22] Ghoshal, A., Swietojanski, P. and Renals, S., "Multilingual training of deep neural networks", In: Proc. ICASSP-2013,Vancouver, Canada, pp. 7319-7323, May 2013.

[23] Seide, F., Li, G., Chen, X. and Yu, D., "Feature engineering in context-dependent deep neural networks for conversational speech transcription", In: Proc. ASRU-2011, Waikoloa, USA, pp. 24-29, Dec. 2011.

[24] Yao, K., Yu, D., Seide, F., Su, H., Deng, L. and Gong, Y., "Adaptation of context-dependent deep neural networks for automatic speech recognition", In: Proc. SLT-2012, Miami, USA,pp. 366-369, Dec. 2012.

[25] Palm, R.B., "Prediction as a candidate for learning deep hierarchical models of data", Technical University of Denmark, 2012*(deepLearnToolbox for MATLAB)*.

[26] Patil, V., Joshi, S. and Rao, P., "Improving the robustness of phonetic segmentation of to accent and style variation with a two-staged approach", In: Proc. INTERSPEECH-2009, *Brighton, UK*, pp. 1761–1765. Sept. 2009.

[27] Wei, S., Hu, G., Hu, Y. and Wang, R. H., "A new method for mispronunciation detection using support vector machine based on pronunciation space models", Speech Communication, 51(10), pp.896-905, 2009.