INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving Sub-phone Modeling for Better Native Language Identification With Non-native English Speech

Yao Qian[1], *Keelan Evanini*[1], *Xinhao Wang*[1], *David Suendermann-Oeft*[1], *Robert A Pugh*[1], *Patrick L Lange*[1], *Hillary R Molloy*[1], *Frank K Soong*[2]

[1]Educational Testing Service Research, USA
[2]Microsoft Research Asia, China

{yqian,kevanini,xwang002,suendermann-oeft}@ets.org, frankkps@microsoft.com

Abstract

Identifying a speaker's native language with his speech in a second language is useful for many human-machine voice interface applications. In this paper, we use a sub-phone-based i-vector approach to identify non-native English speakers' native languages by their English speech input. Time delay deep neural networks (TDNN) are trained on LVCSR corpora for improving the alignment of speech utterances with their corresponding sub-phonemic "senone" sequences. The phonetic variability caused by a speaker's native language can be better modeled with the sub-phone models than the conventional phone model based approach. Experimental results on the database released for the 2016 Interspeech ComParE Native Language challenge with 11 different L1s show that our system outperforms the best system by a large margin (87.2% UAR compared to 81.3% UAR for the best system from the 2016 ComParE challenge).

Index Terms: time delay deep neural network, i-vector, native language identification

1. Introduction

Native Language Identification (NLI) is to identify the native language (L1) of a speaker, based upon his voice or written input in a second language (L2). Accurate L1 detection is critical for many human-machine voice interface applications, e.g., to engage a non-native user with a different language background. In Computer Assisted Language Learning (CALL) systems, the common pronunciation errors made by L1 learners are used to build an L1-L2 phone confusion table, which is useful for designing more customized pronunciation training [1]. Also, L1 information of speakers has been utilized to improve frame phonetic accuracy through multi-task learning and the performance of DNN-based speaker recognition can be significantly improved [2]. In addition, the knowledge of L1 can aid an automatic speech recognition system to build better acoustic and language models, e.g., modeling pronunciation variation mapping between native speakers and L1-specific speakers [3], which can yield reliable recognition performance. It also can facilitate a human-machine dialog system, which can be aware of a user's cultural background suggested by the identified native language.

NLI works under the assumption that a speaker's L2 production patterns are influenced by his L1 origin. Many Chinese ESL learners confuse English /r/ with /l/ and sometimes with /w/ since the distinctions do not exist in quite a few Chinese dialects, or their native L1 languages. Appending an extra vowel to a consonant at the end of a syllable is common among Japanese English speakers. It is due to the fact that Japanese syllable structure only allows a vowel ending except a final /n/. Foreign language learners with different native language background do make different errors in grammar, vocabulary and other usage in learning a new language [4, 5].

The approaches to automatic NLI are usually based on supervised learning where statistical models, e.g., SVM and GMM classifier, are trained on data labelled with corresponding L1 information. The features extracted from ESL learners' writings are generally based on lexicon and syntax, including N-gram features on character, word and part-of-speech (POS), Stanford dependencies [6], and spelling and grammatical errors, while the features obtained from learners' speech for NLI can be frame-level like MFCC, phone-level confusion, and lexical features like language use error. The latter of the two kinds of features needs to be supported by a speech recognizer with either a phone-loop grammar or language model. The NLI shared task [7] hosted by the BEA workshop at NAACL 2013 and Computational Paralinguistics Challenge (ComParE) [8] at INTERSPEECH 2016 show that automatic NLI methods can achieve 84% and 81% accuracy in detecting L1 from 11 different L1 backgrounds based on writing and speech, respectively.

Our study focuses on how to predict L1 from nonnative speakers' English speech input, similar to recognizing the spoken language or speaker. The following approaches: GMM-UBM+MAP, GMM-UBM+SVM, and i-Vector are fairly successful in recognizing speaker and language identity, and can be similarly applied to NLI. Phone-level features, one of the key clues for human listeners to detect a speaker's native language, have been proposed to improve the performance of NLI systems. In [9], a phonetically-inspired UBM by clustering the Gaussian components of the acoustic model built for ASR is integrated into an SVM-based classifier with ASR-based features and the best published results on the Foreign Accented English (FAE) database [10] is reported. The ComParE participants' systems based on the i-Vector approach can all achieve approximately 70% or more on accuracy for identification of 11 L1 languages. Log-likelihood ratios of phone posterior probabilities (PLLR) in four language-dependent phonetic decoders are used to extract the i-vectors in [11]. This PLLR based i-vector system is fused together with the MFCC feature based i-vector system at the scoring level and achieves the best performance in ComParE. Various features at the frame, phone and lexical levels, multiple classifiers, SVM, PLDA and DNN, and fusion at both feature and score levels are investigated in detail [12]. The performance of the final fused system is just slightly worse than the best system.

A language learner's mispronunciation patterns are often concentrated across two or three confusable canonical phones. A finer, sub-phonemic analysis can provide a higher resolution

http://dx.doi.org/10.21437/Interspeech.2017-245

than that of a coarser, phonemic counterpart. In this paper, the sub-phonemic "senone" modeling are adopted for extracting i-vector which, hopefully, can enhance NLI performance. Motivated by the advancement of the phonetically-aware deep neural network based ASR, e.g. DNN, RNN and TDNN, and its advantages in high performance i-vector based speaker and language recognition [13-17], we used TDNN to build UBM in our i-vector based NLI system. To the best of our knowledge, this is the first time exploring TDNN based i-vector for NLI.

2. TDNN based I-Vector for NLI

2.1. I-Vector Front-end

I-vector, a compact representation of a speech utterance in a low-dimensional subspace, is based upon the concept of factor analysis. An i-vector model [18], the *t-th* frame of the *u-th* segment, $\mathrm{x}_{\mathrm{t}}^{(\mathrm{u})}$, is sampled from the following distribution:

$$\mathrm{x}_{\mathrm{t}}^{(\mathrm{u})} \sim \sum_{\mathrm{k}} \gamma_{\mathrm{kt}}^{(\mathrm{u})} N(\mathrm{m}_{\mathrm{k}} + \mathrm{T}_{\mathrm{k}} \omega(\mathrm{u}), \Sigma_{\mathrm{k}}) \tag{1}$$

where m_k and Σ_k are the mean and covariance of the *k-th* Gaussian component if universal background model (UBM) is trained by GMM; T_k, called the total variability, is a low rank rectangular matrix; $\omega(\mathrm{u})$ is the segment-specific standard normal distributed latent vector. The i-vector of segment *u* can then be estimated as

$$\hat{\omega}(\mathrm{u}) = (\mathrm{I} + \mathrm{T}^{\mathrm{t}} \Sigma^{-1} \mathrm{N}(\mathrm{u}) \mathrm{T})^{-1} \mathrm{T}^{\mathrm{t}} \Sigma^{-1} \tilde{\mathrm{F}}(\mathrm{u}) \tag{2}$$

where $\mathrm{N}(\mathrm{u})$ and $\tilde{\mathrm{F}}(\mathrm{u})$ are the zeroth-order and mean shifted first order statistics, respectively.

$$\mathrm{N}_{\mathrm{k}}(\mathrm{u}) = \sum_{\mathrm{t}} \gamma_{\mathrm{kt}}^{(\mathrm{u})} \tag{4}$$

$$\tilde{\mathrm{F}}_{\mathrm{k}}(\mathrm{u}) = \sum_{\mathrm{t}} \gamma_{\mathrm{kt}}^{(\mathrm{u})} (\mathrm{x}_{\mathrm{t}}^{(\mathrm{u})} - \mathrm{m}_{\mathrm{k}}) \tag{5}$$

$\gamma_{\mathrm{kt}}^{(\mathrm{u})}$ is the statistical alignment result of the frame $\mathrm{x}_{\mathrm{t}}^{(\mathrm{u})}$, i.e., posterior probability calculated from a UBM.

2.2. TDNN for Extracting Sufficient Statistics

A DNN can model a long time-span, as well as high dimensional and correlated features. Except using a long acoustic context, e.g., 21 frames MFCC, as input feature representations to model the temporal dynamics, the neural network architecture also can capture the long-term temporal dependencies between the sequential events. A time delay deep neural network (TDNN) [19, 20] has such an architecture designed to work on sequential data. A TDNN is formulated as a feedforward network but it has delays on the layer weights associated with the input weights. The data are represented at different time points by adding a set of delays to the input. This allows the TDNN to have a finite dynamic response to time series input data [21]. A TDNN is similar to convolutional neural networks (CNN), where the convolution is done only along the time axis.

Recently, TDNNs have been shown to outperform DNNs for LVCSR and speaker recognition [22, 14]. The subsampling technology, in which hidden activations are computed at only few time steps, has been proposed and largely speeds up the training time of TDNNs [22]. This allows TDNNs to be more attractive than recurrent neural networks (RNNs), which also can capture temporal dynamics by using internal memory to process arbitrary sequences of inputs, but the training is more time consuming.

In our study, a TDNN is used as UBM and employed to extract Baum-Welch statistics, i.e. the TDNN, replacing the GMM, is used to compute frame posterior probabilities over each of the classes (sub-phones, senones, instead of the components of GMM). Given TDNN-UBM, the $\gamma_{\mathrm{kt}}^{(\mathrm{u})}$ is computed from "soft-max" output of TDNN.

$$\gamma_{kt}^{(\mathrm{u})} = p(s_k \mid \ddot{x}_t^{(\mathrm{u})}) \tag{6}$$

where s_k is the *k-th* senone and $\ddot{x}_t^{(u)}$ is spliced input vector.

A TDNN models phonetic units, senones, in a supervised manner. It allows the comparison among different utterances at the same senone set and then makes it easier to distinguish one L1 from the others by comparing GMM-UBM, in which the classes may be phonetically indistinguishable due to the unsupervised training approach. In addition, even if a TDNN and a GMM are both trained in a supervised manner, the TDNN can capture much more temporal dynamics and estimate model parameters discriminatively, which can lead to more accurate posterior estimation than for a GMM.

2.3. L1 Recognition

L1 recognition is performed with Probabilistic Linear Discriminant Analysis (PLDA) scoring [23]. Given an i-vector $l_u = \hat{\omega}(\mathrm{u}) / \|\hat{\omega}(\mathrm{u})\|$ extracted from *u*-th testing utterance after length normalization [24], the log likelihood ratio (LLR) is calculated for l_u and each i-vector, l_j, of target L1s, $l_j \in \{l_1, ..., l_N\}$ and N is the total number of L1s, then select one with highest value as recognized L1.

$$\mathrm{L1} = \arg\max_{l_j \in \{l_1, ..., l_N\}} \log \frac{p(l_j, l_u \mid \Theta_s)}{p(l_j \mid \Theta_d) p(l_u \mid \Theta_d)} \tag{7}$$

where Θ_{s} is the hypothesis that l_j and l_u share the same L1 identity while Θ_{d} is the hypothesis that l_j and l_u are from different L1 identities.

3. Experiments and Results

3.1. Corpora

Our approach of TDNN based i-vector for L1 recognition is mainly evaluated on a Non-Native Spoken English (NNSE) Corpus [8]. Another two corpora: LibriSpeech [25] and AELP (Assessment of English Language Proficiency) are employed to train the TDNN to create sufficient statistics for i-vector extraction.

NNSE corpus is provided by Educational Testing Service (ETS) as the Native Language (N) sub-challenge corpus for ComParE Challenge at Interspeech 2016 [8]. This corpus consists of spoken responses provided during a global assessment of English language proficiency. It contains 64 hours of speech sampled at 16 kHz from 5,132 non-native speakers of English, with 11 different L1 backgrounds: Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL) and Turkish (TUR). Each L1 is covered

by speech recordings from over 450 speakers. Each speaker's recording (one utterance) is roughly 45 seconds long. The dataset was divided into three partitions: training (3,300 utterances, ~41.3 hours), development (965 utterances, ~12.1 hours) and testing (867 utterances, ~10.8 hours)

Librispeech corpus is a free corpus of read English speech derived from LibriVox's audiobooks containing approximately 1,000 hours of speech sampled at 16 kHz. The speakers' accents are various and not marked, but the majority are US English. The acoustic model trained on this corpus shows a high resolution on U.S. English speech recognition, i.e., less than 7% WER on the test sets of the WSJ corpus [22]. All three training sets, approximately 960 hours, collected from 2,338 speakers, with normalized text of audiobook are employed to train the TDNN.

AELP Corpus consists of 800 hours of speech from 8,700 test-takers over 100 countries. It is also drawn from the English proficiency assessment, administered a different year than the NNSE Corpus. Each speaker has 6 utterances, i.e., 45-second spoken responses to express their opinions on a familiar topic or 60-second spoken responses based on reading and listening to relevant prompt materials, yielding roughly 5 minutes per speaker. All the recordings and corresponding manual transcriptions are used for training the TDNN. The metadata, including country information, is not used for training.

3.2. Experimental Setup

NLI systems are constructed by using the tools from Kaldi [26].

3.2.1. GMM based i-Vector (Baseline)

The front-end for the baseline NLI system contains 20 dimensional MFCCs including C0, extracted from a 20ms hamming window with 10ms time shift along with their first and second derivatives. Non-speech segments within utterances were deleted through an energy-based voice active detection (VAD) method. Utterance-based cepstral mean normalization was performed on the acoustic feature vectors. A GMM and a full covariance matrix was trained as the UBM by using the training set of the NNSE corpus. The same training set was also used to train an i-vector extractor T-matrix as well as PLDA projection matrices. The number of Gaussian components and the dimension of i-vector are optimized on NNSE development set.

3.2.2. TDNN based i-Vector

40-dim MFCCs are used as input features to train the TDNN. Speaker adaptation techniques like transforming acoustic features by fMLLR and appending i-Vector with input features are not employed here. TDNN architecture is similar to the one described in [22]. The context specification of the TDNN is configured as: input layer {-2,2} indicating 5 frames (t-2, t-1, t, t+1, t+2) spliced, then three hidden layers splicing as {-1,2}, {-3,3} and {-7,2}. The output layer of the TDNN has 5,809 and 4,057 nodes for Librispeech and AELP, separately. The output nodes are the "senones" of HMM got by decision-tree based clustering. The input and output feature pairs are obtained through frame alignment for the "senones" with GMM-HMM, which are also trained by LibriSpeech and AELP, separately. The i-vector extractor, T-matrix, and PLDA projection matrices are trained by different corpora to evaluate their performance.

3.3. Results and Analysis

The performance of NLI systems are evaluated in both accuracy (Acc) and Unweighted Average Recall (UAR), same as the metrics used by ComParE.

Table 1 shows the baseline system performance in terms of Acc and UAR with different dimensions of i-Vector and number of Gaussian components on the development set. It indicates that 600-dim i-Vector extracted from the posterior supervector of GMM with 1,024 Gaussian components can achieve the best performance. These results are similar to those that the Challenge systems [11,12] obtained, i.e., approximately 76% for both UAR and Acc by only using corpus, NNSE, provided by ComParE organizers. Thereafter, the dimension of i-vector is fixed to 600.

Table 1: *Acc and UAR obtained by the baseline system on the development set*

Systems	Acc (%)	UAR(%)
iVec400_GMM1024	73.8	73.9
iVec600_GMM1024	75.6	75.7
iVec800_GMM1024	75.3	75.5
iVec600_GMM2048	72.8	72.9

The performance of the NLI systems on the development set is presented in Table 2, which also shows the corpora used to train individual modules: UBM (Supervector), T-matrix and PLDA. Motivated by the performance improvement of using super large datasets to train the T-matrix and PLDA in text-independent speaker recognition, we tried to use the AELP corpus to train the T-matrix by assuming each utterance is from one L1 since no L1 info is available. But the performance is slightly worse than that T-matrix trained by the NNSE corpus.

Table 2: *Acc and UAR obtained by different NLI systems on the development set*

UBM	T-matrix	PLDA	Acc (%)	UAR(%)
NNSE	NNSE	NNSE	75.6	75.7
Librispeech	NNSE	NNSE	84.1	84.3
AELP	NNSE	NNSE	88.6	88.6
AELP	AELP	NNSE	88.1	88.2

Table 3 lists Acc and UAR obtained by different NLI systems on the test set. TDNN based i-Vector significantly outperforms the GMM based i-Vector. The systems, TDNN_Librispeech and TDNN_AELP (i-Vector from TDNN trained by LibriSpeech and AELP corpora) can achieve improvements of 8.1% and 12.5% on Acc, 8.2% and 12.5% on UAR over the baseline system. TDNN_AELP can increase the performance, 4.4% on Acc and 4.3% on UAR, by comparing with TDNN_Librispeech. TDNN trained by AELP, which contains much more sub-phone variation caused by speakers from different countries, should have a much higher sub-phone resolution than TDNN trained by Librispeech recorded dominantly by U.S. English speakers, and thus is more capable of capturing L2 spoken English patterns influenced by speakers' language backgrounds. The UBM trained by the DNN is also listed in Table 3 as a performance comparison between the TDNN and the DNN. DNN also has three hidden layers but each layer consists of 1,024 nodes. The input features is also 40-dim MFCCs but stacked over a 21 frame window (10 frames to either side of the center frame). It shows that TDNN outperforms DNN by over 1% on both Acc and UAR.

To back the above suppositions, we transcribe the test set of NNSE and use TDNN_Librispeech, TDNN_AELP and DNN_AELP as AMs to decode it. The corresponding frame accuracy, which is often used to evaluate the performance of a DNN by isolating the issues caused by the vocabulary and the language model in the ASR system, is shown in Table 4. It indicates that the frame accuracy of models trained by AELP is much higher than that trained by Librispeech. The TDNN outperforms the DNN, i.e., the frame accuracy on the test set is improved from 48.5% to 51.3%.

Table 3: *Acc and UAR obtained by different NLI systems on the test set*

	Acc (%)	UAR(%)
Baseline	74.6	74.7
TDNN_LibriSpeech	82.7	82.9
TDNN_AELP	87.1	87.2
DNN_AELP	85.9	86.1

Table 4: *Frame accuracy of different UBMs on the testing set*

	TDNN_Librispeech	TDNN_AELP	DNN_AELP
Acc (%)	38.7	51.3	48.5

Our best results are at 87.1% in Acc and 87.2% in UAR, which are significantly better than those achieved by the best system (L^2F) in ComParE. L^2F achieved 81.3% UAR, assisted by the ASR systems for European Portuguese, Brazilian Portuguese, European Spanish and American English to calculate PLLR features based i-vector [11]. Our system with TDNN based i-Vector trained by Librispeech can achieve 82.9% UAR, which is 1.6% better than the L^2F system which fuses two i-vector based systems. In [12], Librispeech is also used to train a phone recognizer for extracting L1-pronunciation projection features which are used for i-vector extraction. The fusion system achieves 79.9% Acc and 80.1% UAR. No single system performance on the test set is available in the publications, so it is difficult to conclude whether our system built on Librispeech significantly outperforms each individual system used for fusion. Fusion approach can be further investigated via our approach in the future.

Confusion matrix of the best results on the test set is shown in Table 5. The most confusable classification are between Hindi and Telugu, which are both languages used in India. Similar observations were found in the systems reported in [11, 12], although with a lower count.

3.4. Large-scale L1 Recognition

In order to evaluate the performance of the NLI system in a more realistic context with a broader range of L1s, we further evaluated it on a data set containing 25 L1 languages. The data set with those additional 14 L1 languages is smaller than those 11 L1 languages. The dataset consists of 3,000 responses (120 for each L1) used for training and 750 responses (30 for each L1) employed for testing. Each response is roughly 45 seconds long. There are no overlapped speakers between the training and test sets. Figure 1 depicts the performance in terms of Acc obtained by baseline and our approach for 25 L1s identification along with those of 11 L1s identification with different training sizes. Our approach significantly outperforms the baseline for all the configurations.

With more training data, performance of both the baseline (NLI11_baseline) and our approach (NLI11_TDNN) for 11 L1s recognition is continuously improved. The best performance does not seem to converge with 800 responses, i.e., roughly 10 hours per native language, for training. The performance of our approach on recognizing 25 L1s (NLI25_TDNN) is largely degraded to 60.1% Acc. We think it is mainly caused by limited training data for those 25 L1 NLI. A 10.1% accuracy reduction is observed for 11 L1 NLI when the training size is decreased from 300 to 120 responses per L1. Approximately 60% accuracy for 25 L1 recognition has ever been reported in [9] but the system is evaluated on fisher and the foreign-accented English (FAE) databases. Their system and our system, cannot be compared, since different databases were used for training. However, the relative low performance of large-scale NLI is still challenging and justifies further investigations.

Table 5: *Confusion matrix of the best results (Acc=87.1% and UAR=87.2%) on the test set (rows: references; column: hypotheses)*

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	65	1	3	1	1	1	6	0	1	0	1
CHI	0	67	0	3	0	0	1	3	0	0	0
FRE	3	1	65	2	0	0	0	1	4	1	1
GER	0	0	1	73	0	0	0	0	1	0	0
HIN	1	0	0	0	63	0	0	0	0	18	0
ITA	1	0	1	1	0	58	0	0	7	0	0
JPN	1	0	0	0	0	0	70	3	1	0	0
KOR	0	7	0	0	0	0	5	67	1	0	0
SPA	0	1	2	1	0	2	1	1	69	0	0
TEL	0	0	0	0	16	0	0	0	0	72	0
TUR	2	2	0	0	0	0	0	0	0	0	86

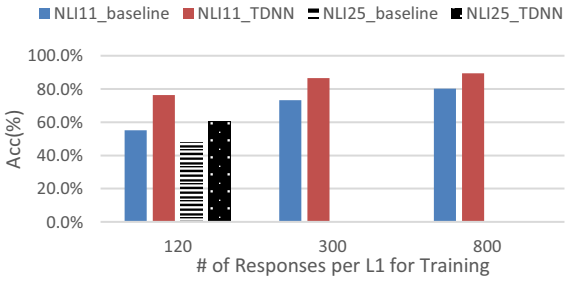

Figure 1: *Performance obtained by baseline and our approach for 25 L1s along with those of 11 L1s with different training sizes.*

4. Conclusions

TDNN and LVCSR corpora collected worldwide are explored for improving sub-phone modeling so as to enhance native language identification performance, based upon a non-native English speaker's speech input. The NLI performance, measured in Acc and UAR, is improved significantly with the help of sufficient statistics extracted from the TDNN trained with the LVCSR corpus collected worldwide. Our future work will be on investigating the L1 recognition of more languages and the fusion of multiple systems with different features and classifiers.

5. References

[1] X. Qian, H. M. Meng, and F. K. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training," *in Proc. of INTERSPEECH,* pp.775-778, 2012.

[2] Y. Qian, J. Tao, D. Suendermann-Oeft, K. Evanini, A. V. Ivanov, V. Ramanarayanan, "Noise and Metadata Sensitive Bottleneck Features for Improving Speaker Recognition with Non-Native Speech Input," *in Proc. of INTERSPEECH*, pp.3122-3126, 2016.
[3] S. H. Yang, M. Na1, M. Chung, "Modeling Pronunciation Variations for Non-native Speech Recognition of Korean Produced by Chinese Learners," in *Proc. of SLaTE,* pp.95-99, 2015.
[4] L.M. Arslan and J. H. L. Hansen, "A study of temporal features and frequency characteristics in American English foreign accent," *The Journal of Acoustical Society of America*, Vol. 102, No.1, PP. 28-40, 1997.
[5] M. Swan and B. Smith, editors, *Learner English: A teacher's guide to interference and other problems*. Cambiridge University Press, 2 edition, 2001.
[6] http://nlp.stanford.edu/software/stanford-dependencies.shtml
[7] J. Tetreault, D. Blanchard and A. Cahill, "A Report on the First Native Language Identification Shared," in *Proc. of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications,* pp. 48–57, 2013.
[8] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of INTERSPEECH*, pp. 2001-2005, 2016.
[9] M. K. Omar and J. Pelecanos, "A Novel Approach to Detecting Non-native Speakers and Their Native Language", in *Proc. of IEEE ICASSP*, pp. 4398~4401, 2010.
[10] "CSLU foreign-accented english corpus," http://www.cslu.ogi.edu/corpora/fae/
[11] A. Abad, E. Ribeiro, F. Kepler, R. Astudillo and I. Trancoso, "Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers", In *Proc. of INTERSPEECH*, pp. 2413-2417, 2016.
[12] P. G. Shivakumar, S. N. Chakravarthula, P. Georgiou, "Multimodal Fusion of Multirate Acoustic, Prosodic, and Lexical Speaker Characteristics for Native Language Identification", In *Proc. of INTERSPEECH*, pp. 2408-2412, 2016.
[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware Deep Neural Networks," in *Proc. of IEEE ICASSP*, pp. 1695–1699, 2014.
[14] D. Snyder, D. Garcia-Romero, D. Povey, "Time Delay Deep Neural Network-based Universal Background Models for Speaker Recognition" in *Proc. of IEEE ASRU,* 2015.
[15] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep Neural Networks for extracting Baum-Welch statistics for speaker recognition," in *Proc. of Odyssey 2014*, pp. 293–298, 2014.
[16] F. Richardson, D. A. Reynolds and N. Dehak, "A unified Deep Neural Network for speaker and language recognition," in *Proc. of INTERSPEECH*, pp.1146-1150, 2015.
[17] H. Zheng, S. Zhang, and W. Liu, "Exploring robustness of dnn/rnn for extracting speaker baum-welch statistics in mismatched conditions," in *Proc. of INTERSPEECH*, pp.1161-1165, 2015.
[18] N. Dehak, P. Kenny, R. Dehak, P. Ouellet, and P. Dumouchel, "Front end factor analysis for speaker verification," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 19, no. 4, pp. 788–798, 2011.
[19] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328–339, 1989.
[20] A. Waibel, "Modular construction of time-delay neural networks for speech recognition," Neural computation, vol. 1, no. 1, pp. 39–46, 1989.
[21] https://www.mathworks.com/help/nnet/ref/timedelaynet.html
[22] V. Peddinti, D. Povey, S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts, " in *Proc. of INTERSPEECH*, pp. 3214-3218, 2015.
[23] S. Ioffe*, "Probabilistic linear discriminant analysis," in *Proc. of ECCV-2006*, pp.531-542, 2006.
[24] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems, " in *Proc. of INTERSPEECH*, pp. 249-252, 2011.
[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in in *Proc. of IEEE ICASSP*, pp. 5206–5210, 2015.
[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. of IEEE ASRU,* 2011.