Turkish Regional Dialect Recognition Using Acoustic Features of Voiced Segments

Baran Uslu
Atilim University, Electrical and Electronics Engineering, Ankara, Turkey
Email: baran.uslu@atilim.edu.tr

Hakan Tora
Atilim University, Electrical and Electronics Engineering / Avionics, Ankara, Turkey
Email: hakan.tora@atilim.edu.tr

***Abstract*—In this study, we aimed to classify Turkish dialects by machine learning methods. The Black Sea, Aegean and Eastern dialects were examined in the scope of this paper. First of all, the speeches belonging to the relevant regions were obtained by recording. Then the speech records were examined in detail by Praat and Matlab software. The attributes that characterize these three regional dialects were selected as pitch, jitter and shimmer. These features were then used for training an artificial neural network classifier. Classification rates of 98.9% for training set and 95.8% for test set were achieved. Also, this study bridges between the two disciplines of Linguistics and Speech Processing for Turkish dialect recognition.**

***Index Terms*—turkish dialect, regional accent identification, jitter, shimmer, neural network**

I. Introduction

It is well known that the factors of speech disorder, age, gender and emotional state play an important role in the way that people speak. In addition, people in a different region of a country speak their own language in a different way. Therefore, regional varieties influence the speaking style as well. These varieties are referred to as dialects. Dialect is mostly related with the speed, loudness and intonation of the speech. In other words, it may be viewed as dancing of utterances in harmony. How are these variations characterized in the speech signal or which features of the signal describe dialect? There have been many studies dealing with these questions in the literature. In [1], Li *et al.* classifies the stress and emotion using the MFCC (Mel Frequency Cepstral Coefficients) and jitter/shimmer parameters together. Lazaridis *et al.* [2] uses an SVM (Support Vector Machine) to identify regional Swiss French dialect employing nine features (amplitude tilt, duration tilt, the number of voiced samples, the number of unvoiced samples, DLOP coefficients, etc.) along with jitter, shimmer and intensity. Ullah and Karray [3], [4], classified two dialect regions of American English (Northern Midland and Western). They introduced a comparison of GMM, HMM, VQ-GMM and their proposed approach which uses features from learning distance metric in an evolutionary-based K-means clustering algorithm [3]. They used MFCC, the first three formant frequencies and energy as features [4]. In [5], Yusnita *et al.* proposes an approach classifying accents in Malaysian English using formants and linear predictive coefficients (LPC). It is well known that MFCCs are mostly used for speech recognition and speaker verification. Although they are employed for accent classification, they are not preferred due to their noise sensitivity. Therefore, in this paper we consider only pitch, jitter and shimmer as features that represent the prosodic variations in speech for dialect recognition. These features, along with some other prosodic features, were also used for speaker recognition in [6] and analysis of mimicked speech in [7].

Turkic language is the fifth among the most spoken languages in the world. The area in which Turkic languages are spoken extends from Eastern Europe, through Turkey and its neighbors, to Eastern Turkistan and farther into China and expands to North and South Siberia. There exist currently twenty Turkic standard languages. The most important ones can be listed as Turkish, Azerbaijanian, Turkmen, Kazak, Karakalpak, Kirghiz, Uzbek, Uyghur, Tuvan, Yakut, Tatar, Bashkir and Chuvash [8]. In this study, only the regional differences of Turkish spoken in Turkey were discussed. As in other languages, Turkish also has a number of different dialects according to the regions. Researchers from linguistics have done many studies on lexical and phonological differences with respect to linguistics [9], [10]. Our study aims to establish a link between the two societies, linguistics and speech processing. In other words, lexical and phonological varieties are described in terms of acoustic attributes. Thus, this study offers researchers from both societies an opportunity to exchange their wide knowledge for a better analysis of Turkish dialects.

Existing studies take a fusion of features from different domains (prosodic and phonetic) into consideration for accent classification. However, we focus only on prosodic features of pitch, jitter and shimmer for Turkish

Manuscript received May 10, 2018; revised June 1, 2018.

doi: 10.18178/ijsps.6.2.17-21

regional dialect recognition in this study. They are used to train a multilayer neural network for recognizing regional dialects. Considering that only these three features are employed for training, our study differs from the others. Additionally, the study herein can be seen as one of the leading works implementing Turkish dialect recognition by machine learning. The map in Figure 1 illustrates the seven geographical regions of Turkey. Within the scope of this study, three regional varieties of Turkish spoken in Turkey were considered. They are Aegean, Black Sea and Eastern dialects.

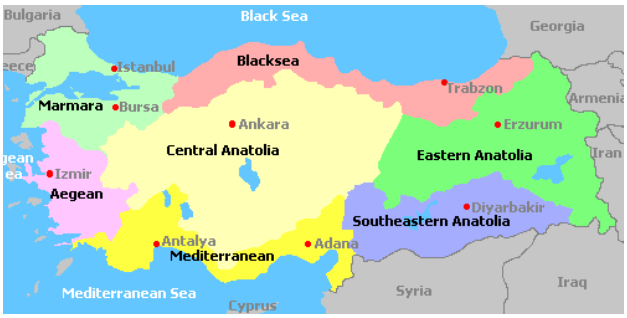

Figure 1. Turkey's regions map.

This paper is organized as follows. Section 2 introduces our approach. Experimental works and results are given in Section 3. Finally, conclusion is presented in Section 4.

II. Methodology

A pattern classification system always consists of three components: data, features and classifier. In our case, voice records from the regions of Aegean, Black Sea and Eastern are collected as input data. Then, the features are extracted from the records. They completely characterize the data. Finally, a classifier is trained to identify the regional dialects using the features.

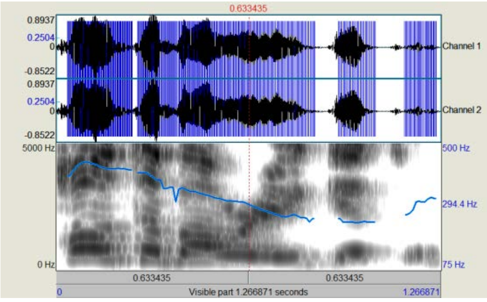

Figure 2. Typical Black Sea dialect with voiced parts and pitch trajectory

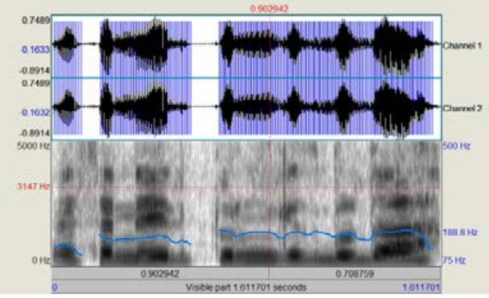

Figure 3. Typical Aegean dialect with voiced parts and pitch trajectory.

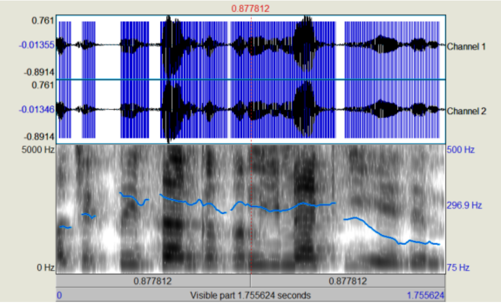

Figure 4. Typical Eastern dialect with voiced parts and pitch trajectory.

Figures 2, 3 and 4 illustrate typical recordings from Black Sea, Aegean and Eastern regions in Turkey, respectively. Stereo records labelled with pitch pulses can be seen on the upper panel, while pitch contours are shown on the lower panel in the figures. There are significant differences between the regions when we examine the pitch contours. The Black Sea dialect usually starts with a high pitch, carrying the stress, at the beginning of the utterance and then decreases gradually. On the other hand, the Aegean dialect shows an almost flat pitch contour. Intonation is usually added to the utterance not with pitch variations but with some specific words in the Aegean dialect. Meanwhile in the Eastern dialect, the pitch contour increases, makes ripples in the body and finally decreases.

The following subsections detail each component of the system.

A. *Collecting Speech Records and Calculating Features*

In order to assess the performance of our approach, we created our own dataset because there is no available Turkish dialect dataset, ready-to-use like TIMIT. Consequently, we made a severe effort to collect the proper dialects from the different regions of interest. This part placed a burden on the study and took much time. The speech records were collected by directly recording from the speaker who was born and grown up in the region to be identified. Cell phones were used for this purpose. Since all the records were sampled at different frequencies, before selecting the features, each one was resampled at 44100 Hz. 10 speakers from each region were asked to speak 10 different sentences. It is also worth noting that the recorded texts from the three regions are not the same, either. Hence, the regional examples are not correlated one another. In respect to forming the training data, we do not place any limitation on the training set. In other words, our methodology is text-independent.

Pitch, jitter and shimmer are chosen as features in our work. Fundamental frequency, also known as pitch frequency, does not remain constant but varies in continuous speech. It corresponds to the bas and timbre parts of the speech. Jitter is the shift that occurs in the pitch frequency. This can be perceived as detuning. Shimmer is the amplitude (loudness) changes occurring in the pitch periods. This can be perceived as distortion in the voice intensity. The sum of the pitch periods results in a measure of the length of the voiced parts in the speech. Therefore, intonation or pitch contour of a speech signal carries information about speaking style. As expected, regional dialects indicate some specific patterns. The different dialects may have a projection on the related pitch, jitter and shimmer parameters. Therefore, we focus on examining the regional effects on these three parameters. For example, bas sounds are more dominant in Eastern Turkey, people in the North usually speak quickly and the ones in some part of the West speak slowly.

The basic formulas of the jitter and shimmer parameters that we used are given as:

$$jitter(abs) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \qquad (1)$$

$$shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1}/A_{i)}| \qquad (2)$$

where T_i denotes the pitch period of the i^{th} cycle, A_i denotes the maximum loudness value in the same cycle and N is the total number of the cycles in a voiced segment of interest.

For every speech file, we selected the voiced segments and calculated the values of pitch, jitter and shimmer for each segment by using the software Praat [8]. Figure 5 illustrates pitch cycles with T_i and A_i variations. Pitch periods and maximum amplitudes in each cycle generally demonstrate slight difference depending on the intonation of speech. Table 1 depicts the values indicated in Figure 5. These variations actually introduce an intonation to speech. Therefore, one can witness some beneficial patterns belonging to the regional accents.

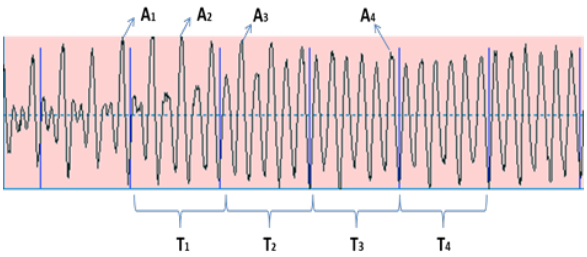

Figure 5. Pitch cycles.

TABLE I. REPRESENTATIVE T AND A VALUES

Pitch Period (ms)	Amplitude
$T_1 = 7.463$	$A_1 = 0.6281$
$T_2 = 7.421$	$A_2 = 0.6331$
$T_3 = 7.505$	$A_3 = 0.6381$
$T_4 = 7.505$	$A_4 = 0.6504$

Table II shows the selected features for our classification. They represent the characteristics of regional speeches. The features of J1 through J5 indicate jitter values, S1 to S6 are shimmer values, and P1 to P5 are pitch values. As a result, we obtained a set of features, 16 in total, whose definitions are given below [11]:

Jitter(local): This is the average absolute difference between consecutive periods, divided by the average period.

Jitter(abs): This is the average absolute difference between consecutive periods, in seconds.

Jitter(rap): This is the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period.

Jitter(ppq5): This is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period.

Jitter(ddp): This is the average absolute difference between consecutive differences between consecutive periods, divided by the average period.

Shimmer(local): This is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

Shimmer(dB): This is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20.

Shimmer(apq3): This is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.

Shimmer(apq5): This is the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude.

Shimmer(apq11): This is the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude.

Shimmer(dda): This is the average absolute difference between consecutive differences between the amplitudes of consecutive periods.

TABLE II. THE SET OF USED FEATURES

Feature Code	Feature Name
F_1	J1- local
F_2	J2-abs
F_3	J3-rap
F_4	J4-ppq5
F_5	J5-ddp
F_6	S1-local
F_7	S2-dB
F_8	S3-apq3
F_9	S4-apq5
F_{10}	S5-apq11
F_{11}	S6-dda
F_{12}	P1-med
F_{13}	P2-mean
F_{14}	P3-std
F_{15}	P4-min
F_{16}	P5-max

B. *Classifier: Multilayer Perceptron*

A neural network (NN) with multilayer perceptron is employed to identify the regional speeches. We build an NN with 16 inputs, one hidden layer with 20 neurons, and an output layer with 3 neurons. Its structure is shown in Figure 6. This network is trained by 16 features defined in previous section. Each output of the network corresponds to a region to which the input speech belongs. Sigmoid as an activation function is used in both hidden and output neurons. The NN is implemented in Matlab.

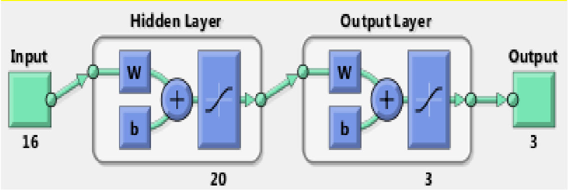

Figure 6. Neural network structure.

III. Experimental Works

Collecting speech samples from the three regions of Turkey has been already discussed in Section II. Our database consists of 10 different utterances from each region. Thus, we have 30 recordings in total. The sentences were selected such that each record includes at least 10 voiced segments. Hence, the number of the voiced parts is set to 10 in order to equalize the number of the feature vectors for each region. As a result, we obtain a set of 300 vectors in total. 60% of them are utilized for the training phase and 40% for the test phase. Note that the feature vector must not be confused with the recorded speech sentences. Indeed, each vector represents a voiced part of the selected record.

TABLE III. Performance Scores in Training

	Black Sea	**Aegean**	**Eastern**
Black Sea	60 (100%)	0	0
Aegean	0	59 (98.3%)	1
Eastern	0	1	59 (98.3%)

TABLE IV. Performance Scores in Test

	Black Sea	**Aegean**	**Eastern**
Black Sea	39 (97.5%)	0	1
Aegean	0	38 (95%)	2
Eastern	1	1	38 (95%)

The results for training and test sets are shown in Table III and Table IV, respectively. As can be seen, speech samples of Black Sea for training set are not confused with the others. However, two samples from the other two regions are misclassified. Overall performance of the system is achieved as 98.9%. On the other hand, for the test set, more misclassifications among the regional samples occur. These obtained preliminary results encourage us to make further studies on Turkish dialects. In this paper, we empirically show that only the three features are enough to distinguish the Turkish three regional speeches from each other.

Since we were unable to reach or find a study on Turkish dialect recognition using the speech processing techniques, unfortunately we can not present a comparison regarding to Turkish regional accents. Therefore, we compare our study with the ones conducted for foreign languages. The study in [2] recognizes four regional dialects in Switzerland. SVM is trained by using two different feature sets. They employed phonetic features in the first set. The second set was constructed by phonetic features along with jitter, shimmer and energy. They achieved 32.90% and 37.13% accuracies for these sets, respectively. In [4], the authors classified two regional dialects in American English. They obtained a classification accuracy rate of 75%. The preceding two studies are similar to ours because they are also handling regional dialects in a language. On the other hand, the work in [5] focuses on the foreign accent classification. The authors identified the three-class accents in Malaysian English. They accomplished 89.33% for Malay accent, 84.22% for Chinese accent and 84% for Indian accent.

IV. Conclusion

Applications where linguistics and computing technology cooperate are missing in the literature for Turkish. This study fills that gap by trying to classify three Turkish regional dialects using only pitch, jitter and shimmer parameters. The preliminary results show that these three parameters are sufficient to classify the regional dialects successfully. To our best knowledge, this is one of the pioneer works on Turkish dialects by utilizing a machine learning algorithm. As a future work, we plan to recognize dialects from Marmara, Central Anatolia and Mediterranean regions.

References

[1] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman, "Stress and emotion classification using jitter and shimmer features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2007, pp. 1081-1084.

[2] A. Lazaridis, J. P. Goldman, M. Avanzi, and P. N. Garner, "Syllable-based regional Swiss French dialect identification using prosodic features," *Nouveaux Cahiers de Linguistique Francaise*, 2014.

[3] S. Ullah and F. Karray, "An evolutionary approach for accent classification in IVR systems," in *Proc. IEEE International Conference on Systems, Man and Cybernetics*, 2008, pp. 418-423.

[4] Ullah and F. Karray, "Speaker accent classification using distance metric learning approach," in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, 2007, pp. 900-905.

[5] M. A. Yusnita, P. M. Pandiyan, Y. Sazali, A. B. Shahriman, and M. N. Fadzilah, "Statistical formant descriptors with linear

predictive coefficients for accent classification," in *Proc. 8th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2013, pp. 906-911.

[6] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Proc. 8th Annual Conference of the International Speech Communication Association ISCA*, Antwerp (Belgium), 2007, pp. 778-781.

[7] L. Mary, K. K. A. Babu, and A. Joseph, "Analysis and detection of mimicked speech based on prosodic features," in *International Journal of Speech Technology*, vol. 153, pp. 407-417, 2012.

[8] L. Johanson, "Discoveries on the Turkic linguistic map," *Swedish Research Institute in Istanbul*, Publications 5, 2001.

[9] M. Karlık and A. Akbarov, "Three regional dialects in Turkey," *International Journal of Language and Linguistics*, vol. 3, no. 6, 2015.

[10] M. Demirci, "Gender differences in the perception of Turkish regional dialects," in *Handbook of Perceptual Dialectology*, 2002.

[11] P. Boersma and D. Weenink, (2012). Praat, version 6.0.29 [Online]. Available: http://www.fon.hum.uva.nl/praat/

Baran Uslu received his B.Sc. and PhD. degrees from Ankara University and M.Sc. degree from Baskent University in Electrical and Electronics Engineering in Turkey. He has been working at Atilim University since 2012 and actively conducting research on speech signal processing, especially Text-to-Speech synthesis systems, prosody analysis, and pattern recognition.

Hakan Tora received his B.Sc. and M.Sc. degrees in Electrical and Electronics Engineering from Anadolu University in Turkey and Ph.D. from Vanderbilt University in USA. He has been working at Atilim University since 2002. His research interests are speech and image processing, neural networks, and machine learning algorithms.