IMPROVING NATIVE LANGUAGE (L1) IDENTIFATION WITH BETTER VAD AND TDNN TRAINED SEPARATELY ON NATIVE AND NON-NATIVE ENGLISH CORPORA

Yao Qian[1], *Keelan Evanini*[1], *Patrick L Lange*[1], *Robert A Pugh*[1], *Rutuja Ubale*[1], *Frank K Soong*[2]

[1] Educational Testing Service Research, USA,
{yqian, kevanini, plange, rpugh, rubale}@ets.org,

[2]Microsoft Research Asia, China
frankkps@microsoft.com

ABSTRACT

Identifying a speaker's native language (L1), i.e., mother tongue, based upon non-native English (L2) speech input, is both challenging and useful for many human-machine voice interface applications, e.g., computer assisted language learning (CALL). In this paper, we improve our sub-phone TDNN based i-vector approach to L1 recognition with a more accurate TDNN-derived VAD and a highly discriminative classifier. Two TDNNs are separately trained on native and non-native English, LVCSR corpora, for contrasting their corresponding sub-phone posteriors and resultant supervectors. The derived i-vectors are then exploited for improving the performance further. Experimental results on a database of 25 L1s show a 3.1% identification rate improvement, from 78.7% to 81.8%, compared with a high performance baseline system which has already achieved the best published results on the 2016 ComParE corpus of only 11 L1s. The statistical analysis of the features used in our system provides useful findings, e.g. pronunciation similarity among the non-native English speakers with different L1s, for research on second-language (L2) learning and assessment.

Index Terms— native language identification, i-vector, time delay deep neural networks (TDNN)

1. INTRODUCTION

Native Language Identification (NLI) is the task of identifying the native language (L1) of a person based upon spoken or written input in a second language (L2). NLI from speech works under the assumption that a speaker's L2 production patterns are influenced by the L1. For example, many French English learners delete the glottal fricative /h/ and replace the dental fricatives /θ/ or /ð/ with /s/ and /z/, respectively, since French does not contain these phonemes. Another example is the epenthetic vowels that are commonly produced in syllable-final position by Japanese speakers of English; this is due to the fact that Japanese syllable structure does not allow syllable-final consonants (except for /n/). NLI from written input reveals common language-usage patterns in specific L1 groups. Foreign language learners with different L1 backgrounds do make different errors in grammar, vocabulary and other areas when learning a new language [4, 5].

In recent years, there has been growing interest in NLI for applications in second-language acquisition and forensic linguistics. The common pronunciation errors made by L2 learners are used to build an L1-L2 phone confusion table, which is useful for designing more L1 customized pronunciation training [1]. Linguistic origin analysis and dialectology can also be a useful tool for criminal intelligence and law enforcement agencies. In addition, L1 information or L1 detection scores can be used to improve (a) the performance of speaker recognition system with a phonetically-aware universal background model (UBM) [2, 9], (b) speech recognition by modeling pronunciation variation between native speakers and L1-specific speakers [3], and (c) other human-machine voice interface applications, e.g., facilitating a spoken dialog system, which can benefit from an awareness of the user's cultural background as suggested by the identified native language.

Most previous work on NLI involves only a limited number of L1s. The NLI shared task [7] hosted by the BEA workshop at NAACL 2013 and Computational Paralinguistics Challenge (ComParE) [8] at INTER-SPEECH 2016 show that automatic NLI methods can achieve 84% and 81% accuracy in detecting L1 from 11 different L1s based on writing and speech, respectively. Our sub-phone TDNN based i-vector approach [27] to NLI on the ComParE corpus outperforms the best system by a large margin (87% UAR, compared to 81% UAR of the best system from the 2016 ComParE challenge). In this paper, we improve our approach of NLI of non-native speakers' English speech input, evaluate it on an L1 detection task (with 25 L1s) and explore its applications to second language learning.

2. RELATED WORK

The approaches to automatic NLI are usually based on supervised learning where statistical models, e.g., SVM and GMM classifiers, are trained on data labelled with corresponding L1 information. The features extracted from English learners' written language are generally based on lexicon and syntax, including N-gram features on characters, words and part-of-speech (POS), Stanford

dependencies [6], and spelling and grammatical errors, while the features obtained from learners' speech for NLI can be based on frame-level information, such as MFCCs, phone-level confusions, and lexical features like language usage errors. The latter two features need to be provided by a speech recognizer with either a phone-loop grammar or a statistically trained language model.

Our study focuses on how to predict the L1 from non-native speakers' English speech input, similar to recognizing the spoken language or speaker. The GMM-UBM+MAP, GMM-UBM+SVM, and i-vector based approaches have been shown to be successful speaker and language recognition tasks, 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 was proposed to cluster the Gaussian components in the acoustic model from the ASR system and then integrated into an SVM-based classifier which achieved the best published results on the Foreign Accented English (FAE) database [10]. The ComParE participants' systems based on i-vectors achieve approximately 70% or higher 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 then fused together with the MFCC feature based i-vector system at the scoring level and achieves 81.3% UAR with the ComParE test set. Various features at the frame, phone and lexical levels, multiple classifiers, SVM, PLDA and DNN, and fusion at both the feature and score levels are investigated in detail in [12]. The performance of the final fused system is 79.9% accuracy and 80.1% UAR.

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 than a coarser, phonemic counterpart. We have proposed to adopt sub-phone modeling for extracting the i-vector. Motivated by the phonetically-aware deep neural network based ASR, e.g. DNN, RNN and TDNN, and its advantages in i-vector based speaker and language recognition [13-17, 35], we used a TDNN to build the UBM in our i-vector based NLI system [27]. The system achieves the best published results on the 2016 ComParE database.

Research shows that in the case of a large-scale L1 identification task involving 25 L1s with limited training data, the performance deteriorated to 60% accuracy [27]. Approximately 60% accuracy for L1 recognition involving 25 L1s was also reported in [9], but the system is evaluated on the Fisher and the FAE databases.

3. LARGE-SCALE L1 RECOGNITION

A schematic diagram of our large-scale L1 recognition system is depicted in Figure 1.

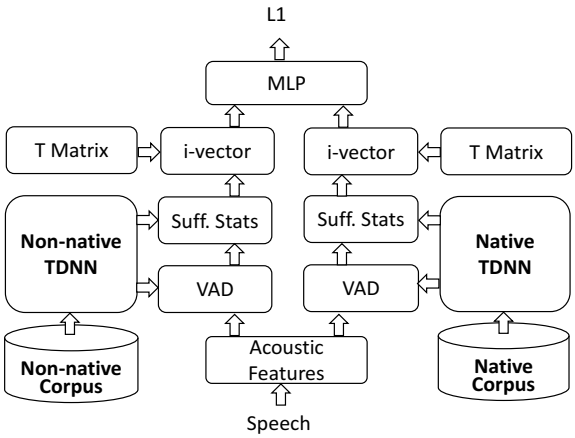

Fig. 1: Block diagram of large-scale L1 recognition system

3.1 Time Delay Deep Neural Networks (TDNN)

A feed-forward DNN can model a long time-span, as well as a high dimensional and correlated stochastic process. However, with only a long acoustic context, e.g. a moving window of many frames to represent the temporal structure of speech, the neural network architecture may not be able to capture the long-time dependencies between sequential events. A time delay deep neural network (TDNN) [19, 20] is more capable than a conventional feed-forward DNN to characterize the long time, sequential structure of a process like speech. TDNN is similarly formulated as a conventional feedforward network, but has delays on the layer weights associated with the input weights. The data can then be characterized at different time points by adding a set of delays to the input. This time-delay structure enables TDNN to have a finite, non-linear response to a time series input [21]. TDNN is similar to a convolutional neural network (CNN), but the convolution is only done along the time axis.

Recently, TDNNs have been shown to outperform DNNs in LVCSR and speaker recognition [22, 14]. A subsampling scheme, where hidden activations are computed at only few time steps, has been shown to be effective in speeding up TDNN training [22]. As a result, TDNN is also more attractive than a recurrent neural network (RNN), which can capture the temporal dynamics with its internal recurrent memory to process input sequences in arbitrary length, but takes a longer time to converge.

3.2 TDNN based i-vector extraction

An i-vector, which is a compact, low dimensional vector representation of speech, is a factor analysis based approach to modeling speech in a text-independent manner. To extract the i-vector, a low-rank, rectangular T-matrix, which represents the total variability in the acoustic space, is estimated with the EM algorithm. The sufficient statistics:

$$N_k(u) = \sum_t \gamma_{kt}^{(u)} \quad (1)$$

$$F_k(u) = \sum_t \gamma_{kt}^{(u)} (x_t^{(u)} - m_k) \quad (2)$$

are computed to learn the basis functions of the total variability subspace [18]. N(u) and F(u) are the 0^{th}-order and mean shifted 1^{st} order statistics for the *u*-th speech segment, respectively. $\gamma_{kt}^{(u)}$ is the stochastic alignment of the *t*-th frame of speech feature vectors, $x_t^{(u)}$, i.e., the posterior probability calculated from a UBM. A TDNN is used as a UBM and employed to extract EM sufficient statistics. Given TDNN-UBM, the $\gamma_{kt}^{(u)}$ is computed from the soft-max output of TDNN.

$$\gamma_{kt}^{(u)} = p(s_k \mid x_t^{(u)}) \quad (3)$$

where s_k is the *k-th* senone of TDNN outputs and $x_t^{(u)}$ is a spliced input vector.

3.3. TDNN based VAD

Separating speech from background non-speech frames has been shown to be a critical preprocessing module for efficient speech recognition and for constructing i-vectors to equalize the speaker-dependent effect in robust speech recognition [28]. A voice activity detection (VAD) method based on energy thresholds, commonly used to remove non-speech segments within a speech utterance, is not capable of performing robust VAD since the energy level is too crude a feature to separate speech from non-speech reliably, particularly when the background noise is non-stationary. There are many studies on using statistical model-based approaches to VAD [36-40]. Given sufficient training data (in this case, several hundred hours of speech accompanied by the corresponding transcriptions), we can detect speech with a much higher confidence using a TDNN over an energy based VAD. A phone posteriorgram, i.e., the posteriors of the phones over time, is generated by summing up the posteriors of senones (the output nodes of TDNN) of the same phone. We can skip those frames where the background (silence) has the highest phone posterior in constructing the supervector, hence the i-vector.

3.4 MLP for L1 recognition

Probabilistic Linear Discriminant Analysis (PLDA) scoring [23], due to its state-of-the-art performance in speaker recognition, has also been successfully applied to L1 recognition [12, 27]. Given a pair of two i-vectors, PLDA calculates the log likelihood ratio that the hypothesis that two i-vectors share the same L1 identity to the hypothesis that two i-vectors are from different L1s. However, a discriminative classifier should be more appropriate to L1 recognition when there is adequately large training data from each L1. We investigate multi-layer perceptron (MLP) based multi-class classification for 25 L1s when i-vectors are used as inputs.

3.5 Native and non-native TDNNs

Two LVCSR corpora are used to train TDNNs as UBMs and to calculate sufficient statistics for i-vector extraction. The first one is a non-native English speech corpus collected worldwide. Another one is a corpus of native US English speech. These two corpora have different degrees of coverage of the phonetic space. The native English corpus contains authentic phone pronunciations and the projection to such a space should be distinguishable for pronunciation errors. The non-native English corpus covers the phone variabilities caused by different accents of non-native English speakers and L1 effects. The projection of the affected English speech to this space can help distinguish different L1s. The i-vectors extracted from the TDNNs trained by the two corpora can be used jointly for proper L1 recognition. The two corresponding i-vectors are augmented as input to the final MLP for L1 identification.

4. EXPERIMENTS AND RESULTS

4.1 Corpora

The proposed approach, based upon TDNN derived i-vectors, is evaluated on two L1 recognition corpora: one with 11 L1s and the other with 25 L1s. The first corpus was used 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 speech sampled at 16 kHz from 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). The second corpus of 25 L1s was collected from the same global English assessment used in the ComParE corpus. This corpus also includes non-native speakers with the 14 additional L1 backgrounds: Bengali (BEN), Farsi (FAS), Gujarati (GUJ), Nepali (NEP), Marathi (MAR), Indonesian (IND), Portuguese (POR), Romanian (RUM), Russian (RUS), Tagalog (TGL), Thai (THA), Urdu (URD), Tamil (TAM) and ENG (test takers who list English as their L1 but are still required to take an English proficiency test for university admissions based on their location of residence). The number of utterances for each data partition in these two corpora is shown in Table 1. One utterance is included for each speaker and each utterance is roughly 45 seconds in duration. Utterances from different L1s are evenly distributed in the 25 L1 corpus.

As mentioned above, a Native English corpus and a non-native English corpus are used to train separate TDNNs

to create sufficient statistics for the corresponding i-vectors. The native corpus is LibriSpeech [25], 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 of them are U.S. English. The acoustic model trained on this corpus shows a high performance on U.S. English speech recognition test, 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, along with the normalized audiobook text are employed to train the TDNN. The non-native English corpus consists of 800 hours of speech from 8,700 test-takers from over 100 countries. It is drawn from the same English proficiency assessment as the NLI corpus. All the recordings and corresponding manual transcriptions are used for training the TDNN. The metadata, including country information, is not used for training.

Table 1: The number of utterances for each data partition

	Train	Dev	Test
11 L1s	3,300	965	867
25 L1s	17,500	2,500	5,000

4.2 Experimental setup

NLI systems are constructed with Kaldi [26], Keras [29] and SKLL [30]. The front-end for the baseline NLI system contains 20-dimensional MFCCs including C0, extracted from a 20 ms Hamming window with a 10 ms time shift, along with their first and second ordered time derivatives. Non-speech segments within utterances were removed with the TDNN based VAD, as described in Section 3.3. Utterance-based cepstral mean normalization was performed on the acoustic feature vectors.

40-dimensional MFCCs are used as input features to train the TDNNs. Speaker adaptation techniques such as transforming acoustic features by fMLLR and augmenting the i-vectors with input features are not employed here. The TDNN architecture is similar to the one described in [22]. The context specification of the TDNN is configured as follows: 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 the native corpus and the non-native corpus, respectively. The output nodes are the senones of the HMM determined by decision-tree based clustering. The input and output feature pairs are obtained through frame alignment for the senones with the GMM-HMM, which are also trained separately for the native and non-native corpora.

Our previous work [27] shows that 600-dimensional i-vectors can achieve the best performance on the 11 L1 corpus. The same dimension of i-vector is used for all experiments in this paper. A fully-connected MLP with 2 hidden layers is used to predict the L1 from the i-vectors; each hidden layer consists of 512 nodes; the rectified linear unit (ReLU) activation function and dropout (with p=0.5) are used for all hidden layers; softmax is used for the output layer and categorical cross-entropy is employed as the training loss function. Different stochastic gradient descent (SGD) optimization algorithms, e.g., RMSprop, Adadelta, and Adam, were evaluated together with the MLP structure on the development data set. The Adam algorithm achieved slightly better performance than other algorithms and is used in this study.

4.3 Results and analysis

Table 2 lists the accuracy obtained by different NLI systems on the test sets of the 11 L1 and 25 L1 corpora. The baseline system, which employs a sub-phone TDNN based i-vector approach, energy-based VAD and PLDA for L1 recognition, can obtain identification accuracies of 82.7% and 75.6% with i-vectors extracted from the native TDNN, and 87.1% and 78.7% with i-vectors extracted from the non-native TDNN on the test sets of 11 L1s corpus and 25 L1s corpus, respectively. The systems with non-native TDNNs for VAD can outperform the baseline systems, i.e., 0.4% and 1.1% accuracy improvements on the 11 L1 and 25 L1 tasks, respectively. However, the performance of the systems with the native TDNN based VAD is inferior to that of the baseline system. We think this is caused by the mismatched recording environments among the corpus used for L1 recognition and the corpus used for TDNN training. The discriminative MLP classifier is only efficient when the number of training samples per L1 is large as in the 25 L1 corpus. The fusion systems with the concatenated i-vectors extracted from both the native and non-native TDNNs can achieve improvements of 1.0% and 3.1% in accuracy over the baseline system for the 11 L1 and 25 L1 tasks. We also adopt score-level fusion by using the L1 posterior outputs generated from the two MLPs as input features to a logistic regression classifier for predicting L1 labels. The results for this approach show that the performance of score-level fusion is on par with that of feature-level fusion.

Table 2: Accuracies obtained by different NLI systems on the test sets of two NLI corpora

	11 L1s	25 L1s
Native TDNN (baseline)	82.7	75.6
Native TDNN (VAD)	82.4	75.3
Native TDNN (MLP)	82.5	**76.5**
Native TDNN (VAD+MLP)	82.2	76.4
Non-native TDNN (baseline)	87.1	78.7
Non-native TDNN (VAD)	**87.5**	**79.8**
Non-native TDNN (MLP)	86.9	**80.1**
Non-Native TDNN (VAD+MLP)	87.3	**80.7**
Native and Non-native TDNNs	**88.1**	**81.8**

Motivated by the good performance of text-based NLI, e.g., essays written by English learners with different L1 backgrounds, we use the recognition hypotheses to enhance

the NLI system. A state-of-the-art DNN based non-native speech recognition system is used to decode the utterances in the L1 recognition corpus. The i-vector based speaker adaptation technique is used in the ASR system to compensate for the mismatch between trained models and testing speakers' data. Since word transcriptions of utterances are not available in the L1 corpus, the recognition accuracy cannot be obtained, but a WER of 18.5% on a monologic non-native speech test set reported in [31] can be used as a reference. A support vector machine (SVM) classifier with word-based *n*-grams and a sophisticated CNN based sentence classifier with a word2vec front-end are used for L1 identification with speech recognition hypotheses. The preliminary results show that it can only achieve less than 30% accuracy on the 25 L1 task. In [12], the authors report an accuracy of 44.6% on the 11 L1 task by using the 1,000 best recognition hypotheses. Automatic transcription does not contain any spelling errors. Although language use and grammatical errors can be captured from recognition hypotheses, the approach still suffers from inaccurate speech recognition results. The strong language model used in ASR automatically corrects certain grammatical errors.

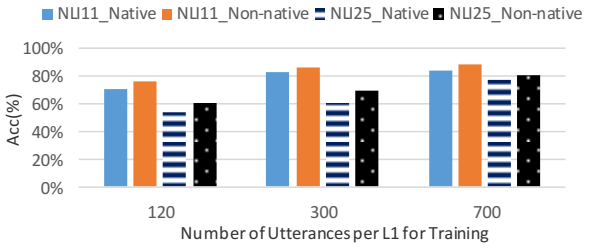

Fig. 2: Accuracy obtained by our approach for 11 L1s and 25 L1s with different training sizes.

Figure 2 depicts the performance in terms of accuracy obtained by our systems with the TDNNs trained from native and non-native corpora for the 11 L1 and 25 L1 tasks with different training set sizes. Here the data used for the 11 L1 task is a subset of the 25 L1 corpus and is different from the 11 L1 corpus introduced in Section 4.1. With more training data, performance for both NLI tasks (11 L1s and 25 L1s) improves continuously. The best performance does not seem to converge with 700 utterances for training, i.e., roughly 9 hours per L1. The performance gap between the 11 L1 and 25 L1 tasks is also smaller with more training data.

The confusion matrix of the best results on the test set with 25 L1s and the corresponding F_1-score of the individual L1s are shown in Figure 3 and Table 3, respectively. The most distinguishable L1s are Chinese (CHI), Thai (THA), Turkish (TUR), German (GER), Portuguese (POR), Korean (KOR), French (FRE) and Japanese (JPN), which can achieve over F_1-scores over 0.9. The performance of Hindi (HIN) recognition is worst (0.39 F_1-score). The most confusable L1s with Hindi are Gujarati (GUJ), Marathi (MAR), and Urdu (URD), which are languages used in India that belong to the same language family as Hindi, i.e., the Indo-Aryan language family. The second worst performance (0.53 F_1-score) is obtained on ENG recognition. This is not surprising since some speakers who live in countries where English is an official language, but typically not the first language that is learned, e.g., the Philippines, India and Hong Kong, claim English as a native language. The presence of various accents in the speech produced by these speakers makes the task more difficult.

	ARA	BEN	CHI	ENG	FAS	FRE	GER	GUJ	HIN	IND	ITA	JPN	KOR	MAR	NEP	POR	RUM	RUS	SPA	TAM	TEL	TGL	THA	TUR	URD
ARA	174	2	0	5	1	5	1	0	0	0	2	0	0	1	0	0	0	1	1	0	1	3	0	2	1
BEN	1	131	0	7	0	0	0	3	16	1	1	2	0	2	10	2	0	1	0	6	3	3	1	0	10
CHI	0	0	191	0	0	0	0	1	0	0	0	2	2	0	0	1	0	0	0	0	0	0	2	0	1
ENG	5	1	7	109	0	2	7	4	16	2	0	5	5	2	1	2	0	2	4	1	8	5	2	1	9
FAS	9	1	0	5	172	0	3	0	0	1	0	1	0	0	0	1	0	0	1	0	0	2	1	0	3
FRE	0	0	0	3	0	188	2	0	0	0	2	0	0	0	0	1	0	1	1	0	0	0	1	1	0
GER	0	0	0	0	0	0	196	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	2	0
GUJ	1	6	0	10	0	0	0	101	25	0	0	0	0	22	6	0	1	0	0	7	5	1	0	0	15
HIN	0	6	0	8	2	0	0	22	78	1	1	0	0	31	4	0	0	0	0	7	17	0	0	1	22
IND	1	0	2	7	0	0	0	0	0	175	0	1	2	0	0	0	0	3	0	0	0	7	2	0	0
ITA	0	0	0	4	0	4	2	0	0	0	177	0	1	0	0	3	1	1	7	0	0	0	0	0	0
JPN	0	0	0	4	0	0	0	0	0	0	0	187	9	0	0	0	0	0	0	0	0	0	0	0	0
KOR	0	0	1	3	0	1	0	0	0	0	0	6	189	0	0	0	0	0	0	0	0	0	0	0	0
MAR	0	4	0	0	0	0	0	18	24	0	0	0	0	123	2	1	0	0	0	9	11	0	0	1	7
NEP	0	9	0	3	0	2	1	1	3	1	0	1	0	0	171	0	0	0	0	0	0	4	0	0	4
POR	0	0	0	5	0	0	0	0	0	0	1	0	1	0	0	188	1	1	1	0	1	0	0	0	1
RUM	0	0	0	1	0	2	0	0	0	2	7	1	0	0	1	4	164	12	1	0	0	2	0	0	3
RUS	0	1	1	4	0	1	3	0	0	0	0	1	0	0	0	0	10	173	1	0	0	2	0	2	1
SPA	0	0	0	5	0	0	3	0	0	0	2	0	0	0	1	3	0	0	178	0	1	5	2	0	0
TAM	0	4	0	8	0	2	0	3	10	0	0	0	0	6	0	0	0	1	0	129	33	0	0	0	4
TEL	0	3	0	4	0	0	0	3	8	0	0	0	0	8	3	1	0	0	0	17	145	0	2	0	6
TGL	0	0	0	5	0	0	0	1	0	2	0	0	0	0	1	1	0	0	2	0	0	188	0	0	0
THA	0	0	0	4	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	192	0	0
TUR	0	0	0	2	6	0	1	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	1	186	1
URD	1	4	1	8	4	0	2	4	19	1	0	1	0	4	3	0	1	1	0	5	12	0	0	1	128

Fig. 3: Confusion matrix of the best results on the test set of 25 L1s (references in rows, predictions in columns)

Table 3: The F_1-score of the individual L1s obtained by the best system for the 25 L1 task

L1s	F_1-score	L1s	F_1-score	L1s	F_1-score
HIN	0.39	RUS	0.87	FRE	0.92
ENG	0.53	RUM	0.87	KOR	0.92
GUJ	0.56	ARA	0.89	POR	0.92
MAR	0.62	FAS	0.89	GER	0.93
URD	0.62	SPA	0.89	TUR	0.94
TEL	0.66	TGL	0.89	THA	0.95
TAM	0.68	IND	0.90	CHI	0.95
BEN	0.70	ITA	0.90		
NEP	0.85	JPN	0.91		

4.4 Using NLI features to explore L1 transfer

Language transfer refers to the phenomenon that a speaker's L2 production is influenced by L1 knowledge, in either speaking or writing. In second-language acquisition, language transfer makes the acquisition process quite different from first-language acquisition. Language transfer can occur in all aspects of linguistic production, including grammar, pronunciation, vocabulary, and discourse [32]. Features for NLI based on writing have been used to study language transfer effects in second-language acquisition [33]. In our study, we focus on the speech features for NLI and investigate language transfer on pronunciation by

projecting the L2 speakers' speech to a L1 speakers' sub-phonemic space, which is modeled by the native TDNN.

I-vectors are compact acoustic-phonetic representations of speech in a subspace, i.e., the supervectors (in the subphonemic space) are projected unto a lower dimensional i-vector space. Due to this projection, each coordinate in the i-vector space has lost its original phonetic meaning. To keep the phonetic labels in the original supervector, we use senone posteriors, or the senone posteriorgram averaged over the whole utterance after removing silence frames. Kullback-Leibler Divergence (KLD), an information-theoretic measure of dissimilarity between two probability distributions, is used to measure the similarity among the English speech spoken by L2-speakers with different L1 backgrounds. We employ the symmetric KLD between discrete probability distributions P and Q, defined as

$$D_{KL}(P,Q) = \sum_i p(\mathrm{s}_i \mid x) \log \frac{p(\mathrm{s}_i \mid x)}{q(\mathrm{s}_i \mid x)} + \sum_i q(\mathrm{s}_i \mid x) \log \frac{q(\mathrm{s}_i \mid x)}{p(\mathrm{s}_i \mid x)} \quad (4)$$

where s_i is the i-th senone; x, the acoustic feature vector; P and Q, senone posterior vectors from different L1s. Each L1 is represented by one senone posterior vector, created by averaging all of the samples associated with that L1.

Table 4: The 10 lowest KLD and the 5 highest KLD of L1 pairs among 600 comparison pairs.

L1	Pairs	KLD
Gujarati	Marathi	0.0048
Portuguese	Spanish	0.0058
Marathi	Hindi	0.0074
Gujarati	Hindi	0.0105
Bengali	Hindi	0.0115
Gujarati	Bengali	0.0123
Arabic	Spanish	0.0123
Bengali	Marathi	0.0130
Tamil	Telugu	0.0136
Italian	Spanish	0.0145
…		
Telugu	Tagalog	0.2611
Nepali	Telugu	0.2682
Japanese	Tamil	0.2833
Thai	Telugu	0.3071
Telugu	Japanese	0.3540

Table 4 lists the L1 pairs of the 10 lowest KLDs and the 5 highest KLDs among 600 possible L1 pairs. A lower KLD score indicates a higher similarity. The nearest neighbors, in the KLD sense, tend to cluster L1s in the same language families and the L1 pairs with high KLDs are those in different language families. Interestingly, while Arabic and Spanish are from different language families, the KLD is still small. It may indicate the fact that Arabic has had a great influence on Spanish. There are ~ 5,000 words in the Spanish language with Arabic origin [34]. We conjecture that English learners in these two L1s share some common pronunciation behaviors.

To investigate the pronunciation of specific English phones by speakers from different L1 backgrounds, we convert the senone posterior vector to a phone posterior vector by summing the posteriors of the senones belonging to the same phones. Consequently, we build a 41*25 matrix, where each L1 has a 41-dimensional vector (i.e., the 41 phones used in this study) and each phone has 25 scores (for 25 L1s). Phones with relative low scores in an L1 compared with scores in other L1s may represent phones for which speakers from that L1 have difficulty pronouncing in a native-like manner. Table 5 lists examples of phones that were detected using this method; e.g., it suggests that Japanese and Thai speakers may have English difficulties pronouncing the phones /l/, /z/, /dʒ/, /tʃ/ and /ər/, etc. Some of the phone-L1 pairs detected by this method correspond to well-known patterns of English foreign accents due to L1 transfer (such as /l/ for Japanese learners of English and /h/ for French learners of English) whereas the cause of some of the other pairs is less clear. Future research will investigate how these phone-L1 pairs can inform research into second language acquisition and language assessment.

Table 5: Some examples of English phones with low scores based on phone posterior vectors by L1 background

Phone	L1s
/h/:	French, Arabic, Tagalog, Italian
/l/:	Japanese, Nepali, Thai
/z/:	Japanese, Thai, Tagalog, Nepali
/tʃ/:	Chinese, Thai, Russian, Japanese,
/dʒ/:	Japanese, Thai, Chinese
/ər/:	Japanese, Thai, Arabic
/ɑ/:	Telugu, Tamil, Hindi, Bengali, Marathi, Gujarati
/ɔ/:	Telugu, Tamil
/aʊ/:	Italian, Urdu, French
…	

5. CONCLUSIONS

TDNNs trained separately on native and non-native English LVSCR corpora are investigated to improve sub-phone based acoustic models to identify 25 native languages based on non-native spoken English. TDNN-based VAD and a highly discriminative classifier are shown to improve L1 recognition. Native language identification accuracy for 25 L1s is improved by 3.1%, from 78.7% to 81.8%. Senone posteriorgrams of a speaker's speech are used to analyze the effects of second-language acquisition and language transfer among L2 English learners. Future work will investigate additional features and applications to human-machine voice interface design.

6. 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*. Cambridge 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.

[27] Y. Qian, K. Evanini, X. Wang, D. Suendermann-Oeft, R. A Pugh, P. L Lange, H. R Molloy and F. K Soong, "Improving Sub-phone Modeling for Better Native Language Identification with Non-native English Speech", *Accepted by INTERSPEECH,* 2017.

[28] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey and S. Khudanpur, "JHU Aspire System: Robust LVCSR with TDNNs, I-vector Adaptation and RNN-LMs", in *Proc. of IEEE ASRU,* 2015.

[29] https://keras.io/

[30] https://github.com/EducationalTestingService/skll

[31] Y. Qian, X. Wang, K. Evanini, D. Suendermann-Oeft, "Self-Adaptive DNN for Improving Spoken Language Proficiency Assessment", in *Proc. of INTERSPEECH*, 2016.

[32] R. Ellis, *The Study of Second Language Acquisition.* Oxford, UK: Oxford University Press. ISBN 978-0-19-442257-4, 2008.

[33] S. Malmasi and D. Mark, "Language Transfer Hypotheses with Linear SVM Weights." In *Proc. of EMNLP*, 2014.

[34] https://en.wikipedia.org/wiki/Talk%3AList_of_Spanish_words_of_Arabic_origin

[35] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "The IBM 2016 speaker recognition system," in Proc. of Odyssey: Speaker and Language Recognition. Workshop, pp. 174–180, 2016.

[36] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies," in *Proc. of ICASSP*, pp. 483–487, 2013.

[37] T. Hughes and K. Mierle, “Recurrent Neural Networks for Voice Activity Detection”, in *Proc. of ICASSP*, pp. 7378-7382, 2013.
[38] X.L. Zhang and D.L. Wang, “Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection”, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol.24, No.2, 2016.
[39] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” Signal Processing Letters, IEEE, vol. 6, no. 1, pp. 1–3, 1999.
[40] M. V. Segbroeck, A. Tsiartas and S. S. Narayanan, “A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice”, In *Proc. of INTERSPEECH*, pp. 704-708, 2013.