Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 19–33

www.elsevier.com/locate/specom

Unsupervised accent classification for deep data fusion of accent and language information

John H.L. Hansen*, Gang Liu

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, Richardson, TX, USA

Received 26 November 2014; received in revised form 13 December 2015; accepted 14 December 2015
Available online 11 January 2016

Abstract

Automatic Dialect Identification (DID) has recently gained substantial interest in the speech processing community. Studies have shown that the variation in speech due to dialect is a factor which significantly impacts speech system performance. Dialects differ in various ways such as acoustic traits (phonetic realization of vowels and consonants, rhythmical characteristics, prosody) and content based word selection (grammar, vocabulary, phonetic distribution, lexical distribution, semantics). The traditional DID classifier is usually based on Gaussian Mixture Modeling (GMM), which is employed as baseline system. We investigate various methods of improving the DID based on acoustic and text language sub-systems to further boost the performance. For acoustic approach, we propose to use i-Vector system. For text language based dialect classification, a series of natural language processing (NLP) techniques are explored to address word selection and grammar factors, which cannot be modeled using an acoustic modeling system. These NLP techniques include: two traditional approaches, including N-Gram modeling and Latent Semantic Analysis (LSA), and a novel approach based on Term Frequency–Inverse Document Frequency (TF-IDF) and logistic regression classification. Due to the sparsity of training data, traditional text approaches do not offer superior performance. However, the proposed TF-IDF approach shows comparable performance to the i-Vector acoustic system, which when fused with the i-Vector system results in a final audio-text combined solution that is more discriminative. Compared with the GMM baseline system, the proposed audio-text DID system provides a relative improvement in dialect classification performance of +40.1% and +47.1% on the self-collected corpus (UT-Podcast) and NIST LRE-2009 data, respectively. The experiment results validate the feasibility of leveraging both acoustic and textual information in achieving improved DID performance.

Keywords: NLP; TF-IDF; Accent classification; Dialect identification; UT-Podcast.

1. Introduction

Automatic Dialect Identification (DID)/Classification has recently gained significant interest in the speech processing community (Hansen et al., 2004; Torres-Carrasquillo, 2004; Ma et al., 2006; Li et al., 2007; Biadsy et al., 2009; Hansen et al., 2010; Liu et al., 2011; Liu et al., 2012; Sangwan and Hansen, 2012; William et al., 2013; Zhang et al., 2014). For dialects of a language, using related material such as lexicons, audio, and text can help, but ground truth knowledge is critical, especially if there is a potential for code-switching[1] between dialects. The ability to leverage additional signal dependent information within the speech audio stream can help improve overall speech system performance (i.e., the use of "Environmental Sniffing" to characterize noise (Akbacak and Hansen, 2007); similarities between classes such as in-set/out-of-set recognition (Angkititrakul and Hansen, 2007); or content based text structure based on latent semantic analysis (Bellegarda, 2000)). In a related domain, DID is important for characterizing speaker traits (Arslan and Hansen, 1997)

* Corresponding author. Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA. Tel.: +1 972 883 2910; fax: +1 972 883 2710.

E-mail address: john.hansen@utdallas.edu (J.H.L. Hansen).

URL: http://crss.utdallas (J.H.L. Hansen)

[1] In linguistics, code-switching is the practice alternating between two or more languages, or language varieties, in the context of a single conversation. For more details, refer to (Muysken, 1995).

http://dx.doi.org/10.1016/j.specom.2015.12.004
0167-6393/

and can help improve speaker verification systems as well. In general, Dialect/Accent is one of the most important factors that influence automatic speech recognition (ASR) performance next to gender (Gupta and Mermelstein, 1982; Huang et al., 2001). Research has shown that traditional ASR systems are not robust to variations due to speaker dialect/accent (Huang et al., 2004). Therefore, the formulation of effective dialect classification for selection of dialect dependent acoustic models is one solution to improve ASR performance. Dialect knowledge could also be used in various components of an ASR system such as pronunciation modeling (Liu et al., 2000), lexicon adaptation (Ward et al., 2002), and acoustic model training (Humphries and Woodland, 1998) or adaptation (Diakoloukas et al., 1997). Dialect classification techniques were used for rich indexing of historical speech corpora as well as providing dialect information for spoken document retrieval systems (Gray and Hansen, 2005). Dialect knowledge could also be directly applied in automatic call center and directory lookup service (Zissman et al., 1996). Effective methods for accent modeling and detection have also been developed, which can contribute to improving speech systems (Angkititrakul and Hansen, 2006).

We note there are some subtle differences in the definition of accent versus dialect. To prevent this study from concentrating on too many details, accent and dialect are used interchangeably here. The term *dialect* is defined as: a pattern of pronunciation and/or vocabulary of a language used by the community of native speakers belonging to some geographical region (Lei and Hansen, 2011). Dialects can be further classified into family-tree and sub-tree dialects, all of which are part of the "language forest" (see Fig. 1). Family-tree dialects are the family sub-branches in the dialect tree, where their parent node is the actual language. As an example, for the English language it is possible to consider broad groups of American, Australian, and United Kingdom branches within the overall family tree. Beneath each main partition would be sub-classification (i.e., Belfast, Bradford, Cardiff, etc. for UK English). Moving upwards in the "language forest", it is possible to realize an English forest, a Spanish forest (this may include Cuban Spanish, Peruvian Spanish, and Puerto Rican Spanish, family-trees), etc. The general speech processing community does not have a well-defined definition of the language space relating to dialects, accents, and languages. In general, a more detailed focused level below family-tree dialects would be called "sub-tree dialects" which would reflect the sub-branches of the family-trees. For example in American English, there are many regional sub-dialects which are estimated to be 56 that include geographical regions: such as Boston/New England, New York City/New Jersey/Philadelphia, New Orleans, Texan, etc. There is much effort dedicated to investigating language identification at the higher language forest level. For example, National Institute of Standards and Technology (NIST) has conducted a number of automatic language recognition evaluations (LRE) since 1996. This has resulted in the introduction of successful algorithms, such as Parallel Phone Recognition and Language Modeling (PPRLM) (Zissman, 1996), Vector Space Modeling (VSM) (Li et al., 2007), and others. However, this research primarily focused at the language forest level with only limited or no attention paid to the lower level family-tree dialects level (NIST LRE, 2007, 2009, 2011, 2015), where the classification is usually more difficult than at the language forest level (for example, English vs. Spanish). In the NIST LREs, some closely related language pairs have been considered (i.e., Russian vs. Ukrainian, Urdu vs. Hindi), as well as dialects (i.e., Arabic Iraqi, Arabic Levantine, Arabic Maghrebi, and modern standard Arabic). Researchers have also explored the differences between automatic versus human assisted classification for speaker recognition (Hansen and Hasan, 2015), and language ID (Zissman and Berkling, 2001). Therefore, this study is positioned to focus at the family-tree dialect level to further research in this domain. Research advancements at this level would not only help boost overall performance of language identification, but shed new light on more subtle challenges stemming from the sub-tree dialect level.

In order to achieve good performance in English dialect classification, it is first necessary to understand how dialects differ. Fortunately, there are numerous studies on English dialectology (Purnell et al., 1999; Trudgill, 1999; Wells, 1982). English dialects differ in the following areas (Wells, 1982):

1. Phonetic realization of vowels and consonants
2. Phonotactic distribution (e.g., rhotic in *farm*: /farm/ vs. /fa:m/)
3. Phonemic system (the number or identity of phonemes used)
4. Lexical distribution
5. Rhythmical characteristics
6. Semantics

The first four areas/items are visible at the word level from both production and perception levels. From a linguistic point of view, a word may be the best unit to classify dialects. However, for an automatic classification system, it is impossible to build models for all words from different dialects. Therefore, many researchers focus on identifying pronunciation differences for dialect classification (Huang and Hansen, 2005; Huang and Hansen, 2006) to address items 1, 2 and 3. Huang and Hansen (2005) addressed dialect classification using word level based modeling, which was termed Word based Dialect Classification, converting the text independent decision problem into a text dependent problem, producing multiple combination decisions at the word level rather than making a single overall decision at the utterance level. Gray and Hansen (2005) considered temporal and spectral based features including the Stochastic Trajectory Model (STM), pitch structure, formant location and voice onset time (VOT) for dialect classification to address items 1 and 5. In order to make this process unsupervised, Huang and Hansen (2006) proposed the use of frame-based selection via Gaussian Mixture Models (GMM) for unsupervised dialect classification. One challenge is that most research studies are based on in-house data and more traditional acoustic modeling approaches. It is not until recently that some groups have begun to employ state-of-the-art technology (i.e., i-Vector) to perform the acoustic