i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition

Hamid Behravan, *Member IEEE*, Ville Hautamäki, *Member IEEE*, Sabato Marco Siniscalchi, *Member IEEE*, Tomi Kinnunen, *Member IEEE*, and Chin-Hui Lee, *Fellow IEEE*

***Abstract*—We propose a unified approach to automatic foreign accent recognition. It takes advantage of recent technology advances in both linguistics and acoustics based modeling techniques in automatic speech recognition (ASR) while overcoming the issue of a lack of a large set of transcribed data often required in designing state-of-the-art ASR systems. The key idea lies in defining a common set of fundamental units "universally" across all spoken accents such that any given spoken utterance can be transcribed with this set of "accent-universal" units. In this study, we adopt a set of units describing manner and place of articulation as speech attributes. These units exist in most spoken languages and they can be reliably modeled and extracted to represent foreign accent cues. We also propose an i-vector representation strategy to model the feature streams formed by concatenating these units. Testing on both the Finnish national foreign language certificate (FSD) corpus and the English NIST 2008 SRE corpus, the experimental results with the proposed approach demonstrate a significant system performance improvement with p-value < 0.05 over those with the conventional spectrum-based techniques. We observed up to a 15% relative error reduction over the already very strong i-vector accented recognition system when only manner information is used. Additional improvement is obtained by adding place of articulation clues along with context information. Furthermore, diagnostic information provided by the proposed approach can be useful to the designers to further enhance the system performance.**

***Index Terms*—Attribute detectors, i-vector system, Finnish corpus, English corpus.**

I. Introduction

AUTOMATIC foreign accent recognition is the task of identifying the mother tongue (L1) of non-native speakers given an utterance spoken in a second language (L2) [1]. The task attracts increasing attention in the speech community because accent adversely affects the accuracy of conventional *automatic speech recognition* (ASR) systems (e.g., [2]). In fact, most existing ASR systems are tailored to native speech only, and recognition rates decrease drastically when words or sentences are uttered with an altered pronunciation (e.g., foreign accent) [3]. Foreign accent variation is a nuisance factor that adversely affects automatic speaker and language recognition systems as well [4], [5]. Furthermore, foreign accent recognition is a topic of great interest in the areas of intelligence and security, including immigration screening and border control sites [6]. It may help officials detect a fake passport by verifying whether a traveler's spoken foreign accent corresponds to accents spoken in the country he claims he is from [6]. Finally, connecting customers to agents with similar foreign accent in targeted advertisement applications may help create a more user-friendly environment [7].

It is worth noting that *foreign accents* differ from *regional accents* (dialects), since the deviation from the standard pronunciation depends upon the influence that L1 has on L2 [8]. Firstly, non-native speakers tend to alter some phone features when producing a word in L2 because they only partially master its pronunciation. To exemplify, Italians often do not aspirate the /h/ sound in words such as *house*, *hill*, and *hotel*. Moreover, non-native speakers can also replace an unfamiliar phoneme in L2 with the one considered as the closest in their L1 phoneme inventory. Secondly, there are several degrees of foreign accent for the same native language influence according to L1 language proficiency of the non-native speaker [9], [10]: non-native speaker learning L2 at an earlier age can better compensate for their foreign accent factors when speaking in L2 [11].

In this study, we focus on automatic L1 detection from spoken utterances with the help of statistical pattern recognition techniques. In the following, we give a brief overview and current state-of-the-art methods before outlining our contributions. It is common practice to adopt automatic *language recognition* (LRE) techniques to the foreign accent recognition task. Indeed, the goal of an LRE system is to automatically detect the spoken language in an utterance, which we can parallel with that of detecting L1 in an L2 utterance. Automatic LRE techniques can be grouped into to main categories: *token-based* (a.k.a, *phonotactic*) and *spectral-based* ones. In the token-based approach, discrete units/tokens, such as phones, are used to describe any spoken language. For example, parallel phone recognition followed by language modeling (PPRLM) [12] approach employs a bank of phone recognizers

Manuscript received December 27, 2014; revised June 15, 2015; accepted September 28, 2015. This project was partially supported by the Academy of Finland projects 253120, 253000 and 283256, Finnish Scientific Advisory Board for Defence (MATINE) project nr. 2500M-0036 and Kone Foundation - Finland. Dr. Hautamäki and Dr. Siniscalchi were supported by the Nokia Visiting Professor Grants 201500062 and 201600008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mohamed Afify.

H. Behravan is with the School of Computing, University of Eastern Finland, Joensuu, Finland. E-mail: behravan@cs.uef.fi

V. Hautamäki is with the School of Computing, University of Eastern Finland, Joensuu, Finland. E-mail: villeh@cs.uef.fi

S. M. Siniscalchi is with the Department of Computer Engineering, Kore University of Enna, Enna, Italy, and with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: marco.siniscalchi@unikore.it

T. Kinnunen is with the School of Computing, University of Eastern Finland, Joensuu, Finland. E-mail: tkinnu@cs.uef.fi

C.-H. Lee is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. E-mail: chl@ece.gatech.edu

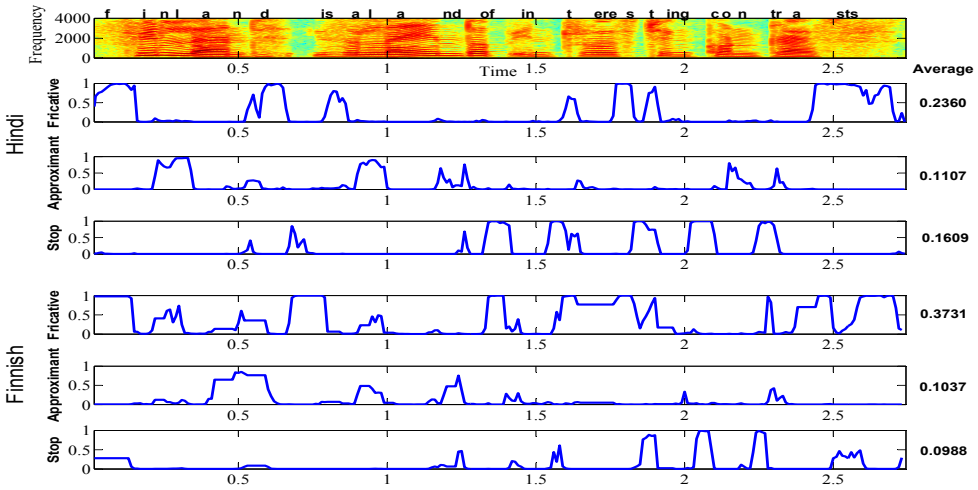


Fig. 1: An example showing the detection score differences in the three selected attributes from a Hindi and a Finnish speaker. Both speakers utter the same sentence 'Finland is a land of interesting contrasts'. Speech segments are time-aligned with dynamic time warping (DTW). The Finnish speaker shows higher level of activity in fricative in comparison to the Hindi speaker. However, in the Hindi speech utterance, the level of activity in stop is higher than in the Finnish utterance.

to convert each speech utterance into a string of tokens. In the spectral-based approach a spoken utterance is represented as a sequence of short-time spectral feature vectors. These spectral vectors are assumed to have statistical characteristics that differ from one language to another [13], [14]. Incorporating temporal contextual information to the spectral feature stream has been found useful in the language recognition task via the so-called *shifted-delta-cepstral* (SDC) features [15]. The long-term distribution of language-specific spectral vectors is modeled, in one form or another, via a language- and speaker-independent universal background model (UBM) [16]. In the traditional approaches [16], [17], language-specific models are obtained via UBM adaptation while the modern approach utilizes UBMs to extract low-dimensional *i-vectors* [18]. I-vectors are convenient for expressing utterances with varying numbers of observations as a single vector that preserves most utterance variations. Hence, issues such as session normalization are postponed to back-end modeling of i-vector distributions.

Table I shows a summary of several studies on foreign accent recognition. In [1], the accented speech is characterized using acoustic features such as frame power, zero-crossing rate, LP reflection coefficients, autocorrelation lags, log-area-ratios, line-spectral pair frequencies and LP cepstrum coefficients. 3-state hidden Markov models (HMMs) with a single Gaussian density were trained from these features and evaluated on spoken American English with 5 foreign accents reporting 81.5% identification accuracy. The negative effects of non-native accent in ASR task were studied in [19]. Whole-word and sub-word HMMs were trained on either native accent utterances or a pool of native and non-native accent sentences. The use of phonetic transcriptions for each specific accent improved speech recognition accuracy. An accent dependent parallel phoneme recognizer was developed in [20] to discriminate native Australian English speakers and two migrant speaker groups with foreign accents, whose L1's were either Levantine Arabic or South Vietnamese. The best average accent identification accuracies of 85.3% and 76.6% for accent pair and three accent class discrimination tasks were reported, respectively. A text-independent automatic accent classification system was deployed in [5] using a corpus representing five English speaker groups with native American English, and English spoken with Mandarin Chinese, French, Thai and Turkish accents. The proposed system was based on stochastic and parametric trajectory models corresponding to the sequence of points reflecting movements in the speech production caused by coarticulation. This system achieved an accent classification accuracy of 90%.

All the previous studies used either suprasegmental modeling, in terms of trajectory model or prosody, or phonotactic modeling to recognize non-native accents. Recently, spectral features with i-vector back-end were found to outperform phonotactic systems in language recognition [18]. Spectral features were first used by [21] in a L1 recognition task. The non-native English speakers were recognized using multiple spectral systems, including i-vectors with different back-ends [21], [23]. The i-vector system outperformed other methods most of the time, and spectral techniques based on i-vector model are thus usually adopted for accent recognition. The lack of large amount of transcribed accent-specific speech data to train high-performance acoustic phone models hinders the deployment of competitive phonotactic foreign accent recognizers. Nonetheless, it could be argued that phonotactic methods would provide valuable results that are informative to humans [24]. Thus, a unified foreign accent recognition framework that gives the advantages of the subspace modeling techniques without discharging the valuable information provided by the phonotactic-based methods is highly desirable.

The *automatic speech attribute transcription* (ASAT) framework [25], [26], [27] represents a natural environment to make these two above contrasting goals compatible, and is adopted here as the reference paradigm. The key idea of ASAT is

TABLE I: Summary of the previous studies on foreign accent recognition and the present study.

Study	Spoken language	#accents	#speakers	#utterances	Features	Model
Hansen and Arslan [1]	American English	4	27	N/A	Prosodic	HMM
Teixeira et al. [19]	British English	5	20	20	Phonotactic	HMM
Kumpf and King [20]	Australian English	3	67	3950	Phonotactic	HMM
Angkititraku and Hansen [5]	English	5	179	N/A	Phoneme sequence	Trajectory-model
Bahari et al. [21]	English	5	265	359	Spectral	GMM supervector
Behravan et al. [22], [10]	Finnish	9	450	1973	Spectral	i-vector modeling
Present study	Finnish (FSD)	8	415	1644	Attributes	i-vector modeling
Present study	English (NIST)	7	348	1262	Attributes	i-vector modeling

to use a compact set of speech attributes, such as *fricative*, *nasal* and *voicing* to compactly characterize any L2 spoken sentence independently of the underlying L1 native language. A bank of data-driven detectors generates attribute posterior probabilities, which are in turn modeled using an i-vector back-end, treating the attribute posteriors as acoustic features. A small set of speech attributes suffices for a complete characterization of spoken languages, and it can therefore be useful to discriminate accents [28]. For example, some sister languages, e.g., Arabic spoken in Syria and Iraq, only have subtle differences that word-based discrimination usually does not deliver good results. In contrast, these differences naturally arise at an attribute level and can help foreign accent recognition. Robust universal speech attribute detectors can be designed by sharing data among different languages, as shown in [29], and that bypasses the lack of sufficient labeled data for designing ad-hoc tokenizers for a specific L1/L2 pair. Indeed, the experiments reported in this work concern detecting Finnish and English foreign accented speech, even though the set of attribute detectors was originally designed to address phone recognition with minimal target-specific training data [29]. Although speech attributes are shared across spoken languages, the statistics of the attributes can differ considerably from one foreign accent to another, and these statistics improve discrimination [30]. This can be appreciated by visually inspecting Figure 1, which shows attribute detection curves from Finnish and Hindi speakers. Although both speakers uttered the same sentence, namely "Finnish is a land of interesting contrasts," differences between corresponding attribute detection curves can be observed: (i) the fricative detection curve tends to be more active (i.e. stays close to 1) in Finnish speaker than in Hindi, (ii) the stop detection curve for the Hindi speaker more often remains higher (1 or close to 1) than that for the Finnish speaker, (iii) approximant detection curve seem instead to show similar level of activity for both speakers.

In this work, we significantly expand our preliminary findings on automatic accent recognition [31] and re-organize our work in a systematic and, self-contained form that provides a convincing case why universal speech attributes are worthwhile of further studies in accent characterization. The key experiments, not available in [31], can be summarized as follows: (i) we have investigated the effect of *heteroscedastic* linear discriminant analysis (HLDA) [32] dimensionality reduction on the accent recognition performance and compared and contrasted it with linear discriminant analysis (LDA), (ii) we have studied training and test duration effects on the overall system performance, and (iii) we have expanded our initial investigation on Finnish data by including new experiments on English foreign accent. Even if the single components have been individually investigated in previous studies, e.g., [30], [33], [18], the overall architecture (combining the components) presented in this paper, as well as its application to foreign accent recognition, are novel. The key novelty of our framework can be summarized as follows: (i) speech attributes extracted using machine learning techniques are adopted to the foreign accent recognition task for the first time, (ii) a dimensionality reduction approach is used for capturing temporal context and exploring the effect of languages, (iii) the i-vector approach is successfully used to model speech attributes. With respect to point (iii), Diez et al. [34], [35] proposed a similar solution but to address a spoken language recognition task, namely they used log-likelihood ratios of phone posterior probabilities within the i-vector framework. Although Diez et al.'s work has some similarities with ours, there are several implementation differences in addition to the different addressed task: (i) we describe different accents using a compact set of language independent attributes, which overcomes high computational issues caused by high-dimension posterior scores, as mentioned in [34], (ii), we introduce context information by stacking attribute probability vectors together, and we then capture context variability directly in the attribute space, and (iii) we carry out i-vector post-processing to further improve accents discriminability. Moreover, useful diagnostic information can be gathered with our approach, as demonstrated in Section IV-D.

Finally in [22], [10], the authors demonstrated that i-vector modeling using SDCs outperforms conventional Gaussian mixture model - universal background model (GMM-UBM) system in recognizing Finnish non-native accents. The method proposed in [10] is here taken to build a reference baseline system to compare with. We evaluate effectiveness of the proposed attribute-based foreign accent recognition system with a series of experiments on Finnish and English foreign accented speech corpora. The experimental evidence demonstrates that the proposed technique compares favorably with conventional SDC-MFCC with i-vector and GMM-UBM approaches. In order to enhance accent recognition performance of the proposed technique, several configurations have been proposed and evaluated. In particular, it was observed that contextual information helps to decrease recognition error rates.

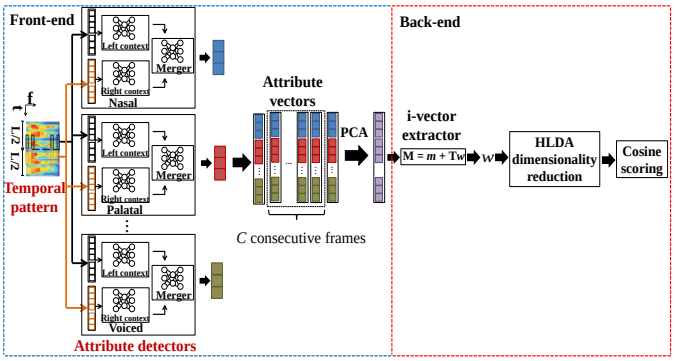


Fig. 2: Block diagram of the proposed system. In the attribute detectors [29], [30], [27], spectral features are fed into left-context and right-context artificial neural networks. A merger then combines the outputs generated by those two neural networks and produce the final attribute posterior probabilities. Principal component analysis (PCA) is then applied on C consecutive frames of these posterior probabilities to create long-term contextual features. We use i-vector approach [33] with cosine scoring [33] to classify target accents.

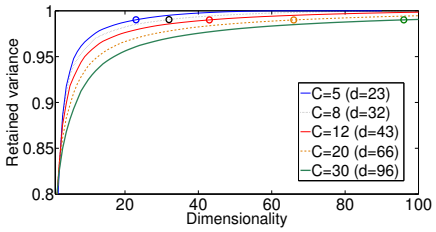


Fig. 3: Remaining variance after PCA. Comparing stacked context sizes (C) 5, 8, 12, 20 and 30 frames for manner attributes. d varies from ∼20 to ∼100, with larger dimensionality assigned to longer context sizes.

II. Foreign Accent Recognition

Figure 2 shows the block diagram of the proposed system. The front-end consists of attribute detectors and building long-term contextual features via principal component analysis (PCA). The features created in the front-end are then used to model target foreign accents using a i-vector back-end. In the following, we describe the individual components in detail.

A. Speech attribute extraction

The set of speech attributes used in this work are acoustic phonetic features, namely, five *manner of articulation* classes (**glide, fricative, nasal, stop, and vowel**), and **voicing** together with nine *place of articulation* (**coronal, dental, glottal, high, labial, low, mid, retroflex, velar**). Attributes could be extracted from a particular language and shared across many different languages, so they could also be used to derive a universal set of speech units. Furthermore, data-sharing across languages at the acoustic phonetic attribute level is naturally facilitated by using these attributes, so more reliable language-independent acoustic parameter estimation can be anticipated [29]. In [30], it was also shown that these attributes can be used to compactly characterize any spoken language along the same lines as in the ASAT paradigm for ASR [27]. Therefore, we expect that it can also be useful for characterizing speaker accents.

B. Long-term Attribute Extraction

Each attribute detector outputs the posterior probability for the target class i, $p(H^{(i)}_{\text{target}}|\boldsymbol{f})$, non-target, $p(H^{(i)}_{\text{anti}}|\boldsymbol{f})$, and noise, $p(H^{(i)}_{\text{noise}}|\boldsymbol{f})$, class given a speech frame $\boldsymbol{f}$. As probabilities, they sum up to one for each frame. A feature vector $\boldsymbol{x}$ is obtained by concatenating those posterior probabilities generated by the set of manner/place detectors into a single vector. The final dimension of the feature vector, $\boldsymbol{x}$, is 18 in the manner of articulation case, for example.

Since language and dialect recognizers benefit from the inclusion of long temporal context [36], [16], it is natural to study similar ideas for attribute modeling as well. A simple feature stacking approach is adopted in this paper. To this end, let $\boldsymbol{x}(t) \in \mathbb{R}^n$ denote the 18-dimensional (6 manner attributes $\times$ 3) or 27-dimensional (9 place attributes $\times$ 3) feature attribute vector at frame t. A sequence of $q = 18\,C$ (or $q = 27\,C$, for place) dimensional stacked vectors $\tilde{\boldsymbol{x}}_C(t) = (\boldsymbol{x}(t)^\top, \boldsymbol{x}(t+1)^\top, \ldots, \boldsymbol{x}(t+C-1)^\top)^\top, t = 1, 2, \ldots,$ is formed, where C is the context size, and $\top$ stands for transpose. PCA is used to project each $\tilde{\boldsymbol{x}}_C(t)$ onto the first $d \ll q$ eigenvectors corresponding to the largest eigenvalues of the sample covariance matrix. We estimate the PCA basis from the same data as the UBM and the T-matrix, after VAD, with 50 % overlap across consecutive $\tilde{\boldsymbol{x}}_C(t)$'s. We retain 99 % of the cumulative variance. As Figure 3 indicates, d varies from ∼20 to ∼100, with larger dimensionality assigned to longer context as one expects.

C. I-vector Modeling

I-vector modeling or total variability modeling, forms a low-dimensional *total variability space* that contains spoken content, speaker and channel variability [33]. Given an utterance, a GMM supervector, $\boldsymbol{s}$, is represented as [33],

$$\boldsymbol{s} = \boldsymbol{m} + \mathbf{T}\boldsymbol{w}, \tag{1}$$

where $\boldsymbol{m}$ is the utterance- and channel-independent component (the universal background model or UBM supervector), $\mathbf{T}$ is a rectangular low rank matrix and $\boldsymbol{w}$ is an independent random vector of distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. $\mathbf{T}$ represents the captured variabilities in the supervector space. It is estimated by the expectation maximization (EM) algorithm similar to estimating the speaker space in joint factor analysis (JFA) [37], with the exception that every training utterances of a given model is treated as belonging to different class. The extracted i-vector is then the mean of the posterior distribution of $\boldsymbol{w}$.

D. Inter-session Variability Compensation

As the extracted i-vectors contain both within- and between accents variation, we used dimensionality reduction technique

to project the i-vectors onto a space to minimize the within-accent and maximize the between-accent variation. To perform dimensionality reduction, we used *heteroscedastic* linear discriminant analysis (HLDA) [32], which is considered as an extension of linear discriminant analysis (LDA). In this technique, i-vector of dimension n is projected into a p-dimensional feature space with $p < n$, using HLDA transformation matrix denoted by $\mathbf{A}$. The matrix $\mathbf{A}$ is estimated by an efficient row-by-row iteration with EM algorithm as presented in [38].

Followed by HLDA, within-class covariance normalization (WCCN) is then used to further compensate for unwanted intra-class variations in the total variability space [39]. The WCCN transformation matrix, $\mathbf{B}$, is trained using the HLDA-projected i-vectors obtained by Cholesky decomposition of $\mathbf{B}\mathbf{B}^\top = \mathbf{\Lambda}^{-1}$, where a within-class covariance matrix, $\mathbf{\Lambda}$, is computed using,

$$\mathbf{\Lambda} = \frac{1}{L}\sum_{a=1}^{L}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{w}_i^a - \overline{\boldsymbol{w}}_a)(\boldsymbol{w}_i^a - \overline{\boldsymbol{w}}_a)^\top, \tag{2}$$

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

$$\hat{\boldsymbol{w}} = \mathbf{B}^\top\mathbf{A}^\top\boldsymbol{w}. \tag{3}$$

E. Scoring Against Accent Models

We used *cosine scoring* to measure similarity of two i-vectors [33]. The cosine score, t, between the inter-session variability compensated test i-vector, $\hat{\boldsymbol{w}}_{\text{test}}$, and target i-vector, $\hat{\boldsymbol{w}}_{\text{target}}$, is computed as the dot product between them,

$$t = \frac{\hat{\boldsymbol{w}}_{\text{test}}^\top \;\; \hat{\boldsymbol{w}}_{\text{target}}}{\|\hat{\boldsymbol{w}}_{\text{test}}\| \;\; \|\hat{\boldsymbol{w}}_{\text{target}}\|}, \tag{4}$$

where $\hat{\boldsymbol{w}}_{\text{target}}$ is the average i-vector over all the training utterances of the target accent, i.e.

$$\hat{\boldsymbol{w}}_{\text{target}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{w}}_i, \tag{5}$$

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

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

$$t'_a = \log\frac{\exp(t_a)}{\frac{1}{L-1}\sum_{k\neq a}\exp(t_k)}, \tag{6}$$

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

III. Experimental Setup

A. Baseline System

To compare the attribute system recognition performance, two baseline systems were built. Both systems were trained using 56 dimensional SDC (49)-MFCC (7) feature vectors and they use the same UBM of 512 Gaussians. The first system is based on the conventional GMM-UBM system with adaptation similar to [16]. It uses 1 iteration to adapt the UBM to each target model. Adaptation consists of updating only the GMM mean vectors. The detection scores are then generated using a fast scoring scheme described in [40] using top 5 Gaussians. The second system uses i-vectors approach to classify accents. The i-vectors are of dimensionality 1000 and HLDA projected i-vectors of dimensionality 180.

B. Corpora

The "stories" part of the OGI Multi-language telephone speech corpus [41] was used to train the attribute detectors. This corpus has phonetic transcriptions for six languages: English, German, Hindi, Japanese, Mandarin, and Spanish. Data from each language were pooled together to obtain 5.57 hours of training and 0.52 hours of validation data.

A series of foreign accent recognition experiments were performed on the *FSD* corpus [42] which was developed to assess Finnish language proficiency among adults of different nationalities. We selected the oral responses portion of the exam, corresponding to 18 foreign accents. Since the number of utterances is small, 8 accents — Russian, Albanian, Arabic, English, Estonian, Kurdish, Spanish, and Turkish — with enough available data were used. The unused accents are, however, used in training UBM and the T-matrix. Each accent set is randomly split into a test and a train set. The test set consists of (approximately) 30% of the utterances, while the training set consists of the remaining 70% to train foreign accent recognizers in the FSD task. The raw audio files were partitioned into 30 sec chunks and re-sampled to 8 kHz. Statistics of the test and train portions are shown in Table II.

The NIST 2008 SRE corpus was chosen for the experiments on English foreign accent detection. The corpus has a rich metadata from the participants, including their age, language and smoking habits. It contains many L2 speakers whose native language is not English. Since the number of utterances in some foreign accents is small, 7 accents — Hindi (HIN), Thai (THA), Japanese (JPN), Russian (RUS), Vietnamese (VIE), Korean (KOR) and Chinese Cantonese (YUH) — with enough available utterances were chosen in this study. These accents are from the short2, short3 and 10sec portions, of the NIST 2008 SRE corpus. We used over 5000 utterances to train the UBM and total variability subspace in the NIST 2008 task. Table III shows the distribution of train and test portions in the English utterances. Speakers do not overlap between training and testing utterances both in the FSD and NIST corpora.

C. Attribute Detector Design

One-hidden-layer feed forward multi-layer perceptrons (MLPs) were used to implement each attribute detector. The

TABLE II: Train and test files distributions in each target accent in the FSD corpus. Duration is reported for only active speech frames.

Accent	#train files (hh:mm)	#test files	#speakers
Spanish	47 (00:26)	25	15
Albanian	60 (00:32)	29	19
Kurdish	61 (00:37)	32	21
Turkish	66 (00:39)	34	22
English	70 (00:37)	36	23
Estonian	122 (01:07)	62	38
Arabic	128 (01:15)	66	42
Russian	556 (03:15)	211	235
Total	1149 (08:46)	495	415

TABLE III: Train and test file distributions in the NIST 2008 SRE corpus. Duration is reported for only active speech frames.

Accent	#train files (hh:mm)	#test files	#speakers
Hindi	80 (03:39)	109	53
Russian	74 (03:32)	84	42
Korean	91 (03:05)	99	41
Japanese	53 (02:02)	73	41
Thai	70 (02:53)	93	52
Cantonese	68 (03:14)	92	50
Vietnamese	127 (04:01)	149	69
Total	563 (22:44)	699	348

number of hidden nodes with a sigmoidal activation function is 500. MLPs were trained to estimate attribute posteriors, and the training data were separated into "feature present", "feature absent", and "other" regions for every phonetic class used in this work. The classical back-propagation algorithm with a cross-entropy cost function was adopted to estimate the MLP parameters. To avoid over-fitting, the reduction in classification error on the development set was adopted as the stopping criterion. The attribute detectors employed in this study were actually just those used in [29].

Data-driven detectors are used to spot speech cues embedded in the speech signal. An attribute detector converts an input utterance into a time series that describes the level of presence (or level of activity) of a particular property of an attribute over time. A bank of 15 detectors (6 manner and 9 place) is used in this work, each detector being individually designed to spot of a particular event. Each detector is realized with three single hidden layer feed-forward ANNs (artificial neural networks) organized in a hierarchical structure and trained on sub-band energy trajectories extracted through 15-band mel-frequency filterbank. For each critical band, a window of $310ms$ centered around the frame being processed is considered and split in two halves: left-context and right-context [43]. Two independent front-end ANNs ("lower nets") are trained on those two halves to generate, left- and right-context speech attribute posterior probabilities. The outputs of the two lower nets are then sent to the third ANN that acts as a merger and gives the attribute-state posterior probability of the target speech attribute.

D. Evaluation Metrics

System performance is reported in terms of *equal error rate* (EER) and average detection cost ($C_{\rm avg}$) [44]. Results are reported per each accent for a cosine scoring classifier. $C_{\rm avg}$ is defined as [44],

$$C_{\rm avg} = \frac{1}{M}\sum_{j=1}^{M} C_{\rm DET}(L_j), \tag{7}$$

where $C_{\rm DET}(L_j)$ is the detection cost for subset of test segments trials for which the target accent is L_j and M is the number of target languages. The per target accent cost is then,

$$\begin{aligned} C_{\rm DET}(L_j) &= C_{\rm miss}P_{\rm tar}P_{\rm miss}(L_a) \\ &+ C_{\rm fa}(1-P_{\rm tar})\frac{1}{J-1}\sum_{k\neq j} P_{\rm fa}(L_j,L_k). \end{aligned} \tag{8}$$

The miss probability (or false rejection rate) is denoted by $P_{\rm miss}$, i.e., a test segment of accent L_i is rejected as being in that accent. On the other hand $P_{\rm fa}(L_i,L_k)$ denotes the probability when a test segment of accent L_k is accepted as being in accent L_i. It is computed for each target/non-target accent pairs. The costs, $C_{\rm miss}$ and $C_{\rm fa}$ are both set to 1 and $P_{\rm tar}$, the prior probability of a target accent, is set to 0.5 following [44].

IV. Results

A. Accent Recognition Performance on the FSD corpus

Table IV reports foreign accent recognition results for several systems on the FSD corpus. The results in the first two rows indicate that i-vector modeling outperforms the GMM-UBM technique when the same input features are used, which is in line with findings in [10], [45]. The results in the last two rows, in turn, indicate that the i-vector approach can be further enhanced by replacing spectral vectors with attribute features. In particular, the best performance is obtained using manner attribute features within the i-vector technique, yielding a $\rm C_{avg}$ of 5.80, which represents relative improvements of 45% and 15% over the GMM-UBM and the conventional i-vector approach with SDC+MFCC features, respectively. The FSD task is quite small, which might make the improvements obtained with the attribute system not statistically different from those delivered by the spectral-based system. We therefore decided to run a proper statistical significance test using a dependent Z-test according to [46]. We applied the statistical test for comparing per target accents EERs between attribute systems and SDC-MFCC i-vector system. In Table V, we indicated in boldface cases where the proposed attribute-based foreign accent recognition techniques outperform the spectral-based one. To exemplify, Z-test results in the second column of Table V demonstrates that the manner system significantly outperforms the SDC-MFCC i-vector system on 7 out of 8 accents. For the sake of completeness, we have also compared manner and place of articulation systems, and we have reported the Z-test results in the third column of Table V.

To verify that we are not recognizing the channel variability, we followed the procedure highlighted in [47], where

TABLE IV: Baseline and attribute systems results in terms of $\mathrm{EER_{avg}}$ and $\mathrm{C_{avg}}$ in the FSD corpus. In parentheses, the final dimensionality of the feature vectors sent to the back-end. In manner system, for 7 out of 8 accents, the difference between EERs is significant at a confidence level of 95% if $Z \geq 1.960$.

Feature (dimensionality)	Classifier	$\mathrm{EER_{avg}}$(%)	$\mathrm{C_{avg}} \times 100$
SDC+MFCC (56)	GMM-UBM	19.03	10.56
SDC+MFCC (56)	i-vector	12.60	6.85
Place (27)	i-vector	10.37	6.00
Manner (18)	i-vector	**9.21**	**5.80**

TABLE V: In the first two columns, the Z-test results per target accent EERs at the EER threshold between the proposed attribute- and spectral-based system performance on the FSD corpus are reported. The difference between EERs is significant at a confidence level of 95% if $Z \geq 1.960$. Boldface values refer to cases in which our solution significantly outperforms the SDC-MFCC system. The third column shows the same Z-test results between manner- and place-based systems, where manner is significantly better than place if the score is in boldface.

Accents	**Place/SDC-MFCC**	**Manner/SDC-MFCC**	**Manner/Place**
Albanian	1.1041	1.6503	**2.2866**
Arabic	**5.9139**	**5.6975**	1.0587
English	**1.9973**	**4.3714**	**3.0224**
Estonian	0.4907	**2.2240**	1.2108
Kurdish	**5.1326**	**3.1453**	**2.2361**
Russian	**2.3955**	**5.2633**	**3.1523**
Spanish	**5.4506**	**2.2105**	**2.3521**
Turkish	**4.9694**	**1.9604**	**3.6600**

the authors performed language recognition experiments on speech and non-speech frames separately. The goal of the authors was to demonstrate that if the system performance on the non-speech frames is comparable with that attained using speech frames, then the system is actually modeling the channel and not language variability. Therefore, we have first split data into speech and non-speech frames. Then we have computed the $\mathrm{EER_{avg}}$ on the non-speech frames, which was equal to 40.51% and 40.18% in manner and place cases, respectively. The $\mathrm{EER_{avg}}$ on the speech frames was instead equal to 8.48% and 14.20% in the manner and place systems, respectively. These results suggest that our technique is not modeling channel effects.

Next we explore different architectural configurations to assess their effect on the recognition accuracy.

1) Effect of i-vector dimensionality on the FSD corpus: In Table IV, we showed that attribute system outperforms the baseline spectral system in foreign accent recognition. Here, we turn our attention to the choice of i-vector dimensionality used to train and evaluate different models. Figure 4 shows recognition error rates on the FSD corpus as a function of i-vector size. Results indicate that better system performance can be attained by increasing the i-vector dimensionality up to 1000, which is inline with the findings reported in [22]. However, further increasing the i-vector dimensionality to 1200, or 1400 degraded the recognition accuracy. For example, $\mathrm{C_{avg}}$ increased to 6.10 and 6.60 from the initial 5.80 for the manner-based foreign accent recognition system with i-vector

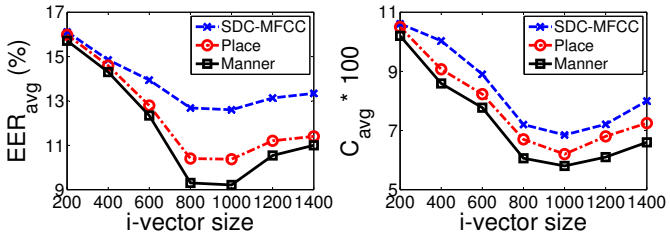


Fig. 4: Recognition error rates as a function of i-vector dimensionality on the FSD corpus.

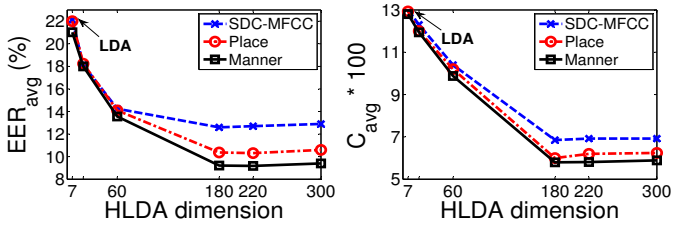


Fig. 5: Recognition error rates as a function of HLDA dimension on the FSD corpus. I-vectors are of dimensionality 1000. For lower HLDA dimensions, i.e., 7, 20 and 60, the systems attain lower recognition accuracies.

dimensionality of 1200 and 1400, respectively.

We also investigated the effect of HLDA dimensionality reduction algorithm on recognition error rates using 6 different HLDA output dimensionalities on the FSD corpus. Figure 5 shows that the optimal HLDA dimension is around 180, yielding C_{avg} of 5.8 and 6 in the manner and place systems, respectively. For lower HLDA dimensions, i.e., 7, 20 and 60, the systems attain lower recognition accuracies as shown. Comparing HLDA results in Figure 5 with LDA, the recognition error rates increase to $\mathrm{EER_{avg}}$ of 21.65% and 21.87% in manner and place systems, respectively. The output dimensionality of LDA is then restricted to maximum of seven.

2) Effect of training set size and testing utterance length on the FSD corpus: To demonstrate the recognition error rates as a function of training set size in this study, we split the Finnish training i-vectors into portions of 20%, 40%, 60%, 80% and 100% of the whole training i-vectors within each model in such a way that each individual portion contains the data from previous portion. Fixing the amount of test data, we experimented with each training data portion to report the recognition error rates as a function of training data size. Results in Figure 6 shows that the proposed attribute-based foreign accent recognition system outperforms the spectral-based system in all the cases (i.e., independently of the amount of training data). Further to see the effect of test data length on recognition error rates, we extracted new i-vectors from the 20%, 40%, 60%, 80% and 100% of *active speech frames* and used them in evaluation. Results in Figure 7, which refers to the FSD corpus, indicate that the proposed attribute-based accent recognition system compares favorably to the SDC-MFCC system in all the cases.

3) Effect of Temporal Context – FSD corpus: In Section II-B, it was argued that temporal information may be beneficial

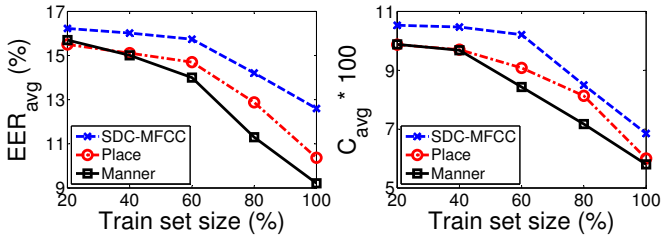


Fig. 6: Recognition error rates as a function of training set size on the FSD corpus. Increasing training set size within each target accent models degrades recognition error rates.

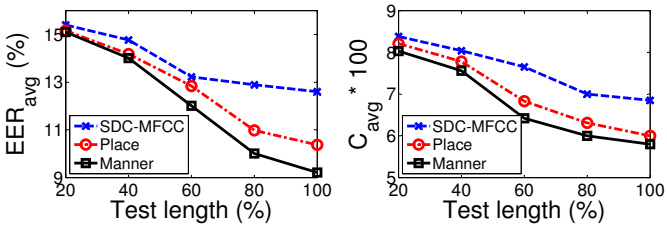


Fig. 7: Recognition error rates as a function of testing utterance length on the FSD corpus. Different portions of active speech segments were used to extract evaluation i-vectors.

to accent recognition. Figure 4 indicates that $\mathrm{C}_{\mathrm{avg}}$ attains minima at context sizes 10 and 20 frames, for the place and manner features, respectively. Optimum for the PCA-combined features occurs at 10 frames. Increasing the context size beyond 20 frames negatively affects recognition accuracy for all the evaluated configurations. In fact, we tested context window spanning up to 40 adjacent frames, but that caused numerical problems during UBM training, leading to singular covariance matrices. Hence, context size in the range of 10 to 20 frames appears a suitable trade-off between capturing contextual information while retaining feature dimensionality manageable for our classifier back-end.

Table VI shows results for several configurations of the proposed technique and optimal context window sizes selected according to Figure 8. Systems using context dependent information are indicated by adding the letters CD in front of their name. The last two rows show the result for context-independent attribute systems for reference purposes. Table VI demonstrates that context information is beneficial for foreign accent recognition. The best performance is obtained by concatenating C=20 adjacent manner feature frames followed by PCA to reduce the final vector dimensionality to d=48. A 14% relative improvement, in terms of $\mathrm{C}_{\mathrm{avg}}$, over the context-independent manner system (last row) is obtained by adding context information.

4) Effect of Feature Concatenation on the FSD corpus: We now turn our attention to the effects of feature concatenation on the accent recognition performance. The first row of Table VII shows that $\mathrm{C}_{\mathrm{avg}}$ of 5.70 is obtained by appending the place features with the SDC+MFCC features, which yields a relative improvement of 5% over the place system (third last row). A 12% relative improvement over the manner system (second last row) is obtained by concatenating the

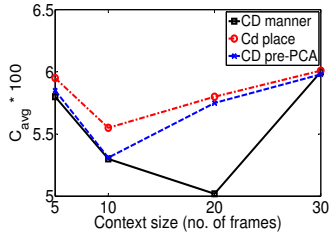


Fig. 8: $\mathrm{C}_{\mathrm{avg}}$ as a function of the context window size on the FSD corpus. Context dependent (CD) manner and place features attain the minimum $\mathrm{C}_{\mathrm{avg}}$ at context sizes 10 and 20 frames, respectively. In pre-PCA, PCA is applied to combined manner and place vectors.

TABLE VI: Recognition results for several attribute systems and different context window sizes. C represents the length of context window, and d the vector dimension after PCA. PCA can be applied either before (pre-PCA) or after (post-PCA) concatenating manner and place vectors.

System (context, dimension)	$\mathrm{EER}_{\mathrm{avg}}$(%)	$\mathrm{C}_{\mathrm{avg}} \times 100$
CD Place (C=10, d=31)	8.87	5.55
CD post-PCA (C=10, d=70)	8.20	5.43
CD pre-PCA (C=10, d=60)	7.97	5.31
CD Manner (C=20, d=48)	**7.38**	**5.02**
Place (27)	10.37	6.00
Manner (18)	9.21	5.80

SDC+MFCC features and the manner features, yielding $\mathrm{C}_{\mathrm{avg}}$ of 5.13 (the second row). If context-dependent information is used before forming the manner-based vector to be concatenated with the SDC+MFCC features, a further improvement is obtained, as the third row of Table VII indicates. Specifically, $\mathrm{C}_{\mathrm{avg}}$ of 4.74 is obtained by using a context of 20 frames followed by PCA reduction down to 48 dimensions (C=20, d=48). The result represents 19% relative improvement over the use of CD manner-only score with the same context window and final dimensionality (last row).

For the sake of completeness, Table VII shows also results obtained by concatenating manner and place attributes, which is referred to as *Manner+Place* system. This system obtains $\mathrm{C}_{\mathrm{avg}}$ of 5.51, which represents 5% and 8% relative improvements over the basic manner and place systems, respectively. In contrast, no improvement is obtained by concatenating context-dependent manner and place systems (see the row labeled CD Manner (C=20, d=48) + CD Place (C=10, d=31)) over context-dependent manner system (last row).

5) Detection Performance versus Target Language – FSD corpus: Table VIII shows language-wise results on the FSD task. The so-called leave-one-speaker-out (LOSO) technique, already used in [10], was adopted to generate these results and to compensate for lack of sufficient data in training and evaluation. For every target accent, each speaker's utterances are left out one at a time while the remaining utterances are used in training the corresponding accent recognizer. The held-out utterances are then used as the evaluation utterances.

The CD manner-based accent recognition system was selected for this experiment, since it outperformed the place-

TABLE VII: Results on the FSD corpus after feature concatenation (+). In parentheses, the final dimension of the feature vectors sent to the back-end.

System (feature dimensionality)	**Performance**	
(SDC+MFCC) Vector + Attribute Vector	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
(SDC+MFCC) + Place (83)	9.14	5.70
(SDC+MFCC) + Manner (74)	7.78	5.13
(SDC+MFCC) + CD Manner (C=20, d=48) (104)	**6.18**	**4.74**
Feature Concatenation (+) within Attributes	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
Manner + Place (45)	8.34	5.51
CD Manner (C=20, d=48) + CD Place (C=10, d=31) (79)	8.00	5.34
Basic Accent Recognition System	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
SDC+MFCC (56)	12.60	6.85
Place (27)	10.37	6.00
Manner (18)	9.21	5.80
CD Manner (C=20, d=48) (48)	7.38	5.02

TABLE VIII: *Per language results in terms of EER % and* $\mathrm{C_{DET}} \times 100$ *on the FSD corpus. Results are reported for the CD Manner (C=20, d=48).*

Accents	EER %	$\mathrm{C_{DET}} \times 100$
English	15.11	7.00
Estonian	14.54	6.33
Russian	13.08	6.30
Kurdish	13.00	6.11
Arabic	12.55	6.10
Albanian	11.43	6.07
Spanish	10.74	5.75
Turkish	8.36	5.52
Total (average)	12.35	6.14

based one. Furthermore, since we have already observed that the performance improvement obtained by combining manner- and placed-based information is not compelling, it is preferable to use a less complex system.

Table VIII indicates that Turkish is the easiest accent to detect. In contrast, English and Estonian are the hardest accents to detect. Furthermore, languages with different subfamily from Finnish, are among the easiest to deal with. Nonetheless, the last row of Table VIII shows an $\mathrm{EER_{avg}}$ and a $\mathrm{C_{avg}}$ higher than the corresponding values reported in Table VI. This might be explained recalling that the unused accents employed to train UBM, T-matrix and the HLDA in LOSO induces a mismatch between model training data and the hyper-parameter training data which degrades the recognition accuracy [10].

It is interesting to study the results of Table VIII a bit deeper to understand which language pairs are easier to confuse. Here we treat the problem as foreign accent identification task. Table IX shows the confusion matrix. The diagonal entries demonstrate that correct recognition is highly likely. Taking Turkish as the language with highest recognition accuracy, out of 30 misclassified Turkish test segments, 10 are classified as Arabic. That seems to be a reasonable result, since Turkey is bordered by two Arabic countries, namely Syria and Iraq. In addition, Turkish shares common linguistic features with Arabic. With respect to Albanian as one of the languages in the middle: 11 out of 26 misclassified test segment are assigned to the Russian class. That might be explained considering

TABLE IX: Confusion matrix on the Finnish accent recognition task. Results are reported for the CD manner (C=20, d=48).

		Predicted label							
		TUR	SPA	ALB	ARA	KUR	RUS	EST	ENG
True label	TUR	**70**	3	1	10	5	5	2	4
	SPA	1	**51**	3	8	2	2	3	2
	ALB	1	3	**62**	3	1	11	5	2
	ARA	12	9	7	**128**	10	9	8	8
	KUR	9	3	3	6	**60**	5	3	4
	RUS	43	30	51	20	16	**379**	25	26
	EST	6	8	8	12	6	13	**120**	13
	ENG	7	10	3	6	3	7	6	**63**

TABLE X: English results in terms of $\mathrm{EER_{avg}}(\%)$ and $\mathrm{C_{avg}}$ on the NIST 2008 corpus. In parentheses, the final dimensionality of the feature vectors sent to the back-end.

Feature (dimensionality)	Classifier	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
SDC+MFCC (56)	GMM-UBM	16.94	9.00
SDC+MFCC (56)	i-vector	13.82	7.87
Place (27)	i-vector	12.00	7.27
Manner (18)	i-vector	11.09	6.70
CD Manner (C=20, d=48)	i-vector	**10.18**	**6.30**

that Russian has a considerable influence on the Albanian vocabulary. Russian is one of the most difficult languages to detect, and 43 samples are wrongly recognized as Turkish. The latter outcome can be explained recalling that Russian has some words with Turkish roots; moreover, the two languages have some similarities in terms of pronunciation.

B. Results on the NIST 2008 corpus

Up to this point, we have focused on the FSD corpus to optimize parameters. These parameters are: the UBM and i-vector size, the HLDA dimensionality, and the context window size. The first three parameters, i.e. UBM size 512, i-vector dimensionality 1000 and HLDA dimensionality 180 were optimized in [10] while the context window was set to $C = 20$ for manner attributes based on our analysis in the present study. We now use the optimized values to carry out experiments on English data.

Table X compares results of the proposed and baseline systems on the NIST 2008 SRE corpus. As above, manner- and place-based systems outperform the SDC+MFCC-based i-vector system, yielding 15% and 8% relative improvements in $\mathrm{C_{avg}}$, respectively. These relative improvements are lower compared to the corresponding results for Finnish, which is understandable considering that the parameters were optimized on the FSD data. The best recognition results are obtained using a context window of C=20 adjacent frames and dimensionality reduction to d=48 features via PCA. Similar to FSD task, different architectural alternatives are now investigated to further boost system performance.

1) Effect of Feature Concatenation on the NIST 2008 corpus: Feature concatenation results on the NIST 2008 task are shown in Table XI. Similar to findings on FSD, accuracy is enhanced by combining SDC+MFCC and attribute features. The largest relative improvement is obtained by combining SDC+MFCC and CD manner features (third row in Table

TABLE XI: Results on the NIST 2008 corpus after feature concatenation (+). In parentheses, the final dimensionality of the feature vectors sent to the back-end.

System (feature dimensionality)	**Performance**	
(SDC+MFCC) Vector + Attribute Vector	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
(SDC+MFCC)+Place (83)	11.20	6.82
(SDC+MFCC)+Manner (74)	10.01	6.24
(SDC+MFCC)+CD Manner (C=20, d=48) (104)	**8.56**	**5.73**
Feature Concatenation (+) within Attributes	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
Manner+Place	10.50	6.40
Basic Accent Recognition system	$\mathrm{EER_{avg}}(\%)$	$\mathrm{C_{avg}} \times 100$
SDC+MFCC (56)	13.82	7.87
Place (27)	12.00	7.27
Manner (18)	11.09	6.70
CD Manner (C=20, d=48)	10.18	6.30

TABLE XII: Per-language results in terms of EER % and $\mathrm{C_{DET}} \times 100$ for the i-vector system in the NIST 2008 corpus. Results are reported for CD manner (C=20, d=48)

Accents	EER %	$\mathrm{C_{DET}} \times 100$
Cantonese	16.48	8.46
Hindi	14.97	7.91
Vietnamese	14.04	7.30
Russian	12.09	7.57
Korean	11.54	6.96
Japanese	10.84	6.62
Thai	10.59	6.35
Total (average)	12.93	7.31

XI), yielding $\mathrm{C_{avg}}$ of 5.73. As for FSD, improvement is also obtained by concatenating manner and place features, with final $\mathrm{C_{avg}}$ of 6.40, which represents 7% relative improvement over the basic configurations in the second and third last rows. Nonetheless, higher accuracy is obtained by the CD manner system, shown in the last row.

2) Detection Performance versus Target Language – NIST 2008 corpus: Table XII shows per-accent detection accuracy on the NIST 2008 task. Similar to the FSD experiments, the LOSO technique is applied to make better use of the limited training and testing data. Cantonese attains the lowest recognition accuracy with $\mathrm{C_{DET}}$ of 8.46; and the easiest accent is Thai with $\mathrm{C_{DET}}$ of 6.35. The confusion matrix is shown in Table XIII. It is obvious that East Asian languages, such as Korean, Japanese, Vietnamese and Thai are frequently confused with Cantonese. For example, Thai is the easiest accent to detect, yet 15 out of the 37 misclassified test segments were classified as Cantonese. Thai and Cantonese are both from the same Sino-Tibetan language family; therefore, these languages share similar sound elements. Furthermore, the same set of numbers from one to ten is used for both languages.

Russian and Hindi are both from the Indo-European language group. Hence these languages have many words and phrases in common. These similarities might explain why 12 out of 36 misclassified Russian segments were classified as Hindi. Similarly, 14 out of 48 misclassified Hindi segments were assigned to the Russian language.

TABLE XIII: Confusion matrix of the English results corresponding to Table XII. Results are reported for CD manner (C=20, d=48)

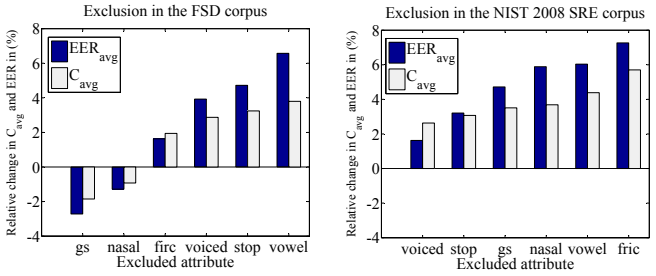
		Predicted label						
		THA	JPN	KOR	RUS	VIE	HIN	CAN
True label	THA	**126**	3	4	3	4	8	15
	JPN	3	**98**	4	2	7	2	10
	KOR	3	5	**145**	6	7	5	17
	RUS	4	3	3	**120**	4	12	10
	VIE	10	16	6	6	**200**	4	33
	HIN	4	4	6	14	5	**128**	15
	CAN	15	10	11	6	14	6	**96**

C. Effect of Individual Attribute on Detection Performance

We now investigate the relative importance of each individual manner attribute and the voiced attribute on both FSD and NIST 2008. We selected manner-based system as it outperformed place-based system both in both FSD and NIST 2008 (Tables IV and X). A 15-dimensional feature vector is formed by leaving out one of these attributes one at a time. The full i-vector system is then trained from scratch using the feature vectors without the excluded attribute. By comparing the change in $\mathrm{EER_{avg}}$ and $\mathrm{C_{avg}}$ of such system relative to the system utilizing all the 15 features allows us to quantify the relative importance of that attribute. When no context information is used, $\mathrm{EER_{avg}}$ and $\mathrm{C_{avg}}$ are 9.21% and 5.80, respectively.

Figure 9a reveals that excluding vowel, stop, voiced, or fricative attributes increases both $\mathrm{C_{avg}}$ and $\mathrm{EER_{avg}}$, indicating the importance of these attributes. In contrast, nasal and glide are not individually beneficial, since both $\mathrm{C_{avg}}$ and $\mathrm{EER_{avg}}$ show a negative relative change. Finnish has a very large vowel space (with 8 vowels) including vowel lengthening. Non-native Finnish speakers may thus have troubles when trying to produce vowels in a proper way, and that shows the L1 influence. This may explain why vowels are individually useful in foreign accent recognition for Finnish.

Figure 9b shows that *all* speech attributes are individually useful in detecting L2 in an English spoken sentence. We recall that $\mathrm{EER_{avg}}$ and $\mathrm{C_{avg}}$ are 11.09% and 6.70, respectively, when no context information is used. Hence, leaving out any of these attributes from the final feature vector, increases the error rates. Fricative and vowel are individually most important, while, voiced and stop attributes are less important. It is known that pronouncing English fricatives is difficult for some L2 speakers [48], [49]. For example, some Russian speakers pronounce dental fricatives /ð/ and /θ/ as /t/ and /d/, respectively [50]. With respect to the vowel class, some East Asian speakers find it difficult to pronounce English vowels, thus producing L1 influence. For example, English contains more vowel sounds than Chinese languages [51]. This may cause Chinese learners of English to have difficulties with pronunciation. Koreans may also have also difficulty pronouncing the sound /ɔ/ which does not exist in Korean language and is frequently substituted with the sound /o/ in Korean [52].



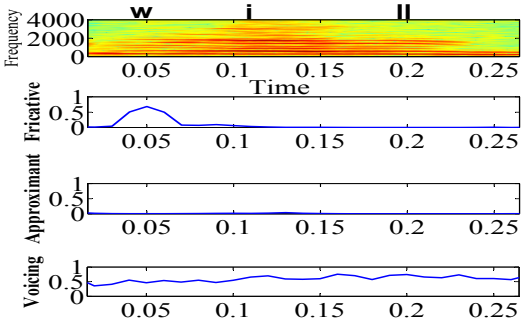
(a) FSD task (b) NIST 2008 task

Fig. 9: Exclusion experiment: relative change in the error rates as one attribute is left out. Positive relative change indicates increment in the error rates.

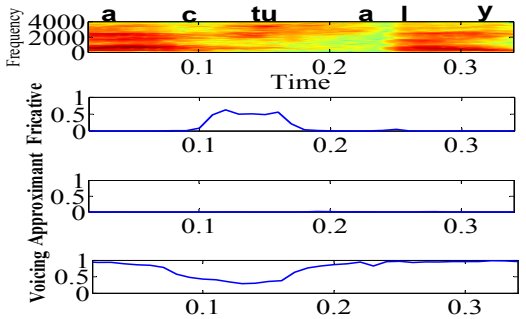
D. Diagnostic Information of Attribute Features

Besides improving the accuracy of state-of-the-art automatic foreign accent recognizer, the proposed technique provides a great deal of diagnostic information to pinpoint *why* it works well in one instance and then fail badly in another. To exemplify, Figure 10 shows analysis of two different spoken words uttered by native Russian and Cantonese speakers in the NIST 2008 SRE corpus on which the proposed attribute-based technique was successful, but the spectral-based SDC+MFCC technique failed. Figure 10a shows the spectrogram along with fricative and the approximant detection curves for the word "will" uttered by a native Russian speaker. Although /w/ belongs to the approximant class, the corresponding detection curve is completely flat. In contrast, a high level of activity is seen in the fricative detector. This can be explained noting that Russian does not have the consonant /w/, and Russian speakers typically substitute it with /v/ [53], which is a fricative consonant. Figure 10b, in turn, signifies that consonant sounds, except nasals and semivowels, are all voiceless in Cantonese [54]. Although /c/ (pronounced as a /k/) and /tu/ (pronounced as a /tʃ/) are voiced consonants in English, voicing activity is less pronounced in the time frame spanning the /c/ and /tu/ consonants, which is a specific feature of Cantonese speakers [54].

Incidentally, such information could also be useful in computer-assisted language learning system to detect mispronunciations and give some proper feedback to the user.



(a) Native Russian speaker substitutes approximant /w/ with fricative /v/.



(b) Consonants in Cantonese are all voiceless.

Fig. 10: The informative nature of the proposed accent recognition system for two spoken utterances from native Russian and Cantonese speakers. For these utterances, attribute-based technique has been successful but the spectral-based technique has failed.

V. Conclusion

In this paper, an automatic foreign language recognition system based on universal acoustic characterization has been presented.

Taking inspiration from [30], the key idea is to describe any spoken language with a common set of fundamental units that can be defined "universally" across all spoken languages. Phonetic features, such as manner and place of articulation, are chosen to form this unit inventory and used to build a set of language-universal attribute models with data-driven modeling techniques.

The proposed approach aims to unify within a single framework phonotactic and spectral based approach to automatic foreign accent recognition. The leading idea is to take the advantages of the subspace modeling techniques without discharging the valuable information provided by the phonotactic-based methods. To this end, a spoken utterance is processed through a set of speech attribute detectors in order to generate attribute-based feature streams representing foreign accent cues. These feature streams are then modeled within the state-of-the-art i-vector framework.

Experimental evidence on two different foreign accent recognition tasks, namely Finnish (FSD corpus) and English (NIST 2008 corpus), has demonstrated the effectiveness of the proposed solution, which compares favourably with state-of-the-art spectra-based approaches. The proposed system based on manner of articulation has achieved a relative improvement of 45% and 15% over the conventional GMM-UBM and the i-vector approach with SDC+MFCC vectors, respectively, on the FSD corpus. The place-based system has also outperformed the SDC+MFCC-based i-vector system with a 8% C_{avg} relative improvement. The difficulty at robust modeling of place of articulation causes that smaller relative improvement. It was also noticed that context information improves system performance.

We plan to investigate how to improve the base detector accuracy of place of articulation. In addition, we will investigate phonotactic [55] and deep learning language recognition systems [56] in the foreign accent recognition task. Especially, we are interested to find out whether in terms of classifier

fusion complementary information exist in those systems and our proposed method.

REFERENCES

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

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

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

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

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

[6] GAO, *Border Security: Fraud Risks Complicate States Ability to Manage Diversity Visa Program*. DIANE Publishing, 2007. [Online]. Available: http://books.google.com/books?id=PfmuLdR66qwC

[7] F. Biadsy, "Automatic dialect and accent recognition and its application to speech recognition," Ph.D. dissertation, Columbia University, 2011.

[8] J. Nerbonne, "Linguistic variation and computation (invited talk)," in *Proc. of EACL*, 2003, pp. 3–10.

[9] J. Flege, C. Schirru, and I. MacKay, "Interaction between the native and second language phonetic subsystems," *Speech Communication*, vol. 40, no. 4, pp. 467–491, 2003.

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

[11] Asher, J. J., and R. Garcia, "The optimal age to learn a foreign language," *Modern languages*, vol. 38, pp. 334–341, 1969.

[12] M. Zissman, T. Gleason, D. Rekart, and B. Losiewicz, "Automatic dialect identification of extemporaneous conversational latin American Spanish speech," in *Proc. of ICASSP*, 1995, pp. 777–780.

[13] W. M. Campbell, J. P. Campbell, and D. A. Reynolds, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20 (2-3), no. 2-3, pp. 210–229, 2005.

[14] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker recognition," *IEEE Signal Processing Letters*, vol. 13 (5), no. 5, pp. 308–311, 2006.

[15] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. of ICSLP*, 2002, pp. 89–92.

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

[17] G. Liu and J. H. Hansen, "A systematic strategy for robust automatic dialect identification," in *Proc. of EUSIPCO*, 2011, pp. 2138–2141.

[18] E. Singer, P. Torres-Carrasquillo, D. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. Sturim, "The MITLL NIST LRE 2011 language recognition system," in *Proc. of Odyssey*, 2012, pp. 209–215.

[19] C. Teixeira, I. Trancoso, and A. J. Serralheiro, "Recognition of non-native accents," in *Proc. of EUROSPEECH*, 1997, pp. 2375–2378.

[20] K. Kumpf and R. W. King, "Automatic accent classification of foreign accented Australian English speech," in *Proc. of ICSLP*, 1996, pp. 1740–1742.

[21] M. Bahari, R. Saeidi, H. Van hamme, and D. van Leeuwen, "Accent recognition using i-vector, Gaussian mean supervector, Guassian posterior probability for spontaneous telephone speech," in *Proc. of ICASSP*, 2013, pp. 7344–7348.

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

[23] A. DeMarco and S. J. Cox, "Iterative classification of regional British accents in i-vector space," in *Proc. of SIGML*, 2012, pp. 1–4.

[24] N. F. Chen, W. Shen, and J. P. Campbell, "A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models," in *Proc. of ICASSP*, 2010, pp. 5014–5017.

[25] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. of INTERSPEECH*, 2004, pp. 109–112.

[26] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

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

[28] M.-J. Kolly and V. Dellwo, "Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition," *Journal of Phonetics*, vol. 42, no. 1, pp. 12–23, 2014.

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

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

[31] H. Behravan, V. Hautamäki, S. M. Siniscalchi, T. Kinnunen, and C.-H. Lee, "Introducing attribute features to foreign accent recognition," in *Proc. of ICASSP*, 2014, pp. 5332–5336.

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

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

[34] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "Dimensionality reduction of phone log-likelihood ratio features for spoken language recognition," in *Proc. of INTERSPEECH*, 2013, pp. 64–68.

[35] M. Díez, A. Varona, M. Peñagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "New insight into the use of phone log-likelihood ratios as features for language recognition," in *Proc. of INTERSPEECH*, 2014, pp. 1841–1845.

[36] M. F. BenZeghiba, J.-L. Gauvain, and L. Lamel, "Phonotactic language recognition using MLP features," in *Proc. of INTERSPEECH*, 2012.

[37] D. Matrouf, N. Scheffer, B. G. B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Proc. of INTERSPEECH*, 2007, pp. 1242–1245.

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

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

[40] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[41] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI multi-language telephone speech corpus," in *Proc. of ICSLP*, 1992, pp. 895–898.

[42] "Finnish national foreign language certificate corpus," http://yki-korpus.jyu.fi.

[43] P. Schwarz, P. Matějaka, and J. Cernock, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of ICASSP*, 2006, pp. 325–328.

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

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

[46] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication," in *Proc. of Odyssey*, 2004, pp. 237–244.

[47] H. Bořil, A. Sangwan, and J. H. L. Hansen, "Arabic dialect identification - 'is the secret in the silence?' and other observations," in *Proc. of INTERSPEECH*, 2012, pp. 30–33.

[48] M. Timonen, "Pronunciation of the English fricatives: Problems faced by native Finnish speakers," Ph.D. dissertation, University of Iceland, 2011.

[49] L. Enli, "Pronunciation of English consonants, vowels and diphthongs of Mandarin-Chinese speakers," *Studies in Literature and Language*, vol. 8, no. 1, pp. 62–65, 2014.

[50] U. Weinreich, *Languages in Contact*. The Hague: Mouton, 1953.

[51] D. Deterding, "The pronunciation of English by speakers from China," *English World-Wide*, vol. 27, no. 2, pp. 157–198, 2006.
[52] B. Cho, "Issues concerning Korean learners of English: English education in Korea and some common difficulties of Korean students," *English World-Wide*, vol. 1, no. 2, pp. 31–36, 2004.
[53] I. Thompson, "Foreign accents revisited: The English pronunciation of Russian immigrants," *Language Learning*, vol. 41, no. 2, pp. 177–204, 1991.
[54] T. T. N. Hung, "Towards a phonology of Hong Kong English," *World Englishes*, vol. 19, no. 3, pp. 337–356, 2000.
[55] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.
[56] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *Proc. of ICASSP*, 2014, pp. 5337–5341.

Hamid Behravan received the B.Sc. degree in Electrical Engineering from the Semnan University in 2010. He received the M.Sc. degree in Computer Science from the University of Eastern Finland in 2012. He is currently a Ph.D. student in Computer Science in the same university. From 2013 to 2015, he has worked as a project researcher for the University of Turku, funded by Kone Foundation. His research interests are in the area of speech processing, with current focus on automatic language and foreign accent recognition. In addition, he is interested in nonlinear analysis of speech signals.

Ville Hautamäki received the M.Sc. degree in Computer Science from the University of Joensuu (currently known as the University of Eastern Finland), Finland in 2005. He received the Ph.D. degree in Computer Science from the same university in 2008. He has worked as a research fellow at the Institute for Infocomm Research, A*STAR, Singapore. In addition, he has worked as a post-doctoral researcher in University of Eastern Finland, funded by Academy of Finland. Currently he is working as a senior researcher in the same university. His current research interests consists of recognition problems from speech signals, such as speaker recognition and language recognition. In addition, he is interested in application of machine learning to novel tasks.

Sabato Marco Siniscalchi is an Associate Professor at the University of Enna "Kore" and affiliated with the Georgia Institute of Technology. He received his Laurea and Doctorate degrees in Computer Engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2006, he was a Post Doctoral Fellow at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he joined the Norwegian University of Science and Technology, Trondheim, Norway, as a Research Scientist at the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. In 2010, he was a Researcher Scientist at the Department of Computer Engineering, University of Palermo, Italy. He acts as an an associate editor in IEEE/ACM Transactions on Audio, Speech and Language Processing. His main research interests are in speech processing, in particular automatic speech and speaker recognition, and language identification.

Tomi Kinnunen received the Ph.D. degree in computer science from the University of Eastern Finland (UEF, formerly Univ. of Joensuu) in 2005. From 2005 to 2007, he was an associate scientist at the Institute for Infocomm Research (I2R) in Singapore. Since 2007, he has been with UEF. In 2010-2012, his research was funded by the Academy of Finland in a post-doctoral project focusing on speaker recognition. He is the PI of a 4-year Academy of Finland project focusing on speaker recognition and a co-PI of another Academy of Finland project focusing on audio-visual spoofing. He chaired the latest Odyssey 2014: The Speaker and Language Recognition workshop, acts as an associate editor in IEEE/ACM Transactions on Audio, Speech and Language Processing and Digital Signal Processing. He also holds the honorary title of Docent at Aalto University, Finland, with specialization area in speaker and language recognition. He has authored about 100 peer-reviewed scientific publications in these topics.

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Dr. Lee received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, in 1973, the M.S. degree in Engineering and Applied Science from Yale University, New Haven, in 1977, and the Ph.D. degree in Electrical Engineering with a minor in Statistics from University of Washington, Seattle, in 1981.

Dr. Lee started his professional career at Verbex Corporation, Bedford, MA, and was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Between 1986 and 2001, he was with Bell Laboratories, Murray Hill, New Jersey, where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. His research interests include multimedia communication, multimedia signal and information processing, speech and speaker recognition, speech and language modeling, spoken dialogue processing, adaptive and discriminative learning, biometric authentication, and information retrieval. From August 2001 to August 2002 he was a visiting professor at School of Computing, The National University of Singapore. In September 2002, he joined the Faculty of Engineering at Georgia Institute of Technology.

Prof. Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society (SPS), and International Speech Communication Association (ISCA). In 1991-1995, he was an associate editor for the IEEE Transactions on Signal Processing and Transactions on Speech and Audio Processing. During the same period, he served as a member of the ARPA Spoken Language Coordination Committee. In 1995-1998 he was a member of the Speech Processing Technical Committee and later became the chairman from 1997 to 1998. In 1996, he helped promote the SPS Multimedia Signal Processing Technical Committee in which he is a founding member.

Dr. Lee is a Fellow of the IEEE, and has published close to 400 papers and 30 patents. He received the SPS Senior Award in 1994 and the SPS Best Paper Award in 1997 and 1999, respectively. In 1997, he was awarded the prestigious Bell Labs President's Gold Award for his contributions to the Lucent Speech Processing Solutions product. Dr. Lee often gives seminal lectures to a wide international audience. In 2000, he was named one of the six Distinguished Lecturers by the IEEE Signal Processing Society. He was also named one of the two ISCA's inaugural Distinguished Lecturers in 2007-2008. He won the IEEE SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition". He was one of the four plenary speakers at IEEE ICASSP, held in Kyoto, Japan in April 2012. More recently, he was awarded the 2012 ISCA Medal for "pioneering and seminal contributions to the principles and practices of automatic speech and speaker recognition, including fundamental innovations in adaptive learning, discriminative training and utterance verification."