INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Multimodal Fusion of Multirate Acoustic, Prosodic, and Lexical Speaker Characteristics for Native Language Identification

Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, Panayiotis Georgiou

University of Southern California, Los Angeles, CA, USA
{pgurunat,nallanch}@usc.edu, georgiou@sipi.usc.edu

Abstract

Native language identification from acoustic signals of L2 speakers can be useful in a range of applications such as informing automatic speech recognition (ASR), speaker recognition, and speech biometrics. In this paper we follow a multistream and multi-rate approach, for native language identification, in feature extraction, classification, and fusion. On the feature front we employ acoustic features such as MFCC and PLP features, at different time scales and different transformations; we evaluate speaker normalization as a feature and as a transform; investigate phonemic confusability and its interplay with paralinguistic cues at both the frame and phone-level temporal scales; and automatically extract lexical features; in addition to baseline features. On the classification side we employ SVM, i-Vector, DNN and bottleneck features, and maximum-likelihood models. Finally we employ fusion for system combination and analyze the complementarity of the individual systems. Our proposed system significantly outperforms the baseline system on both development and test sets.

Index Terms: language nativity detection, i-vectors, VTLN, Phoneme-level prosodic features, phonemic log-likelihood features, Deep neural network, bottleneck features, L1, fMLLR

1. Introduction

Speech signals, in addition to the explicitly expressed lexical content, contain a diverse range of information about speakers such as age, emotions, speaker identity, environment characteristics, language background of the speaker *etc.*. Capturing and describing such diverse information enables adaptation and improved performance of speech processing systems. One of these important characteristics to capture is the native language of the speaker.

Identification of the native language (L1) of a non-native English speaker from English (L2) speech is a challenging research problem. Knowledge of the native language can aid Automatic Speech Recognition systems through specifically tuned models, can provide culturally aware machine-human interfaces and can provide cues towards more accurate speaker recognition, speech biometrics and speech forensics by effectively modeling the phonotactic variability of speakers across various languages.

There has been relatively less research in the area of native language detection. Most of the research involves study with 2 to 4 way classification. In [1], a support vector machine (SVM) was used to classify 8 native languages using ASR based features under a universal background model (UBM) framework. Shriberg *et al.* [2] used multiple approaches based on lexical systems by using phone and word N-gram language models (LM) to show that the word based N-gram LM was more effective than a phone based one. Several studies have shown prosodic information like energy, duration, pitch, and formant based functionals to be effective features [2–4]. The native language identification task was found to be particularly difficult for spontaneous speech [3]. On the acoustic front, Gaussian Mixture Models (GMM) have been used to train a model specific to different accents [5]. For training such GMMs front-end acoustic features in the form of Cepstral based features, like Perceptual Linear Prediction (PLP) [5] and Mel Frequency Cepstral Coefficients (MFCC) [3], and second and third formant features [4], have been employed. Different training techniques like Maximum Mutual Information (MMI) [5] and Minimum Phone Error (MPE) [1] were found to be useful. Stochastic trajectory models (STM) based on phonemes were successfully applied to capture the dynamics of accents specific to each phones [3]. An in-depth analysis of temporal characteristics of accents were performed in [6], showing significant differences between foreign accented English, hinting at the potential of the duration based features towards accent classification.

In this paper, we use acoustic features, MFCC and PLP of different time scales, in an i-Vector framework with probabilistic linear discriminant analysis (PLDA) to model the acoustic information. Deep neural networks (DNN) are used to derive bottleneck features, which in turn are used to train the i-vectors to boost the discriminative power of the frame level acoustic features. We introduce a *Pronounciation-Projection* (L1-ProP) feature by projecting acoustics in the English-language pronounciation space via an ASR, that can capture L1-specific phonemic mismatch. We also propose novel phoneme-level features in terms of *Phonemic Confusion* (PC) and *Phoneme Specific Prosodic Features* (PSPS) which are designed to capture the confusability and the short term prosody-dynamics on phone level. On the lexical front, the grammatical variations on word level persistent in specific languages are exploited. Finally, the introduced features are fused together along with the baseline features for classification. The experimental results are presented on the ETS corpus of non-native spoken English comprising of 11 distinct L1 speakers, as a part of Interspeech Native Language Sub-Challenge [7].

The rest of the paper is organized as follows. First, the database and baseline system are briefly described in Section 2. We then describe the features employed in Section 3 and the classification algorithms in Section 4. We provide a brief description of our fusion method in Section 5 before we proceed to analysis of our results in Section 6. We conclude and provide future directions in Section 7.

2. Database and Baseline System

2.1. Database

The Educational Testing Service (ETS) corpora used in this work is built on the spontaneous speech of non-native English

Copyright © 2016 ISCA

http://dx.doi.org/10.21437/Interspeech.2016-1312

speakers taking the TOEFL IBT exam. The corpora consists of 5,132 speakers from 11 L1 backgrounds with approximately 64 hours of speech (45s per speaker). The 11 L1 categories were Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish. Additional details of division of data, speakers and L1 classes, for training, development and testing corpora is available in [7].

2.2. Baseline System

The baseline for our system is trained using utterance level statistics of the acoustics such as spectral (e.g., formants), cepstral (e.g., MFCC, RASTA), tonal (e.g., CHROMA, CENS) and voice quality (e.g., jitter, shimmer), for a total of 6373 dimensions, extracted using OpenSMILE [8]. The features are used to train a support vector machine (SVM) to classify among the 11 L1 categories and details can be found in [7].

3. Features

In our proposed system, we use multirate information from acoustic, prosodic, phoneme-confusability, phoneme-level prosodic, and lexical streams to train complementary multiple expert systems. The features were tailored to capture (i) discriminative information between the 11 non-native L1 language (ii) discriminative variability with respect to native English speaking patterns.

3.1. Frame-level Features

On the acoustic front-end, we use MFCC, PLP and log power-spectral features due to their success in prior work [3, 5]. We use multiple streams of acoustic features to capture variability in terms of multiple temporal (25ms to 150ms frame size) and spectral resolutions (23-69 mel-filterbanks with 13 to 39 MFCC). The delta and delta-delta features were computed and mean normalized.

VTLN: To reduce inter-speaker variability we can employ speaker normalization techniques such as Vocal Tract Length Normalization (VTLN) [9], Maximum Likelihood Linear Regression (MLLR) [10], and Speaker Adaptive Training techniques (SAT) [11]. In our work we employ linear-VTLN via an affine transformation to approximate the non-linear warping of the frequency axis similarly to the method in [12]. It is unclear however if such normalization also removes L1 specific features, something we intend to investigate.

3.2. Bottleneck features

Bottleneck features were shown to be useful for speaker recognition [13] and language identification [14] task. We generate bottleneck features via a DNN with a 23 frame context input of 257-dimensional log-spectra that mirror the human auditory system [15]. The DNN thus has a 5911 dimensional input and 3 hidden layers with 2000, 50 and 500 neurons.The 50-dimensional bottleneck features along with their delta and delta-delta features are mean normalized and used to train the total variability matrix of the i-vector framework.

3.3. Phoneme-level Features

Past studies have demonstrated the influence of L1 backgrounds on L2 speakers' pronunciation of English vowels and consonants [16–19]. Different backgrounds are associated with specific perceptual errors in recognition between different phonemes. For instance, strong confusion has been observed between Japanese speakers' pronunciation of $/l/$ and $/r/$ phonemes [20] and between $/n/$ and $/l/$ for Chinese speakers [21]. Wiltshire et al observed Gujarati and Tamil influences on pitch accents and slopes, similar to those that Arslan et al observed with Mandarin and German [6, 22]. Phoneme durations have been shown to be a prominent feature characterizing accents and dialects [6] as well.

Such traits are likely complementary to the frame-level acoustic features. Capturing such traits involves a projection of the speaker characteristics on the English-language space and the analysis of this projection. This can be practically implemented as the projection to the likelihood space of each phoneme via a speech recognizer. We can also employ this projection in several ways:

3.3.1. L1-Pronounciation Projection (L1-ProP)

The L1-ProP features are designed to capture the pronunciation variability between the L1 English speakers and the L2 speakers. Since different languages employ a different phonetic inventory, we hypothesize that this will create specific responses in the phonemic projection of L2 English speech on the native English speech space. To obtain a compact projection we used a mono-phone phoneme recognizer trained on native English speakers [23] using the Kaldi toolkit [24]. The frame level log-likelihood score is obtained from the ASR monophone model using the following criterion:

$$LL_p = \max_{s \in S_p} log(P(f|s)) \qquad \forall p \in P \tag{1}$$

where p is a phone from set of phones, P, s is the state from the set of states, S_p, specific to phoneme p, f is the frame. For each frame, we get a 41 dimensional vector corresponding to log-likelihoods for 39 non-silence and 2 silence phones. In short we select the best match per phoneme for all the various states belonging to that phoneme. We further explored projection on a range of different languages.

3.3.2. Phonemic Confusion

To obtain the phoneme confusion features, we used the phoneme likelihoods described in Sec. 3.3.1. We want to investigate phoneme confusion so we generated a pairwise-confusion matrix from the cross-product of the 39 dimensional confusion log likelihoods. We then vectorize the lower-triangular elements and obtain the average confusion vector per phoneme from its instances as determined by the ASR. Finally, we average this vector over all phonemes to obtain a 780-dimensional feature per file.

3.3.3. Phoneme Specific Prosodic Features (PSPS)

Prosodic variability has been shown to be useful in native language identification. The baseline features employ prosody with success. We also hypothesize that phone-specific prosodic variability can provide useful information. Based on phoneme alignments obtained by the ASR above we compute the mean, standard deviation, median, min, max and range of phoneme duration, short-time energy and pitch (only for voiced). We then average over each phoneme type (*i.e.,* over all "AA" phonemes, over all "B" phonemes *etc.*). This results in a 1062-length feature vector over all phonemes (30 features $\times$ 30 voiced phonemes, 18 $\times$ 9 unvoiced). In case a phoneme is not observed in a session, we impute its features using the global averages from other train sessions where it was observed.

3.4. Lexical features

We believe that lexical channel can capture 2 types of information: 1. the style of expression and language use errors will vary according to the native language of the speaker; and 2. an ASR transcript will contain consistent errors based on consistent mispronounciations resulting from L1 specific phonemic confusability. Given the limited lexical data and the error associated with recognizing L2 speech we decide to employ the 1000 n-best list of each utterance of each file as our lexical representation of each speaker. Decoding was done using a DNN-ASR system trained on the Fisher corpus.

3.5. fMLLR Transform based Features

Feature-space Constrained Maximum Likelihood Linear Regression (fMLLR) is a linear transformation used for speaker and environment adaptation in modern ASRs such that it maximizes the observation data likelihood given the model [11]. While it removes a lot of this variability it may also remove native language specific information, we thus decide to investigate whether the 39×40 dimensional fMLLR transform conveys native language information and employ it as a feature.

4. Classification Techniques

4.1. i-Vector

Recently, i-vector modeling was introduced in application to the task of speaker verification [25]. Its excellent state-of-the-art performance gained significant research interest among the signal processing community. The total variability modeling of i-vectors have since been applied to various tasks like language recognition [26], speaker recognition [27], speaker age recognition [28,29]. For our work we use total variability i-vector modeling. We train a full covariance GMM-UBM on the ETS Corpus training dataset. The UBM was trained using 2048 gaussian mixtures. The zeroth and the first order baum welch statistics are computed from the training data and the total variability matrix is estimated using Expectation-Maximization. Finally, we extract mean and length normalized i-vectors.

4.2. PLDA

For scoring, we use probabilistic linear discriminant analysis (PLDA), due to its state-of-the-art results in speaker recognition domain [27]. Given a pair of i-vectors, PLDA evaluates the ratio of probability that the two i-vectors belong to the same native background to the probability that the two i-vectors are from different native backgrounds [30]. The log-likelihood scores obtained after PLDA scoring are used for classification.

4.3. SVM based phoneme-level feature classification

We implemented the phonemic confusability and prosodic features as described in Secs. 3.3.2 and 3.3.3. The session-level features were trained and tested using the same parameters as the baseline system using PolyKernel SVM and Weka [31].

4.4. Maximum Likelihood Lexical Classification

Given the limited lexical data we decided to use a simple *Maximum Likelihood* (ML) classification framework. We considered alternatives, such as a word2vec front end, however the embedings may preserve the lexical similarity but not necessarily the actual word biases of L2 speakers that we desire to capture. Models were smoothed with background data to ensure robustness and to boost the importance of domain-salient words. For transcript we used the 1000 best of each utterance in the test file similarly to [32].

5. Fusion

Both feature and score level fusion techniques were explored in this work. Feature level fusion was used to emphasize the complementarity of the presented features to the baseline. Whereas, the score level fusion was employed for multiple combinations of all the presented modalities to improve performance.

Feature-level fusion: Features from different standalone systems were evaluated by concatenating them to the baseline features and training a SVM directly. Fusion on i-vector level was also tried by applying linear discriminant analysis (LDA) on individual systems first and then on the fused i-vector features. The fused i-vectors are used to train the PLDA system for obtaining the log-likelihood scores.

Score-level fusion: For score-level fusion, logistic regression is performed over the log-likelihood scores obtained from multiple systems using the Bosaris toolkit [33]. For i-vector based systems, the log-likelihood scores are directly obtained from the PLDA scoring. Whereas, for the SVM/DNN and lexical classifiers, the posteriors and perplexity are converted to log-likelihoods respectively. For training the fusion systems, we perform k-fold cross-validation on training data to obtain new set of perturbed log-likelihoods, which is more representative of the errors the i-vector framework makes on testing data.

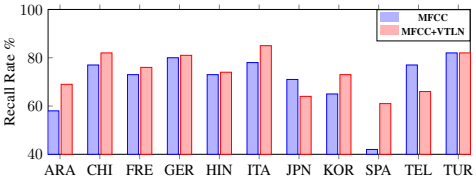

Figure 1: Effect of VTLN on recall rates of languages

6. Experimental results & Discussions

We present the results for individual systems first, and then finally we evaluate the fusion performance of multiple systems.

6.1. Standalone system performance

Table 1 gives the summary of performance for different standalone systems.

Acoustic i-vector modeling: We observe typical PLP and MFCC based acoustic features to be reliable giving the best individual-system results. We find that PLP outperforms the MFCC features by approximately 2% absolute both in terms of Accuracy and UAR.

Effect of VTLN: Figure 1 demonstrates the effect of VTLN on MFCC features. The recall of 11 different languages are plotted for raw MFCCs and VTLN-MFCCs. We see that the VTLN gives consistent improvement to most of the languages except Japanese and Telugu. We obtain a significant increase of absolute 19% recall for Spanish. Overall, we find VTLN to be useful providing 3.6% absolute increase in accuracy and recall rates.

L1-ProP and i-vector: We find that using VTLN-MFCCs to extract the log-likelihood features does not significantly improve the performance. Further, gaussianization of features and PCA dimension reduction (23 dimensions) were found to be useful providing a boost of 9% absolute. Overall, the phoneme confusability log-likelihood features prove to be less reliable compared to the acoustically trained i-vectors. L1-ProP features on other foreign languages like Spanish, Hindi, Telugu, Arabic, French and German were also experimented with and were seen to give similar performance to the Spanish. We retain the system for fusion to extract complementary information.

Bottleneck features: We observe that the bottleneck features never approach the performance of other acoustic features (MFCC or PLP). Since they are based on the same modality as MFCC and PLP they also do not provide complementary information thus we do not pursue these further.

Phoneme level features: While both the prosodic and confusability features fail to beat the baseline performance, the prosodic features are observed to provide complementarity to the baseline. Since they also perform similar to the baseline despite using elementary statistics, this supports the need for better phoneme-level modeling.

Lexical features: Lexical features provide performance similar to the baseline and given the different modality we expect them to provide complementary information.

fMLLR features: We see from the result that the raw fMLLR transforms inherit certain L1 characteristics and could be used

	Results on Development		
	Features ➞ Classifier	Accuracy	UAR
45s	Baseline	45.00%	45.10%
25ms	MFCC ➞ iVector ➞ PLDA	70.90%	70.90%
25ms	MFCC-VTLN ➞ iVector ➞ PLDA	74.20%	74.20%
25ms	PLP ➞ iVector ➞ PLDA	72.30%	72.50%
25ms	PLP-VTLN ➞ iVector ➞ PLDA	**76.40%**	**76.40%**
25ms	Bottleneck features on log power spectrogram ➞ iVector ➞ PLDA	36.40%	36.70%
45s	fMLLR ➞ SVM	42.30%	42.70%
Word	Lexical ➞ Maximum Likelihood w/ smoothing	44.60%	41.00%
25ms	L1-ProP ➞ iVector ➞ PLDA (English ASR)	60.50%	60.70%
25ms	L1-ProP ➞ Gaussianiazation ➞ PCA iVector ➞ PLDA (English ASR)	69.60%	69.80%
25ms	L1-ProP ➞ Gaussianiazation ➞ PCA iVector ➞ PLDA (Spanish ASR)	66.00%	66.30%
25ms	L1-ProP + VTLN ➞ iVector ➞ PLDA	60.90%	61.30%
~80ms	Phone Confusability Distribution ➞ SVM	25.50%	25.80%
~80ms	Phoneme Specific Prosodic Signature (PSPS) ➞ SVM	40.70%	41.10%
	Feature level fusion	Accuracy	UAR
25ms	Bottleneck + MFCC-VTLN ➞ iVector ➞ PLDA	46.40%	46.80%
	Baseline + Phone Confusability Distribution ➞ SVM (English ASR)	44.40%	44.50%
	Baseline + Phoneme Specific Prosodic Signature ➞ SVM	51.50%	51.70%
	Score level Fusion via Logistic Regression	Accuracy	UAR
	Baseline + (Bottleneck & MFCC-VTLN)	48.20%	48.60%
	Baseline + fMLLR	48.10%	48.30%
	Baseline + Lexical	52.10%	52.10%
	Baseline + Lexical + L1-ProP (English ASR)	66.50%	66.60%
	Baseline + Lexical + MFCC-VTLN	76.90%	77.00%
	Baseline + Lexical + PLP-VTLN	77.80%	77.90%
	Baseline + Lexical + PLP-VTLN + MFCC-VTLN + L1-ProP-VTLN (English ASR)	**78.50%**	**78.60%**
	+ PSPS	64.30%	65.40%
	+ Phone Confusion	74.70%	74.90%
	+ PSPS + Phone Confusion	74.90%	75.10%
	+ fMLLR	78.10%	78.20%
	Leave One Out (From best system) via Logistic Regression		
	Baseline + Lexical + PLP-VTLN + MFCC-VTLN + L1-ProP-VTLN	**78.50%**	**78.60%**
	- MFCC-VTLN	76.80%	76.90%
	- PLP-VTLN	75.60%	75.70%
	- Baseline	75.10%	75.30%
	- Lexical	76.70%	76.80%
	Results on Test		
	MFCC-VTLN + PLP-VTLN + Baseline + Lexical (Submission 3)	79.93%	80.13%

Table 1: Results of the various systems as described in text.

as a potential feature for L1 identification. It was also found to provide some complementarity to the baseline features.

6.2. Fusion Performance

Feature-level fusion: We attempted feature-level fusion for our lowest-performing features to increase performance. We can see from Table 1 that all three improve marginally above baseline, but not significantly so.

Score-level fusion: Analyzing the performance of multiple score level fusion combinations for i-vectors, on the acoustic front, we find that PLP and MFCC exhibit acoustic complementarity. Fusion of acoustic features with the baseline and lexical systems provide further improvements. Even-though the L1-ProP i-vector system doesn't provide noticeable increase in performance when fused with acoustic features, we see improvements when used along with the lexical and baseline features. However, we observe that the Phonemic Confusability (PC) and Phoneme Specific Prosodic Signature (PSPS) do not improve the overall performance of the system. We believe that the noise in the feature extraction may be responsible for the low performance and we intend to investigate further. We also believe that these features can provide improvements for discriminability of specific language pairs. The fMLLR features did not affect the performance of our best system significantly. We believe that the information captured by fMLLR features is redundant with the combination of other features.

Our best performing system is a combination of acoustic (MFCC-PLP), lexical, prosodic (Baseline), and L1-ProP. The best performing system achieves an Accuracy of 78.5% and UAR of 78.60% on the development test.

We perform leave-one-out from the best system to analyze the importance of each feature. We find PLP and Baseline features to be significant contributors in terms of complementary information giving approximately 3% improvements, whereas, MFCC and Lexical features contribute around 2%. Finally L1-ProP features improves the overall system by a small margin.

	ARA	CHI	FRE	GER	HIN	ITA	JPN	KOR	SPA	TEL	TUR
ARA	54	1	5	0	1	4	8	2	2	1	2
CHI	0	62	0	2	0	1	2	6	0	1	0
FRE	4	3	56	3	2	3	1	0	5	1	0
GER	0	0	4	69	0	0	0	0	2	0	0
HIN	0	0	0	0	64	1	0	0	1	16	0
ITA	1	1	2	0	0	59	1	0	2	0	2
JPN	1	1	0	0	0	0	71	2	0	0	0
KOR	2	5	0	1	0	0	12	60	0	0	0
SPA	1	0	4	2	2	8	2	2	54	1	1
TEL	0	0	0	0	19	0	0	0	0	69	0
TUR	7	1	2	1	0	1	1	0	2	0	75

Table 2: Confusion matrix of the best results on test, corresponding to an Accuracy = 79.93% and UAR = 80.13%

Across the modalities, we observe different features providing discriminability between specific language pairs. In future, we intend to employ a hierarchical classification method to exploit such properties.

6.3. Inter-class confusion analysis

Figure 2 shows the confusion matrix obtained for our best performing system on the development set. Italian and Turkey are the least confused languages and French is the most confused. The matrix shows inter-language confusions between Hindi - Telugu and Japanese - Korean languages correlating with demographics between the languages. In our human-analysis, that included three Indian speakers, we couldn't separate most of the confusable development set Hindi and Telugu pairs. Overall, comparing with the baseline system, we find the confusion to be significantly more sparse suggesting not only better performance but also less confusion among language pairs with our improved system.

6.4. Results on the test

For testing, we used the score level fusion MFCC-VTLN, PLP-VTLN i-vector system, Baseline and Lexical features to achieve a performance of **79.93%** Accuracy and **80.13%** UAR. We believe that inclusion of other systems and further calibration during fusion on per-language level basis rather than global 11 class classification metrics could boost the performance. Due to time constraints, we were unable to try further combinations and didn't incorporate L1-ProP features with Gausianization.

7. Conclusion

In this work, we have addressed a challenging research problem of detecting the L1 native language from spontaneous speech on 11 different L1 language categories. We exploit different modalities, multiple feature rates, and a range of methods towards robust classification. Each modality was shown to improve the performance of the baseline system when fused with the baseline features, demonstrating the complementarity of the proposed features. We also showed the effectiveness of speaker normalization. We successfully demonstrate that some L1 information exists in the normalization (fMLLR) feature and could be used as a potential feature for L1 detection. While the phoneme confusability and phoneme-level prosodic features did not improve the overall system performance, they were shown to be effective in improving the baseline. Different fusion techniques were applied to extract complementary information across various modalities.

By analyzing the confusion in the system, we observed inherent correlations with the demographics among certain languages. From an unscientific sampling of human listeners our system seems to face similar challenges to humans especially for the highly confusable language pairs. In short, we present an accurate multimodal, multirate L1 identification system via a range of feature, classification, and fusion methods.

8. References

[1] M. K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4398–4401.

[2] E. Shriberg, L. Ferrer, S. S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak, "Detecting nonnative speech using speaker recognition approaches." in *Odyssey*. Citeseer, 2008, p. 26.

[3] S. Gray and J. H. Hansen, "An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*. IEEE, 2005, pp. 35–40.

[4] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*. IEEE, 2005, pp. 139–143.

[5] G. Choueiter, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4265–4268.

[6] L. M. Arslan and J. H. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *The Journal of the Acoustical Society of America*, vol. 102, no. 1, pp. 28–40, 1997.

[7] 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*. Proceedings INTERSPEECH 2016, ISCA, San Francisco, USA, 2016.

[8] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[9] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 346–348.

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

[11] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[12] D. Kim, S. Umesh, M. Gales, T. Hain, and P. Woodland, "Using vtln for broadcast news transcription," in *Proc. ICSLP*, vol. 4, 2004.

[13] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of dnn." in *INTERSPEECH*, 2013, pp. 3661–3664.

[14] K. Vesely, M. Karafiát, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[15] F. Xie and D. C. Van, "A family of mlp based nonlinear spectral estimators for noise reduction," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–53.

[16] T. Piske, I. R. MacKay, and J. E. Flege, "Factors affecting degree of foreign accent in an l2: A review," *Journal of phonetics*, vol. 29, no. 2, pp. 191–215, 2001.

[17] J. E. Flege, O.-S. Bohn, and S. Jang, "Effects of experience on non-native speakers' production and perception of english vowels," *Journal of phonetics*, vol. 25, no. 4, pp. 437–470, 1997.

[18] R. K. Bansal, "The pronunciation of english in india," *Studies in the pronunciation of English: A commemorative volume in honour of AC Gimson*, pp. 219–230, 1990.

[19] J. E. Flege, "Assessing constraints on second-language segmental production and perception," *Phonetics and phonology in language comprehension and production: Differences and similarities*, vol. 6, pp. 319–355, 2003.

[20] A. Sheldon and W. Strange, "The acquisition of/r/and/l/by japanese learners of english: Evidence that speech production can precede speech perception," *Applied Psycholinguistics*, vol. 3, no. 03, pp. 243–261, 1982.

[21] H. Meng, Y. Y. Lo, L. Wang, and W. Y. Lau, "Deriving salient learners mispronunciations from cross-language phonological comparisons," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 437–442.

[22] C. R. Wiltshire and J. D. Harnsberger, "The influence of gujarati and tamil l1s on indian english: A preliminary study," *World Englishes*, vol. 25, no. 1, pp. 91–104, 2006.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[24] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

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

[26] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction." in *INTERSPEECH*, 2011, pp. 857–860.

[27] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.

[28] M. H. Bahari, M. McLaren, D. A. van Leeuwen *et al.*, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99–108, 2014.

[29] P. G. Shivakumar, M. Li, V. Dhandhania, and S. S. Narayanan, "Simplified and supervised i-vector modeling for speaker age regression," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4833–4837.

[30] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition." in *INTERSPEECH*, 2011, pp. 25–28.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] P. G. Georgiou, M. P. Black, A. Lammert, B. Baucom, and S. S. Narayanan, ""That's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proceedings of Affective Computing and Intelligent Interaction (ACII), Lecture Notes in Computer Science*, October 2011.

[33] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.