Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 66 (2015) 118–129

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish

Hamid Behravan [a,b,*], Ville Hautamäki [a], Tomi Kinnunen [a]

[a] School of Computing, University of Eastern Finland, Box 111, FIN-80101 Joensuu, Finland
[b] School of Languages and Translation Studies, University of Turku, Turku, Finland

Received 22 December 2013; received in revised form 19 September 2014; accepted 15 October 2014
Available online 23 October 2014

Abstract

i-Vector based recognition is a well-established technique in state-of-the-art speaker and language recognition but its use in dialect and accent classification has received less attention. In this work, we extensively experiment with the spectral feature based i-vector system on Finnish foreign accent recognition task. Parameters of the system are initially tuned with the CallFriend corpus. Then the optimized system is applied to the *Finnish national foreign language certificate* (FSD) corpus. The availability of suitable Finnish language corpora to estimate the hyper-parameters is necessarily limited in comparison to major languages such as English. In addition, it is not immediately clear which factors affect the foreign accent detection performance most. To this end, we assess the effect of three different components of the foreign accent recognition: (1) recognition system parameters, (2) data used for estimating hyper-parameters and (3) language aspects. We find out that training the hyper-parameters from non-matched dataset yields poor detection error rates in comparison to training from application-specific dataset. We also observed that, the mother tongue of speakers with higher proficiency in Finnish are more difficult to detect than of those speakers with lower proficiency. Analysis on age factor suggests that mother tongue detection in older speaker groups is easier than in younger speaker groups. This suggests that mother tongue traits might be more preserved in older speakers when speaking the second language in comparison to younger speakers.

Keywords: Foreign accent recognition; i-Vector; Language proficiency; Age of entry; Level of education; Where second language is spoken

1. Introduction

Foreign spoken accents are caused by the influence of one's first language on the second language (Flege et al., 2003). For example, an English–Finnish bilingual speaker may have an English accent in his/her spoken Finnish because of learning Finnish later in life. Non-native speakers induce variations in different word pronunciation and grammatical structures into the second language (Grosjean, 2010). Interestingly, these variations are not random across speakers of a given language, because the original mother tongue is the source of these variations (Witteman, 2013). Nevertheless, between-speaker differences, gender, age and anatomical differences in vocal tract generate within-language variation (Witteman, 2013). These variations are nuisance factors that adversely affect detection of the mother tongue.

Foreign accent recognition is a topic of great interest in the areas of intelligence and security including immigration and border control sites. It may help officials to detect travelers with a fake passport by recognizing the immigrant's actual country and region of spoken foreign accent (GAO, 2007). It has also a wide range of commercial

* Corresponding author at: School of Computing, University of Eastern Finland, Box 111, FIN-80101 Joensuu, Finland.
E-mail addresses: behravan@cs.uef.fi (H. Behravan), villeh@cs.uef.fi (V. Hautamäki), tkinnu@cs.uef.fi (T. Kinnunen).

http://dx.doi.org/10.1016/j.specom.2014.10.004
0167-6393/

applications including services based on user-agent voice commands and targeted advertisement.

Similar to spoken language recognition (Li et al., 2013), various techniques including *phonotactic* (Kumpf and King, 1997; Wu et al., 2010) and *acoustic* approaches (Bahari et al., 2013; Scharenborg et al., 2012; Behravan et al., 2013) have been proposed to solve the foreign accent detection task. The former uses phonemes and phone distributions to discriminate different accents; in practice, it uses multiple phone recognizer outputs followed by language modeling (Zissman, 1996). The acoustic approach in turn uses information taken directly from the spectral characteristics of the audio signals in the form of *mel-frequency cepstral coefficient* (MFCC) or *shifted delta cepstra* (SDC) features derived from MFCCs (Kohler and Kennedy, 2002). The spectral features are then modeled by a "bag-of-frames" approach such as *universal background model* (UBM) with adaptation (Torres-Carrasquillo et al., 2004) and *joint factor analysis* (JFA) (Kenny, 2005). For an excellent recent review of the current trends and computational aspects involved in general language recognition tasks including foreign accent recognition, we point the interested reader to (Li et al., 2013).

Among the acoustic systems, total variability model or *i-vector* approach originally used for speaker recognition (Dehak et al., 2011a), has been successfully applied to language recognition tasks (González et al., 2011; Dehak et al., 2011b). It consists of mapping speaker and channel variabilities to a low-dimensional space called *total variability space*. To compensate intersession effects, this technique is usually combined with *linear discriminant analysis* (LDA) (Fukunaga, 1990) and *within-class covariance normalization* (WCCN) (Kanagasundaram et al., 2011).

The i-vector approach has received less attention in dialect and accent recognition systems. Caused by more subtle linguistic variations, dialect and accent recognition are generally more difficult than language recognition (Chen et al., 2010). Thus, it is not obvious how well i-vectors will perform on these tasks. However, more fundamentally, the i-vector system has many data-driven components for which training data needs to be selected. It would be tempting to train some of the hyper-parameters on a completely different out-of-set-data (even different language), and leave only the final parts – training and testing a certain dialect or accent – to the trainable parts. This is also motivated by the fact that there is a lack of linguistic resources available for languages like Finnish, comparing to English for which corpora from NIST[1] and LDC[2] exist.

The i-vector based dialect and accent recognition has previously been addressed in (DeMarco and Cox, 2012; Bahari et al., 2013). DeMarco and Cox (2012) addressed a British dialect classification task with fourteen dialects, resulting in 68% overall classification rate while (Bahari et al., 2013) compared three accent modeling approaches in classifying English utterances produced by speakers of seven different native languages. The accuracy of the i-vector system was found comparable as compared to the other two existing methods. These studies indicate that the i-vector approach is promising for dialect and foreign accent recognition tasks. However, it can be partly attributed to availability of massive development corpora including thousands of hours of spoken English utterances to train all the system hyper-parameters. The present study presents a case when such resources are not available.

Comparing with the prior studies including our own preliminary analysis (Behravan et al., 2013), the new contribution of this study is a detailed account into factors affecting the i-vector based foreign accent detection. We study this from three different perspectives: parameters, development data, and language aspects. Firstly, we study how the various i-vector extractor **parameters**, such as the UBM size and i-vector dimensionality, affect accent detection accuracy. This classifier optimization step is carried out using the speech data from the CallFriend corpus (Canavan and Zipperle, 1996). As a minor methodological novelty, we study applicability of *heteroscedastic linear discriminant analysis* (HLDA) for supervised dimensionality reduction of i-vectors. Secondly, we study **data**-related questions on our accented Finnish language corpus. We explore how the choices of the development data for UBM, i-vector extractor and HLDA matrices affect accuracy; we study whether these could be trained using a different language (English). if the answer turn out positive, the i-vector approach would be easy to adopt to other languages without recourse to the computationally demanding steps of UBM and i-vector extractor training. Finally, we study **language aspects**. This includes three analyses: ranking of the original accents in terms of their detection difficulty, study of confusion patterns across different accents and finally, relating recognition accuracy with four affecting factors such as Finnish language proficiency, age of entry, level of education and where the second language is spoken.

Our hypothesis for the Finnish language proficiency is that recognition accuracy would be adversely affected by proficiency in Finnish. In other words, we expect higher accent detection errors for speakers who speak fluent Finnish. For the age of entry factor, we expect that the younger a speaker enters a foreign country, the higher the probability of fluency in the second language. Thus, we hypothesize that it is more difficult to detect the speaker's mother tongue in younger age groups than in older ones. This hypothesis is reasonable also because older people tend to keep their mother tongue traits more often than younger people (Munoz, 2010). Regarding the education factor, we hypothesize that mother tongue detection is more difficult in higher educated speakers than in lower educated ones. Finally, We also hypothesize that mother tongue detection is more difficult for the person who consistently use their second languages for social interaction

[1] http://www.itl.nist.gov/iad/mig/tests/spk/.
[2] http://www.ldc.upenn.edu/.

as compared to the speakers who do not use their second language in regular basis for social interaction.

2. System components

Fig. 1 shows the block diagram of the method used in this work. The i-vector system consists of two main part: front-end and back-end. The former consists of cepstral feature extraction and UBM training, whereas the latter includes sufficient statistics computation, training of the T-matrix, i-vector extraction, dimensionality reduction and scoring.

2.1. i-vector system

i-Vector modeling (Dehak et al., 2011a) is inspired by the success of *joint factor analysis* (JFA) (Kenny et al., 2008) in speaker verification. In JFA, speaker and channel effects are independently modeled using *eigenvoice* (speaker subspace) and *eigenchannel* (channel subspace) models:

$$\mathbf{M} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{U}\mathbf{x}, \tag{1}$$

where $\mathbf{M}$ is the speaker supervector, $\mathbf{m}$ is a speaker and channel independent supervector created by concatenating the centers of UBM and low-rank matrices $\mathbf{V}$ and $\mathbf{U}$ represent, respectively, linear subspaces for speaker and channel variability in the original mean supervector space. The latent variables $\mathbf{x}$ and $\mathbf{y}$ are assumed to be independent of each other and have a standard normal distributions, i.e. $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Dehak et al. (2011a) found that these subspaces are not completely independent, therefore a combined total variability modeling was introduced.

In the i-vector approach, the GMM supervector ($\mathbf{M}$) of each accent utterance is decomposed as (Dehak et al., 2011a),

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{2}$$

where $\mathbf{m}$ is again the UBM supervector, $\mathbf{T}$ is a low-rank rectangular matrix, representing between-utterance variability in the supervector space, and $\mathbf{w}$ is the i-vector, a standard normally distributed latent variable drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The $\mathbf{T}$ matrix is trained using a similar technique which is used train $\mathbf{V}$ in JFA, except that each training utterance of a speaker model is treated as belonging to different speakers. Therefore, in contrast to JFA, the $\mathbf{T}$ matrix training does not need speaker or dialect labels. To this end, i-vector approach is an unsupervised learning method. The i-vector $\mathbf{w}$ is estimated from its posterior distribution conditioned on the Baum–Welch statistics extracted from the utterance using the UBM (Dehak et al., 2011a).

The i-vector extraction can be seen as a mapping from a high-dimensional GMM supervector space to a low-dimensional i-vector that preserves most of the variability. In this work, we use 1000-dimensional that are further length normalized and whitened (Garcia-Romero and Espy-Wilson, 2011).

Cosine scoring is commonly used for measuring similarity of two i-vectors (Dehak et al., 2011a). The cosine score t of the test i-vector, $\mathbf{w}_{\text{test}}$, and the i-vectors of target accent a, $\mathbf{w}^a_{\text{target}}$, is defined as their inner product $\langle \mathbf{w}_{\text{test}}, \mathbf{w}^a_{\text{target}} \rangle$ and computed as follows:

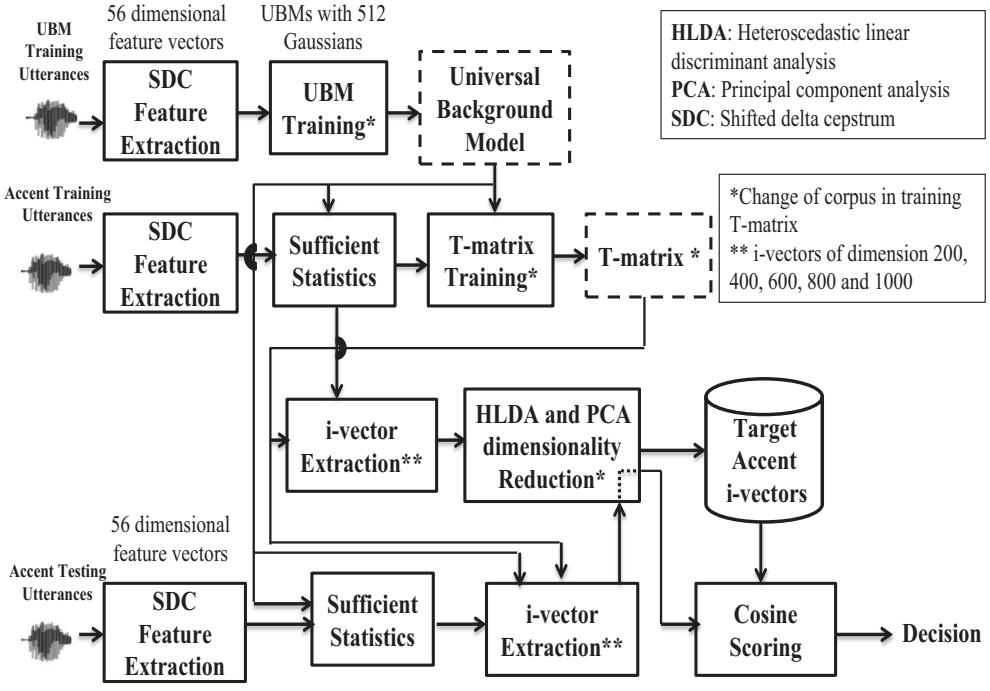


Fig. 1. The block diagram of the method used in this work.

$$t = \frac{\hat{\mathbf{w}}_{\text{test}}^{\mathrm{T}} \; \hat{\mathbf{w}}_{\text{target}}^{a}}{\|\hat{\mathbf{w}}_{\text{test}}\| \; \|\hat{\mathbf{w}}_{\text{target}}^{a}\|}, \tag{3}$$

where $\hat{\mathbf{w}}_{\text{test}}$ is,

$$\hat{\mathbf{w}}_{\text{test}} = \mathbf{A}^{\mathrm{T}}\mathbf{w}_{\text{test}}, \tag{4}$$

and $\mathbf{A}$ is the HLDA projection matrix (Loog and Duin, 2004) to be detailed below in Section 2.2. Further, $\hat{\mathbf{w}}_{\text{target}}^{a}$ is the average i-vector over all the training utterances in accent a, i.e.

$$\hat{\mathbf{w}}_{\text{target}}^{a} = \frac{1}{N_a}\sum_{i=1}^{N_a}\hat{\mathbf{w}}_i^a, \tag{5}$$

where N_a is the number of training utterances in accent a and $\hat{\mathbf{w}}_i^a$ is the projected i-vector of training utterance i from accent a, computed the same way as (4).

Obtaining the scores $\{t_a,\ a = 1,\ldots,L\}$ for a particular test utterance compared with all the L target accent models of accent a, those scores are further post-processed as (Brümmer and van Leeuwen, 2006):

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

where $t'(a)$ is the detection log-likelihood ratio or final score used in the detection task.

2.2. Reducing the i-vector dimensionality

As the extracted i-vectors contain both intra- and between-accent variations, the aim of dimensionality reduction is to project the i-vectors onto a space where between-accent variability is maximized and intra-accent variability is minimized. Traditionally, LDA is used to perform dimensionality reduction where, for R-class classification problem, the maximum projected dimension is $R-1$.

As (Loog and Duin, 2004) argue, these $R-1$ dimensions do not necessarily contain all the discriminant information for the classification task. Moreover, LDA separates only the class means and it does not take into account the discriminatory information in the class covariances. In recent years, an extension of LDA, *heteroscedastic* linear discriminant analysis (HLDA), has gained popularity in speech research community. HLDA, unlike LDA, deals with discriminant information presented both in the means and covariance matrices of classes (Loog and Duin, 2004).

HLDA was originally introduced in (Kumar, 1997) for auditory feature extraction, and later applied to speaker (Burget et al., 2007) and language (Rouvier et al., 2010) recognition with the purpose of reducing dimensionality of GMM supervectors and acoustic features, respectively. In this work, we also use it to reduce the dimensionality of extracted i-vectors. For completeness, we briefly summarize the HLDA technique below.

In the HLDA technique, the i-vectors of dimension n are projected into first $p < n$ rows, $d_{j=1\ldots p}$, of $n \times n$ HLDA transformation matrix denoted by $\mathbf{A}$. The matrix $\mathbf{A}$ is estimated by an efficient row-by-row iteration method (Gales, 1999), whereby each row is iteratively estimated as,

$$\hat{\mathbf{d}}_k = \mathbf{c}_k\mathbf{G}^{k-1}\sqrt{\frac{N}{\mathbf{c}_k\mathbf{G}^{k-1}\mathbf{c}_k^{\mathrm{T}}}}. \tag{7}$$

Here, $\mathbf{c}_k$ is the kth row vector of the co-factor matrix $\mathbf{C} = |\,\mathbf{A}\,|\,\mathbf{A}^{-1}$ for the current estimate of $\mathbf{A}$ and

$$\mathbf{G}^k = \begin{cases} \sum_{j=1}^{J}\frac{N_j}{\mathbf{d}_k\widehat{\boldsymbol{\Sigma}}^{(j)}\mathbf{d}_k^{\mathrm{T}}}\widehat{\boldsymbol{\Sigma}}^{(j)} & k \leqslant p, \\ \frac{N}{\mathbf{d}_k\widehat{\boldsymbol{\Sigma}}\mathbf{d}_k^{\mathrm{T}}}\widehat{\boldsymbol{\Sigma}} & k > p, \end{cases} \tag{8}$$

where $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}^{(j)}$ are estimates of the class-independent covariance matrix and the covariance matrix of the jth model, N_j is the number of training utterances of the jth model and N is the total number of training utterances. To avoid near-to-singular covariance matrices in HLDA training process, principal component analysis (PCA) is first applied (Loog and Duin, 2004; Rao and Mak, 2012) and the PCA-projected features are used as the inputs to HLDA. The dimension of PCA is selected in such a manner that most of the principal components are retained and within-models scatter matrix becomes non-singular (Loog and Duin, 2004).

2.3. Within-class covariance normalization

To compensate for unwanted intra-class variations in the total variability space, within-class covariance normalization (WCCN) (Hatch et al., 2006) is applied to the extracted i-vectors. To this end, a within-class covariance matrix, $\boldsymbol{\Lambda}$, is first computed using,

$$\boldsymbol{\Lambda} = \frac{1}{L}\sum_{a=1}^{L}\frac{1}{N_a}\sum_{i=1}^{N_a}(\mathbf{w}_i^a - \overline{\mathbf{w}}_a)(\mathbf{w}_i^a - \overline{\mathbf{w}}_a)^{\mathrm{T}}, \tag{9}$$

where $\overline{\mathbf{w}}_a$ is the mean i-vector for each accent a, L is the number of target accents and N_a is the number of training utterances for the accent a. The inverse of $\boldsymbol{\Lambda}$ is then used to normalize the direction of the projected i-vectors in the cosine kernel. This is equivalent to projecting the i-vector subspace by the matrix $\mathbf{B}$ obtained by Cholesky decomposition of $\boldsymbol{\Lambda}^{-1} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$.

3. Experimental setup

3.1. Corpus

We use *Finnish national foreign language certificate* (FSD) corpus (University of Jyväskylä, 2000) to perform foreign accent classification task. The corpus consists of official language proficiency tests for foreigners interested in Finnish language proficiency certificate for the purpose of applying for a job or citizenship. All the data has been recorded by language experts. Generally, the test is intended for evaluating test-takers' proficiency in listening

Table 1
Grades within different levels in the FSD corpus.

Levels	Grades		
Basic	0	1	2
Intermediate	3	4	
Advanced	5	6	

comprehension, reading comprehension, speaking, and writing. This test can be taken at basic, intermediate and advanced levels. The test-takers choose the proficiency level at which they wish to participate. The difference between the levels is the extent and variety of expression required. At the basic level, it is important that test-takers convey their message in a basic form, while in the intermediate level, richer expression is required. More effective and natural expressions should be presented in the advanced level. However, communication purposes, i.e. functions and questions, are more or less the same at all levels. Table 1 shows the grading scale at each level of the tests in this corpus.[3]

For our purposes, we selected Finnish responses corresponding to 18 foreign accents. Unfortunately, as the number of utterances in some accents was not large enough, a limited number of eight accents – Russian, Albanian, Arabic, English, Estonian, Kurdish, Spanish, and Turkish – with enough data were chosen for the experiments. However, the unused accents were utilized in training the hyper-parameters of the i-vector system, the UBM and the T-matrix.

To perform the recognition task, each accent set is randomly partitioned into a training and a test subset. To avoid speaker and session bias, the same speaker was not placed into the test and train subsets. The test subset corresponds to (approximately) 40% of the utterances, while the training set corresponds to the remaining 60%. The original audio files, stored in MPEG-2 Audio Layer III (mp3) compressed format, were decompressed, resampled to 8 kHz and partitioned into 30-s chunks. Table 2 shows the distribution of train and test files in each target accent.

The NIST SRE 2004[4] corpus was chosen as the out-of-set-data for hyper-parameter training. For our purposes, 1000 gender-balanced utterances were randomly selected from this corpus to train the UBM and T-matrix. We note that this is an American English corpus of telephone-quality speech.

Unlike UBM and T-matrix, training the HLDA projection matrix requires labeled data. Since accent labels are not represented in the NIST corpus, we use the *CallFriend* corpus (Canavan and Zipperle, 1996) to train HLDA. This corpus is a collection of unscripted conversations of 12 languages recorded over telephone lines. It includes two dialects for each target language available. All utterances are organized into training, development and evaluation subsets. For our purposes, we selected all the training utterances from dialects of English, Mandarin and Spanish languages and partitioned them into 30-s chunks, resulting in approximately 4000 splits per each subset. All audio files have 8 kHz sampling rate.

Table 2
Train and test files distributions in each target accent in the FSD corpus.

Accent	No. of train files	No. of test files	No. of speakers
Spanish	47	25	15
Albanian	56	29	19
Kurdish	61	32	21
Turkish	66	34	22
English	70	36	23
Estonian	122	62	38
Arabic	128	66	42
Russian	556	211	235
Total	1149	495	415

3.2. Front-end configuration

The front-end consists of concatenation of MFCC and SDC coefficients (Kohler and Kennedy, 2002). To this end, speech signals framed with 20 ms Hamming window with 50% overlap are filtered by 27 mel-scale filters over 0–4000 Hz frequency range. RASTA filtering (Hermansky and Morgan, 1994) is applied to log-filterbank energies. Seven first cepstral coefficients (c0–c6) are computed using discrete cosine transform. The cepstral coefficients are further processed using utterance-level cepstral mean and variance normalization (CMVN) and vocal tract length normalization (VTLN) (Lee and Rose, 1996), and converted into 49-dimensional *shifted delta cepstra* (SDC) feature vectors with 7-1-3-7 configuration parameters (Kohler and Kennedy, 2002). These four parameters correspond to, respectively, the number of cepstral coefficients, time delay for delta computation, time shift between consecutive blocks, and number of blocks for delta coefficient concatenation. Removing non-speech frames, the 7 first MFCC coefficients (including c0) are further concatenated to SDCs to obtain 56-dimensional feature vectors.

In a preliminary experiment on our evaluation corpus FSD (Behravan, 2012), the combined feature set is shown to give a relative decrease in EER of more than 30% as compared to the only SDC feature based technique.

3.3. Objective evaluation metrics

System performance is reported in terms of both average *equal error rate* (EER_{avg}) and *average detection cost* (C_{avg}) (Li et al., 2013). EER indicates the operating point on detection error trade-off (DET) curve (Martin et al., 1997) at which false alarm and miss rates are equal. EER per target accent is computed in a manner that other accents serve as non-target trials. Average *equal error rate*

[3] The FSD corpus is available by request from http://yki-korpus.jyu.fi/. Filelists used in this study are available by request from the first author.
[4] http://catalog.ldc.upenn.edu/LDC2006S44.

(EER_{avg}) is computed by taking the average over all the L target accent EERs.

C_{avg}, in turn, is defined as follows (Li et al., 2013),

$$C_{avg} = \frac{1}{L}\sum_{a=1}^{L} C_{DET}(L_a), \tag{10}$$

where $C_{DET}(L_a)$ is the detection cost for subset of test segments trials for which the target accent is L_a:

$$C_{DET}(L_a) = C_{miss}P_{tar}P_{miss}(L_a) + C_{fa}(1-P_{tar}) \times \frac{1}{L-1}\sum_{m\neq a} P_{fa}(L_a, L_m). \tag{11}$$

P_{miss} denotes the miss probability (or false rejection rate), i.e. a test segment of accent L_a is rejected as not being in that accent. $P_{fa}(L_a, L_m)$ is the probability when a test segment of accent L_m is detected as accent L_a. It is computed for each target/non-target accent pairs. C_{miss} and C_{fa} are costs of making errors and are set to 1. P_{tar} is the prior probability of a target accent and is set to 0.5.

4. Results

We first optimize the i-vector parameters in the context of dialect and accent recognition tasks. For this purpose, we utilize the CallFriend corpus. The results are summarized in Table 3.

In Fig. 2, we show EER as a function of HLDA output dimension. We find that the optimal dimension of the HLDA projected i-vectors is 180 and too aggressive reduction in dimension decreases accuracy. We also find that accuracy improves with the increase of i-vector dimensionality as Table 4 shows. Furthermore, our results showed that the UBM with smaller size outperforms larger UBM as Table 5 shows. Based on these previous findings, UBM size, i-vector size and output dimensionality are set to 512, 1000 and 180, respectively.

4.1. Effect of development data on i-vector hyper-parameters estimation

Table 6 shows the results on the FSD corpus when the hyper-parameters are trained from different datasets. Here, WCCN and score normalization are not applied. By considering the first row with matched language as a baseline (13.37% EER_{avg}), we observe the impact of each of the hyper-parameter training configurations as follows:

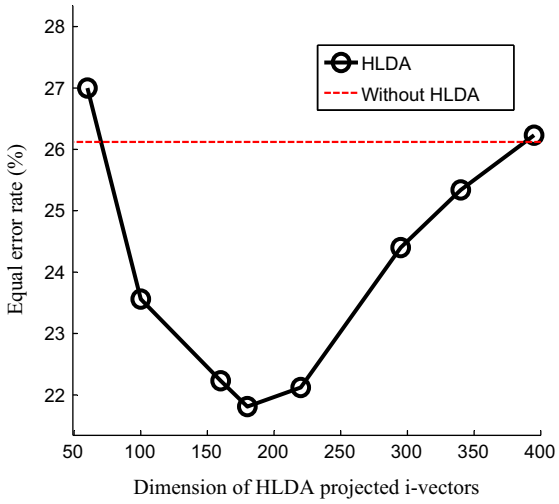


Fig. 2. Equal error rates at different dimensions of the HLDA projected i-vectors in the CallFriend corpus as reported in (Behravan et al., 2013).

- Effect of HLDA (row 1 vs row 2): EER_{avg} increases to 18.28% (relative increase of 37%).
- Effect of T-matrix (row 1 vs 3): EER_{avg} increases to 20.98% (relative increase of 57%).
- Effect of UBM (row 1 vs 4): EER_{avg} increases to 23.85% (relative increase of 78%).
- Effect of UBM and T-matrix (row 1 vs 5): EER_{avg} increases to 26.76% (relative increase of 101%).

In the light of these findings, it seems clear that the 'early' system hyper-parameters (UBM and T-matrix) have a much larger role and they should be trained from as closely matched data as possible; we see that when all the hyper-parameters are trained from the FSD corpus, the highest accuracy is achieved. The most severe degradation (101%) is attributed to the joint effect of UBM and T-matrix and the least severe (37%) to HLDA, T-matrix (57%) and UBM (78%) falling in between. It is instructive to recall the order of computations: sufficient statistics from UBM → i-vector extractor training → HLDA training. Since all the remaining steps depend on the "bottleneck" components, i.e. UBM and T-matrix, it is not surprising that they have the largest relative effect.

The generally large degradation relative to the baseline set-up with matched data is reasonably explained by the

Table 3
The i-vector system's optimum parameters as reported in (Behravan et al., 2013).

i-vector parameters	Search range and **optima**
UBM size	**256**, **512**, 1024, 2048, 4096
i-vector dimensionality	200, 400, 600, 800, **1000**
HLDA dimensionality	50, 100, 150, **180**, 220, 300, 350, 400

Table 4
Performance of the i-vector system in the CallFriend corpus for selected i-vector dimensions (EER in %, form). UBM has 1024 Gaussians as reported in (Behravan et al., 2013).

i-vector dim.	English	Mandarin	Spanish
200	23.20	20.49	20.87
400	22.60	19.11	20.21
600	21.30	18.45	19.63
800	19.83	16.31	18.63
1000	**18.01**	**14.91**	**16.01**

Table 5
Performance of the i-vector system in the CallFriend corpus for five selected UBM sizes (EER in %, form). i-vectors are of dimension 600 as reported in (Behravan et al., 2013).

UBM size	English	Mandarin	Spanish
256	**21.12**	17.93	**19.00**
512	21.61	**17.91**	19.15
1024	21.30	18.45	19.63
2048	23.81	21.15	22.01
4096	23.89	21.57	22.66

Table 6
EER_{avg} and $C_{avg} \times 100$ performance for effect of changing datasets in training the i-vector hyper-parameters. (WCCN and score normalization turned off.)

UBM	T matrix	HLDA	EER_{avg}%	$C_{avg} \times 100$	$\mathbf{Id_{error}}$%
Database used for training					
FSD	FSD	FSD	**13.37**	**7.04**	**33.65**
FSD	FSD	CallFriend	18.28	7.49	38.29
FSD	NIST	FSD	20.98	7.83	40.30
NIST	FSD	FSD	23.85	8.15	42.91
NIST	NIST	FSD	26.76	8.41	44.67

Table 7
Effect of score normalization on the recognition performance. (HLDA and WCCN turned on and off, respectively.)

Score normalization	EER_{avg}%	$C_{avg} \times 100$	Id_{error}%
No	13.37	7.04	33.65
Yes	**13.01**	**6.94**	**32.85**

large differences between type of data of evaluation corpus (FSD) and hyper-parameter estimation corpora (NIST SRE and CallFriend). FSD consists of Finnish language data recorded with close-talking microphones in a classroom environment. Even though speech is very clear, background babble noise from the other students is evident in all the recordings. This is contrast to the NIST SRE and CallFriend corpora where most of the speech files are recorded over telephone line and babble noise is less common.

The results of Table 6 were computed with WCCN and score normalization turned off. Let us now turn our attention to these additional system components. Firstly, Table 7 shows the effect of score normalization when all the hyper-parameters are trained from the FSD corpus (i.e., row 1 of Table 6). EER_{avg} decreases from 13.37% to 13.01%, which indicates a slightly increased recognition accuracy when the scores are normalized in the backend.

Secondly, Table 8 shows the joint effect of WCCN and HLDA on the recognition performance when all the hyper-parameters are trained from the FSD corpus (i.e., row 1 of Table 6). In addition to that, score normalization is also applied. EER_{avg} decreases from 17.10% to 12.60% when both HLDA and WCCN are applied. The worst case is when HLDA is turned off and WCCN is turned on. This is because turning off HLDA leads to inaccurate estimation of covariance matrix in higher dimensional i-vector space.

4.2. Comparing i-vector and GMM-UBM systems

In order to have a baseline comparison between the i-vector approach and the classical accent recognition systems, we used conventional GMM-UBM system with MAP adaptation similar to the work presented in (Torres-Carrasquillo et al., 2004). GMM-UBM system is simpler and computationally more efficient in comparison to the i-vector systems. Map adaptation consists of single iteration for adapting the UBM to each dialect model using SDC + MFCC features. It requires updating only centers of UBM. The testing is a fast scoring process described in (Reynolds et al., 2000) to score the input utterance to each adapted foreign accent models by selecting top five Gaussians per speech frame.

Table 9 shows the result of GMM-UBM system with four different UBM sizes. Increasing the number of Gaussians results in higher recognition accuracy. Table 10 further compares the best recognition accuracies achieved by both recognizers. In the i-vector system, the best recognition accuracy, i.e. EER_{avg} of 12.60%, is achieved with all the hyper-parameters trained from the FSD corpus and HLDA, WCCN and score normalization being turned on. On the other hand, the best GMM-UBM recognition accuracy, EER_{avg} of 17.00%, is achieved with UBM order 2048 when score normalization is applied. The results indicate that the i-vector system outperforms the conventional GMM-UBM system with 25% relative improvements in terms of EER_{avg} at the cost of higher computational time and additional development data.

4.3. Detection performance per target language

In the previous section, we analyzed the overall average recognition accuracy. Now, here we focus on performance for each individual foreign accent. In order to compensate the lack of sufficient development data in reporting these results, we used the previously unused accents in the FSD corpus to train UBM, T-matrix and HLDA. These unused accents are Chinese, Dari, Finnish, French, Italian, Somali, Swedish and Misc[5] corresponding to 210 speakers and 1110 utterances in total. Further, to increase the number of test trials in the classification stage, we report the results using a leave-one-speaker-out (LOSO) protocol. As demonstrated in the pseudo code below, for every accent, each speaker's utterances are held out one at a time and the remaining utterances are used in modeling the $\hat{\mathbf{w}}_{target}$ as in Eq. (5). The held-out utterances are used as the evaluation utterances.

[5] Refers to those utterances in which the spoken foreign accent is not clear.

Table 8
The joint effect of WCCN and HLDA on the recognition accuracy. (Score normalization turned on.)

HLDA	WCCN	EER_{avg}%	$C_{avg} \times 100$	Id_{error}%
No	No	17.70	7.04	39.58
Yes	No	13.01	6.94	32.85
No	Yes	19.00	7.31	41.55
Yes	Yes	**12.60**	**6.85**	**30.85**

Algorithm 1. Leave-one-speaker-out (LOSO)

Let $A = \{a_1, a_2, \ldots, a_L\}$ be the set of L target accents
Let $S(a_i)$ be the set of speakers in target accent a_i
$\hat{\mathbf{w}}^a_{\text{target}}$ defines the i-vectors of target accent a after HLDA and WCCN.
for $a_i \in A$ **do**
  for $s_j \in S(a_i)$ {Held-out test speaker} **do**
    Let $S' = S(a_i) - s_j$ {Remove the speaker being tested}
    Form $\hat{\mathbf{w}}^a_{\text{target}}$ using the i-vectors in set S', Eq. (5)
    Compute cosine scores $\langle \mathbf{w}^{s_j}_{\text{test}}, \mathbf{w}^a_{\text{target}} \rangle$ {$\mathbf{w}^{s_j}_{\text{test}}$ are the test i-vectors of speaker s_j}
  end for
end for
Normalize scores per each target accent, Eq. (6)

Table 11 shows the language wise results. The results suggest that certain languages which do not belong to the same sub-family as Finnish are easier to detect. Turkish achieves the highest recognition accuracy, whereas English shows highest error rate. The recognition accuracy is consistent among Albanian, Arabic, Kurdish and Russian languages. C_{avg} is bigger than the results already given in Table 10. Note that in Table 11, the unused accents are used to train UBM, T-matrix and HLDA. This induces mismatch between model training data and the hyper-parameter training data. Which is not the case in Table 10.

Fig. 3 further exemplifies the distribution of scores for three selected languages of varying detection difficulties. The histograms are plotted with the same number of bins, 50. For visualization purposes, the width of bins in the non-target score histogram was set smaller than in the target score histogram. The score distribution explains the differences between EERs. For example, in case of Turkish as the easiest and English as the most difficult detected accent,

Table 9
Recognition performance of GMM-UBM system with different UBM sizes.

UBM size	EER_{avg}%	$C_{avg} \times 100$
256	19.94	11.02
512	19.03	10.56
1024	18.20	10.12
2048	**17.00**	**9.46**

Table 10
Comparison between the best recognition accuracy in the GMM-UBM and i-vector system. (Score normalization turned on for the both cases.)

Recognition system	EER_{avg}%	$C_{avg} \times 100$	Id_{error}%
GMM-UBM	17.00	9.46	43.65
i-vector	**12.60**	**6.85**	**30.85**

the overlap between the target and the non-target scores is higher in the latter.

Here, the problem is treated as foreign accent identification task. Table 12 displays the confusion matrix corresponding to Table 11. In all the cases, majority of the detected cases corresponds to the correct class (i.e., the entries in the diagonal). Taking Turkish as the language with the highest recognition accuracy, out of the 11 misclassified Turkish test segments, 7 were misclassified as Arabic. This might be because Turkey is bordered by two Arabic countries, Syria and Iraq, and Turkish shares common features with Arabic. Regarding Spanish, out of the 27 misclassified test segments, 9 were detected as Arabic. It is possibly due to the major influence of Arabic on Spanish. In particular, numerous words of Arabic origin are adopted in the Spanish language.

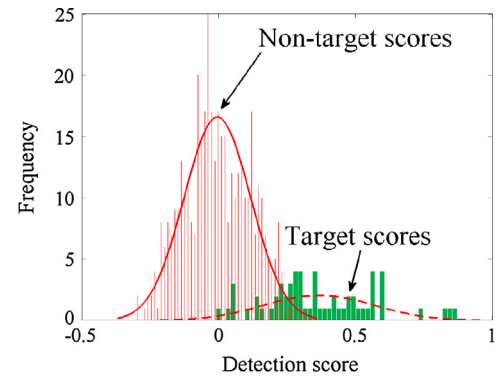
To analyze further reasons why some languages are harder to detect, we first compute the average target language score on a speaker-by-speaker basis. To measure the degree of speaker variation, we show the standard deviation of these average scores in Table 13, along with the corresponding EER and C_{DET} values. The results indicate that languages with more diverse speaker populations, having speaker-dependent biases in the detection scores, are more difficult to handle. It does not yet explain why certain languages, such as Russian, have a larger degree of speaker variation, but suggests that there will be space for further research in speaker normalization techniques.

4.4. Factors affect foreign accent recognition

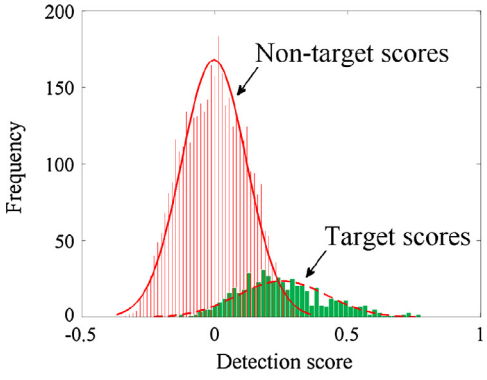
We are interested to find out what factors affect the foreign accent recognition accuracies. The rich metadata available in the FSD corpus includes language proficiency, speaker's age, education and the place where the second language is spoken. In the following analysis, we used the

Table 11
Per language results in terms of EER% and $C_{DET} \times 100$ for the i-vector system.

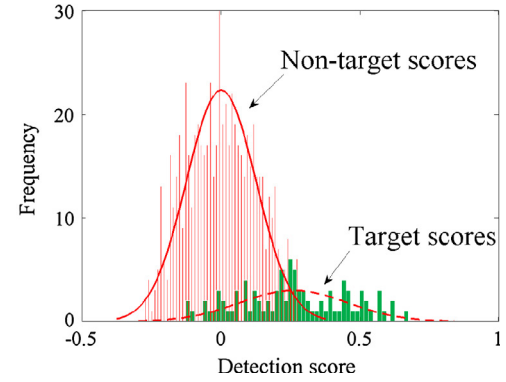
Accents	EER%	$C_{DET} \times 100$
Turkish	11.90	6.35
Spanish	16.49	6.92
Albanian	18.76	7.00
Arabic	18.98	7.17
Kurdish	19.37	7.19
Russian	19.68	7.21
Estonian	20.05	7.52
English	23.60	8.00



(a) Turkish



(b) Russian



(c) English

Fig. 3. Distribution of scores for Turkish, Russian and English accents.

whole set of scores from the LOSO experiment and grouped them to different categories according to each metadata variable at a time.

Language proficiency

To find out the impact of language proficiency, we take the sum of spoken and written Finnish grades in the FSD corpus as a proxy of the speaker's Finnish language proficiency. The objective was to find out how speakers' language proficiency and their detected foreign accent are related. Fig. 4 shows C_{avg} for each grade group. As hypothesized, the lowest C_{avg} is attributed to speakers with the lower grade (5) and the highest accuracy to speakers with the higher grade (8). This indicates that detecting the foreign accents from speakers with higher proficiency in Finnish is considerably more difficult than speakers with lower proficiency.

In addition, we looked at language proficiency across different target languages. We study the average language proficiency grade across the speakers in different languages (Table 14). For the three most difficult languages to detect, Russian, Estonian and English, the average language proficiency grades are higher than the rest of languages, supporting the preceding analysis.

Age of entry

Age is one of the most important effective factors in learning a second language (Krishna, 2008). The common notion is that younger adults learn the second language more easily than older adults. (Larsen-Freeman, 1986) argues that during the period of time between birth and the age when a children enters puberty, learning a second language is quick and efficient. In the second language acquisition process, one of the affecting factors relates to the experience of immigrants, such as the age of entry and the length of residence (Krishna, 2008). We analyze the relationship between the age of entry and the foreign accent recognition results. To analyze the effect of age to foreign accent detection, we categorized the detection scores into six age groups with 10 years age interval (Fig. 5). Our hypothesis was that mother tongue detection is easier in older people than younger ones. The results support this hypothesis. C_{avg} decreases from 5.30 (a relative

Table 12
Confusion matrix of the results corresponding to Table 11.

	Predicted label							
	Turk.	Span.	Alba.	Arab.	Kurd.	Russ.	Esto.	Engl.
True label								
Turk.	50	0	1	7	0	1	0	2
Span.	1	58	1	11	2	3	7	2
Alba.	1	0	61	9	1	5	11	1
Arab.	4	2	14	110	7	7	12	4
Kurd.	5	1	1	5	50	6	3	6
Russ.	51	21	51	26	2	369	13	28
Esto.	5	5	7	15	1	6	117	15
Engl.	7	3	3	6	3	7	9	59

Table 13
The standard deviation of the average target language score on a speaker-by-speaker basis along with the corresponding EER and C_{DET} results.

Accents	Standard deviation	EER%	$C_{DET} \times 100$
Turkish	0.1205	11.90	6.35
Spanish	0.1369	16.49	6.92
Albanian	0.1380	18.76	7.00
Arabic	0.1505	18.98	7.17
Kurdish	0.1392	19.37	7.19
Russian	0.1402	19.68	7.21
Estonian	0.1621	20.05	7.52
English	0.1667	23.60	8.00

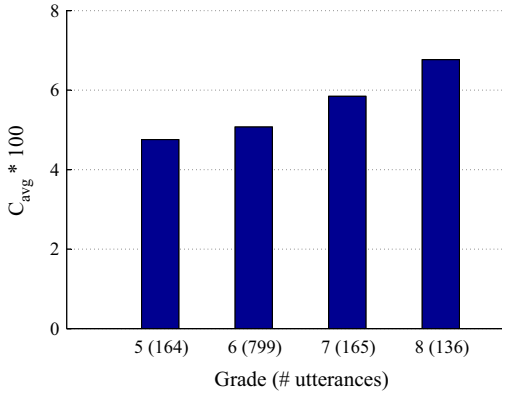


Fig. 4. $C_{avg} \times 100$ for different grade groups in the language proficiency measurement.

Table 14
The average language proficiency grade across the speakers in different languages along with the corresponding EER and C_{DET} results.

Accents	Grade	EER%	$C_{DET} \times 100$
Turkish	6.09	11.90	6.35
Spanish	6.20	16.49	6.92
Albanian	5.78	18.76	7.00
Arabic	5.73	18.98	7.17
Kurdish	5.71	19.37	7.19
Russian	6.30	19.68	7.21
Estonian	7.02	20.05	7.52
English	6.34	23.60	8.00

decrease of 16%) to 4.45 from the age group [11–20] to [61–70]. This indicates that the mother tongue detection in older age groups could be easier than in the younger age groups.

Level of education

According to Gardner's socio-educational model (Gardner, 2010), intrinsic motivation to learn a second language is strongly correlated to educational achievements. The objective was to find out how speakers' level of education and their detected foreign accent might be related. To analyze the effect of education, we categorized the detection scores into different levels of education groups. We hypothesized that people with higher level of education

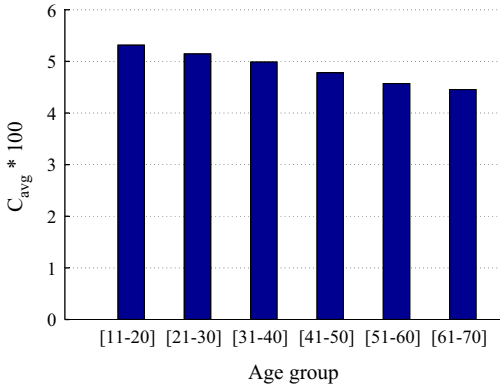


Fig. 5. $C_{avg} \times 100$ for different age groups. Age refers to age of entry to foreign country. Number of utterances for the age group [11–20], [21,30],..., [61–70] is 46, 342, 535, 239, 100, 12, respectively.

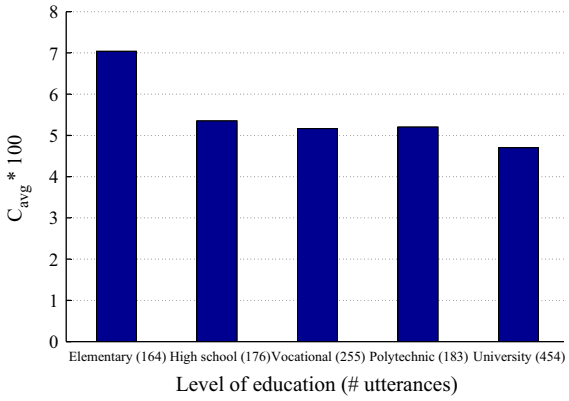


Fig. 6. $C_{avg} \times 100$ for different level of education groups.

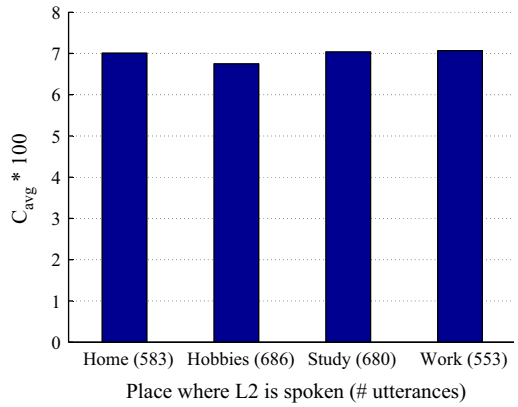


Fig. 7. $C_{avg} \times 100$ for different places where the second language is spoken.

speak the second language more fluently than lower educated people. As a consequence, mother tongue detection for higher educated people is relatively difficult. But the results in Fig. 6 in fact show the opposite; the highest C_{avg} belongs to elementary school and the lowest to university education. However, C_{avg} is somewhat similar for the high school, vocational school, and polytechnic level of education.

Where second language is spoken

Finally, we were also interested to observe whether the place or situation, where the second language is spoken, affects foreign accent detection or not. To this end, we categorized the scores into four groups based on the level of social interaction: home, hobbies, study and work. We hypothesized that the places with more social interactions between people, the mother tongue traits will be less in the second spoken language, therefore making it more difficult to detect the mother tongue. Fig. 7 shows the C_{avg} for different places where the second language is spoken. The results indicate no considerable sensitivity to the situation where the second language is spoken.

5. Conclusion

In this work, we studied how the various i-vector extractor parameters, data set selections and the speaker's language proficiency affects foreign accent detection accuracy. Regarding **parameters**, highest accuracy was achieved using UBMs with 512 Gaussians, i-vector dimensionality of 1000 and HLDA dimensionality of 180. These are similar to those reported in general speaker and language recognition literature, except for the higher-than-usual i-vector dimensionality of 1000.

Regarding **data**, we found that the choice of the UBM training data is the most critical part, followed by T-matrix and HLDA. This is understandable since the earlier system components affect the quality of the remaining steps. In all cases, the error rates increased unacceptably high for mismatched sets of hyper-parameter training. Thus, our answer to the question whether hyper-parameters could be reasonably trained from mismatched language and channel is negative. The practical implication of this is that the i-vector approach, even though producing reasonable accuracy, requires careful data selection for hyper-parameter training – and this is not always feasible.

Applying within-class covariance normalization followed by score normalization technique further increased the i-vector system performance by 6% relative improvements in terms of C_{avg}. We also showed that the i-vector system outperforms the conventional GMM-UBM system by 28% relative decrease in terms of C_{avg}.

In our view, the most interesting contribution of this work is the analysis of **language aspects**. The results, broken down by the accents, clearly suggested that certain languages which do not belong to the same sub-family as Finnish are easier to detect. Turkish was the easiest (C_{DET} of 6.35) while for instance Estonian, a language similar to Finnish, yielded C_{DET} of 7.52. The most difficult language was English with C_{DET} of 8.00. In general, confusion matrix revealed that phonetically similar languages are more often confused.

Our analysis on affecting factors suggested that language proficiency and age of entry affect detection performance. Specifically, accents produced by fluent speakers of Finnish are more difficult to detect. Speaker group with the lowest language grade 5 yielded C_{avg} of 4.75 while the group with grade 8 yielded C_{avg} of 6.76. Analysis of the age of entry, in turn, indicated that mother tongue detection in older speakers is easier than younger speakers. The age group [61–70] years yielded C_{avg} of 4.45 while the group with age interval [11–20] years old yielded C_{avg} of 5.31.

After optimizing all the parameters, the overall EER_{avg} and C_{avg} were 12.60% and 6.85, respectively. These are roughly an order of magnitude higher compared to state-of-the-art text-independent speaker recognition with i-vectors. This reflects the general difficulty of the foreign accent detection task, leaving a lot of space for future work on new feature extraction and modeling strategies. While these values are unacceptably high for security applications, the observed correlation between language proficiency and recognition scores suggests potential applications for automatic spoken language proficiency grading.

Acknowledgements

We would like to thank Ari Maijanen from University of Jyväskylä for an immense help with the FSD corpus. This work was partly supported by Academy of Finland (projects 253000, 253120 and 283256) and Kone Foundation – Finland.

References

Bahari, M.H., Saeidi, R., hamme, H.V., Leeuwen, D.V., 2013. Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, May 26–31. Vancouver, BC, Canada, pp. 7344–7348.

Behravan, H., 2012. Dialect and Accent Recognition. Master's Thesis, School of Computing, University of Eastern Finland, Joensuu, Finland.

Behravan, H., Hautamäki, V., Kinnunen, T., 2013. Foreign accent detection from spoken Finnish using i-Vectors. In: INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25–29, pp. 79–83.

Brümmer, N., van Leeuwen, D., 2006. On calibration of language recognition scores. In: IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, June 28–30, pp. 1–8.

Burget, L., Matejka, P., Schwarz, P., Glembek, O., Cernocký, J., 2007. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans. Audio, Speech Lang. Process. 15 (7), 1979–1986.

Canavan, A., Zipperle, G., 1996. CallFriend Corpus. <http://yki-korpus.jyu.fi/> (Accessed 04.07.13).

Chen, N.F., Shen, W., Campbell, J.P., 2010. A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Sheraton Dallas Hotel, Dallas, Texas, USA, March 14–19, pp. 5014–5017.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech Lang. Process. 19 (4), 788–798.

Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R., 2011b. Language recognition via i-vectors and dimensionality reduction. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 857–860.

DeMarco, A., Cox, S.J., 2012. Iterative classification of regional British accents in I-vector space. In: Machine Learning in Speech and Language Processing (MLSLP), Portland, OR, USA, September 14–18, pp. 1–4.

Flege, J.E., Schirru, C., MacKay, I.R.A., 2003. Interaction between the native and second language phonetic subsystems. Speech Commun. 40 (4), 467–491.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Speech Audio Process. 7 (3), 272–281.

GAO, 2007. Border Security: Fraud Risks Complicate States Ability to Manage Diversity Visa Program. DIANE Publishing.

Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of i-vector length normalization in speaker recognition systems. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 249–252.

Gardner, R.C., 2010. Motivation and Second Language Acquisition: The Socio-educational Model. Peter Lang, New York.

González, D.M., Plchot, O., Burget, L., Glembek, O., Matejka, P., 2011. Language recognition in ivectors space. In: INTERSPEECH 2011: 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 861–864.

Grosjean, F., 2010. Bilingual: Life and Reality. Harvard University Press.

Hatch, A.O., Kajarekar, S.S., Stolcke, A., 2006. Within-class covariance normalization for SVM-based speaker recognition. In: INTERSPEECH 2006, ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17–21, pp. 1471–1474.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2 (4), 578–589.

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., Mason, M., 2011. i-vector based speaker recognition on short utterances. In: INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27–31, pp. 2341–2344.

Kenny, P., 2005. Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Technical Report CRIM-06/08-13.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of interspeaker variability in speaker verification. IEEE Trans. Audio, Speech Lang. Process. 16 (5), 980–988.

Kohler, M.A., Kennedy, M., 2002. Language identification using shifted delta cepstra. In: 45th Midwest Symposium on Circuits and Systems, vol. 3, pp. III-69–72.

Krishna, B., 2008. Age as an affective factor in second language acquisition. Engl. Specif. Purp. World 21 (5), 1–14.

Kumar, N., 1997. Investigation of Silicon-auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition. Ph.D. Thesis, Baltimore, Maryland.

Kumpf, K., King, R.W., 1997. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In: Fifth European Conference on Speech Communication and Technology, EUROSPEECH, Rhodes, Greece, September 22–25, pp. 2323–2326.

Larsen-Freeman, D., 1986. Techniques and Principles in Language Teaching. Oxford University Press, New York.

Lee, L., Rose, R.C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proceedings of the Acoustics, Speech, and Signal Processing, May 7–10, pp. 353–356.

Li, H., Ma, B., Lee, K.-A., 2013. Spoken language recognition: from fundamentals to practice. Proc. IEEE 101 (5), 1136–1159.

Loog, M., Duin, R.P.W., 2004. Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion. IEEE Trans. Pattern Anal. Mach. Intell. 26 (6), 732–739.

Martin, A.F., Doddington, G.R., Kamm, T., Ordowski, M., Przybocki, M.A., 1997. The DET curve in assessment of detection task performance. In: EUROSPEECH 1997, 5th European Conference on Speech Communication and Technology, Rhodes, Greece, September 22–25, pp. 1895–1898.

Munoz, C., 2010. On how age affects foreign language learning. Adv. Res. Lang. Acquisit. Teach., 39–49.

Rao, W., Mak, M.-W., 2012. Alleviating the small sample-size problem in i-vector based speaker verification. In: 8th International Symposium on Chinese Spoken Language Processing, Kowloon Tong, China, December 5–8, pp. 335–339.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10 (1–3), 19–41.

Rouvier, M., Dufour, R., Linarès, G., Estève, Y., 2010. A language-identification inspired method for spontaneous speech detection. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, September 26–30, pp. 1149–1152.

Scharenborg, O., Witteman, M.J., Weber, A., 2012. Computational modelling of the recognition of foreign-accented speech. In: INTERSPEECH 2012: 13th Annual Conference of the International Speech Communication Association, September 9–13, pp. 882–885.

Torres-Carrasquillo, P.A., Gleason, T.P., Reynolds, D.A., 2004. Dialect identification using Gaussian mixture models. In: Proceeding Odyssey: The Speaker and Language Recognition Workshop, May 31–June 3, pp. 757–760.

University of Jyväskylä, 2000. Finnish National Foreign Language Certificate Corpus, University of Jyväskylä, Centre for Applied Language Studies. <http://yki-korpus.jyu.fi/>.

Witteman, M., 2013. Lexical Processing of Foreign-accented Speech: Rapid and Flexible Adaptation. Ph.D. Thesis.

Wu, T., Duchateau, J., Martens, J., Compernolle, D., 2010. Feature subset selection for improved native accent identification. Speech Commun. 52 (2), 83–98.

Zissman, M.A., 1996. Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. 4 (1), 31–44.