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.

© 2015 Elsevier B.V. All rights reserved.

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-switching1 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).

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).
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. /fəm/)
3. Phonemic system (the number or identity of phonemes used)
4. Lexical distribution
5. Rhythmic 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
modeling based English, Finnish, Swiss French accent identification (Bahari et al., 2013, Behravan et al., 2013, Lazaridis et al., 2014). But the corpus used in those research are relative homogenous and small. This study is an extension on this direction by validating on two different large corpora.

Alternatively, limited emphasis has been placed on the lexical distribution for dialect classification. In particular, word selection and grammar are additional factors, which cannot be modeled directly using frame based GMM or i-Vector acoustic information. For example, word selection differences between UK and US dialects can be distinguished with examples such as - “lorry” vs. “truck”, “lift” vs. “elevator”, “dickie” vs. “shirt”, etc. Australian English has its own lexical terms such as Tucker (food), outback (wilderness), etc. (Layer, 1994). It is also common to observe the use of present perfect (UK) vs. simple past tense (US) which reflects differences at the grammar level - “I’ve already eaten.” vs. “I already ate.”. One additional factor in which dialects differ is in semantics. For example, momentarily which means for a moment duration (UK) vs. in a minute or any minute now (US). The sentence “This flight will be leaving momentarily” could represent different time periods therefore in US vs. UK dialects (Layer, 1994). In this study, some traditional natural language processing (NLP) techniques (such as N-Gram and LSA), as well as novel NLP techniques (TF-IDF logistic regression) are employed to address these two factors (highlighted as items 4 and 6 above).

It is also noted that the lack of extensive dialect corpora in the field is one reason for the slow research progress, so in this study we also wish to explore a potential solution for this challenge. As such, this paper is organized as follows. The next section begins with a description of dialect corpora and their statistics, followed by a baseline system description in Section 3. The proposed system includes three Sub-Systems described in Section 4. In Section 5, an experimental evaluation is reported and analyzed. The study is concluded in Section 6 with a summary and directions for future research.

2. Database description

It has previously been shown that it is more probable to observe semantic differences in spontaneous text and speech rather than formal written newspapers or prompted/read speeches (Antoine, 1996; Hansen, 2004; Hasegawa-Johnson and Levinson, 2006). Simply reading prepared text does not in general convey actual dialect content of a language (Huang and Hansen, 2007a; Liu et al., 2010b). There is no publicly available corpus for audio across dialects of common languages in the speech community, and in particular on containing spontaneous speech to address the problems discussed in Section 1. Therefore, the first challenge in this study is the collection of a database which is both spontaneous and realistic. We propose a cost-effective solution by using web based online podcasts of interviews where subjects speak spontaneously, and for which it is possible to determine the ground truth dialect in the family-tree from web content and/or metadata. Since this audio is cost free, and easy to find through web based podcasts from various parts of the world, this is an effective way of obtaining a diverse range of dialect specific data. Hence, a process was initiated to crawl podcasts of interviews from various websites. Although there are many English dialects, the focus here was on three broad family-tree dialects of English: American English (denoted as US), Australian English (denoted as AU), and United Kingdom English (denoted as UK). Since most subjects in these interviews are well-known, it is possible to identify the speaker’s dialect based on their demographic information. This corpus is named as UT-Podcast in this study for convenience (see Chitturi and Hansen, 2007; Chitturi and Hansen, 2008; Chitturi, 2008 for preliminary corpus discussion).
Text transcriptions are decoded from audio with the CMUspinhx ASR engine (Walker et al., 2004). The pre-trained US English HUB4 based language model and acoustic model are used in the ASR engine (Sourceforge, 2015). It is noted that the pre-trained ASR has a non-negligible word error rate, which means the ASR engine cannot transcribe all English audio files 100% correctly. Our focus here instead is to employ a reasonable ASR engine to extract the corresponding text as an input feature, which should allow for some text based noise for a practical deployment. Another potential question is why a specific English dialect (i.e., US English in this study) based ASR engine is chosen. It is because that this pre-trained ASR engine can be taken as a common audio-to-text transformation system. The resultant text will be used as feature for further processing. During the transformation, different dialects, due to the above-mentioned (Section 1) unique pronunciation characteristics, are expected to demonstrate different error during audio-to-text projection and could potentially contribute to dialect identification after proper modeling. Also, due to the high similarity among different English dialects, we expect many words can be correctly decoded no matter what specific English dialect based ASR engine is actually employed, such as “lorry” vs. “truck” can be safely taken as good indicator of UK English vs. American English. Although other English dialect ASR engine is possible, this will entail tremendous more effort for language modeling and acoustic modeling, which will severely reduce the popularity of this approach. The experiment results from text approach in latter section (Section 5.1.2) of this study also support this claim.

The acoustic and text statistics of the assembled podcast database are described in Section 2.1. To make this study more general, a second database based on the NIST LRE-2009 is introduced in Section 2.2.

### 2.1. Podcast dataset: UT-Podcast

In an effort to demonstrate actual performance for open test conditions, training and test data are collected from entirely different websites within each dialect region. To increase the representativeness of the data, we collect data from 12, 5, and 8 websites for AU, UK, and US English, respectively. One benefit from this approach is that the bias caused by channel/noise/recording can be dramatically mitigated in actual dialect classification performance. Also to make the data more realistic, we include podcasts which cover a wide range of topics, such as news report, scientific report, religion, social life. Compared with the sound booth collection, this approach produces more spontaneous and representative data.

We note that data collected from online podcasts is spontaneous and in general not well structured. Therefore, it is expected that some non-speech audio content such as music, silence, etc. will be presented in the audio streams. To address this, energy based voice activity detection (VAD) is employed (Sohn et al., 1999). Since the collected audio files range in duration from 30–60 min, after pre-processing of the audio as described above, the audio data is subsequently segmented into much smaller audio segment files for either training or test. With the help from previous VAD processing, the segmentation is performed in a manner so as to avoid abrupt cuts within the middle of a conversation, which follows the same approach as that used by LDC in the NIST LRE data preparation. In general, each segmented audio file is of duration 17 s and contains on average 46 words. Further details, such as file numbers, duration, word count, etc. are summarized in Table 1, where both acoustic and text statistics are provided.

It is noted that the corpus is released for public research usage under a free license. To give an idea concerning the structure of these podcasts, sample text extracts are provided here:

**Extract 1:** “He’s laying on the ground in front of the hotel next to my office. And there’s a truck parked right there. In the picture that I showed, there’s a truck, pickup truck, parked right there. He was stood up in front of that pickup truck and shot in the back five more times. It was shown on CNN”.

**Extract 2:** “Jackie: ‘Well, according to the report, alcohol is the main cause of antisocial behavior in Britain. … Let’s listen to Elena. Is she worried about anti-social behavior?’ Elena: ‘My name is Elena and I live in North West London...’”.

From these two examples, it is apparent that some keywords may carry more dialect-dependent information than others. For example, some words are more American English dialect dependent — “truck”, “apartment” — versus others which may be more likely to be UK dialect — “lorry”, “flat”. It should be noted that both word selection and pronunciation/prosody reflect dialect. Extract 1 and 2 were drawn from American English and UK English, respectively. Apparently, this salient information is not directly reflected in the acoustic modeling approach. However, a language model based approach that can weigh the dialect-sensitive word selection information properly can be more helpful in dialect identification.

### 2.2. LRE-2009 dataset

In addition to UT-Podcast, the NIST LRE-2009 dataset is also considered. While this is a rich, diverse and extensive corpus, one drawback is that it is only available at no cost to
those members who agreed to and participated in the NIST LREs. In spite of this limitation, we still wish to consider it in order to validate the proposed algorithms as a comparison with other sites. For the LRE-2009 dataset, we limit the language set to American English (noted as US) and Indian English (noted as IN) since the focus of this study is on dialect/accent identification. The training data for American English includes broadcast news programs from Voice of America (VOA), and conversation telephone speech (CTS). The training data for Indian English is purely based on conversation telephone speech. The American test data include both VOA and CTS audio, while the Indian test data only consist of CTS audio. VOA data is pre-processed using the data purification procedure proposed in (Liu et al., 2012). Audio files are segmented into blocks, which on average are 19 s in duration and contain 49 words. Further details such as file numbers, duration, word count, etc. are summarized in Table 2.

3. Baseline: GMM-based DID system

A number of past research studies have explored the use of spectral based features such as MFCC (Hansen et al., 2004; Huang and Hansen, 2006) for the purpose of dialect classification. As such, MFCC features are employed here. Before feature extraction, VAD must be performed to set aside silence and music in all audio files. Next, acoustic features are extracted with an analysis window size of 20 ms and a skip-rate of 10 ms, resulting in a frame rate of 100 frames per second. The entire feature set includes MFCCs and their shifted delta cepstrum (SDC) features (Bielefeld, 1994) to represent the audio sequence. The SDC features can incorporate additional temporal information into the feature vector, which is proved to be effective in utterance based identification such as did dialect and emotion identification (Liu et al., 2010a). The classical SDC parameter configuration N-d-P-k for language identification is 7-1-3-7, which combines with MFCC to produce an overall 56-dimensional feature vector on a frame-by-frame basis. To be specific, the windowed signal is filtered through Mel-scale filterbank over the band 300–4000 Hz, producing 26 log-filterbank energies. Then the log-filterbank energies are processed via DCT and producing seven cepstral coefficients (0-c6). The cepstral coefficients are further normalized using cepstral mean and variance normalization (CMVN). Then by following the 7-1-3-7 SDC configuration, 56-dimension feature vectors are derived.

Once we extract all the features, we will use Gaussian Mixture Modeling (GMM) to train the model and make the decision accordingly.

4. Proposed DID system

The proposed system for DID consists of three Sub-Systems – an acoustic Sub-System (i-Vector based DID), a language or text-based classification Sub-System (noted as Sub-System 2), and a score fusion Sub-System (noted as Sub-System 3). Sub-System 1 (acoustic system) and 2 (text system) is discussed in Sections 4.1 and 4.2, respectively. Sub-System 3 is simply an equal weight fusion solution. The overall proposed DID system is illustrated in Fig. 2.

4.1. Audio based dialect identification: Sub-System #1

The MFCC-SDC features derived in Section 3 are also referred to as raw features since they can be further processed and converted into more refined features such as i-Vectors. i-Vector, initially introduced for speaker recognition (Dehak et al., 2010; Liu et al., 2014a; Liu et al., 2014b; Yu et al., 2014a; Yu et al., 2014b; Zhang et al., 2015), is now widely adopted in the field of speech processing such as emotion identification (Xia and Liu, 2012; Liu and Hansen, 2014) and language identification (Martinez et al., 2011). However, it is seldom explored in dialect identification and therefore warrants further exploration here. i-Vector is extracted by following the factor analysis (Dehak et al., 2010):

\[ M = m + T_ω \]  

where \( T \) is referred to as the Total Variability matrix and \( ω \) is the resulting i-Vector, \( m \) is the universal background model (UBM) mean supervector, and \( M \) is the super-vector derived from raw features. The extraction converts frame length-varied spectral features matrix into dimension-fixed features vector for each speech utterance.

All available training data are employed to train both the UBM and total variability matrix with five iterations of Expectation–Maximization. Next, the i-Vectors for both training and test sets are extracted with the total variability matrix.

Once the i-Vectors are ready, we rely on a Gaussian Backend algorithm for dialect modeling. The MFCC-SDC + i-Vector + Gaussian Backend is the state-of-the-art system for language identification task (Torres-Carrasquillo et al., 2010; Dehak et al., 2011) and is adopted for DID task here. Contrary to the popular GMM approach (Shen and Reynolds, 2008; Huang and Hansen, 2006; Huang and Hansen, 2007b), the performance of i-Vector approach (Behravan et al., 2013) dramatically surpass GMM as well as other popular approaches (Choueiter et al., 2008, Larcher et al., 2011; Liu et al., 2011, Brümmer et al., 2011). So we also want to explore its performance on the English dialect scenario. The flowchart of the proposed i-Vector based DID system is illustrated in Fig. 3.

In this study, the mixture number for UBM is 32 and the dimension of iVector is 100, which is whitened and length normalized before using Gaussian Backend.
4.2. Text based dialect identification: Sub-System # 2

Next, we focus on Sub-System 2 (text system), which is illustrated in Fig. 4. As stated in Section 2, an ASR engine is used to derive the estimated text sequence from the input audio stream. Next, three text categorization approaches are explored, which are introduced in the following three subsections.

4.2.1. N-Gram based dialect identification: Text DID system # 1

It is assumed that a given text document associated with the corresponding audio which consists of many sentences. Each sentence can be regarded as a sequence of words W. The probability of generating the word sequence W can be measured as,

\[ P(W|D) = P(W_1, W_2, \ldots, W_n|D) = P(W_1|D)P(W_2|W_1, D) \ldots P(W_n|W_1, \ldots, W_{n-1}, D) = \prod_{i=1}^{n} P(W_i|W_{i-1}, \ldots, W_1, D) \]

where \( n \) is the number of words in \( W \), \( W_i \) is the \( i \)th word, and \( D \) is the dialect specific language model. For example, in UT-Podcast corpus, \( D \in \{AU, UK, US\} \) (corresponds to Australia English, United Kingdom English, and American English, respectively). The final equation comes from the N-Gram definition. The N-Gram probabilities are calculated via occurrence
Fig. 4. (a) Text based dialect identification system. Text is derived from ASR. Three text based sub-systems follow the similar workflow, but differ in implementation of individual modules. (b) TF-IDF logistic regression based Text DID system. See Section 4.2 for more details.

counting. The final result is then decided by

$$C = \arg \max_D \prod_{W \in \psi} P(W|D),$$

where $\psi$ is the set of sentences in a given document stemming from the corresponding input audio stream. The prior probabilities are dropped in this equation since different classes are assumed to have equal priors for simplification. In this study, we use the derivative measure of the cross entropy, known as the test set perplexity for dialect classification. If the word sequence is sufficiently long, the cross entropy of the word sequence $W$ is approximated as,

$$H(W|D) = -\frac{1}{n} \log_2 P(W|D),$$

where $n$ is the length of the test word sequence $W$, measured in words. The perplexity of the test word sequence $W$ as it relates to the language model D is,

$$PP(W|D) = 2^{H(W|D)} = P(W|D)^{-\frac{1}{n}}.$$  

The perplexity of the test word sequence is therefore the generalization capability of the given language model. The smaller the perplexity, the better the language model generalizes to the test word sequence. The final classification decision is then written as,

$$C = \arg \min_D \prod_{W \in \psi} P(W|D)^{-\frac{1}{n}}.$$  

By comparing Eqs. 3 and 6, it is noted that the perplexity measure is actually the normalized probability measure, where the normalization factor is simply the word sentence length $n$.

4.2.2. Latent semantic analysis based dialect classification: text DID System # 2

One approach used to address topic classification problems has been latent semantic analysis (LSA), which was first explored for document indexing in (Deerwester, 1990). That work focused on addressing the issues of synonymy - many ways to refer to the same idea, and polysemy – words having more than one distinct meaning. These two issues present problems for dialect classification because two conversations about the same topic need not contain the same words and conversely two conversations about different topics may in fact contain the same words but with different intended meanings. There are other semantic differences as previously described in Section 1, which can be explored in the classification of dialect.

In order to find an alternate feature space which avoids these problems, singular value decomposition (SVD) is performed to derive an orthogonal vector representation of the documents. SVD uses eigen-analysis to derive linearly independent directions of the original term via a document matrix $A$ whose columns correspond to the number of dialects, while the rows correspond to the words/terms in the entire text database. SVD decomposes this original term document matrix $A$ into three other matrices: $A = USV^T$, where the columns of $U$ are the eigenvectors of $AA^T$ (called the left eigenvectors), $S$ is a diagonal matrix whose diagonal elements are the singular values of $A$, and the columns of $V$ are the eigenvectors of $A^TA$ (called the right eigenvectors). An example is shown in Figs. 5 and 6 to illustrate this concept for text/language information SVD decomposition.
In any language, there are many words like: “a”, “the”, “and” and many other high frequency words that convey little or no meaning, but are indispensable as a feature for our dialect text categorization tasks. There are two empirical intuitions in using TF-IDF: (a) the more frequent a term (say $t_i$) occurs in a document $d_j$, the more important it is for the term vector coordinates in this lower dimensional space are the rows of $V$. The coordinates (noted as $d_j$) of the projection of query file $q$ in the lower dimensional space are given by $q_1 = q^T \cdot U \cdot S^{-1}$. The test utterance is then classified as a particular dialect based on the scores, given by the cosine similarity measure (also called as Cosine Distance Scoring) as shown in Eq. 7,

$$d_{best} = \arg \max_{d_j} \frac{(q_1 \cdot d_j)}{|q_1||d_j|},$$

where $d_j$ is one of dialect classes.

4.2.3. TF-IDF – logistic regression based dialect classification: Text DID system # 3

Both N-Gram and LSA are traditional popular NLP approaches, which directly work on text sequence data to find potential semantic cues to achieve dialect classification. Another approach is using a keyword spotting approach for dialect classification. The underlying assumption behind this is that each dialect should demonstrate some unique word set, which could help distinguish. The text sequences need to be pre-processed before performing feature extraction. For feature extraction, a weighting factor algorithm called Term Frequency–Inverse Document Frequency (TF-IDF) is adopted (Jones, 1972). Next, logistic regression is employed to perform dialect modeling and classification.

4.2.3.1. Text pre-processing. In any language, there are many words that convey little or no meaning, but are indispensable for the grammar structure of the language; these words are defined as “stop words”. For example, in the English language words like: “a”, “the”, “and” and many other high frequency words are considered as stop words (Greenberg, 1997). It is common practice to exclude stop words from the feature vector since they are not expected to be class/dialect sensitive.

A stop word list is employed for this procedure. Another text pre-processing procedure is called stemming, which means to derive a morphological root for the words. For example, if an English document would contain 5 instances of the word “play”, 2 instances of the word “plays” and 1 instance of the word “playing”, then a stemming reduction for these three separate features would reduce them to one word that describes the set of words with the root similar to “play” occurred in the document 8 ($=5 + 2 + 1$) times.

After applying these text processing procedures, we can perform an N-Gram operation (we need to decide how many word grams should be united, which is stated in Section 4.2.1. It effectively groups the text information into some basic unit. Bigger N can store more context. If we have sufficient training data, we can set N to a high number in the N-Gram operation. However, if the data is sparse, a large N can render the resulting model ineffective due to the sparsity of specific word sequence occurrences. Due to the data scarcity of the data in this study, we only investigate two scenarios: $N = 1$ and $N = 2$, which is also called unigram and bigram, respectively.

The collection of these pre-processing procedures are used in a sequential manner, with combinations listed in Table 3.

4.2.3.2. Feature generation. Once the raw text from ASR is pre-processed in the pre-processing phase, there are several alternative feature extraction approaches, which are detailed in Table 4.

For Feature Method 0 and 1, their implementation is self-evident. Further explanation is needed for Method 2, that is, the TF-IDF approach. TF-IDF is a numerical statistic which reflects how important a term/word is to a document in a dataset (Salton and Buckley, 1988). It is often used as a weighting factor for information retrieval and text mining (Hansen et al., 2005; Kim and Hansen, 2007). We use it here as a feature for our dialect text categorization tasks. There are two empirical intuitions in using TF-IDF: (a) the more frequently a term (say $t_i$) occurs in a document $j$ (denoted as $d_j$), the more important it is for $d_j$ (the term frequency intuition);
Table 4
Feature representation variations.

<table>
<thead>
<tr>
<th>Feature Method</th>
<th>Feature representation approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Binary feature: use 0 or 1 to signify whether some word/item is present or not.</td>
</tr>
<tr>
<td>1</td>
<td>Term frequency: how often a word occurred in a document.</td>
</tr>
<tr>
<td>2</td>
<td>TF-IDF: Term Frequency–Inverse Document Frequency, reflecting how important a word is to a document in a collection.</td>
</tr>
</tbody>
</table>

(b) the more documents $t_i$ occurs in, the less discriminative it is (i.e., the smaller its contribution is in characterizing the semantics of a document in which it occurs) (the inverse document frequency intuition). In another words, the TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word across the entire corpus, which reflects its discriminative ability and helps control for the fact that some words are generally more common than others. The weighting factor computed by TF-IDF techniques is often normalized so as to adjust for the tendency of TF-IDF to emphasize longer duration documents. To be more specific, TF-IDF can be calculated as:

$$TFIDF(t_i, d_j) = TF(t_i, d_j) \cdot IDF(t_i, D)$$ (8)

The first term in the right hand side of Eq. 8 is defined as:

$$TF(t_i, d_j) = \frac{n_{i,j}}{\sum_h n_{k,j}} = \frac{Number \ of \ times \ term \ t_i \ appears \ in \ document \ d_j}{Total \ number \ of \ terms \ in \ document \ d_j},$$ (9)

where the numerator, $n_{i,j}$, is the number of occurrences of the specific term $t_i$ in document $d_j$, and the denominator, the total number of terms in document $d_j$, is used for normalization. Note, to simplify the notation, the subscript, $i$ and $j$, are dropped in Eq. 8. The second term in the right hand side of Eq. 8 is defined as:

$$ IDF(t_i, D) = \log \frac{|D|}{1 + ||d : t_i \in d||},$$ (10)

where $|D|$ is the total number of documents in the Dataset $D$, and $||d : t_i \in d||$ is the number of documents where term $t_i$ actually appears. To avoid any potential divide-by-zero, $1 + ||d : t_i \in d||$ is used as a minimum floor value instead.

This feature extraction procedure essentially assigns some weighting to the terms derived from the pre-processing step (see Section 4.2.3.1 ). Having addressed the text pre-processing and feature generation steps of Fig. 4(b), we turn to model development.

4.2.3.3. Text Categorization Modeling. Logistic regression (LR) is a type of discriminative probabilistic classification model that operates over real-valued vector inputs. It has been explored in large text classification tasks (Genkin et al., 2007), but limited studies have investigated its potential application in dialect text classification. Therefore, LR is investigated here.

Given a set of instance-label pairs $\{x_i, y_i\}_{i=1}^l$, where $x_i \in \mathbb{R}^N$, and $y_i \in \{-1, 1\}$, LR binary classification assumes the following probability model:

$$P(y = \pm 1|x, w) = \frac{1}{1 + \exp(-y(w^T x + b))}.\quad (11)$$

To consider a simpler derivation without the bias term $b$, one often augments each instance with an additional dimension as follows:

$$x_i^T \leftarrow [x_i^T, 1], \quad w^T \leftarrow [w^T, b].\quad (12)$$

With this, Eq. 11 can be updated as follows:

$$P(y = \pm 1|x, w) = \frac{1}{1 + \exp(-y w^T x)}.$$ (13)

The only LR model parameter, $w$, can now be estimated by maximizing the probability as follows:

$$\max \prod_{i=1}^l \frac{1}{1 + \exp(-y_i w^T x_i)},$$ (14)

or by minimizing the negative log-likelihood as follows:

$$\min_w \sum_{i=1}^l \log(1 + \exp(-y_i w^T x_i)).$$ (15)

Moreover, to avoid over-fitting, one adds a regularization term $w^T w/2$; thus, in this study, we consider the following unconstrained optimization problem as regularized logistic regression:

$$\min_w \left\{ \frac{1}{2} w^T w + C \sum_{i=1}^l \log(1 + \exp(-y_i w^T x_i)) \right\},$$ (16)

where $C > 0$ is a penalty parameter used to balance the two terms in Eq. 16. The trust region Newton method is employed to learn the LR model (i.e., $w$, by solving Eq. 16), which is also referred to as the primal form of the logistic regression.

5. Experiments

As discussed in Section 2 and shown in Fig. 2, the input to the DID system is audio data only. As such, we rely on an ASR engine to derive the necessary parallel text information, which is then used for text modeling and classification. Next as Fig. 2 shows we leverage both audio and text modeling to boost the final dialect classification system performance. The proposed algorithms are evaluated on the two corpora (e.g., UT-Podcast and NIST LRE-2009) described in Section 2 to fully assess the validity of the proposed audio-text solution. To measure performance, two metrics are used in this study. The first one is unweighted average recall (UAR)(Liu et al., 2014). In the binary class case (‘X’ and ‘NX’), it is defined as:

$$UAR = \frac{Recall(X) + Recall(NX)}{2}$$ (17)
Our study relies on unweighted average recall rather than weighted average (WA) recall or conventional classification accuracy (Bahari et al., 2013) since it is also meaningful for highly unbalanced data. Eq. 17 can be easily generalized to multi-class case. The advantage of the metric of UAR is that it is application independent. To assist the potential comparison, another popular language recognition evaluation metric, average detection cost, or Cavg, is also adopted here (Li, 2013), another popular language recognition evaluation metric, it is application independent. To assist the potential comparison, another popular language recognition evaluation metric, average detection cost, or Cavg, is also adopted here (Li, et al., 2013). It is application dependent and follows the definition of LRE2009$^3$.

5.1. UT-Podcast Corpus

5.1.1. Audio system

Both GMM baseline system (Section 3) and i-Vector system is purely based on audio (see Fig. 2), and therefore named as “Audio System” to distinguish from “Text System”. The 3-way classification confusion matrix of GMM and i-Vector are summarized in Table 5. As can be seen, the GMM baseline system offers decent performance in the dialect of AU and US but faces huge challenge in modeling the dialect of UK due to the sparsity of training data for UK English (see Table 1). However, i-Vector can successfully model all the three dialects and offers an absolute +14.2% improvement over the GMM baseline system in terms of UAR.

5.1.2. Text system

Next, we consider the performance of various text processing based DID solutions.

5.1.2.1. N-Gram based DID system

For the N-Gram text categorization system, one important parameter is “N”, which represents the number of grams. As we can see from Table 6, a good choice for N is 2. However, due to the limited text content duration (i.e., 1, 2, or 3 word sequences), overall classification performance is weak. In reality, the performance of this N-Gram text processing is very close to random guessing.

5.1.2.2. LSA based DID system

As another popular text categorization approach, LSA Cosine Distance Scoring approach is also evaluated. The corresponding results from UT-Podcast data is detailed in Table 7. Compared with the N-Gram approach, LSA offers much better results, but is still far below the i-Vector audio system as a stand-alone classifier. Once again, this is because the sparse training data cannot support an effective combined feature representation and modeling solution for the text sequence.

5.1.2.3. TF-IDF based DID system

For the TF-IDF approach, since the overall feature extraction includes pre-processing and various feature representation approach, we would like to first investigate optimal feature configuration. Performance of the pre-processing step will depend on whether employing stopword removal, stemming, unigram or not, with these various text pre-processing options, there are a total of 8 ($2^3$) combinations which are detailed in Table 3 (noted as Step 1). Step 2 (i.e., feature representation) has four variations as listed previously from Table 4 (Section 4.2.3.2). To find the optimal feature configuration, we consider all combinations presented by the above two steps and summarize the corresponding results in Table 8. Based on Tables 3, 4, and 8, we can observe that the optimal text feature configuration is: Pre-processing = 0; Feature representation = 2. That is, a unigram model without stopword removal and stemming should be the combination employed as the final pre-processing operation. Next, the pre-processed text files are used to extract TF-IDF weighting information, which is used as the final feature. Usually pre-processing techniques, such as stopword removal, stemming and high order N-Grams, should present more discriminative information. However both training and test text data are very limited in this study, so to keep as much information as possible in this case is more beneficial. Therefore, we will use this optimal configuration for the remainder of this paper.

---

30s
10s
3s
Mixed

Cavg (%)
Audio System
Text System
Fusion System (Audio + Text)

Fig. 7. Dialect identification system performance comparisons across three testing duration scenarios of English dialects from NIST LRE-2009 corpus. Here the text system is the TF-IDF logistic regression system, and audio system is the i-Vector solution. The duration including 4 scenarios: 30 s, 10 s, 3 s and ‘Mixed’ (i.e., the mixture of the 3 duration scenarios). The metric is Cavg.

Table 8
Optimization of pre-processing and feature representation for text-based dialect identification. 8 pre-processing options from Tables 3, and 4 feature representations from Table 4 are considered. Optimal configuration is: Pre-processing = 0; Feature representation = 2.

Cavg (%) Feature representation

Pre-processing 0 1 2
0 34.7 33.1 31.0
1 34.2 33.5 33.2
2 33.5 31.4 32.0
3 34.4 33.2 32.2
4 35.6 35.2 34.6
5 36.0 36.6 35.3
6 36.2 37.1 37.5
7 36.2 37.1 36.3

Table 9
Performance comparison of different dialect identification approaches on UT-Podcast. AU, UK and US correspond to Australian English, United Kingdom English, and American English, respectively. ‘UAR’ is the unweighted average recall.

System # Classification system Recall (%) UAR (%) Cavg (%)

<table>
<thead>
<tr>
<th>System #</th>
<th>Classification system</th>
<th>Recall (%)</th>
<th>UAR (%)</th>
<th>Cavg (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Audio System (Baseline: GMM)</td>
<td>85.5</td>
<td>32.6</td>
<td>62.9</td>
</tr>
<tr>
<td>1</td>
<td>Audio System (i-Vector)</td>
<td>78.0</td>
<td>61.8</td>
<td>83.8</td>
</tr>
<tr>
<td>2</td>
<td>Text System (N-Gram Perplexity)</td>
<td>22.6</td>
<td>53.9</td>
<td>30.4</td>
</tr>
<tr>
<td>3</td>
<td>Text System (LSA Cosine Distance Scoring)</td>
<td>55.4</td>
<td>48.3</td>
<td>51.7</td>
</tr>
<tr>
<td>4</td>
<td>Text System (TF-IDF logistic regression)</td>
<td>83.1</td>
<td>32.6</td>
<td>60.4</td>
</tr>
<tr>
<td>5</td>
<td>Proposed Audio-Text system (Fusion of System 1 and 4)</td>
<td>86.1</td>
<td>60.7</td>
<td>82.1</td>
</tr>
</tbody>
</table>
Table 10
Comparison of different dialect identification approaches on the NIST LRE-2009 corpus. US and IN correspond to American English and Indian English, respectively. ‘UAR’ is the unweighted average recall.

<table>
<thead>
<tr>
<th>System #</th>
<th>Classification system</th>
<th>Recall (%)</th>
<th>UAR (%)</th>
<th>Cavg (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>US</td>
<td>IN</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>Audio System (Baseline: GMM)</td>
<td>32.4</td>
<td>94.8</td>
<td>63.6</td>
</tr>
<tr>
<td>1</td>
<td>Audio System (i-Vector)</td>
<td>60.0</td>
<td>94.4</td>
<td>77.2</td>
</tr>
<tr>
<td>2</td>
<td>Text System (TF-IDF logistic regression)</td>
<td>73.8</td>
<td>61.3</td>
<td>67.6</td>
</tr>
<tr>
<td>3</td>
<td>Proposed Audio-Text system (Fusion of Systems 1 and 2)</td>
<td>64.9</td>
<td>91.6</td>
<td>78.3</td>
</tr>
</tbody>
</table>

i-Vector acoustic system map an utterance with different phoneme distribution into a lower dimensional space, it ignore the time/context relationship among the phonemes and words of the utterance. This also suggests that i-Vector system is more appropriate to capture global image of the utterance in the dialect space, but may have difficulty to mode the local details which can help identification. TF-IDF text system focuses on the words level and can capture the semantic details to assist modeling. Due to the high similarity among dialects, many words may be the same and the resulted system may not be very effective. However, if the salient dialect cue do appear in some utterance, the text approach can capture it and do an efficient classification. These two different approaches can complement each other and explain away the gain in the fusion. The proposed audio-text system offers a relative +40.1% improvement against the GMM baseline system in terms of Cavg.

It is noted that the text based approach is computationally very efficient on UT-Podcast (similar results are also observed in LRE-2009 corpus). On our hardware, Intel(R) Xeon(R) CPU X5650, the text approach only takes 2 s, while i-Vector audio approach requires 922 s.

5.2. NIST LRE-2009 corpus

To further verify the proposed DID approach, we also want to validate its performance on the English dialect portion of the NIST LRE-2009 corpus. The results are summarized in Table 10. First of all, we noticed that the GMM baseline system cannot well model US English and the overall performance is the worst. i-Vector system can provide far more balanced performance in each dialect and is specially good at modeling Indian English. Compared with i-Vector system, the TF-IDF logistic regression based text DID system, demonstrate different decision tendency: favor more US English. This is caused by the accent mismatch between the US English based Engine and Indian English accent and the difference in acoustic space is automatically translated into textual space. This can compensate the best individual system (i-Vector system) and their fusion offer a relative +47.1% improvement over the GMM baseline system in terms of Cavg.

By comparing with results on UT-Podcast data in Section 5.1 (Table 9), we can observe that i-Vector based audio system offers best individual system performance consistently. Due to the heterogeneous nature of Text system, the fusion of it with i-Vector audio system can consistently boost the performance at trivial computation cost.

Since the duration of the test data of LRE-2009 has three variations: 30 s, 10 s, and 3 s. It is also of interest to more closely look at performance in the three test duration scenarios, where results are illustrated in Fig. 7. From Fig. 7 we observe duration affects the performance in a significant way. In the scenario of 30 s, the text DID approach has comparable performance with audio as an individual system. However, in shorter durations (i.e., 10 s and 3 s), the performance/gain rapidly decreases. Overall, however, the text approach can consistently help boost DID performance in all the three scenarios as seen when audio and text based solutions are fused. These results are also highlighted in Table 10.

6. Conclusions

This study has presented an integrated audio and text based solution for dialect identification (DID). Various progressing options were considered for both audio and text sub-systems. The proposed solution provides a cost-effective method to collect spontaneous and realistic dialect data, which was harvested by downloading from Podcast websites. A range of important factors were addressed to determine how dialects differ across a range of speech/language characteristics. The solution focused on investigating this based on two main approaches: acoustic and text based dialect classification. The acoustic classification system was a state-of-the-art i-Vector system, based on MFCC features and a Gaussian Backend which was shown to consistently outperform the GMM baseline system. Next, a text modeling approach was systematically explored. Two traditional text categorization approaches (i.e., N-Gram perplexity and LSA cosine similarity) were adopted but failed to provide sufficient individual performance gains as stand-alone solutions. Motivated by these results, a new text based DID approach was investigated using TF-IDF logistic regression, which not only individually demonstrated comparable performance to the audio system (i-Vector system), but also demonstrated a degree of complementarity performance, making the final integrated audio-text system more discriminative. The proposed system was shown to improve the Cavg performance by +40.1% and +47.1% relatively on the UT-Podcast and LRE-2009 corpora, respectively. By comparing with the best individual system (i-Vector system), the audio-text system boosts performance by +6.8% and +4.4%
on the UT-Podcast and LRE-2009 corpora, respectively. This benefit was achieved at a far less computational cost than the best individual system (more than 400 times faster than the i-Vector system). This observation is meaningful since it suggests that it is possible to leverage broader internet text data at a trivial cost to further boost overall DID performance. It was also observed that comparable performance can be offered by TF-IDF text approach when the duration of test file is 30 s. For shorter test duration scenarios, the TF-IDF approach is less competitive but can still offer consistent complementary gains. All results demonstrate that a multi-level DID approach which leverages information sources including audio and text is beneficial. The effective performance gains from the innovative combination of audio and text information suggest that deep data mining of the available data can help boost performance on machine learning systems.

The algorithms developed in this study have led to advancements in the area of dialect identification. While these algorithms have resulted in effective performance for English dialects, it is not claimed that the resulting solution is optimal or the final contribution in dialect classification. Other strategies could be considered in order to improve overall dialect classification rates, such as deep neural networks (Montavon 2009; Liu, 2010; Glorot et al., 2011; Ngiam et al., 2011; Hinton et al., 2012) could also be applied to both audio and text at the feature learning time. It is suggested that these be considered in future studies, in the context of the algorithms developed in the current study.

Finally, in an effort to contribute to the research community in dialect recognition, the collected and organized UT-Dallas-DID corpus will be freely shared with those interested, given that users complete a no-cost license agreement which prevents commercial/corpus distribution.

Acknowledgment

This project was supported by AFRL through contract FA8750-15-1-0205, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen. The authors wish to acknowledge that preliminary work on leveraging audio and text conducted by Rahul Chitturi as part of his MSCS Thesis at CRSS-UTDallas, and published in (Chitturi and Hansen, 2007; Chitturi and Hansen, 2008; Chitturi, 2008). The work to collect and organize the UT-Podcast corpus stemmed from these earlier studies.

References


