Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 70–81

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Native vs. non-native accent identification using Japanese spoken telephone numbers

Kanae Amino *, Takashi Osanai

National Research Institute of Police Science, 6-3-1 Kashiwanoha, Kashiwa, Chiba 277-0882, Japan

Received 25 December 2012; received in revised form 26 July 2013; accepted 29 July 2013
Available online 11 August 2013

Abstract

In forensic investigations, it would be helpful to be able to identify a speaker's native language based on the sound of their speech. Previous research on foreign accent identification suggested that the identification accuracy can be improved by using linguistic forms in which non-native characteristics are reflected. This study investigates how native and non-native speakers of Japanese differ in reading Japanese telephone numbers, which have a specific prosodic structure called a bipodic template. Spoken Japanese telephone numbers were recorded from native speakers, and Chinese and Korean learners of Japanese. Twelve utterances were obtained from each speaker, and their F0 contours were compared between native and non-native speakers. All native speakers realised the prosodic pattern of the bipodic template while reading the telephone numbers, whereas non-native speakers did not. The metric rhythm and segmental properties of the speech samples were also analysed, and a foreign accent identification experiment was carried out using six acoustic features. By applying a logistic regression analysis, this method yielded an 81.8% correct identification rate, which is slightly better than that achieved in other studies. Discrimination accuracy between native and non-native accents was better than 90%, although discrimination between the two non-native accents was not that successful. A perceptual accent identification experiment was also conducted in order to compare automatic and human identifications. The results revealed that human listeners could discriminate between native and non-native speakers better, while they were inferior at identifying foreign accents.

Keywords: Foreign accent identification; Non-native speech; Spoken telephone numbers; Prosody; Forensic speech science

1. Introduction

Globalisation has provided us with more opportunities to communicate with people from all around the world. This also leads to more chances to hear foreign accents. Investigation of foreign accents is important not only for second language (L2) acquisition research and language teaching, but also for technologies such as speech recognition, speaker recognition, and accent identification. The term "accent" can be defined as speech properties that indicate which country, or which part of a country, the speaker originates from. Accent identification is commonly used to identify a speaker's mother dialect (D1) by using speech samples spoken in D1 or other dialects (D2). For foreign accent identification, a speaker's first language (L1) is identified using speech in L2 or a later language. Applications of accent identification include preprocessing for automatic speech recognition and language support for L2 speakers. The performance of a speech recognition system can be improved by applying accent identification in advance and then using a dialect or language model in which the accent colour is taken into consideration (e.g., Brousseau and Fox, 1992; Blackburn et al., 1993; Arslan and Hansen, 1996; Fung and Kat, 1999). This is also useful for assisting L2 speakers when call routing is needed for emergency operators or in multi-lingual voice-controlled information retrieval systems (Muthusamy et al., 1994; Zissman, 1996). Furthermore, in forensic situations, when there is

* Corresponding author. Tel.: +81 (0)4 7135 8001.
E-mail addresses: amino@nrips.go.jp (K. Amino), osanai@nrips.go.jp (T. Osanai).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.07.010

a possibility that obtained speech samples were spoken by a D2/L2 speaker, identifying the speaker's accent, and consequently his/her nationality and/or hometown, can often lead to important clues with regard to the suspect.

A speech technology similar to accent identification is language identification. However, compared to language identification, in which the language spoken by a native speaker is identified, accent identification is considered to be a more challenging task. One reason is that the traits of a speaker's D1/L1 are carried into D2/L2 speech in various ways. These traits, often called language transfers, may appear on the segmental level, for instance, the substitution of unfamiliar phonemes with similar sounds from the D1/L1, or on the supra-segmental (prosodic) level, e.g., erroneous word accents, clumsy rhythm, and inappropriate intonation. What makes accent identification more difficult is the fact that language transfer is not unique to one target dialect/language or speaker but depends on the speaker's D1/L1, the language-typological distance between D1/L1 and D2/L2, and various individual factors. For example, different phonemic inventories and phonotactics will bring about different articulatory errors, and different accentuation systems will cause different prosodic problems. Also, the degree of language transfer is reported to depend on each speaker's age of learning (or age of arrival), amount of exposure and interactive contact with native speakers (e.g., Flege, 1988; Flege and Fletcher, 1992), experience of learning other foreign dialects or languages (Mehlhorn, 2007; Wrembel, 2009), and the individual's language talent (Markham, 1999); there are also several reports that disclaim the effects of the former two factors (Mackay and Fullana, 2009; Fullana and Mora, 2009).

Previous research on accent identification can be classified into three groups: that based on segmental and articulatory features (Arslan and Hansen, 1996; Kumpf and King, 1996; Teixeira et al., 1996; Berkling et al., 1998; Yanguas et al., 1998), that based on prosodic features (Itahashi and Yamashita, 1992; Itahashi and Tanaka, 1993; Hansen and Arslan, 1995; Mixdorff, 1996; Piat et al., 2008), and that based on both (Piat et al., 2008; Arslan and Hansen, 1997; Vieru-Dimulescu et al., 2007). Kumpf and King (1996) identified three accents of Australian English: Lebanese, Vietnamese, and native. They used a system based on a hidden Markov model (HMM) trained using 2000 sentences recorded from 16 speakers, and identified more than 50 utterances produced by 63 speakers using 12th-order mel-frequency cepstral coefficients (MFCC), log energy and the delta of both as the acoustic features. Their system performed 85.3% correctly in pair-wise identification on average and 76.6% in the identification of the three accents. Similarly, Teixeira et al. (1996) identified six accents of English (Portuguese, Danish, German, British, Spanish, and Italian) using a HMM-based system. They used a speech corpus that contained 200 English isolated words, and calculated linear predictive coding (LPC) cepstra and their delta as the acoustic features. Their system obtained a 65.5% correct identification rate. An example of using prosodic cues was described by Itahashi and Tanaka (1993). They analysed a Japanese passage read by speakers of 14 regional dialects, and extracted 19 acoustic parameters related to the F0. A principal component analysis was performed on these 19 components and the results showed that the 14 dialects could be classified into six groups that approximately corresponded to the regions that the dialects belonged to. Finally, Piat et al. (2008) carried out a study that involved identification of four accents (French, Italian, Greek, and Spanish) of English. They compared the identification performance of their HMM-based system using 1-dimensional duration, 3-dimensional energy, 36-dimensional MFCC, and other prosodic features. The results showed that the MFCC yielded the highest identification rate of 82.9%, whereas duration and energy yielded rates of 67.1% and 68.6%, respectively. They thus concluded that MFCC provided a superior identification rate, although the computational cost was higher.

It is not easy to compare the identification results of the above studies, as they used different speech corpora and different comparison methods; however, these previous studies indicate that accent identification performance improves by using linguistic knowledge of the target languages effectively. This can be, for example, knowledge of linguistic forms for which non-native speakers saliently differ from native speakers, or knowledge of how to detect these linguistic forms in running speech. Blackburn et al. (1993) suggested a method for classifying non-native English accents using features related to phonological differences between the accents. They exploited knowledge on segmental differences among Arabic-accented, Mandarin-accented and Australian (native) English, and extracted features such as the phoneme duration of the sibilants, the voice onset time of the plosives, and the formant frequencies of the vowels. With their system, which was based on a neural network, 96% of Australian English, 35% of Arabic, and 62% of Mandarin male speech were correctly identified using voiced segments. Cleirigh and Vonwiller (1994) developed a phonological model of Australian English that included information on English syllable structure and the distribution of phonemes within a syllable. Berkling et al. (1998) applied this model for the identification of Vietnamese-accented and Lebanese-accented Australian English. They conducted two accent identification experiments, one using the linguistic model and the other not using the model. When they incorporated the linguistic model into their system, the performance improved by 6–7% (84% for the English–Lebanese pair and 93% for the English–Vietnamese pair) compared to the system not using the model (78% for the English–Lebanese pair and 86% for the English–Vietnamese pair). Zissman (1996) built a speech corpus for testing accent identification (using the term "dialect identification") systems for conversational Latin-American Spanish. He also built an accent identification system using HMM-based phoneme recognition. By applying N-gram language modeling, his system