Int J Speech Technol (2015) 18:277–286
DOI 10.1007/s10772-014-9268-y

Automatic analysis of dialect/language sets

Mahnoosh Mehrabani · John H. L. Hansen

Received: 2 December 2013 / Accepted: 9 December 2014 / Published online: 14 January 2015

Abstract Dialect variations of a language represent considerable challenges for sustained performance of speech systems. In a given language space, estimation of similarity or diversity between multiple dialects provides valuable information for speech researchers. In the present study, fundamental differences between dialects or closely related languages are explored based on available speech data from those dialects/languages. First, a method is proposed to measure spectral acoustic differences between dialects based on a volume space analysis within a 3D model using log likelihood score distributions derived from traditional Mel Frequency Cepstral Coefficient features and Gaussian Mixture Models. Next, text-independent prosody features based on pitch and energy contour primitives are proposed to study excitation structure differences between dialects. The proposed dialect proximity measures are evaluated and compared on a corpus of Arabic dialects, as well as a corpus of South Indian languages, which are closely related languages. The presented measures are shown to be consistent and repeatable.

Keywords Dialect separation · Dialect recognition · Prosody · Language recognition

M. Mehrabani · J. H. L. Hansen (✉)
Center for Robust Speech Systems (CRSS), Department of Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, P.O. Box 830688, EC33, 800 West Campbell Road, Richardson, TX 75080-3021, USA
e-mail: John.Hansen@utdallas.edu
URL: http://crss.utdallas.edu
URL: http://www.utdallas.edu/∼John.Hansen/

Assessing the proximity between multiple dialects of a language is an interesting yet challenging research topic on which little if any work has been done. Such assessment shows how confusable/distinguishable dialects are in the given language space. The goal of this study is to develop a consistent measurement strategy of dialect differences which helps study groups of dialects within new languages. Assessing the proximity of dialects also allows speech researchers to conserve resources, if additional acoustic/language models are not needed (e.g., can an American English trained ASR system perform satisfactory for UK English? or Cardiff, Bradford, or such sub-dialects of UK English?).

From an assessment perspective, one could consider separation based on (i) physical speech production differences, (e.g., measuring and comparing phoneme or vowel spaces, etc.), (ii) linguistic speech formulation differences, (e.g., studying the history of dialect evolution), (iii) perception assessment traits, and (iv) automatic speech system classifier differences. All four are viable, and one would easily expect that differences which are statistically significant in one domain, may not carry over to another. Here, we focus on automatic methods to help speech scientists, engineers, and linguists develop better understanding of dialects. Note that the fundamental challenge in this study is the lack of ground-truth on what is actually the difference between dialects/languages.

Similarities between different languages have been studied in the literature for a number of reasons such as: adapting the speech recognition system of one language for use in other languages (Sooful and Botha 2001), and leveraging differences in the acoustic phone space based on multi-lingual phoneme modeling (Kohler 1996). Kohler (1996) applied the log likelihood measure as a numerical distance to determine sound similarities between languages. This would help in the formulation of multi-lingual phoneme models to be

employed in multi-lingual speech recognition systems. Yin et al. (2007) measured language differences in order to automatically cluster similar languages for hierarchial language identification. Similarities between languages have also been studied from a linguistic point of view. Bradlow (2008) rated the similarity of various languages to English based on native and non-native listeners' perceived distance from English. Walter (2009) classified languages based on their perceptual sound structure similarity to English.

Phonetic distances between different dialects of the same language have been computed, as well. A dialect is a variety of a language that is used by a group of speakers belonging to some geographical region. Dialects of a language are different in phonetic, grammatical, and lexical features. The distinction between a dialect and a language is sometimes contradictory. Mutual comprehensibility is a primary criterion for distinguishing a dialect from a language. Unlike speakers of different languages, speakers of different dialects of a language generally understand each other, even with some difficulty (Curzan and Adams 2006). Dialectology is the study of dialects, as well as the different features that focuses on geographic variations within a language.

Recently, there has been interest in developing computational dialect comparison and classification methods in order to divide geographical maps into dialect areas. In the literature, considerable linguistic work has been performed on the calculation of pronunciation differences between dialects (Heeringa 2004; Heeringa and Hinskens 2012; Nerbonne et al. 1996; Wieling et al. 2011). These are non-probabilistic approaches based on average string distances between corresponding words, or phones pronounced in different dialects. Various string distances are used for this purpose, such as Levenshtein, Euclidean, and Manhattan distance. The study by Heeringa et al. (2006) explored and evaluated string distance algorithms for modeling dialect distances. A related study Nerbonne and Heeringa (2001), considered a number of methods for measuring phonetic distance between dialects. Dialect differences in the vowel spaces and their impact on the perception of vowels in different dialects have also been studied by Faber et al. (1994). A number of linguistic studies have applied statistical data driven approaches to quantify differences between phonetic features of dialects (Shackleton 2007; Wieling et al. 2013; Nerbonne 2009).

Apart from linguistic approaches, little if any work has been done to perform a meaningful proximity assessment between dialects. In this study, we present a procedure for assessing the separation or proximity between dialects based on only the available un-transcribed training data. The proposed automatic dialect assessment framework shows how accurately the dialects can be distinguished. Therefore, it provides some sense of the resulting dialect classification system performance: an important property when new dialects are introduced for training and model construction. Note that the proposed method is not intended to provide an accurate prediction of the final classification performance, but more of a measure for analysis of the dialect/language separation. The proximity measurement framework presented in this study, is intended to provide a built-in self-test, and therefore, the training data is used for this measurement.

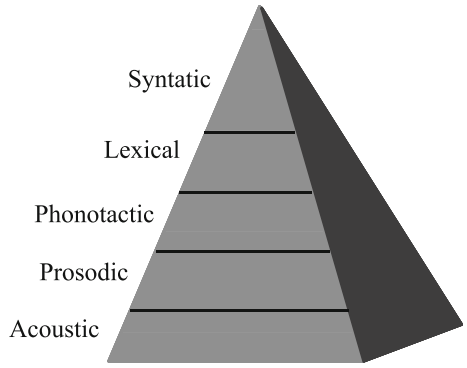

Fig. 1 Hierarchial pyramid for different levels of speech features

In addition, recent studies (Nallasamy et al. 2011; Biadsy et al. 2012) show the impact of dialectal variations of a language on performance of speech recognition systems. A dialect proximity measure can be applied to predict speech recognition performance for new dialects of a language when training data is not available for those dialects. Note that separation between dialects in different feature spaces may vary. Here, we consider acoustic and prosodic feature spaces, and we present and compare results of dialect proximity assessment in these two feature domains.

Figure 1 shows a hierarchial pyramid of different speech features based on which, dialect separation assessment can be performed. We start by comparing statistical models using lower level acoustic, and prosodic features. Future research could also consider higher level features such as word selection, or grammatical differences between dialects. The majority of linguistic studies have focused on higher level structure for specific dialects, generally with text transcript knowledge.

Fundamental differences between dialects are explored in this study. Our intent here is to formulate an effective solution, which will serve as a foundation for further research studies. Since little prior research has focused on automatic quantitative assessment of dialects, and ground-truth knowledge of dialect separation within languages is not known, the advancements here can only identify what is believed to be effective and repeatable (i.e., it is not possible to develop the "optimum" method, or to compare with existing proven methods). The proposed dialect proximity measures are also shown to be effective for language separation assessment, especially for closely related languages.

First, traditional Mel Frequency Cepstral Coefficient (MFCC) features are used to measure the spectral acoustic differences. Gaussian Mixture Model (GMM) output score

histograms are compared and analyzed in a 3D space to obtain a measure of dialect separation. Next, differences in excitation structure between dialects are studied using two types of low level prosody features: pitch, and energy contour primitives. These text-independent prosody features are exploited as building blocks for training statistical models of pitch/energy change in each dialect. It should be noted that other methods may be more suitable if text knowledge is available. However, the underlying constraint for this framework is that text knowledge is not provided. Starting with a small size three sample set, basic contour change patterns are modeled. Next, by means of N-gram modeling, longer temporal based prosodic patterns are compared. The resulting proposed proximity assessment methods are evaluated on a corpus of Arabic dialects, as well as a corpus of South Indian languages. South Indian (Dravidian) languages are closely related languages from the same language family, and bare some similarities to pure dialect separation assessment. The proximity measures from different proposed methods are compared for both corpora. Consistency of the proposed measures is also studied with changes in the training data used for assessment. Finally, a subjective listener assessment is performed to illustrate the relation between automatic system results and human perception.

1 Dialect proximity measurement based on log likelihood score distributions

This section considers traditional MFCCs as a means to represent spectral based differences between dialects. Since we pursue a statistical approach, extracted features are modeled using GMMs. The proposed method of assessing proximity is based on comparing the log likelihood score statistics. An earlier version of this method was presented by Mehrabani and Hansen (2008). First, MFCC features are extracted from the available audio data. Each dialect is modeled using traditional GMM training. In this study, 64-mixture GMMs and 12-dimensional MFCCs (excluding the 0th cepstrum coefficient) were used. The same number of mixtures was applied to model each dialect/language.

Next, a closed-set dialect test is performed (i.e., the same train data for each dialect is tested against all dialect models in order to obtain score distributions from matched and mismatched train data models). Score distributions, derived from the score histograms are the basis for the proposed dialect separation measure. In order to compare two dialects D_1 and D_2, D_1 training data is tested against D_1 and D_2 GMM models to obtain two sets of log likelihood scores: S_{11} and S_{12}, respectively. S_{21} and S_{22} are obtained in a similar manner. Next, a histogram is formulated for each score set, and is approximated with a probability distribution function. We used the generalized extreme value (GEV) probability distribution (Kotz and Nadarajah 2000) to model the score histograms. The GEV distribution is a flexible three-parameter model that combines three types of extreme value issues into the distribution model.

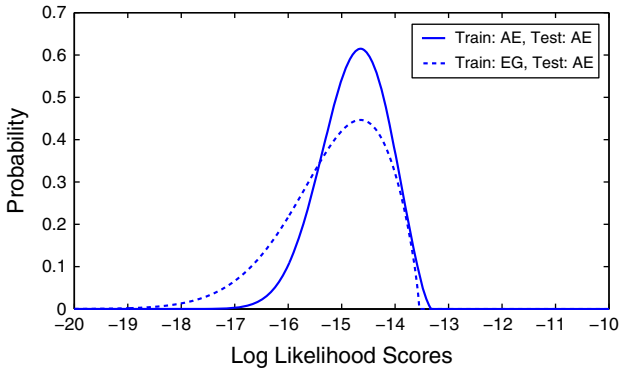

Fig. 2 Comparison between two score distributions when testing dialects of Arabic: AE data versus AE (S_{11} distribution) and EG (S_{12} distribution) GMMs

The GEV has the following PDF with location parameter μ, scale parameter σ, and shape parameter k:

$$f(x) = \begin{cases} \frac{1}{\sigma} \exp\left(-(1+kz)^{-1/k}\right)(1+kz)^{-1-1/k}, k \neq 0 \\ \frac{1}{\sigma} \exp\left(-z - \exp\left(-z\right)\right), k = 0 \end{cases} \tag{1}$$

where $z = \frac{x-\mu}{\sigma}$, and $\sigma > 0$. The range of definition of the distribution depends on the shape parameter, and for $k \neq 0$: $1 + kz > 0$. The parameters of the GEV score distribution are estimated based on maximum likelihood estimation.

The core idea here is that the more distinguishable the distributions are for the score sets S_{11} and S_{12}, the greater the distance from D_1 to D_2 or $d(D_1, D_2)$. Similarly, comparing the S_{21} and S_{22} score distributions yields the distance from D_2 to D_1 or $d(D_2, D_1)$, which is not necessarily equal to $d(D_1, D_2)$. Figure 2 shows S_{11} and S_{12} score distributions when D_1 and D_2 represent two Arabic dialects of AE (United Arab Emirates) and EG (Egypt), respectively. The underlying assumption is that the true spectral mismatch between two dialects can accurately be measured using MFCC features and GMMs.

Here, we employ the mean square error criterion to compare score distributions, and obtain an estimate of the proximity between each dialect pair. Our initial experiments showed that when two score distributions do not have considerable overlap, the corresponding dialects are far apart and can be well classified. However, the inverse is not always true. In other words, there are cases where the two distributions have measurable overlap, but the classification accuracies are still high. The reason is that when comparing score distributions, the score statistics of the entire data set are compared. Alternatively, in a classification task, only two scores at a time

are compared which correspond to the same test stream. The solution to this problem is to move the analysis from a 2D surface to a 3D space.

In order to assess the separation between dialects $D1$ and $D2$, represented as $d(D_1, D_2)$, an estimate is formed to determine how well D_1 can be identified from D_2. This estimation involves building a 3D score distribution based on two 2D score distributions via testing D_1 data against the D_1 and D_2 models. Let us refer to these distribution functions as f_1 and f_2, respectively. The joint PDF of the two sets of scores is calculated as follows:

$$f(x, y) = f_1(x) \times f_2(y). \tag{2}$$

This 3D distribution has the partial distributions of f_1 and f_2 in the XZ and YZ planes, respectively. Note that the S_{11} and S_{12} score distributions are not generally independent. Therefore, f is not exactly the distribution for the pairs of scores which result from testing each original train (now test) token in D_1 against the D_1 and D_2 models. However, it does represent a good approximation that reflects the separation of the dialects. Next, the volume under f which lies between the YZ and the (X = Y)Z planes is calculated. This volume corresponds to the accumulated amount of dialect D_1's correctly classified tokens for which the score against D_1 model is higher than the outside dialect D_2 model, and yields an estimation for the separation between D_1 and D_2:

$$\mathrm{d}(D_1, D_2) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y)\mathrm{d}x\mathrm{d}y. \tag{3}$$

Here, a primary aim is to keep the dialect separation assessment process as simple as possible. Therefore, this procedure can serve as an initial step prior to the actual classification, in order to give the user an estimate of how effective the results of dialect classification might be. All score analysis in the proposed method is based on the training data and therefore represents a close-set test (for D_1 against D_1 GMM, but an open test when D_1 data is tested against D_2 GMM). Figure 3 shows the 3D score distribution for the 2D distributions depicted in Fig. 2. The figure shows the contour distribution of the log likelihood scores from pairs of Arabic dialects. Contours of equal likelihood are projected onto the XY plane in the lower portion, and a bi-secting plane is constructed to determine a decision surface to obtain a dialect separation volume measure.

To calculate $\mathrm{d}(D_2, D_1)$, S_{21} and S_{22} are applied in a similar manner. Next, $\mathrm{d}(D_1, D_2)$ and $\mathrm{d}(D_2, D_1)$ are averaged to obtain one combined proximity measure for the pair of dialects:

$$\overline{D}_{12-21} = \frac{1}{2}\left(\mathrm{d}(D_1, D_2) + \mathrm{d}(D_2, D_1)\right). \tag{4}$$

The range of the proposed measure is [0.5,1]. The smallest distance between two dialects is 0.5, which is the worst case scenario for a dialect identification task. It occurs when the dialects are identical. However, if the data used in training the dialect models is open or separate from that used to assess the separation, the resulting distance may be slightly larger than 0.5, which is due to the variability within a dialect. The greatest distance is 1 which reflects two completely distinct and separated dialects. If the distance is less than 0.5, it implies that D_1 data points are more likely to have the same dialect as D_2. This can occur due to the impurity of the training data. Since one of the objectives of this study is assessment of the training data for dialects, if the distance is lower than 0.5, it means that the majority of D_1 training data is in fact from D_2 and the resulting model is not reliable for classification purposes.

2 Prosodic proximity between dialects

Human perception tests indicate that prosodic cues can be employed to distinguish one language or accent from another (Muthusamy et al. 1994; Kumpf and King 1997). However, prosodic features have only briefly been considered for language identification (LID) systems (Zissman and Berkling 2001), and even less for dialect classification. Thyme-Gobbel and Hutchins (1996) explored the utility of syllable based parameters extracted from pitch and amplitude contours in LID tasks. Their results showed that prosodic cues alone render results comparable to many non-prosodic systems for some language pairs. Tong et al. (2006) integrated different levels of features for language identification, including prosodic features. Their study showed that different levels of discriminative features provide complementary cues for LID. In this section, we explore prosodic differences between dialects/languages, with the primary focus on pitch movement differences. Prosodic features including fundamental frequency patterns, have a suprasegmental nature (i.e., they cannot be associated with a single phone-sized segment) (Wightman and Ostendorf 1994). Therefore, modeling prosody is still an open ended problem (Rouas 2007). Syllable based speech units have been used previously for prosodic feature extraction. However, these approaches use segmentation as front-end processing. Manual segmentation of the speech signal can be costly in terms of time for large corpora. Alternatively, automatic segmentation methods introduce errors which can bias overall results. Rouas (2007) used pseudosyllables as the prosodic units, which are automatically extracted from input audio stream speech data. Adami

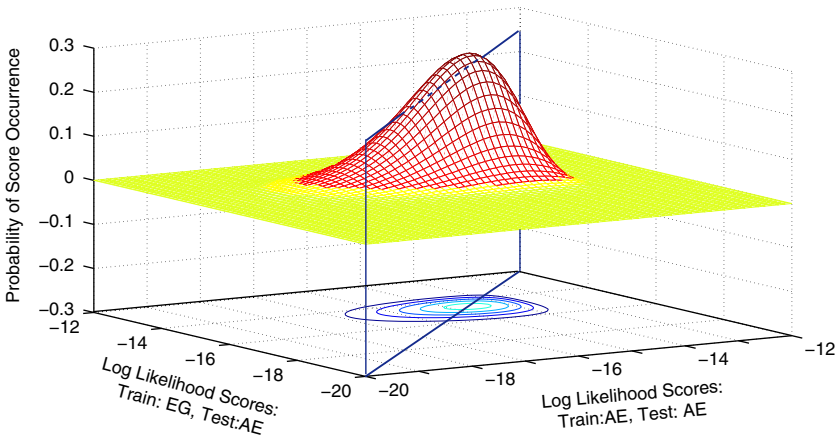

Fig. 3 3D score distribution for likelihood scores of pair-wise dialects in Arabic

et al. (2003) modeled pitch and energy contour dynamics for speaker recognition using linear piecewise stylization.

This study, concentrates on conversational data without any manual labeling or transcription. During conversations between speakers of the same dialect, speakers generally use more dialect specific language compared to directed read data. This has been observed for accent classification by Angkititrakul and Hansen (2006). The method employed here focuses on local variations of the pitch and energy contours which are compared among dialects. The proposed text-independent prosody features do not require segmentation.

A method for dialect separation assessment is proposed which compares statistical models of pitch contour details. An earlier version of this method was presented by Mehrabani et al. (2010). As a first step, a single pitch vector per utterance is obtained by first extracting pitch frequencies from every utterance of each dialect. The robust algorithm for pitch tracking (RAPT), proposed by Talkin (1995) is used for pitch extraction. RAPT is based on the normalized cross-correlation function (NCCF), and applies dynamic programming as a post-processing technique to select the best F_0 and voicing state candidates at each frame. Next, 3-Dimensional feature vectors are generated from groups of three consecutive nonzero pitch values. To obtain a representation of the pitch contour microstructure, rather than speaker/utterance dependent absolute pitch values, pitch slopes are subsequently extracted from the 3-Dimensional pitch vectors. Since the step size in pitch extraction is fixed (10 ms), a feature directly proportional to pitch slope is calculated as the difference between consecutive pitch values, transforming the pitch vector $[F_{0_1}\ F_{0_2}\ F_{0_3}]$ into a 2D vector $[(F_{0_2}-F_{0_1})\ (F_{0_3}-F_{0_2})]$. For the remainder of this study, this extracted feature will be referred to as pitch slope. Figure 4 shows the example feature extraction from a pitch contour. As shown, an analysis window slides along the pitch contour, extracting three nonzero pitch values at a time. There is overlap of two samples between adjacent windows.

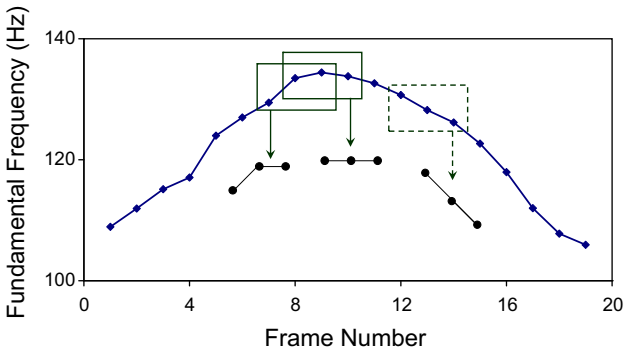

Fig. 4 Extraction of proposed text-independent pitch features from pitch contour

In the next step, the pattern of changes in every three consecutive pitch values is determined, using the 2D pitch slopes. A positive slope implies an increase in pitch, and alternatively, a negative slope represents a decrease. The absolute value of the slope or pitch change is also important. Pitch slopes close to zero, independent of the sign, represent almost flat fragments of the pitch contour. However, steep slopes correspond to abrupt changes in pitch.

In order to obtain a codebook of pitch patterns for 2D pitch slope vectors, a threshold is set for pitch slopes based on studying pitch slope histograms for all dialects. For each dialect, 2D pitch slope feature vectors extracted from every speaker and utterance of the dialect's data are used to build a 3D histogram as the statistical representation of pitch change in that dialect. Figure 5 shows an example of a 3D pitch slope histogram. Each 2D pitch slope vector corresponds to a point on the XY plane.

A set of 9 distinct patterns are considered for each dialect, depicted in Fig. 6. If the absolute change of pitch is less than 3 Hz, the pitch is considered unchanged. However, for absolute pitch slopes greater than 3 Hz, two options are considered: positive and negative.

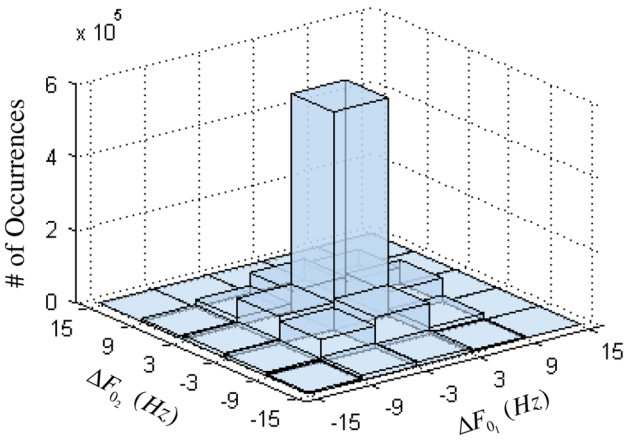

Fig. 5 3D histogram for Egyptian dialect pitch slopes

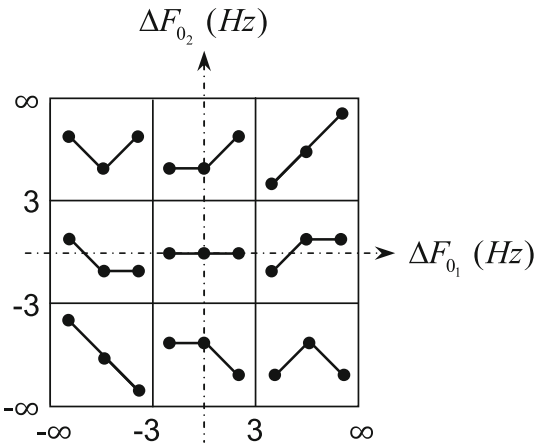

Fig. 6 Dialect assessment based codebook of potential three-frame pitch patterns

After classifying all 2D pitch slope features as one of the 9 pitch patterns, the next step is to model pitch changes in each dialect. These models are later compared to obtain pitch movement differences between different dialects of a language. Statistical models are used here with discrete probability distributions. Each distribution shows the probability of occurrence for each pattern in the given dialect, and can be described by the probability matrix $P(3 \times 3)$. The variability of these distributions reflects overall differences in the excitation structure between dialects.

Next, the obtained pitch pattern model profiles for 3D pitch vectors are exploited to build statistical models for longer temporal pitch patterns by means of N-gram modeling. The codebook of 9 pitch patterns from Fig. 6 is considered as a dictionary of different words: $\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$. The unigram counts for this dictionary are already calculated, which are the probabilities of occurrence for each word (pattern). Conditional probabilities are computed from the N-gram frequency counts:

$$P\left(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}\right) = \frac{C\left(w_{i-(n-1)}, \ldots, w_{i-1}, w_i\right)}{C\left(w_{i-(n-1)}, \ldots, w_{i-1}\right)}, \tag{5}$$

where C represents the count of the word sequences. Using the conditional probabilities, the probability of different sequences of pitch contour patterns can be calculated as,

$$\begin{aligned} P\left(w_1, \ldots, w_m\right) &= \prod_{i=1}^{m} P\left(w_i \mid w_1, \ldots, w_{i-1}\right) \\ &\approx \prod_{i=1}^{m} P\left(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}\right). \end{aligned} \tag{6}$$

The same approach is also performed on energy contours to compare the statistics of log energy contour patterns among dialects, also using an entry codebook such as that in Fig. 6, with a slope threshold of 0.05 in place of the 3 Hz value used for pitch. The KL divergence is used to compare pitch/energy pattern models. If P and Q are two discrete probability distributions, the KL divergence of Q from P is:

$$D_{KL}\left(P \| Q\right) = \sum_i P(i) \log \frac{P(i)}{Q(i)}. \tag{7}$$

Having formulated the statistical models for pitch contour and energy contour distributions, along with their temporal language models in this section, combined with the statistical MFCC based method in the previous section, we now turn to an evaluation of dialect assessment in the next section.

3 Experimental results and evaluation

Results from the proposed approaches for dialect proximity assessment are presented and compared for three Arabic dialects: United Arab Emirates (AE), Syria (SY), and Egypt (EG), as well as three South Indian (Dravidian) languages: Kannada (KAN), Tamil (TAM), and Telugu (TEL). Approximately 3 h of conversational speech from 32 male speakers for each dialect, and 6 h of conversational speech from 74 male speakers for each language was used as train data. Conversations were held between two speakers of the same dialect/language. Each speaker's part of the conversation was recorded separately. The entire training data set was used for this evaluation.

Each assessment consists of three measurements corresponding to the three pairwise dialect/language comparisons. Dialect proximity assessment scores from different methods will naturally have their own numerical ranges. Therefore, the numerical scores in each set are normalized in order for an effective comparison. Figure 7 schematically shows three distances $d_1 \geq d_2 \geq d_3$ between pairs of a set of three

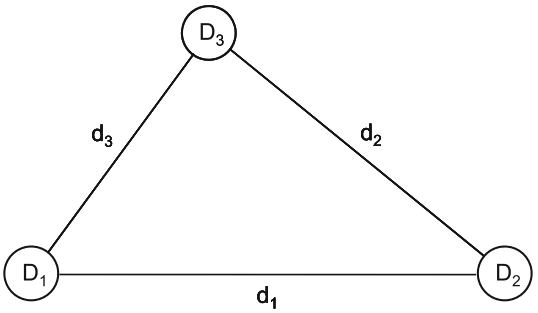

Fig. 7 Schematic distance triangle for a set of three dialects/languages: D_1, D_2, D_3. Here d_1 reflects the combined distance of D_1 to D_2, and D_2 to D_1

dialects/languages, represented as D_1, D_2, D_3. The normalization process consists of subtracting the minimum distance from each distance and dividing by the dynamic range of the distances. Next, each normalized distance is multiplied by 5 and added to 5 in order to map the largest distance (d_1) to 10, and the smallest distance (d_3) to 5:

$$d_i{}' = \frac{d_i - \min\limits_j \left(d_j\right)}{\max\limits_j \left(d_j\right) - \min\limits_j \left(d_j\right)} \times 5 + 5 \tag{8}$$

where $d_i{}'$ represents the final normalized distance, with $i = 1, 2, 3$, using $j = 1, 2, 3$.

Figure 8 shows the normalized measures from the log likelihood score distribution and pitch pattern methods for the 3-way Arabic dialect set. For each method, a set of three distances are depicted as a triangle, which reflects three scores for the dialect pairs: (AE,SY), (AE,EG), (SY,EG). Each vertex represents a dialect, and length of each side is proportional to the distance between dialects represented at the vertices. Since the distances are normalized as explained, in each triangle the length of the largest side is 10 and the length of the smallest side is 5. As shown in the figure, for all three methods, the largest triangle side corresponds to distance between AE and SY, and the smallest side corresponds to distance between SY and EG. In other words, among these three Arabic dialects, AE and SY are the most separate pair, while SY and EG are the closest dialect pair. This means that the log likelihood, pitch pattern unigram, and pitch pattern bigram methods yield the same order of the proximity scores for this corpus. However, the results from pitch pattern unigram comparison are closer to the log likelihood method.

Figure 9 compares the normalized distances from the log likelihood score distribution, pitch pattern bigram, and energy pattern bigram methods for pairs of three South Indian languages. As shown in the figure, KAN and TAM have the largest distance among these language pairs with three methods. Next, repeatability and consistency of the proposed measures and the amount of data required for a reliable proximity

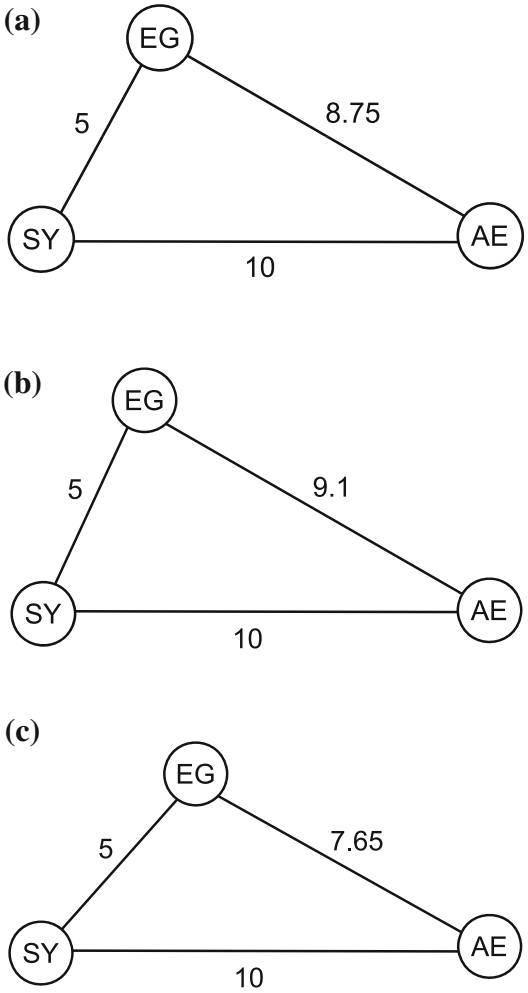

Fig. 8 Comparison between normalized proximity measures for Arabic dialect pairs from **a** log likelihood scores distribution, **b** pitch pattern unigram, and **c** pitch pattern bigram methods. Each triangle corresponds to one assessment method. Triangle sides are proportional to the normalized measures

assessment are evaluated for South Indian languages based on the increased amount of data for this corpus compared to the dialect database.

3.1 Consistency

In this section, the proposed proximity measures are shown to be consistent and repeatable. As mentioned, the dialect separation assessment framework is based on scores derived from the available training data. However, if adequate data is used for the assessment, the resulting dialect proximity measure will be resistant to differences in training data. In order to show this, we performed individual dialect assessment a total of 20 times for the South Indian corpus, where for each pass 24 randomly selected speakers were used out of a total possible 74 speakers. Mean and standard deviation of the results

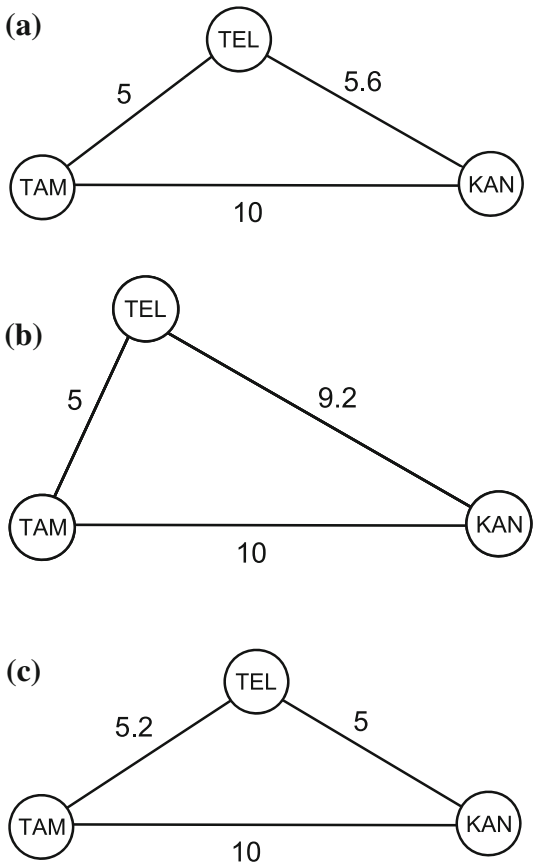

Fig. 9 Comparison between normalized proximity measures for South Indian language pairs from **a** log likelihood scores distribution, **b** pitch pattern bigram, and **c** energy pattern bigram methods

for log likelihood, pitch pattern bigram, and energy pattern bigram measures are shown in Table 1. Note that these results are not normalized. As seen, the Standard Deviation (SD) in the three assessment measures show a minimal change, with SD values ranging from 0.003 to 0.041. This confirms both repeatability and consistency of the assessment methods for dialect separation.

3.2 How much data?

This section explores the changes in proximity measure scores obtained from a set of speakers, as the number of speakers is decreased. Starting with the entire data set consisting of 74 male speakers for each language, with approximately 5 min. conversation data per speaker (approximately 6 h of conversational data per language), the number of speakers are reduced to 60, 48, 36, 24, and finally to 12. The amount of speech per speaker is kept the same. Therefore, the total number of speakers used for each model corresponds to 5, 4, 3, 2, and 1 h of speech data per language, respectively. Changes in log likelihood, pitch pattern bigram, and energy pattern bigram proximity measures when reducing data size from 6 to 1 h are shown in Fig. 10. The three scores from language pairs: (KAN,TAM), (KAN,TEL), and (TAM,TEL) are shown in the same figure for a comparison of the relative changes.

For log likelihood and pitch pattern bigram there is a steady increase in scores as the data size is decreased down to 2 h, while the language pairs with the largest and the smallest proximity scores remain the same. For the energy pattern bigram, the proximity scores are flat until the data drops from 3 to 2 h, suggesting that a minimum of 3 h is required for this method.

3.3 Perceptive evaluation

Finally, in order to assess the correlation between the objective dialect proximity framework and actual dialect separation, a formal listener evaluation is performed. A subjective distance is obtained for three South Indian languages using conversational data. The subjective test consists of 30 experiments. In each experiment, three audio files are presented from these three different languages: KAN, TAM, and TEL. Each audio file is part of a conversation. One of the three audio files is represented as the reference in each experiment. Listeners are asked to compare the other two samples to the reference and decide which sample sounds more like the reference. Each listener, is asked to provide two distances in each experiment on a scale of 1 (similar to the reference) to 10 (completely different). The reference language changes in a random way among experiments. In order to remove any bias based on knowledge/familiarity of the language in the listener group, an equal number of subjects with their native language of Kannada, Tamil, and Telugu are used. However, many native speakers of one South Indian language, speak or understand other south Indian languages. Walter (2009) showed how the listener's native language or their familiarity

Table 1 Means and standard deviations of 20 measures for South Indian language pairs, where for each pass only using 24 randomly selected speakers out of 74 speakers

	Log likelihood		Pitch pattern bigram		Energy pattern bigram	
	Mean	SD	Mean	SD	Mean	SD
(KAN,TAM)	0.816	0.034	0.013	0.005	0.061	0.016
(KAN,TEL)	0.760	0.041	0.015	0.005	0.040	0.015
(TAM.TEL)	0.761	0.021	0.007	0.003	0.056	0.028

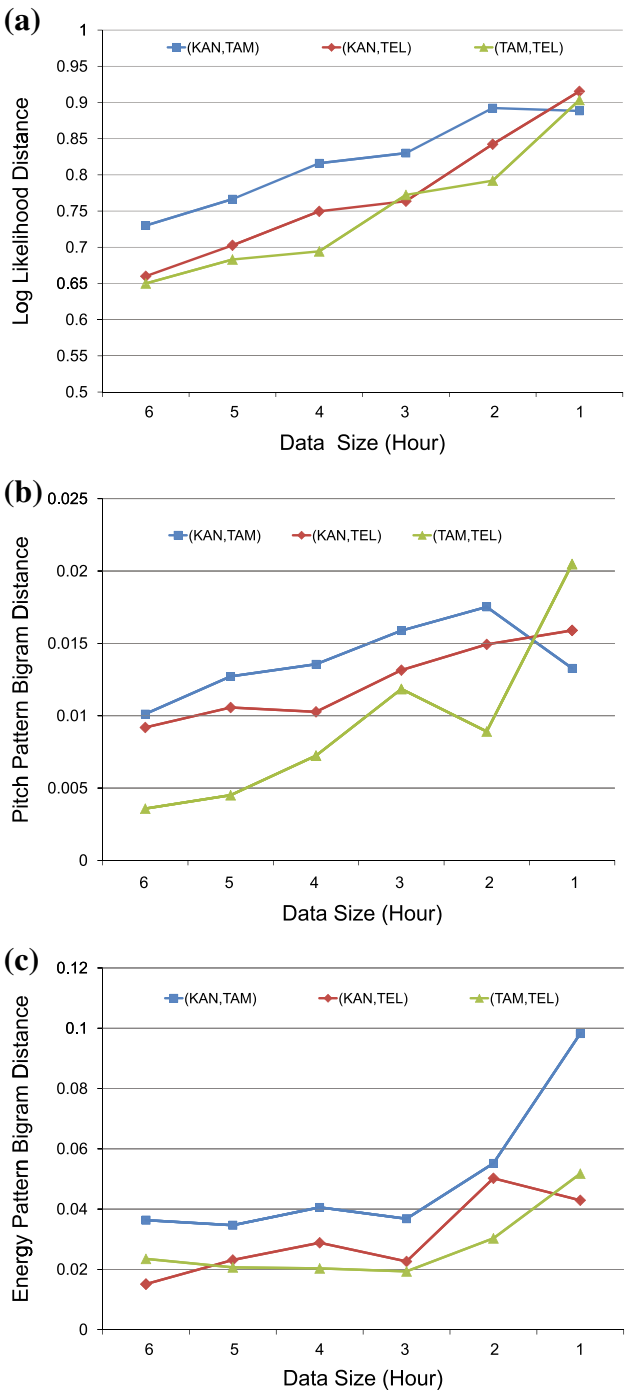

Fig. 10 Changes in **a** log likelihood **b** pitch pattern bigram and **c** energy pattern bigram proximity measures for South Indian language pairs, by reducing the available data size from 6 to 1 h

with another language affects their perception of sound based similarity between languages. The total number of subjects is 15. For each language pair L_1 and L_2, the average subjective distance is calculated as the mean of all the distances: d(L_1, L_2) and d(L_2, L_1) from all the experiments and listeners with native language either L_1 or L_2. The resulting subjective distances for (KAN,TAM), (KAN,TEL), (TAM,TEL) are 6.13, 5.13, 5.75, respectively. While the number of listeners is limited, the 3-way perceptual scores reflect a relative separation for the south Indian languages. Between these three language pairs, KAN and TAM have the largest perceptual distance, which is consistent with the proposed objective measures. Clearly, further listener evaluations would be necessary to draw a statistical measure of significance.

4 Conclusions

In this study, the goal of assessing dialect/language proximity in a 3-way set was considered. Intrinsic differences between dialects were studied, including spectral acoustic, as well as excitation structure differences. First, a method for measuring dialect separation was proposed based on a volume space analysis in a 3D model for GMM output score distributions. Next, prosody-based proximity measures were proposed, comparing statistical models for pitch/energy contour movement patterns between dialects. The proposed measures were evaluated on a corpus of Arabic dialects and a corpus of South Indian languages. The proposed dialect proximity assessment was shown to be consistent and repeatable. Future advances could consider lexical differences including word selection, grammar structure, or wider suprasegmental differences. Further high level linguistic knowledge concerning the evolution of specific dialects could also be considered for future dialect assessment strategies.

Acknowledgments This project was supported in part by the AFRL under Contract FA8750-12-1-0188 and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen.

References

Adami, A. G., Mihaescu, R., Reynolds, D. A., & Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition. In *Proceedings of theIEEE ICASSP* (Vol. 4, pp. 788–791).

Angkititrakul, P., & Hansen, J. H. L. (2006). Advances in phone-based modeling for automatic accent classification. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 634–646.

Biadsy, F., Moreno, P. J., & Jansche, M. (2012). Google's cross-dialect arabic voice search. In *Proceedings of the IEEE ICASSP* (pp. 4441–4444).

Bradlow, A. (2008). Sound structure and function of english as a global language. *The Journal of the Acoustical Society of America*, *123*, 3879(A).

Curzan, A., & Adams, M. (2006). *How english works : A linguistic introduction*. New York: Pearson Education Inc.

Faber, A., Best, C., & Di Paolo, M. (1994). Dialect differences in vowel production and perception. *The Journal of the Acoustical Society of America*, *96*, 3283(A).

Heeringa, W. (2004). *Measuring dialect pronunciation differences using levenshtein distance* (pp. 121–277). Netherlands: Groningen.

Heeringa, W., & Hinskens, F. (2012). The measurement of dutch dialect change in the sound components. *Dialectological and Folk Dialectological Concepts of Space: Current Methods and Perspectives in Sociolinguistic Research on Dialect Change*, *17*, 250.

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006, Jul.). Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances* (pp. 51–62). Sydney.

Kohler, J. (1996, Oct.). Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proceedings of the icslp* (Vol. 4, p. 2195–2198). Philadelphia.

Kotz, S., & Nadarajah, S. (2000). *Extreme value distributions: Theory and applications* (pp. 61–95). London: Imperial College Press.

Kumpf, K., & King, R. W. (1997). Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Eurospeech* (pp. 2323–2326). Rhodes, Greece.

Mehrabani, M., Boril, H., & Hansen, J. (2010). Dialect distance assessment method based on comparison of pitch pattern statistical models. In *Proceedings of the IEEE ICASSP* (pp. 5158–5161). Dallas, TX.

Mehrabani, M., & Hansen, J. H. L. (2008, Sep.). Dialect separation assessment using log-likelihood score distribution. In *Proceedings of the interspeech* (pp. 747–750). Brisbane.

Muthusamy, Y. K., Jain, N., & Cole, R. A. (1994). Perceptual benchmarks for automatic language identifications. In *Proceedings of the ICASSP* (Vol. 1, pp. 333–336). Adelaide.

Nallasamy, U., Garbus, M., Metze, F., Jin, Q., Schaaf, T., & Schultz, T. (2011). Analysis of dialectal influence in pan-arabic asr. In *Interspeech* (pp. 1721–1724).

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass*, *3*(1), 175–198.

Nerbonne, J., & Heeringa, W. (2001). Computational comparison and classification of dialects. *Journal of the International Society for Dialectology and Geolinguistics*, *9*, 69–83.

Nerbonne, J., Heeringa, W., van den Hout, E., van der Kooi, P., Otten, S., & Van De Vis, W. (1996). Phonetic distance between Dutch dialects. In *Clin vi, papers from the sixth clin meeting* (pp. 185–202).

Rouas, J.-L. (2007). Automatic prosodic variations modeling for language and dialect discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(6), 1904–1911.

Shackleton, R. G. (2007). Phonetic variation in the traditional english dialects a computational analysis. *Journal of English Linguistics*, *35*(1), 30–102.

Sooful, J. J., & Botha, E. C. (2001). An acoustic distance measure for automatic cross-language phoneme mapping. In *Proceedings of the prasa* (pp. 99–102).

Talkin, D. (1995). *Speech coding and synthesis* (pp. 495–518). Amsterdam: Elsevier.

Thyme-Gobbel, A. E., & Hutchins, S. E. (1996). on using prosodic cues in automatic language identification. In *Proceedings of the ICLSP* (Vol. 3, pp. 1768–1772). Philadelphia.

Tong, R., Ma, B., Zhu, D., Li, H., & Chng, E. S. (2006). Integrating acoustic, prosodic and phonotactic features for spoken language identification. In *Proceedings of the ICASSP* (pp. 205–208). Toulouse.

Walter, M. (2009). Cross-linguistic variation in language similarity classification. *The Journal of the Acoustical Society of America*, *125*, 2756(A).

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS one*, *6*(9), e23613.

Wieling, M., Shackleton, R. G., & Nerbonne, J. (2013). Analyzing phonetic variation in the traditional english dialects: Simultaneously clustering dialects and phonetic features. *Literary and Linguistic Computing*, *28*(1), 31–41.

Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 31–44.

Yin, B., Ambikairajah, E., & Chen, F. (2007). Hierarchical language identification based on automatic language clustering. In *Proceedings of the interspeech* (pp. 178–181). Antwerp.

Zissman, M. A., & Berkling, K. M. (2001). Automatic language identification. *Speech Communication*, *35*(1–2), 115–124.