DIALECT DISTANCE ASSESSMENT METHOD BASED ON COMPARISON OF PITCH PATTERN STATISTICAL MODELS

Mahnoosh Mehrabani, Hynek Bořil, John H.L. Hansen[*]

Center for Robust Speech Systems, Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, USA

ABSTRACT

Dialect variations of a language have a severe impact on the performance of speech systems. Therefore, knowing how close or diverse dialects are in a given language space provides useful information to predict, or improve, system performance when there is a mismatch between train and test data. Distance measures have been used in several applications of speech processing. However, apart from phonetic measures, little if any work has been done on dialect distance measurement. This study explores differences in pitch movement microstructure among dialects. A method of dialect distance assessment based on pitch patterns modeled progressively from pitch contour primitives is proposed. The presented method does not require any manual labeling and is text-independent. The KL divergence is employed to compare the resulting statistical models. The proposed scheme is evaluated on a corpus of Arabic dialects, and shown to be consistent with the results from the spectral-based dialect classification system. Finally, it is also shown using a perceptive evaluation that the proposed objective approach correlates well with subjective distances.

Index Terms— dialect distance assessment, pitch patterns

1. INTRODUCTION

Dialect is a variety of a language that is used by a group of speakers belonging to some geographical region. Dialects of a language differ in phonetic, grammatical, and lexical features. Like other speaker variations, dialect impacts the performance of speech systems. Therefore, efficient dialect classification algorithms will contribute to improved speech recognition, speaker identification, speech coding, or spoken document retrieval systems. Compared to language identification in which a dictionary and set of language rules are known, dialect identification (ID) is more challenging. In a dialect ID task, dialect-dependent models are trained, and during the test phase, the model which is most likely to produce the test utterance is identified. For both train and test phases, feature vectors are extracted from audio files. The availability of data transcription influences the design of a dialect ID system. For unsupervised dialect classification, systems based on Gaussian Mixture Models (GMMs) have proven to be successful [1].

In this study, our focus is on estimating the proximity or separation between different dialects of the same language by means of a distance measurement. Distance measures have been applied in different fields of speech processing. In speech recognition, from measuring the distortion between input and output [2], [3] to speaker adaptation and speaker clustering [4], measures of similarity have played a significant role in improving system performance. Other areas of speech processing, such as speech coding, enhancement, and synthesis have exploited distances as an objective measure of assessing speech quality [5]. Phonetic distance between dialects has been calculated in several linguistic studies using various string distances including Levenshtein, Euclidean, and Manhattan distance [6]. The obtained distances have been applied in order to divide geographical maps into dialect areas. Apart from linguistic approaches, little if any work has been done on finding a meaningful distance measure between dialects.

Previous studies have shown that pitch movement can provide a cue for accent classification [7] (perceptual study), dialect classification [8] (analysis of pitch contours in English dialects), and language identification [9] (automatic language ID). In this paper, we propose a probabilistic method to compare statistical models for pitch movement patterns between dialects. We start by proposing text-independent pitch features based on pitch contour primitives, and train a model for each dialect based on the proposed features. Next, the pitch primitive models are used as the building blocks to model longer-term pitch patterns in different dialects by means of *N*-grams. There are two points to be noted here. First, the proposed features are text-independent, allowing dialect comparison in an unsupervised manner based on available un-transcribed train data. Second, our efforts are to keep the assessment system as computationally inexpensive as possible, therefore it can serve as an initial step prior to actual classification.

The proposed dialect distance assessment framework suggests how accurately the dialects can be distinguished. Therefore, it provides some sense of the resulting dialect classification system performance, and takes an initial step towards dialect purity assessment. Furthermore, the performance of a dialect-dependent speech recognition system for a new dialect can be estimated based on the distance between dialects. In a previous study [10], we assessed dialect separation comparing log-likelihood score distributions. GMMs were applied as statistical models for each dialect, and Mel Frequency Cepstral Coefficients (MFCCs) were used as extracted features from audio files. In this study, we show that the pitch pattern based separation assessment is consistent with the log-likelihood score distribution distance for the same corpus. We present the proposed distance measure in Sec. 2. In Sec. 3, evaluation on a corpus of Arabic dialects is discussed. We show the repeatability of presented measure, and its correlation with human perception. Conclusions are drawn in Sec. 4. For the remainder of this paper, we use ”dialect distance” and ”dialect separation” interchangeably, and the word ”distance” is not used in the strict sense of metric spaces.

2. PROPOSED METHOD

Human perception tests indicate that prosodic cues, including pitch movements, can be employed to distinguish one language or accent from another [7], [11]. However, prosodic features have only briefly been considered in language ID systems [12], and even less in dialect classification. In the present study, we hypothesize that variation of pitch contours are dialect-dependent and can provide an efficient cue for dialect separation assessment. The presented scheme does not require any prior knowledge, manual segmentation, or pre-processing.

[*] This project was supported in part by USAF under a subcontract to RADC, Inc. under FA8750-05-C-0029. Approved for public release; distribution unlimited.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

2.1. Text-independent Pitch Features

As noted, our objective is to develop an unsupervised system that automatically assesses the separation between dialects based on available train data. The system's input is un-transcribed conversational audio, and the task is to compare different dialects on the basis of pitch movements. Our approach statistically models details of the pitch contour in voiced speech data for each dialect. As a first step, pitch frequencies are extracted from every utterance of each dialect to obtain a single pitch vector per utterance. The pitch is obtained using Robust Algorithm for Pitch Tracking (RAPT), proposed in [13]. RAPT is based on the Normalized Cross-Correlation Function (NCCF), and applies dynamic programming as the 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 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 msec.), 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, the extracted feature is referred to as pitch slope vector. Note the difference from the traditional meaning where pitch slope refers to the long-term trend of a pitch contour.

Fig. 1 shows feature extraction from a pitch contour example. As shown, a window slides along the pitch contour, extracting three nonzero pitches at a time. Note that there is overlap of two samples between adjacent windows. Each 3D pitch vector yields a 2D pitch slope vector which is then classified as one of 9 pitch patterns, which are introduced in the next subsection.

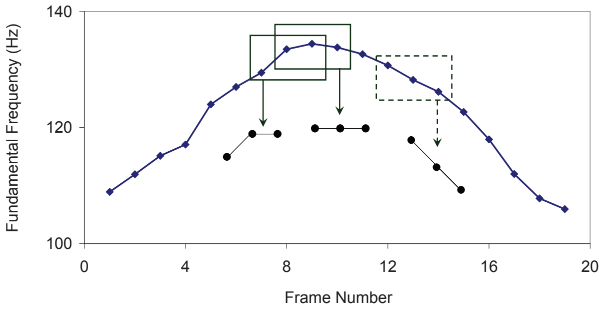


Fig. 1: *Extraction of proposed text-independent pitch features from pitch contour example.*

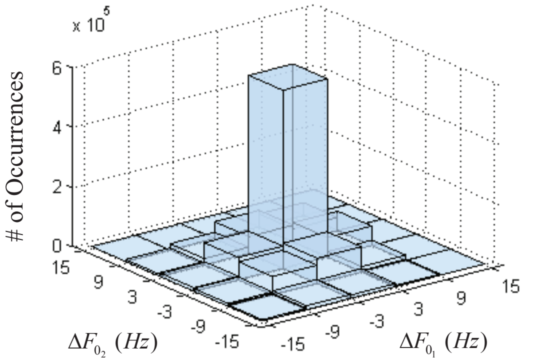


Fig. 2: *Histogram for Egyptian dialect pitch slope vectors.*

2.2. Pitch Patterns Codebook

After extracting 2D pitch slope vectors for each dialect, the pattern of changes from every three consecutive pitches is determined. A positive slope means an increase in pitch, and alternatively, a negative slope represents a decrease. In addition, the absolute value of the slope or pitch change is important. For example, steep slopes correspond to abrupt changes in pitch while slopes close to zero (independent of the sign) reflect steady fragments of the pitch contour. Based on the inspection of pitch slope vector histograms for all dialects , we set thresholds for 2D pitch slope vectors to obtain a codebook of pitch patterns. For each dialect, 2D pitch slope vectors extracted from every speaker and utterance of dialect's data are used to build a 3D histogram as statistical representation of pitch change in that dialect. Fig. 2 shows an example of a pitch slope vector histogram. Each 2D pitch slope vector corresponds to a point on the XY plane. A set of 9 different patterns are considered for each dialect, depicted in Fig. 3. If the absolute change of pitch is less than 3 Hz, the pitch is considered steady. However, for absolute pitch slopes greater than 3 Hz, two options are considered: positive and negative. A pitch contour description utilizing a set of simple shapes was previously used in [14]. 7 patterns were applied in manual labeling of 100 msec.–2 sec. speech segments in an analysis of infant speech production. Note that in our approach, the pitch patterns represent fine details of the pitch contour (20 msec. resolution) and are extracted in a fully automatic manner.

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

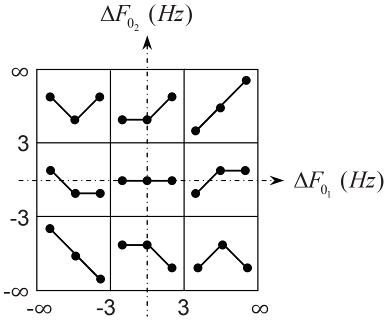


Fig. 3: *Codebook of pitch patterns.*

Furthermore, the obtained pitch patterns from 3D pitch vectors can be extended to higher dimensions (i.e., longer temporal patterns) by means of *N*-gram modeling. First, consider the codebook of 9 pitch patterns as a dictionary of different words: $\{w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9\}$. We already have the unigrams for this dictionary which are the probabilities of occurrence for each word (pattern). Next, conditional probabilities are computed from *N*-gram frequency counts:

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

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

$$\begin{aligned} &P\left(w_1,\ldots,w_m\right)=\prod_{i=1}^{m} P\left(\left.w_i\right| w_1,\ldots,w_{i-1}\right) \\ &\approx \prod_{i=1}^{m} P\left(\left.w_i\right| w_{i-(n-1)},\ldots,w_{i-1}\right). \end{aligned} \tag{2}$$

Here, we calculate the probabilities for different pairs of words which correspond to 3D pitch slope vector patterns. These results, combined with the uni-gram counts, are applied to compute conditional probabilities:

$$P\left(\left.w_i\right| w_j\right)=\frac{C\left(w_j,w_i\right)}{C\left(w_j\right)}, i,j=1,2,...,9. \tag{3}$$

Next, bi-gram models are used to calculate probability distribution for sequences of three words, or 5D pitch vectors.

2.3. Distance Between Pitch Pattern Models

The Kullback Leibler (KL) divergence or relative entropy [15] is a non-commutative measure of similarity/dissimilarity between distributions or statistical models. If P and Q are two discrete probability distributions, the KL divergence of Q from P is:

$$D_{KL}\left(P\|\,Q\right)=\sum_i P(i)\log\frac{P(i)}{Q(i)}. \tag{4}$$

In Sec. 2.2, we modeled pitch patterns for each dialect as a 2D discrete distribution. The next step is to compare dialect models for a measure of dialect distance. We used KL divergence for comparing the distributions, which in this case has a closed form. The distance of dialect2 (D_2) from dialect1 (D_1) is:

$$d(D_1,D_2)=\sum_{i=1}^{3}\sum_{j=1}^{3} P_1(i,j)\log\frac{P_1(i,j)}{P_2(i,j)}, \tag{5}$$

where $P_1(i,j)$ and $P_2(i,j)$ are discrete pitch pattern distributions for dialect1 and dialect2, respectively. Note that $d(D_1,D_2)$ is not necessarily equal to $d(D_2,D_1)$. Therefore, we average the two distances to obtain a separation assessment between two dialects. Fig. 4 shows the block diagram of proposed dialect distance assessment method.

3. EXPERIMENTAL RESULTS AND EVALUATION

3.1. Evaluation

Here, the distance assessment scheme is investigated for a corpus of three Arabic dialects: AE (United Arab Emirates), EG (Egypt), and SY (Syria). Our focus is to keep the training data balanced, (i.e., for each dialect almost 5 hours of conversational speech from 32 male speakers is used). The computed distances using all available data are as follows: d(AE,EG) = 0.0036, d(AE,SY)=0.0043, d(EG,SY) = 0.00018. The distances show that AE and SY have the widest separation, while EG and SY are the closest dialects. This is the same observation that resulted from previously proposed log-likelihood distances [10]. In the next step, we evaluate the consistency of the proposed distance measure. Distances for 10 non-overlapping subsets extracted from the original data are calculated. Each subset contains data from all speakers. The results are summarized in Table 1. The first row shows distances using the entire available data. In the second row, mean and standard deviation of distances obtained from 10 subsets are shown. It can be seen that the mean distances extracted from the subsets are comparable to the distances obtained from the whole set. The low values of standard deviation compared to the distance means confirm the consistency of the distance estimation across the subsets. In [10], we evaluated log-likelihood score distribution distances with results from an open-set GMM-based dialect classification task. 600 mixtures and 26-dimensional MFCC features were used for classification. A confusion score was defined between each dialect pair D_1 and D_2 as the sum of percentages of D_1 classified as D_2 and vice versa. We discussed that the greater the dialect distances, the less the confusion score. Therefore, in this study, inverse confusion is used as a reference to show how well distance measures can predict dialect classification system performance. Fig. 5 shows pitch pattern distances for pairs of three Arabic dialects, and compares them with log-likelihood distances, and inverse confusion scores from dialect ID system. Since each set of scores by nature has a different range, normalized values are presented in the figure. Normalization process for each set of three distances consists of subtracting the mean and dividing by the standard deviation. Next, minimum normalized distance in each set is used as a bias to obtain nonnegative distances:

Set	d(AE,SY) (x10^{-3})	d(AE,EG) (x10^{-3})	d(EG,SY) (x10^{-3})
Whole Set	4.3	3.6	0.18
10 Subsets	4.4 (σ = 0.6)	3.7 (σ = 0.5)	0.22 (σ = 0.07)

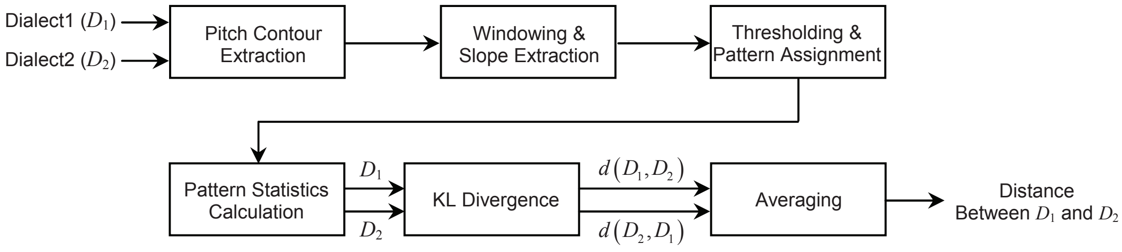
Table 1: Means and variances of 10 distance measures using subsets of the whole data set

$$d_i'=\frac{d_i-\mu}{\sigma}-\frac{\min\limits_i\left(d_i\right)-\mu}{\sigma}=\frac{d_i-\min\limits_i\left(d_i\right)}{\sigma}, \tag{6}$$

where d_i and d_i' are original and normalized distances, respectively, $i=1,2,3$. μ is the mean of three distances in each set, and σ is the standard deviation. It is shown in Fig. 5 that log-likelihood score distribution distances are closest to inverse confusion. This is due to the fact that these distances are derived from GMM score distributions of dialect classification system. In addition, MFCCs are the features used for both classification, and log-likelihood distance measure. Distances from pitch pattern comparison, which represent excitation differences between dialects, display good correlation with inverse confusion scores. This confirms that automatic dialect ID system performance can be well estimated using proposed pitch pattern distances. Note that this study's approach is only based on statistical analysis of pitch contour details and does not require any information from dialect ID system. Extending length of the pitch patterns by means of bi-gram modeling results in distances closer to inverse confusion.

3.2. Perceptive Evaluation

Finally, in this section correlation of objective dialect distance with human perception is investigated. For this experiment, two Egyptian subjects are used. Each subjective test consists of 30 sessions. In each session, three 15 sec. conversations are presented from three different dialects (AE, EG, SY). One audio file is used as the reference. Listeners were asked to compare the two other utterances to the reference and on a scale of 1 (similar to the reference) to 10 (completely different from the reference) give two perceptual distances for each session. The reference dialect in each session is chosen in a random way. To make the decisions as speaker independent as possible, different speakers are used for each session. The perceived distances between each two dialects are averaged across sessions to obtain one



Dialect1 (D_1)
Dialect2 (D_2)
Pitch Contour Extraction
Windowing & Slope Extraction
Thresholding & Pattern Assignment
Pattern Statistics Calculation
D_1
D_2
KL Divergence
$d(D_1, D_2)$
$d(D_2, D_1)$
Averaging
Distance Between D_1 and D_2

Fig. 4: *Block diagram of proposed method.*

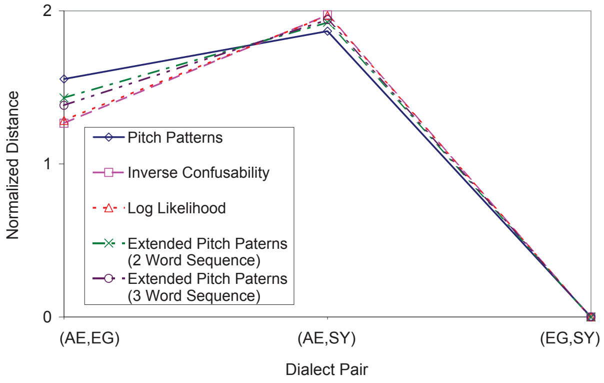


Fig. 5: *Comparison between dialect distance measures.*

perceptive distance per listener. The resultant subjective distances from both listeners show that perceptually, SY is closer to EG than AE to EG. This is the same order of distances that we obtained from the proposed objective distance measure. The average distance from 2 listeners between SY and EG is 6.4, while between AE and EG is 7.2. We clearly recognize that this perceptive test is limited, and more listeners are needed to show the correlation of the proposed distance measure with human perception. In addition, the number of listeners from each dialect should be balanced. Here, the listener group is biased based on knowledge/familiarity of the dialects. Since the native dialect of the subjects is Egyptian their judgment on comparing the other two dialects with their native dialect is more reliable.

4. CONCLUSIONS

In this study, a method for assessing dialect separation based on comparing pitch movement patterns was proposed. 2D pitch slope vectors were classified into 9 patterns of pitch change. Statistical pitch pattern models were compared to obtain dialect distances. Using bigrams, models for longer temporal patterns were derived. The proposed distance measure was evaluated for three Arabic dialects. The results showed that AE dialect's pitch movements are completely distinguishable from the other two dialects (EG and SY), while EG and SY are more confusable. Dialect classification system performance for these three dialects confirms the result of the proposed distance measure. The correlation of the distances with human perception was also investigated in a listener test. The proposed method of measuring dialect distance has applications in dialect classification, performance prediction, as well as dialect data purity assessment. Moreover, it is believed that the newly established statistical modeling of pitch contour used side-by-side with spectral-acoustic features will benefit automatic dialect identification.

References

[1] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, “Dialect identification using Gaussian mixture models,” in *ODYSSEY: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 2977–300.

[2] D. Mansour and B. H. Juang, “A family of distortion measures based upon projection operation for robust speech recognition,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1659–1671, Nov. 1989.

[3] B. A. Carlson and M. A. Clements, “A projection-based likelihood measure for speech recognition in noise,” *IEEE Trans. on Speech and Audio Proc.*, vol. 2, no. 1, pp. 97–102, Jan. 1994.

[4] M. Padmanabhan, L. R. Bahl, D. Nahamoo, and M. A. Picheny, “Speaker clustering and transformation for speaker adaptation in speech recognition systems,” *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 71–77, Jan. 1998.

[5] S. R. Quackenbush, T. P. Barnelwell III, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, 1988.

[6] W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne, “Evaluation of string distance algorithms for dialectology,” in *Proc. of Workshop on Linguistic Distances*, Jul. 2006, pp. 51–62.

[7] K. Kumpf and R. W. King, “Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks,” in *Eurospeech*, 1997, pp. 2323–2326.

[8] D. Bolinger, *Intonational and its uses. Melody in Grammar and Discourse*, Stanford Univ. Press, 1989.

[9] A. E. Thyme-Gobbel and S. E. Hutchins, “On using prosodic cues in automatic language identification,” in *International Conference on Spoken Language Processing*, 1996, vol. 3.

[10] M. Mehrabani and J. H. L. Hansen, “Dialect separation assessment using log-likelihood score distribution,” in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008.

[11] Y. K. Muthusamy, N. Jain, and R. A. Cole, “Perceptual benchmarks for automatic language identifications,” in *Proc. ICASSP*, 1994, vol. 1.

[12] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, “Integrating acoustic, prosodic and phonotactic features for spoken language identification,” in *Proc. ICASSP*, 2006, pp. 205–208.

[13] D. Talkin, *Speech Coding and Synthesis*, chapter A Robust Algorithm for Pitch Tracking (RAPT). Kleijn and Paliwal (Eds.), pp. 495–518, Elsevier, Amsterdam, Netherlands, 1995.

[14] R. D. Kent and A. D. Murray, “Acoustic features of infant vocalic utterances at 3, 6, and 9 months,” *The Journal of the Acous. Soc. of America*, vol. 72, no. 2, pp. 353–365, 1982.

[15] S. Kullback, *Information Theory and Statistics*, Dover Publications Inc., New York, 1968.