Language and Dialect Identification: A Survey

A. Etman
DSPRL, ECE Department
Virginia Tech
Blacksburg, USA
asma@vt.edu

A. A. (Louis) Beex
DSPRL, ECE Department
Virginia Tech
Blacksburg, USA
beex@vt.edu

***Abstract*—Automatic Dialect Identification has attracted researchers in the field of speech signal processing. Dialect can be defined as the language characteristics of a specific community. As such, dialect can be recognized by a speaker's phonemes, pronunciation, and traits such as tonality, loudness, and nasality. Being able to detect dialect accurately helps improve certain applications and services, such as Automatic Speech Recognition, remote access, e-health, e-learning, etc. Hence, working on dialect identification is more challenging than just recognizing a specific language or speech. For that reason, Dialect Identification became an important and attractive research topic. In this paper, we survey work done in dialect identification, i.e., what has been achieved in this area, and then provide an idea of what can be expected within the next few years.**

***Keywords*—ADI; SRS; phonotactic modeling; acoustic modeling; PRLM; GMM; HMM**

I. INTRODUCTION

Most references define dialect as the accent. However, the accent refers to the speaker's pronunciation, while dialect is the speaker's grammatical, lexical, and phonological variation in pronunciation [1]. For our work, we will consider dialect as the pronunciation pattern or the language vocabulary used by a specific community of native speakers [2], those who are usually based in a certain geographical region. Dialect represents an important characteristic of a speaker's voice signature, as it can provide information about the speaker's origin, gender, age, and health status.

Automatic Dialect Identification (ADI) has attracted both academia and industry for its promising positive impact on society. Robust ADI is expected to improve Speech Recognition Systems (SRS), which exist in most of today's electronic devices; ADI is also expected to enhance human computer interaction applications and secure remote access communication. In addition, ADI will help in providing new services for e-health and telemedicine, especially important for older and/or homebound people. Dialect Identification is assumed to be challenging due to its sensitivity to language changes, and regional limitations [3].

The approaches to dialect identification are similar to those used in language identification. These approaches can be classified into two modeling classes: acoustic and phonotactic. Acoustic modeling - as the word implies - deals with spectral feature modeling, while the phonotactic approach deals with speech via phone recognition, language models, and their subsequent scoring. The following subsections highlight both of these approaches.

VT-MENA and National Telecommunication Institute in Egypt

The rest of this paper is organized as follows: Section II focuses on dialect identification modeling schemes, while Section III covers most of the existing databases in the area of speech recognition. Section IV discusses work done on dialect identification from a linguistic and methodology point of view, and finally Section V gives a brief summary and potential topics to consider for future work.

II. MODELING SCHEMES

A. Phonotactic Modeling

In any language, words consist of different sets of phonemes [4-6]. A phone recognizer tokenizes speech into phonemes, which represent the phonetic transcription of any unknown word. In other words, the phonetic transcription is a description of a word as a sequence of known phonemes. Researchers always look for the phone recognizer with the highest accuracy. One of the popular approaches is the Phone Recognition followed by Language Modeling (PRLM)). In the latter approach, a single phone recognizer is employed to recognize the target dialect. The phone recognizer produces phone sequences that will be used to train the n-gram LMs for each dialect. In the recognition process, each utterance will be tokenized using the phone recognizer and the dialect indexed with the LM that yields the highest score is the hypothesized dialect [4]. The second approach is Parallel Phone Recognition followed by Language Modeling (PPRLM), which can be viewed as an extended PRLM. PPRLM uses multiple front-end phone recognizers instead of a single phone recognizer. The given utterances are fed to a bank of phone recognizers that are trained on the target dialects. The outputs from the phone recognizers will be scored using a bank of Language Models (LMs). Both PRLM and PPRLM are recommended when transcribed data is limited. On the other hand, when plenty of transcribed data is available a separate phone recognizer can be used with its own LM [5]. During recognition, a parallel set of phone recognizers is used, and since each recognizer has its own LM the resulting phone sequence will be optimal; this describes the Parallel Phone Recognition (PPR) method. A drawback of the PPR method is the need to have transcribed data for all of the target dialects. The three approaches above, together with the Gaussian Mixture Model (GMM) approach, have been compared for automated language identification [4, 5].

B. Acoustic Modeling

An alternative approach to Phonotactic Modeling is the Gaussian Mixture Model (GMM). The GMM has proven successful for usage in dialect recognition applications, and is

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

in fact at the core of most acoustic modeling approaches. Using the GMM approach one extracts the discriminative features from the acoustic data. Features derived from short-term spectra, prosodic information, such as the fundamental frequency and its time trajectory, loudness on particular parts of the speech and its time evolution, and intonation, can all be taken into account. Gaussian Mixture Models are part of state of the art systems [6-10] that capture acoustic events in the language and different acoustic events across recognized languages.

For example, thousands of Gaussian components can be trained for each language on features with proven discriminability between languages, usually the same features as used in speech recognition. A test utterance is declared to be from the language represented by the GMM with the highest likelihood [5]. Today researchers are trying to add new features and methods that improve robustness and accuracy of the identification system. Some of the other techniques that were employed to model acoustic information for Dialect Identification (DI) were Support Vector Machines (SVM) and Neural Networks (NNs) [6].

III. Databases

A number of databases is publicly available. Some of these were used by researchers in the field of speech, speaker, language, and dialect identification. This section gives a summary description of the most popular ones [11, 14]. Figure 1 summarizes the database taxonomy.

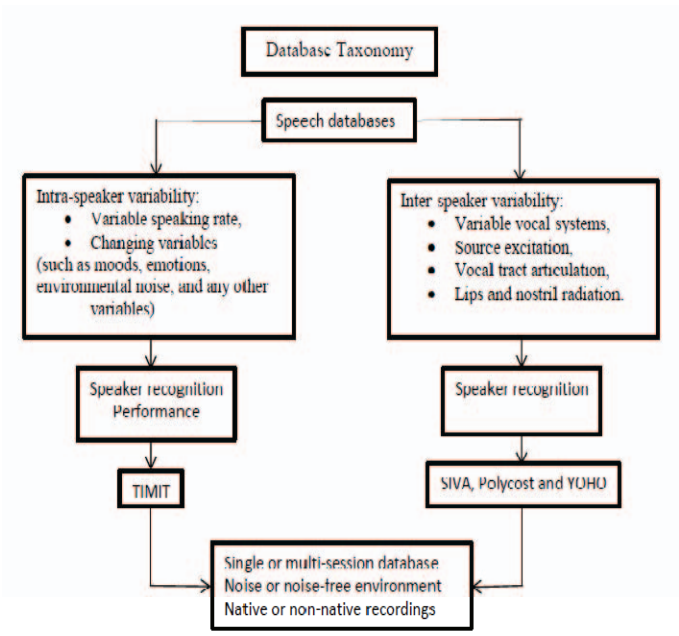

Fig. 1. Database taxonomy

Databases can be divided into two main classes: intra-speaker variability and inter-speaker variability [15]. The former deals with out-of-tract variations - such as speaking rate, emotions, and environmental noise, while the latter deals with in-tract variations - such as variable vocal systems, source excitations, and vocal tract articulation. The intra-speaker variability measurements are generally used when the concern is speaker recognition performance, whereas the inter-speaker variability measurements are mainly used when the concern is identifying speakers. The TIMIT database is an example of an intra-speaker database, and SOHO, Polycost, and YOHO are examples of inter-speaker databases. Both types can be recorded in a single session or more, with or without noise, and include native and non-native speakers.

In Table I, characteristics of some of the common databases are listed.

IV. Dialect Identification in the Literature

A. From a linguistic point of view

Dialect is the most complex aspect of speech recognition, as it pertains to the language characteristics of a specific regional community. Research in the field of dialect linguistics is still limited due to the unavailability of databases and the time consuming analysis process [16]. As technology has advanced a robust speech recognizer - one which is able to handle unstable conditions, such as noise and accent variation - has become an urgent need.

Dialects are not static; they change with place and time. In other words, as dialects change with geographic boundaries they may also vary due to language changes over time. Today, we see that new generations use phrases and vocabulary that were not in existence or use in the past. Therefore, languages change over time and their dialects change as well. Also, any one language includes several dialects. For example the United States has one dialect boundary that exists between the North and the South along what is called the Mason-Dixon Line. However, American English includes more than one dialect boundary. Different dialects exist in the fifty states and each state may encompass more than one dialect.

1) DI Early Studies

Studying dialects started early in 1877, when George Wenker was conducting a series of surveys to identify dialect regions [1, 17]. Then Bailey [18] carried out one of the first attempts to define the Midland dialect and whether it actually exists or not.

The conclusion of this study was that identifying dialects should not depend on vocabulary, as this may vary according to community or class within the same geographic region [18]. Following the same route, Davis and Houck [19] were also trying to find out whether the Midland region can be treated as a separate region of dialect or not. Their study was successful in extracting the phonological and lexical features among 11 cities that lie on a north-south line [19]. The conclusion was that the Midland region cannot be considered to be a transition region and that there is a linear relationship between the distant South and Southern dialects [17]. In opposition to the latter hypothesis, Johnson showed that combining phonological and lexical features was wrong, as doing so affects the data patterns negatively and thus yields incorrect results [20]. Using some words, proof was produced that there is a clear difference between both the North and Midland dialects and the South and Midland dialects, but not between the dialects of the North and the South.

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

TABLE I. CHARACTERISTICS OF DATABASES

Database	*Participants*	*Data type*	*Recording environment*	*Speech style*	*Sampling rate*	*Transcription based on*
TI DIGITS	*326*	*>2500 numbers*	*quiet room*	*reading*	*20*	*word*
TIMIT	*630*	*6300 sentences*	*quiet room*	*reading*	*16*	*phones*
NTIMIT	*630*	*6300 sentences*	*telephone*	*reading*	*8*	*phones*
RM1	*144*	*15024 sentences*	*quiet room*	*reading*	*20*	*sentence*
RM2	*4*	*10608 sentences*	*quiet room*	*reading*	*20*	*sentence*
ATIS0	*36*	*10722 utterances*	*office*	*reading spontaneous*	*16*	*sentence*
SWITCHBOARD (credit card)	*69*	*35 dialogues*	*telephone*	*conversation spontaneous*	*8*	*word*
TI46	*16*	*19136 words*	*quiet room*	*reading*	*12.5*	*sentence*
SWITCHBOARD	*543*	*2400 dialogues*	*telephone*	*conversation spontaneous*	*8*	*word*
ATC	*100*	*30000 dialogues*	*radio frequency*	*spontaneous*	*8*	*sentence*
ATIS2	*351*	*12000 utterances*	*office*	*spontaneous*	*16*	*sentence*
MACROPHONE	*5000*	*200000 utterances*	*telephone*	*conversation*	*8*	*sentence*
YOHO	*138 speakers (106 M - 32 F)*	*24-phrases 4-sessions*	*quiet room*	*reading*	*8*	*sentence*
KING-92	*51 males*	*510 files*	*telephone*	*conversation spontaneous, or taking tasks.*	*8*	*sentence*
TRAINS	*34 speakers*	*98 dialogues*	*collected recordings*	conversational		*word*
BRAMSHILL	police officers	*dialogues*	*office*	conversational	10	*sentence*
NYNEX Phonebook	1358	93,667 utterances	telephone	reading	8	*word*

The conclusion was that the Midland region is a separate region, one that differs from both North and South. These were the first steps towards identifying, and classifying, dialects. After defining dialects, the second stage was to determine how much dialects correlate to each other. What are the common features that exist between dialects of the same language, what are the differences, and how can they be measured?

2) DI using Vowels

One of the first attempts was that of Peterson and Barney [21], who worked on the vowel spacing characteristic. Through their study they found that vowels that are perceptually different occupy different regions in formant space; also the same vowel pronounced by different people appears in different positions in formant space. Finally, the conclusion drawn was that participants' ability to produce and perceive a certain vowel is affected by their background. This work was significant in the area of recognition and was the first to introduce the importance of dialects [21]. The one drawback of this work was heterogeneity of participants, which made it difficult to get accurate specification of vowel spacing changes [17]. For this reason, Hillenbrand, Clark and Wheeler [22] repeated the same experiments in order to be able to obtain more accurate results for vowel spacing changes with one homogeneous class. More vowels were added to those Peterson and Barney included in their work, as well as diphthongs.

In addition, the spectral changes, duration, and steady state F_1, F_2 for each vowel were measured. All these efforts were spent to provide more accurate and updated data. The participants were from Michigan. The findings were that the participants have similar vowel positions, but with reduced vowel separation in the vowel space. The study ended with the conclusion that decreasing the spacing between vowels does not affect their perception. Moreover, these findings coincided with the previous conclusions, which stated that F_1 and F_2 measurements are not enough to characterize vowel spacing.

The work of Peterson, Barney, Clark and Wheeler [21, 22] was repeated by Hajiwara [23]. This time the participants were from Southern California. Hajiwara was looking for formant changes across this new dialect. Through Hajiwara's work, it was found that a casual dialect such as Southern California rarely produces a full rounded vowel, and this was the main reason for having higher F_2 for some vowels. This work identified the new dialect (South Californian) that can be used in comparison studies with other dialects. In addition, Hajiwara believed that more research should be done for variable dialects within one community by increasing the number of both gender participants in future work.

3) DI using Consonants

Consonants on the other hand were recognized as dialect information identifiers as they can reveal foreign accents and social class [17]. This motivated William Labov [24-26] to

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

investigate accent differences based on sociolinguistics. He used the rhoticity: the pronunciation of "r" when it comes after a vowel (as in bar, sort, churn) and which is known as post-vocalic "r" as a new measurement metric that reflects personal origin [24]. Rhoticity was considered a low-prestige feature in Britain, while it was a prestigious feature of pronunciation in the USA. This way it was easy to differentiate between American English and British English. For that reason, Labov made the first sociolinguistic study through which he demonstrated that the New York accent variations revolve around postvocalic "r" usage [24, 25]. He started with a small survey to check the reliability of the new testing methods. Labov departed from interviewing speakers as other researchers had; instead he walked around 3 NYC department stores in Manhattan and pretended to be a customer. At first, he checked what items were on the fourth floor of each store; then he asked the sales assistants about where to find these items. He repeated the experiment in each store. He selected the fourth floor because it includes 2-tokens of postvocalic "r", and by pretending not to hear he got each informant to pronounce the two words twice, once spontaneously and once carefully, so that 4 tokens were obtained from each informant. Analysis of the collected data proved that Labov's hypothesis was moving in the right direction, and that the postvocalic "r" varies according to social class, speech style, and linguistic context associated with the clients of each store [24]. Through this study, higher and lower ranked employees were easily distinguished by studying the post-vocalic "r" pronunciation. Labov showed that the rhotic use of "r" reflected social class and aspiration, and that it was more widespread in younger speakers.

Figure 2 shows that the use of "r" was corresponding to the higher class of a store, and that rhoticity increases in careful speech. What was really interesting is that the greatest increase was obvious in the middle class store (Macy's). Labov explained that it is the norm at which the majority of Macy's employees aim, and not the one they use often. In addition, he found that the percentage of r-usage was much greater on the upper quiet floors of Saks, which contained the more expensive products, when compared with the busy ground floor [24].

According to Labov's study, there was a change in the New York accent by the end of World War II. In addition, American English includes dialects which do not pronounce the post-vocalic "r". This method was a good and easy way to identify the American dialects. Labov was the first to establish new measuring methods to identify dialects through sociolinguistic studies.

To sum up we can say that the previously mentioned works were trials to assign the main building blocks for Dialect Identification (DI) science. The studies were successful to a great extent in proving that both vowel spacing and consonants represent two important features in classifying dialects. After finding that the acoustic features change among specific dialects, the next steps were studies to learn how these changes vary among regional dialects and also among different genders with different ages. This motivated Byrd to study variations in the TIMIT corpora and to classify utterance according to region and gender [28].

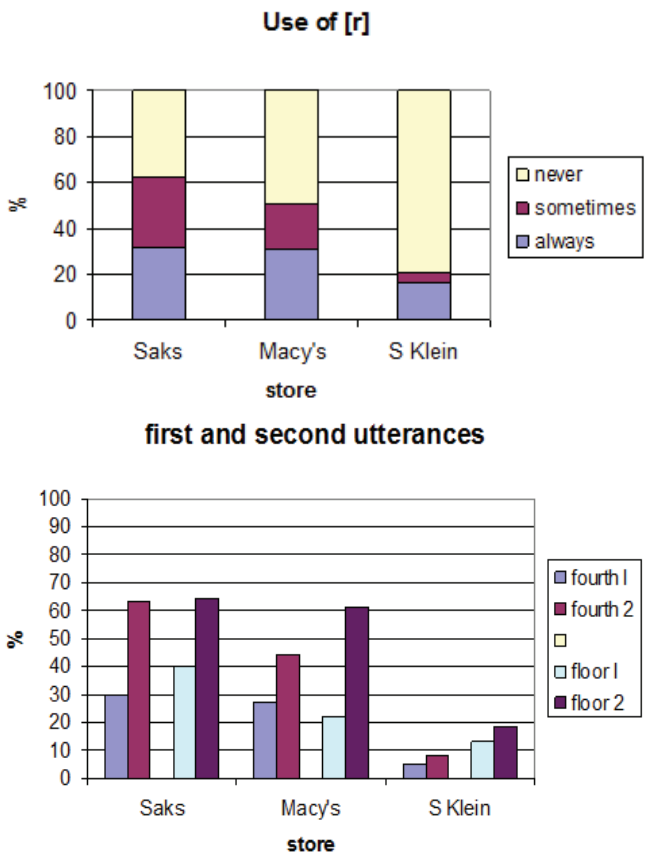

Fig. 2. "r" usage versus class [27]

TIMIT includes 630 speakers of different genders representing the eight American dialects. The study aimed to pick up the differences in glottalization, number of flaps, palatalization, central vowels, and speaking rate among different genders. The results showed that released stop, glottal stop, and vowel reduction rates differ among dialects, while the number of flaps does not differ. In addition, it was shown that men tend to speak faster than women, with less stressed vowels. However, there was no significant change among dialects and a suggestion was made that this may be due to having insufficient data for both genders. This clarifies that there are many parameters/features that can affect dialect variations. Extracting these parameters is not an easy task. In addition, it is difficult to specify the exact feature which is responsible for a certain dialect change as this is a database dependent problem. A good database must include participants from all ages, for both genders, and that includes different affiliations. However, this good database still lacks other hidden conditions that can vary with dialect, such as mood and health status.

4) DI using Acoustic and Phonetic Features

There were some successful trials in which Clopper distinguished and identified six American dialects [29, 30]. At first, the acoustic features of a dialect were studied followed by studying the perceptive features of listeners. In this work some words were picked, that differ among dialects, and their acoustic-phonetic properties were studied. Here, it is worth defining some important parameters before continuing the

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

summary of both papers. The fricatives are consonants produced by forcing air through a narrow channel made by placing two articulators close together. The articulators are the lower lip against the upper teeth, in the case of [f]. This turbulent airflow is called frication. Another important parameter is the vowel back-ness, which is named for the position of the tongue during the articulation of a vowel relative to the back of the mouth. However, vowels are defined as back or front not according to actual articulation, but according to the relative frequency of the second formant (F_2). The higher the F_2 value, the fronter the vowel; the lower the F_2 value, the more retracted (back) the vowel. Di-pthongization, or vowel breaking is the change of a mono-phthong into a di-phthong or tri-phthong. The change into a di-phthong is also known as di-phthongization [31]. Nagy and Zhang were concerned with new parameters such as the r-fullness to identify rhotic and non-rhotic dialects, the vowel brightness to check "r" insertion in a word, fricative voicing and duration to identify if people pronounce words with fricatives [32]. After measuring and identifying the previously mentioned parameters, experiments were done to find out which features listeners use to identify a dialect. The outcome of these experiments was that listeners always use four parameters in perception and dialect identification. The four parameters were the r-fullness, backness, vowel brightness, and di-pthongs. Listeners also were able to recognize the South and New England as being different from other dialects. The ability of a listener to identify a dialect depends on where he/she had lived during his/her life.

Later, in another study the aim was to make a comprehensive characterization of major American English dialects using the acoustic-phonetic features extracted earlier. The surprise here was varying features within the same dialect (again as [28] highlighted). It is difficult to generalize feature variation on a certain dialect group. Unexpected features among Southern males such as fronting and rising of "u" were observed, in addition to merged vowels sometimes. Finally, a conclusion was drawn stating that there are phonological differences among dialects and that we can identify a dialect based on well-chosen groups of lexical sets that are highly correlated. Human perception of dialects uses the same methods, but sometimes fails due to variation in dialect within the same group.

5) DI using Words and Lexical Sets

Supporting the same hypothesis that some specific words are the best dialect identifiers, Wells selected some words and described them as sets having a static pronunciation pattern [31]. The study focused on short words. A group of keywords was built, each containing a lexical set. A lexical set is a set of words where the vowels are pronounced the same way. During experiments it was found that within an accent two lexical sets merged. So, the conclusion was that a dialect can be discriminated using particular mergers plus lexical set pronunciation.

Zhang and Nagy [32] looked at the mutual information between different lexical sets in order to cluster dialects based on their phonological features. This method was tested using 168 binary features describing the pronunciation of vowels and consonants of English for speakers from 35 countries [32]. The clusters produced by this method were similar to those produced by the traditional ways of clustering. Also, a comparison between clustering methods was made, and the differences in clustering outcomes were studied. This work was significant in providing a brief description for different English dialects, as it listed all the variants among lexical sets.

Another method of clustering by analyzing the phonetic transcription was that of Kessler [33]. The method was based on computing linguistic distance between a pair of sites, similar to the isoglosses method. The isoglosses were the traditional method to identify dialect boundaries. This was achieved by drawing boundary lines separating regions where there is a group dialect; people pronouncing the same way. The drawback of isoglosses is the inconsistent boundary lines due to the chosen lexical feature variation within the same region. The new distance metrics to group phonetic metrics were the baseline metric, and the Levenshtien metric. In the baseline metric, two places on the map had "0" distance for similarly equal phone strings, and "1" otherwise. In the Levenshtien metric, the Levenshtein distance between phonetic strings was measured, where the Levenshtien distance is defined as the least expensive set of insertions, deletions, and substitutions needed to transfer one phone string into another [17]. The phonetic and feature string comparisons were used; in the former each operation changes one phone string into another phone string, while in the latter each operation changes the distinctive feature. The distance was measured by the distance between varied features, and averaged across all features. The final conclusion was that the phone string is better for automatic grouping of dialect identification. In addition, the word was defined as a unit for ADI and that can be divided into strings to get accurate measurements.

The last but not least DI method is that provided by Huckvale [34], which presented a new metric for the quantitative assessment of a speaker's accent similarities called ACCDIST. The idea of ACCDIST is based on recording the inter-segment distance measures for each speaker in a table, and then finding the autocorrelation measures between the recorded data and the ones under test; the author was thus able to group accents of the speakers. The ACCDIST technique to classify speakers allows capturing a speaker's pronunciation traits rather than voice characteristics [34].

A database of 20 sentences uttered by 20 males representing 14 dialect regions from the British Isles Speech corpus was employed. In order to verify accuracy of the ACCDIST, a comparison with the formant frequency metric and spectral envelope metric was made. In the formant frequency technique, each vowel is divided into two halves in time, the median value of the first four formants in each half combined to create the eight dimensional vectors for classification. This was done for each vowel in the 20 sentences uttered by the group except for the speaker under test. The accent was identified after computing the mean Euclidean distance between the test speaker's vowels and the accent group's mean vowels. This procedure requires no phonological labeling of the vowels as vowels in the same words were matched with each other. This was done for each

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

speaker. In the spectral envelope metric a 19-channel auditory filter bank was used to analyze each sentence. Each vowel was divided into two halves in time; the mean spectral envelopes in each half were then combined to create the "40-dimensional" vector for classification. This process was followed for each vowel in the 20 sentences for 19 speakers, and followed by classification according to the mean Euclidean distance to the accent group mean vowels. The ACCDIST results were better than both the formant frequency and the spectral envelope metric results, because gender changes never affect its accuracy as it is a pronunciation technique that depends on the context (uses the same text for both genders) in which vowels were produced instead of working on phonetic transcriptions.

Our previous discussion highlighted the gradual development of DI science as an important topic in the field of speaker and speech recognition. In the beginning, researchers were concerned with the features that appear to identify dialects. However, the deeper the study of features, the harder it was to solve the problem and the more time was needed to get results.

For that reason, researchers started to look at the context to achieve better results. Studying the acoustics and phonetics features of speech cannot be ignored, but at the same time they should not be considered to be a standalone solution as their derivation is time consuming. In other words, this kind of study was unable to provide a generalized solution. So, the direction now is to look at approaches that imitate human perception in classifying and identifying dialects.

B. From a methodology point of view

Most if not all of the methods that were used for language identification research were employed for dialect identification too. The Gaussian Mixture Model (GMM) is the first and simplest method. GMM models utterances and considers dialects as random processes that have mean and variance. By computing the likelihood of these processes, the process with the highest likelihood determines the identified dialect [35].

1) GMM

A few studies have used GMM for dialect identification. One of these was that described by Chen, Chang, and Wang to classify Chinese accents [36]. GMMs based on MFCC feature vectors for the identification of Beijing, Shanghai, Guangdong, and Taiwan accents of Mandarin were employed. The GMMs were trained to identify the most likely accent given test utterances. The identified accent type can then be used to select an accent-dependent model for speech recognition.

Another attempt was that of Faria in which a GMM classifier was used to recognize native and non-native speech [37]. This attempt aimed to distinguish between Native American English and non-native English using speech from Russian, Spanish, French, German, Chinese, Indian, and other speakers. This experiment verified that the lexical features from speech transcriptions can provide significant evidence of a speaker's accent.

TABLE II. CONFUSION MATRIX TERMINOLOGY

		Predicted		Terms
		N	P	Accuracy=(a+d)/(a+b+c+d)
Actual	N	a	b	True negative=a/(a+b) False negative=b/(a+b)
	P	c	d	True positive=d/(c+d), False positive=c/(c+d)

Experiments were done using the Fisher corpus (diverse accents), and a conclusion was drawn that accent classifiers that employ acoustic and lexical features achieved 84.5% classification accuracy for native accents. However, this classifier could not achieve more than 7.2% accuracy in recognizing non-native accents. The accuracy percentage is usually computed based on the confusion matrix depicted in Table II [38].

The GMM was a simple and easy way to identify dialects, but unfortunately the GMM was not capable of providing information on the temporal relationships due to lack of memory (to access information for future decisions). For example, if an "f" was pronounced through some statistics there will be an expectation for what will come next based on the current phone. As a result, researchers looked at HMM to resolve this problem because HMM involves memory.

2) HMM

The HMM (as does the GMM) assume dialect to be random processes, that can move from one state to another with different probabilities. Each state is associated with a process, which generates a new state with a new probability. It is called Hidden as the states are hidden and recognition allows finding those states.

The HMM is always trained for a specific feature. In case of dialects, there will be a separate model for each dialect. Forming models that are tuned to different dialects is called dialect specific phone modeling [17]. The test utterance is then compared to each model, resulting in a score. The model which has the highest score is chosen as the likely utterance source. The HMM is one of the well-known techniques for modeling individual phones for automatic speech recognition. The same phone models can be applied for dialect identification.

A novel design for an Arabic Dialect Identification system was designed by Alorifi [39]. An ergodic HMM that has two types of states was built, with the first type of state representing sounds that Arabic dialects have in common, and the second type representing sounds unique to a specific dialect. The common states with same sounds were connected among all models. Different speech features such as time derivatives, energy, and the shifted delta cepstra were utilized to model the phonetic differences between Egyptian and Gulf Arabic. A detailed comparison of the designed Arabic dialect identification system using the different speech feature combinations was presented. The designed system achieved accuracy of 99.45% for equally balanced training databases (for a mixture of unique/common state = 256/512). The high recognition accuracy was due to combining both specific dialect unique sounds and speech features. Specifying dialect

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

unique sounds allows the system to classify both dialects precisely. However, when the speech sound falls within common states, the system has to find the difference by extracting phonetic measures.

In [40], A. du Toit used various GMM and ergodic HMM configurations to identify five South African English dialects: Afrikaans English, Black South African English, Cape Flats English, White South African English, and Indian South African English. Seven tests were used to evaluate the configurations, with different speech duration for each string. The findings were that the system which used the second order ergodic HMMs was the best performing and the identification accuracies were 52.80%, 58.92%, 70.81%, 86.13%, 94.50%, 98.16%, and 98.67% for the 2, 4, 10, 30, 60, 120, and 300 second test segments respectively. A. du Toit explained that the obtained results showed a linear relationship between accuracy and length of test segment, and that the longer the speech duration the higher the accuracy achieved.

Another accent classifier for foreign accented continuous speech was developed by Kumpf and King [41]. The new approach was based on phoneme segmentation. A Parallel Phoneme Recognition (PPR) system was developed. The classifier was designed to process continuous speech and to distinguish between native Australian English (AuE) speakers and two migrant speaker groups with foreign accents, whose first languages are Lebanese Arabic (LA) and South Vietnamese (SV). The database used contained 3650 utterances for Australian (Au) English speakers, 1450 for Lebanese Arabic (LA) speakers, and 1350 utterances for South Vietnamese (SV) English speakers. The system was novel, requiring no manually labeled accent data, and can be automated. The test utterances were processed in parallel by the three accent recognizers. The recognizers employed accent specific HMMs and phone bigram language models to produce accent classification likelihood scores. The results of this approach showed that the likelihood scores obtained by accent-specific phoneme recognizers can be used to distinguish between speaker accents. The identifier was able to differentiate between the three accents; the identification accuracy achieved was 57.3% for Au English, 41.9% for LA English, and 46.7% for SV English. The average phoneme accent classification was 49.5% (the number of features was 48). King explained that the accent specific phoneme language models did not provide the expected significant contribution due to the limited training data. In addition, the classifier was very sensitive to the available data and to speaker variations. The future work suggestion was to expand the classification algorithm to include more discriminative classification techniques. Moreover, the author suggested creating benchmarks for speaker accent classification based on human perception studies for future work.

Following along the same lines, other studies using new sensitive word lists were executed by Arslan and Hensen [42]. Three models were employed for accent classification, the first was to apply HMM on isolated words for individual accents, the second was HMM on mono-phone models with continuous speech, and the third was using the HMM again on mono-phone models but considering the ones that actually exist in the utterance. After extensive study of the American English language education system, Arslan picked 20 isolated words and four test sentences. The selected database was built using head mounted microphone and telephone recordings. Fourty-eight male speakers from Duke University participated in the experiment to represent neutral American, German, Turkish, and Chinese English accents. Arslan used 5 tokens of 12 speakers for each accent group (60 tokens). This was done for 5 words, giving an overall number of trials equal to 60x5=300. Their algorithm outperformed the average human listener. The best model was the first one, which applied HMM on isolated words; its accuracy was 74.5%. Coming next the HMM that processed the mono-phone models within the actual utterance, which achieved accuracy equal to 68.3%. The last in performance was the HMM employing mono-phone models with continuous speech as its accuracy was 61.3%. Arslan and Hensen found that as the test utterance length increased, higher classification accuracy was achieved (coincident with A. du Toit‘s conclusion). An accent classification rate of 93% was achieved by using isolated word strings of 7-8 words uttered by the speaker among the four different language accents [42]. In addition, it was found that some words are better accent discriminators than others.

Towards the same objective of employing more discriminative classifiers, Ma, Zhu, and Tong used multi-dimensional pitch flux and MFCC features to discriminate between three Chinese dialects [43]. The multidimensional pitch flux features decrease the error rate by 30%. The duration of the test utterances was in the range of 3 to 15s, and the system was able to recognize the three dialects with an accuracy of 90%. Zhu and Tong concluded that increasing the duration of a test segment has a positive impact on both accuracy and error rate, as increasing the test segment duration leads to an increase in accuracy (this again coincides with conclusions by A. du Toit [40] and Arslan [42]) and a decrease in error rate.

Other experiments were done by Peter and Gilles to identify the German dialect using intonational cues [44]. In both experiments, listeners were subjected to regional intonational contours of German. In this study, listener performance varied with their linguistic experience. Also, listeners with non-local variety in their personal experience performed better than those who were familiar with local variety only. At the same time, listeners who were not familiar with Hamburg German and Berlin German were more successful in the identification test. The conclusion drawn was that the overall success rates depend on true recognition of local contours and also that overall success can be improved using an elimination strategy. In addition, the choice of the speaker generating the carrier utterances played an important factor that may affect recognizer performance. The identification results were improved using intonational cues.

Searching for more reliable cues for Automatic Dialect Identification (ADI), a study was done to explore the utility of prosodic features in language identification tasks and to check their reliability in the discrimination of Arabic dialects [45]. Results showed that prosodic acoustic cues were useful in classifying dialects. Furthermore, Arabic listeners were successful in identifying the Arabic dialects in their natural and synthesized forms. The identification rate was higher for

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

listeners having the same dialect, in other words listeners achieved higher rate of identification for their own dialect. This study was considered a first step towards the determination of reliable cues for automatic Arabic dialect identification.

Based on the same concept of reliable cues, Hamdi and Barakat-Defradas used the speech rhythm in Arabic dialects that had been described as stress-timed compared with other languages with different rhythm categories [46]. Preliminary results revealed that listeners use rhythm cues in the identification process, and listeners were able to distinguish between North Africa Arabic and Middle East Arabic. Acoustic domain investigations, measuring vowel duration and percentage of intervocalic intervals (as they reflect the rhythmic analysis), were carried out to prove their necessity for discrimination. Experiments prove that all Arabic dialects are still clustered within the stress-timed languages and exhibit a different distribution from languages belonging to other rhythmic categories, such as French and Catalan. By using this method, the complexity of the syllabic structure of the dialect (as in the word cat in the English language which will be identified as consonant vowel consonant), and vowel reduction (acoustic features that may lead to shorten vowel pronunciation) were extracted. Since rhythmic analysis is one of the current issues another study was done by Ramus [47]. The rhythmic analysis mentioned previously is based on the vocalic and intervocalic intervals and their variability. This method achieved limited success due to lack of sufficient data (corpora). Ramus' work was concerned with proposals for cross-linguistic control of speech rate, and statistical analysis improvements. The conclusion was drawn that both the complexity of syllabic structure and the existence of vowel reduction correlates with the rhythmic structure of the dialect and represents an important key factor in dialect identification.

In [16], Chen, Shen, and Campbell proposed a novel approach to adapt bi-phones for dialect identification. Both supervised and unsupervised learning algorithms were applied. Using these algorithms, dialect discriminating phonetic rules were extracted. Also, dialect discriminating bi-phones which are compatible with the linguistic literature were discovered, and at the same time improved a baseline mono-phone system by 7.5% (relative). The proposed dialect discriminating bi-phone system produced relative gains of 14.6-29.3% when fused with PRLM. The work of Chen and Campbell was considered a first step towards a linguistically-informative dialect recognition system, and - as they stated - if word transcriptions were provided, the proposed system can create more dialect-specific rules.

In [48], a PPR method to identify non-native English accents for six European accents was developed. The employed corpus was small; it included 200 isolated English words repeated two times for each accent. The data set was divided into training and test sets, with 60% for the training and 40% for the test set. The test utterance was recognized using accent-specific phone HMMs; an accuracy of 65.48% for average accent identification was achieved. The experiments were done on a comparison basis, accent-specific speech recognition was used and the accuracy achieved was 80.36%, similar to that of the oracle system where accuracy equals 80.78%. Teixeira, Trancoso, and Serralheiro believed that this method was successful and can provide better results if the corpus is larger.

In [49], a PRLM configuration for dialect identification was developed. The work aimed to identify Spanish dialect from Cuban and Peruvian dialects. The phone recognizer was trained on the TIMIT corpus (6300 utterances, 630 speakers). The recognizer used phone loop grammar. The Phonotactic LMs trained on Cuban and Peruvian speech. The duration of each utterance was 3 minutes each, recorded from 40 Cuban and 20 Peruvian speakers. The test sets durations were the same as the training set durations, 3 minutes each, taken from 40 Cuban and 20 Peruvian speakers. The accuracy achieved with this approach of dialect identification was 84%.

In [50], a PPRLM and a logistic back-end classifier to identify five Arabic dialects were employed. The five Arabic dialects were Egyptian, Gulf, Iraqi, Levantine, and Modern Standard Arabic. Different phone recognizers such as: Modern Standard Arabic, English, German, Japanese, Hindi, Mandarin, and Spanish were used. The phone recognizers' role was training of phonotactic LMs. Using 30 second test utterances the recognizer was able to identify the dialects with accuracy reaching 81.6 %. The training sets used were 34.96 hours of Gulf Arabic speech, 18.4 hours of Iraqi Arabic, 68.79 hours of Levantine Arabic, 47 hours of Egyptian Arabic, and 35.54 hours of Modern Standard Arabic. The test sets consisted of 6.06 hours of Gulf Arabic, 7.33 hours of Iraqi Arabic, 10 hours of Levantine Arabic, 10 hours of Egyptian Arabic, and 12.06 hours of Modern Standard Arabic. The best accuracy was 84% for the longer duration of 120s.

In [51], the aim was to distinguish between three South African English accents. South African English was classified into two groups: white South African and black South African English. The black South African group included two accents: Nguni and Sotho. The PPR method was applied for the automatic classification process. The main objective of this work was to determine whether black South African English has to be treated separately when used in ASR or should be treated as belonging to the single variety of South African English. The results were as follows: the system was able to classify the two groups, white and black South African English; however it was unable to distinguish between Nguni and Sotho speakers. Finally, the decision was made to consider Nguni and Sotho native speakers English as a single variety during the ASR development phase. The training set was 2.6 hours of speech for each dialect and the test set was about 13 minutes of speech for each dialect.

3) SVM

The Support Vector Machine (SVM) is a recent alternative discriminant classifier. It can discriminate between two groups by forming a boundary line which separates the two groups with enough space margin.

The main advantage of SVM is that if there is no possibility to divide any groups linearly, these groups can be transformed, by using a kernel function, to another space where they can then be discriminated. The original SVM works for two way classification, however there is a multiclass version of SVM as well.

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

The SVM was used to distinguish between Canadian, Indian, and Chinese English dialect [52]. Two SVMs were used. The pairwise SVM, and the Directed Acyclic Graph SVM (DAGSVM). The DAGSVM was employed for better performance. The word intonation duration, F_2 and F_3 contour, and the word final stop closure duration were the measured parameters for classification. The two SVM models were compared to an HMM model. The SVM results were comparable to that of HMM with 81.2% accuracy for the pairwise technique, while the DAGSVM accuracy was 93.8%.

4) Statistical Covariance

Another effective approach for DI is the cluster technique, which is a good choice when we are looking for a discriminative method. There are some dialect features that vary systematically or in a known way for native speakers (such as the post-vocalic "r"). A method of clustering dialects using their common phonological features was developed [32]. The work aimed at examining the statistical co-variation of dialects and their dependency. The method was based on the Mutual Information (MI) concept, for example if in a specific dialect the "r" is vocalized then there exists a cot/caught merger in this dialect. These variables seem to be independent of each other. However, they do exhibit statistical dependence. The method was tested by exploring a data set of 168 binary features describing the pronunciation of vowels and consonants of English speakers from 35 countries and regions. These dialects were grouped according to patterns of shared features described by Schneider's MI [32]:

$$I(x,y)=\sum_{x}\sum_{y}p(x,y)\log\left\{\frac{p(x,y)}{p(x)\,p(y)}\right\} \quad (1)$$

where $I(x,y)$ is the MI between x and y. The MI is defined by the marginal probability distribution function for the two variables x and y, and their joint probability. Therefore, if x and y are independent then $I(x,y)=0$; in other words, x and y are uncorrelated and if we have any information about x this cannot provide any information about y. This was the reason to search for the highly correlated features. This idea was successful in reflecting the historical relationships between dialects. A proof by experiment that accents with historical relationships may share common acoustics and phonological features was provided (they are correlated). The results of this method of categorizing dialect varieties by binary pronunciation features were compared to traditional groupings based on external features. The clusters produced by both methods were similar. Moreover, the recommended approach provided a highly complex pattern for different dialects among the world English language. This work will help provide the universal pattern variation for the English dialects.

5) GPUs

One of the recent works that focus on the importance of speedy recognizers with new technologies was that of Hanani [53]. The work explored the application of Graphics Processing Units (GPUs) to speech pattern processing to provide substantial processing power at low cost. In addition, it demonstrated the principles for using GPUs such as algorithm selection and effective coding. In the demonstration process, the NIST LRE 2003 standard language identification task was used. The focus was on two parts of the system: the acoustic classifier and the acoustic feature extraction. The acoustic classifier was based on 2048 GMMs, while for the acoustic feature extraction a comparison was done on FFT-based analysis with IIR and FIR filter banks in terms of their ability to exploit the GPU architecture and LID performance. The GPU-based system did not increase error rate, and using an FIR front-end completed the NIST LRE 2003 task in 16 hours instead of 180 hours for the conventional FFT-based systems on a standard CPU (with a significant processing time reduction of 61% in the front-end). The conclusion was that the GPU implementation is a better solution when processing time is a concern.

6) Neural Networks

Looking for new effective models that can replace humans in the recognition process there was an exploration of several Neural Network (NN) models that can be used in speech systems [54]. The NN-based prosody models were demonstrated to capture the prosodic information specific to speaker, language, and sound unit categories [54]. The results were promising, and the developed prosodic models were suggested to be explored with the conventional models for improving speech, speaker, and language recognition system performance. In addition, it was found that the developed prosodic models can enhance the quality of the synthesized speech. Furthermore, mapping functions were developed and evaluated subjectively and objectively. After that, the auto-associative NN models for capturing the emotion specific information from speech using spectral and prosodic features were explored. The final experiment was to employ NN models to distinguish five Hindi dialects using spectral and prosodic features. The accuracies of the dialect identification systems developed using duration, pitch, and energy features were as follows; 80%, 77%, 74%, 86%, and 73% for the Chattisgharhi, Bengali, Marathi, General, and Telugu dialects respectively.

V. SUMMARY

Arslan showed that working with isolated words is better than working with phone models when dealing with local dialects [42]. Zhang worked on words derived from lexical sets to find the Mutual Information between different dialects [32]. Combining phonotactic constraints with HMM improves ADI as the phonotactic constraints reflect more information than that of a word. The SVM is a promising technique for ADI; it gave comparable results to the HMM with high accuracy approaching 95%, however there is no study to indicate how this accuracy will scale with an increase in the number of accents. More research is needed to highlight the power of SVM with scaling (with increasing number of models). Several works stressed the linear relationship between the length of test segments and accuracy, the longer the test segment the higher the accuracy obtained.

Through a broad view of the dialect identification problem we have seen that both temporal and prosodic features are very important as they can reflect more detailed information about the different patterns a word may have across one dialect, and how much a word varies with context. Clustering

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

Mutual Information among dialects can provide universal pattern variations for a specific dialect. More efforts should be spent towards database availability.

Studying the human perceptual system became an urgent need to be able to imitate how listeners discriminate between dialects. Neural Networks are a promising solution for Automatic Dialect Identification. The final point to be considered in future research of dialect identification is to provide speedy solutions that can benefit from today's advanced technology [53]. In Table III the most significant DI methods found in the literature are summarized.

ACKNOWLEDGMENT

A. E. thanks the VT-MENA program for its support.

TABLE III. SUMMARY OF DIALECT IDENTIFICATION METHODS

Methodology	Author, languages	Accuracy	observations	Method description
GMM	Faria, English accent (Fisher corpus) Native & non-native American English using speech from Russian, Spanish, French, German, Chinese, Indian, and other speakers.	accuracy= 84.5% (could not achieve more than 7.2 % in recognizing non-natives)	lexical features from speech transcriptions can provide significant evidence of a speaker's accent	GMM employs acoustic & lexical features
Ergodic HMM	Arabic dialect identification (CALLHOME corpus+ recordings)	accuracy= 96.67%.		An Ergodic HMM that has two types of states was built, the first type of state represents the Arabic dialect common sounds, and the second type represents the specific dialect unique sounds. The common states with same sounds were tied among all models. Different speech features such as time derivatives, energy, and the shifted delta cepstra were utilized to model the phonetic differences between Egyptian and Gulf Arabic.
GMM and Ergodic HMM	A. du Toit 5-way South African English dialects: Afrikaans English, black South African English, Cape Flats English, white South African English, and Indian South African English.	52.8% for 2 s. 58.92% for 4s. 70.81% for 10s. 86.13% for 30s. 94.50% for 60s. 98.16% for 120s. 98.67% for 300s.	The second order ergodic HMM was the best performer. longer speech duration= higher accuracy.	The recognizers employed accent specific HMMs and phone bigram language models to produce accent classification likelihood scores.
A parallel phoneme recognition (PPR)	Kumpf and King native Australian English (AuE) speakers and two foreign accents Lebanese Arabic (LA) and South Vietnamese (SV).	76.6%	Data was not enough to obtain better accuracy. More discriminative classification techniques should be employed.	A classifier designed to process continuous speech.
HMM (isolated words, monophones with continuous speech, monophones using same utterances)	Arslan and Hensen (recordings using mic. & telephone Duke's participants representing neutral English, Turkish, German & Chinese accents)	74.5% (isolated words), 61.3% monophone with CS 68.3% monophone using same utterrance	Longer test utterance = higher accuracy	Isolated word strings of 7-8 words result in an accent classification rate of 93% among four different language accents
multi-dimensional pitch flux and MFCC features	Zhu and Tong 3-way Chinese dialects	accuracy = 90%. error rate reduced by 30%.		
prosodic features in language identification	Barakat and Ohala	listeners achieved higher rate of identification for their own dialect	Prosodic and acoustic cues were useful in classifying dialects. Arabic listeners were able to identify Arabic dialects in their natural and synthesized forms.	
PPR	non-native English accents for six European accents corpus was small. 200 isolated words repeated 2 times	65.48% for average accent identification 80.36% for specific-accent recognition similar to Oracle		

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

	for each accent	system 80.78%,		
PRLM (phone recognizer and Phonotactic Language Model)	Zissman identify Spanish dialect from Cuban and Peruvian dialects	84%		The phone recognizer was trained on the TIMIT corpus. The recognizer used phone loop grammar. The Phonotactic LMs trained on Cuban and Peruvian speech. The duration of each utterance was 3 minutes each, recorded from 40 Cuban and 20 Peruvian speakers.
PPRLM and a logistic back-end classifier	Biadsy and Habash identify five Arabic dialects	Recognizer accuracy=81.6 to 84% for the longer duration of 120s.		The phone recognizers role was training of phonotactic LMs.
SVM	Tang and Ghorbani Canadian, Indian, and Chinese English dialects	81.2% accuracy for the pairwise technique while the DAGSVM accuracy was 93.8%.	The SVM results were comparable to that of HMM	Pairwise SVM, and Directed Acyclic Graph SVM (DAGSVM)
Acoustic classifier and acoustic feature extraction using GPU at low cost	Hanani	The GPU-based system did not increase error rate, and using an FIR front-end completed the NIST LRE 2003 task in 16 hours instead of 180 hours of the conventional FFT-based systems on a standard CPU	GPU implementation is a better solution when processing time is a concern.	The acoustic classifier was based on 2048 GMMs and the acoustic feature extraction was done on FFT-based analysis with IIR and FIR filter banks in terms of their ability to exploit the GPU architecture and LID performance.

REFERENCES

[1] J. K. Chambers and P. Trudgill, “Dialectology”, chapter one, pp. 4-9, 2nd edition, Cambridge University press, 1998.

[2] L. Gang, H. John, and L. Hansen, “A Systematic Strategy for Robust Automatic Dialect Identification”, 19th European Signal Processing Conference (EUSIPCO 2011), pp. 2138-2141, 2011.

[3] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, “Dialect identification using Gaussian mixture models", Proceedings of the Speaker and Language Recognition Workshop (Odyssey '04), pp. 41-44, Toledo, Spain, 2004.

[4] M. A. Zissman and K. M. Berkling, “Automatic language identification", speech communication, vol. 35, pp. 115-124, 2001.

[5] M. A. Zissman, “Comparison of four approaches to automatic language identification of telephone speech", IEEE Transactions on speech and audio processing, vol. 4, pp. 31-44, 1996.

[6] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, “Acoustic, phonetic, and discriminative approaches to automatic language identification", proceedings of the European conference on speech communication and technology (Eurospeech 2003), Geneva, Switzerland, pp. 1345-1348, 2003.

[7] National Institute of Science and Technology (NIST), “The 1996 language recognition evaluation plan", Lang_Rec.04, pp. 1-5, April 1996.

[8] A. F. Martin, and M. A. Przybocki, “NIST 2003 language recognition evaluation", proceedings of the European conference on speech communication and technology (Eurospeech 2003), Geneva, Switzerland, pp. 1341-1344, 2003.

[9] A. F. Martin and A. N. Lee, ”The current state of language recognition: NIST 2005 evaluation results", proceedings of the speaker and language recognition workshop (Odyssey'06), San Juan, Puerto Rico, pp. 1-6, 2006.

[10] NIST, “The 2007 language recognition evaluation plan", LRE07EvalPlan-v8b.doc, pp. 1-5, 2007.

[11] Jankowski et al., "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database", proceedings ICASSP-90, pp. 109-112, April 1990.

[12] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition", proceedings of the international conference on acoustics, speech and signal processing, pp. 651- 654, 1988.

[13] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “SWITCHBOARD: a telephone speech corpus for research and development”, proceedings of ICASSP, vol. I, pp. 517-520, 1993.

[14] J. Bernstein, K. Taussig, and J. Godfrey, ”Macrophone: an American English telephone speech corpus for the POLYPHONE project", proceedings of ICASSP, vol.1, pp. 81-84. 1994.

[15] M. A. Anusuya and S. K. Katti, “Speech recognition by machine: a review”, International Journal of Computer Science and Information Security (IJCSIS), vol.6, No. 3, pp. 181-205, 2009.

[16] N. F. Chen, W. Shen, and J. P. Campbell, “A linguistically-informative approach to dialect recognition using dialect discriminating context-dependent phonetic models”, ICASSP 2010, pp. 5014-5017, 2010.

[17] A. A. Nti, “Studying dialects to understand Human Languages”, M.S. thesis Massachusetts Institute of Technology, 2009.

[18] C. N. Bailey, ”Is there a Midland Dialect?” [Washington, D.C.] : Distributed by ERIC Clearinghouse, 1968. Website: americanspeech.dukejournals.org/content/78/3/307.refs

[19] L. M. Davis and C. L. Houck, “Is There a Midland Dialect Area?—Again”, American speech, vol. 67, no. 1, pp. 61-70, 1992.

[20] E. Johnson, “yet again, the Midland Dialect”, American speech, vol. 69, no. 4, pp. 419-430, 1994.

[21] G. E. Peterson and H. Barney, “Control methods used in the study of the vowels”, journal of the acoustical society of America, vol. 24, pp. 175–184, 1952.

[22] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, “Acoustics characteristics of American English vowels”, journal of the acoustical society of America, vol. 97, pp. 3099-3111, 1995.

[23] R. Hajiwara, “Dialect variation and formant frequency: The American English vowels revisited “journal of acoustical society of America, vol. 102, pp. 655-658, 1997.

[24] http://eprints.aston.ac.uk/439/1/studying_language_2a.pdf, chapter one, pp. 21, Aston University.

[25] W. Labov, “sociolinguistic patterns Philadelphia: university of Pennsylvania”, pp. 43—54, press1973.

[26] W. Labov, C. Boberg, and B. Sharon, “The Atlas of North American English”, chapter seven, 2006.

[27] personalpages.manchester.ac.uk/staff/Harold.Somers/Labov.ppt

[28] D. Byrd, ‘Preliminary results on speaker dependent variations in the TIMIT database”, journal of the acoustical society of America, vol. 92, pp. 593-596, 1992.

[29] C. G. Clopper and D. B. Pisoni, “Some acoustic cues for the perceptual categorization of the American English dialects”, journal of phonetics, vol. 32, pp. 111-140, 2004.

[30] C. G. Clopper, D. B. Pisoni, and K. de Jong “Acoustical characteristics of the vowel systems for six regional varieties of American English”,

978-1-4673-7606-8/15/$31.00 ©2015 IEEE

journal of the acoustical society of America, vol. 118, pp. 1661-1676, 2005.

[31] J. C. Wells, "Accents of English", An Introduction, Cambridge University press, pp. 1–278, 1982.

[32] N. Nagy, X. Zhang, G. Nagy, and E. W. Schneider, 'Clustering dialects automatically: A mutual information approach", working papers in linguistics from NWAV 34, pp. 145-158, 2006.

[33] B. Kessler, "Computational dialectology in Irish Gaelic", proceeding of the European ACL, pp. 60-67, 1995.

[34] M. Huckvale, "ACCDIST: a metric for comparing speakers' accents", proceedings of Interspeech, Korea, pp. 29-32, 2004.

[35] M. A. Zissman, "Automatic language identification using Gaussian Mixture and Hidden Markov Models", proceedings IEEE International Conference for Acoustics, Speech, Signal Processing (ICASSP), pp. 399–402, 1993.

[36] T. Chen, C. Huang, E. Chang, and J. Wang, "Automatic accent identification using Gaussian Mixture Models", Automatic Speech Recognition and Understanding (ASRU '01), pp. 343-346, Italy, 2001.

[37] A. Faria, "Accent classification for speech recognition", proceedings of the Second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI '05), UK, 2005.

[38] R. Kohavi, and F. Provost, "Glossary of terms, Machine Learning", vol. 30, no. 2/3, pp. 271-274, 1998.

[39] F. S. Alorifi, "PhD Dissertation: automatic identification Of Arabic dialects using Hidden Markov Models", University of Pittsburgh, 2008.

[40] A. du Toit, "Automatic classification of spoken South African English variants using a transcription-less speech recognition approach", M.S. thesis, Stellenbosch University, 2003.

[41] K. Kumpf and R. King, "Automatic accent classification of foreign accented Australian English speech", proceedings of ICSLP '96, pp. 1740-1743, 1996.

[42] L. M. Arslan, and J. H. L. Hensen, "Language accent Classification in America English", speech communication, vol. 18, no. 4, pp. 353-367, 1996.

[43] B. Ma, D. Zhu, and R. Tong, "Chinese dialect identification using tone features based on pitch flux", proceedings of ICASP'06, I: pp. 1029-1032, 2006.

[44] J. Peters, P. Gilles, P. Auer, and M. Selting, "identification of regional varieties by intonational cues: an experimental study on Hamburg and Berlin German", vol. 45(2): pp. 115–139, 2002.

[45] M. Barkat, J. Ohala, and F. Pellegrino, "prosody as a distinctive feature for the discrimination of Arabic dialects", proceedings of Eurospeech '99, pp. 395-398, 1999.

[46] R. Hamdi, M. Barkat-Defradas, E. Ferragne, and F. Pellegrino, "Speech timing and rhythmic structure in Arabic dialects: a comparison of two approaches", proceedings of Interspeech'04, pp. 1, 2004.

[47] F. Ramus, "acoustic correlates of linguistic rhythm: perspectives", speech prosody, pp. 115-120, 2002.

[48] C. Teixeira, I. Trancoso, and A. Serralheiro, "accent identification", proceedings of the International Conference on Spoken Language Processing (ICSLP '96), PA, pp. 1784-1787, 1996.

[49] M. A. Zissman, T. P. Gleason, D. M. Rekart, and B. L. Losiewicz, " Automatic dialect identification of extemporaneous, conversational, Latin American Spanish speech", proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96), Atlanta, GA, pp. 777-780, 1996.

[50] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic dialect identification using phonotactic modeling", proceedings of the EACL Workshop on Computational Approaches to Semitic Languages (CASL '09), Greece, pp. 53-61, 2009.

[51] F. De Wet, P. Louw, and T. R. Niesler, "Human and automatic accent identification of Nguni and Sotho Black South African English", South African Journal of Science, vol. 103, no. 3/4, pp. 159-164, 2007.

[52] H. Tang, and A. A. Ghorbani, "Accent classification using Support Vector Machine and Hidden Markov Models", proceedings 16th Canadian conference on Artificial Intelligence AI'03, pp. 629 – 631, 2003.

[53] A. Hanani, "Human and computer recognition of regional accents and ethnic groups from British English speech", chapter one, PhD dissertation, University of Birmingham, March 2012.

[54] K. S. Rao, "Role of neural network models for developing speech systems", Sadhana vol. 36, Part 5, Indian academy of sciences, pp. 783–836, 2011.

978-1-4673-7606-8/15/$31.00 ©2015 IEEE