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/© 2014 Elsevier B.V. All rights reserved.

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/.