Automatic Identification of Speech Disorders for Arabic Children Speakers

Abualsoud Hanani[1], Mays Attari[1], Atta' Farakhna[1], Aseel Joma'a[1], Mohammed Hussein[1]

, Stephen Taylor[2]

[1]Birzeit University, Palestine
[2]Fitchburg state university, USA
ahanani@birzeit.edu,{attarimays,frakhna.atta,aseelmohammad066}@gmail.com,
staylor@fitchburgstate.edu

Abstract

Abstract—Automatic identification of speech disorders in children's speech is very important for the diagnosis and monitoring of speech therapy. In this work, acoustic features (MFCC) have been used with the two most commonly used classification techniques in the speaker and language identification area, GMM-UBM and I-vector, for identifying three error types of speech production associated with phoneme [r] from Arabic children's speech. The sound [r] has been selected as it is the most common pronunciation problem that Arabic speaking children suffer from. The impact of [r] location in a word on the speech disorders has been investigated by considering words with [r] in the beginning, middle and end. The best performance of our 4-class systems, is 75% accuracy with our I-vector system and 61% for our GMM-UBM system. Performance of these two systems are improved to 92.5% and 83.4%, respectively, when the three disorder classes are combined into one class versus normal class (2-class systems).

Index Terms: Automatic diagnosis of Arabic speech disorders, speech disorders, articulation disorders, GMM, i-vector.

1. Introduction

Articulation is the process of producing speech sound. It is the movement of mouth muscles and articulators (lips, tongue, teeth, jaw, soft palate, hard palate, and the rear part of the roof of the mouth) to make the sounds of speech. Each configuration of articulators generates different sounds. Airflow is initially produced by respiratory system and passes through larynx. The vocal folds may or may not vibrate depending on whether the produced sound is voiced (e.g. [i] vowel) or unvoiced (e.g. [s] fricative consonant). The way of moving articulators for producing specific sound depends on the context and the "surrounding context" (i.e., the other sounds which are produced before and after the sound being considered). Speakers make small adjustments (known as co-articulations) to adjust for this context. All speakers do this but in fairly similar and relatively predictable ways. Some of these adjustments are rate dependent, meaning different kinds of adjustments are made in slow rate vs. typical rate vs. fast rates of production. There are certainly differences from speaker to speaker but these don't reflect differences in the manner of production, but rather differences in the anatomy.

There is no one standard way for the correct pronunciation of a phoneme in a specific language. However, sounds, syllables and words must be recognized correctly by humans.

Because the correct pronunciation depends on many physiological and contextual factors, many articulation problems can occur. Some of these problems occur due to difficulty a speaker is experiencing which reflects challenges with physical production.

Errors in what we hear in an individual's output may also reflect an incomplete or incorrect understanding of the way in which sounds are used within a language. For example, the child who says "white" instead of "light" and at the same time also says "white" instead of "right" may not yet appreciate that [w], [l], and [r] are different sounds used to create independent meaning within the language. In addition) they may or may not notice the difference between these sounds. But even if they notice the difference they may not realize that using those three sounds in separate ways results in differences in meaning (i.e., they may think that the three sounds can be easily interchanged with no consequences for meaning).

A child who doesn't notice the difference between particular sounds is said to have a perceptual disorder. A child who doesn't understand how all the sounds in the language are used to create different meaning is said to have a phonological disorder. The child who has failed to learn how to physically produce the sound is said to have an articulation disorder. This three-way distinction of speech disorder is not always possible to sort out. To avoid confusion between these, the field of speech-language pathology groups all three under the generic term "speech sound disorders".

Articulation disorders may become more obvious in rapid sound production in a normal conversation. Articulation disorders occur for many reasons ranging from physical handicaps, such as cerebral palsy, cleft palate, or hearing loss, to other problems in the mouth, such as dental problems or tongue-tie (ankyloglossia). Speech disorders may range from a mild lisp to nearly unintelligible speech. Although articulation problems affect both children and adults, young children often experience such problems during their language development. Bilingual children are more likely to have speech disorder in their second language [1]. According to [1], five to eight percent of preschool children in USA experience speech disorders, making this the most common category of childhood disabilities. On the other hand, childhood speech disorders are the most treatable disabilities when detected early.

Many studies have focused on the quality assessment of speech disorder. There are two main methods for disorder assessment, namely, perceptual assessment and objective analysis assessment [2-3]. In the former method, a clinician listens to the patient's voice to describe and measure the speech disorder. This remains the most common method used by

http://dx.doi.org/10.21437/WOCCI.2016-6

clinicians and it is widely discussed in the literature. In contrast, in the objective analysis assessment method, acoustic features are extracted from speech and are used with an automatic classification system for making a decision. Most of the published work in this field is trying to recognize whether speech is normal or has some disorder without identifying the precise category of speech disorder. Some other studies have focused on improvement of the performance of these systems [4-8].

In this work, we are proposing an automatic system for not only classifying disorder and normal speech, but also to identify the speech error type from a given speech segment. More specifically, proposed systems identify whether speech is normal or has one of the speech error types: distortion, substitution, deletion and addition.

In contrast to other languages, very little research in the field of speech disorders detection has focused on the Arabic language [9-10]. This paper describes our proposed system for detecting speech disorder types from Arabic children's speech using the state of the art techniques applied in the field of speaker, language and accent identifications such as Gaussian Mixture Models (GMM) and the I-vector technique [11-12].

In this study, we specifically focused on the pronunciation disorders of the Arabic phoneme [ر] (corresponding to [r] in the International Phonetic Alphabet (IPA)), as it is considered as one of the hardest and the most common Arabic speech disorders among pre-school children. We perform the study combining the three Arabic vowels with this phone and for the same phone in the word at different positions (start of the word, middle of the word, and end of the word). This is because the speech disorders for [r] are different depending on its place in the word. For example, some children have deletion disorder of [r] when it comes at the end of a word, but not in the beginning or in the middle.

The rest of the paper is organized as follows: types of articulation disorders are illustrated in section 2. A description of collected speech data and systems overview are presented in sections 3 and 4. Experiments and results are presented and discussed in section 5. Section 6 describes experiments and results when combining disorder classes into one class against normal class. Paper conclusion and references are presented in sections 7 and 8.

2. Children speech disorder

Speech disorders can be classified based on the outward appearance of the sound, such as: childhood apraxia of speech, dysarthria, articulation and phonological processes, stuttering, spastic speech, voice disorders, fluency disorders and organic disorders. In this study, we are focusing on articulation and phonological non-organic speech disorder for children. These kinds of speech disorder are usually treated by correcting linguistic behavior of the child and pronunciation exercises. Discovering speech disorder for children early will help much in the treatment.

Generally, error types of speech production among children can be classified as one of the following four categories:

A. Distortion

Voice distortion includes sound pronunciation which is slightly different from the correct sound. It often appears in the sound of certain phonemes such as ([s], [ʃ]), where [s] is accompanied by a sound of a long whistle, or when one pronounces a sound of [ʃ] from the side of the mouth and tongue.

For example, children with this kind of disorder pronounce the word /sajara/ (a car) as /ɵajara/.

This may occur as a result of tooth loss, the tongue is not put in its proper position during pronunciation, deviation of teeth position or loss of teeth on both sides of the lower jaw, which makes air go to both sides of the jaw so that the child cannot pronounce sounds like [s] and [z].

B. Deletion

In this type of speech disorder, children drop one or more phonemes from a word making their speech very difficult to understand. This kind of disorder is more common for preschool children than school children. Children experiencing this kind of disorder often drop phonemes from the end of a word rather than beginning or middle of a word.

For example, the word /bader/ (moon) is pronounced as /bade/.

C. Substitution

Substitution occurs when children pronounce an inappropriate sound for a phoneme instead of desired sound. For example, the word /faraʃe/ (butterfly) is pronounced as /faɣaʃe/. In this word, the sound [r] ر is substituted by sound [ɣ], غ.

D. Addition

In this speech disorder category, children add extra sound (phone) to the word; the sound was heard like the one repeated, for example, the word /radʒʊl/ (man) is pronounced as /rrradʒʊl/.

In some cases, distortion and substitution could be considered as one type of disorder, for example, pronouncing /faraʃe/ as /faɣaʃe/ can be considered as a distortion where the phone [r] was changed to [ɣ]; if it occurs in every word containing [r] then it is also considered as a substitution.

This study focused on four classes of speech disorders associated with sound [r] in its three positions in a word (beginning, middle, and end). Disorder classes include; substitution [r] by [l], substitution of [r] by [ɣ], deletion of [r] and adding extra [r].

3. Data description

In order to get samples for each error type of [r] sound, we visited different clinics specialized in children speech disorder including audiology and speech disorder center at Birzeit University.

We have recorded around 260 school children ages 5-12 years whose native language is Arabic from three primary schools. We explained teachers the purpose of this study, and asked them to help in selecting children for recording. Each of the selected children was asked to pronounce three carefully selected words presented with an image shown on a card, containing phoneme [r] ('ر') in the beginning, middle and the end. Pronunciations of the three selected words are /rʔas/ (head; رأس), /faraʃhe/ (butterfly; فراشة), and /bader/ (moon; بدر), respectively. Two speech therapists, from audiology and speech disorder center at Birzeit University, were used as experts to evaluate and categorize error type of each recorded child. Each therapist listen to all recorded words and does the evaluation independently. Words with disagreement between two therapists were discussed and re-evaluated by the two therapist to agree on one category. Samples with low-quality

	Normal	Sub. by /l/	Sub. by /ɣ /	Del.	Total
/**r**ʔas/	96	50	53	50	249
/fa**r**aʃhe/	95	51	53	50	249
/bade**r**/	96	58	60	52	266
Sum	287	159	166	152	**764**

Table 1: *data set description*

recording are excluded from our dataset. Moreover, Samples with addition error type are very few, and are, therefore, excluded from the following experiments.

Out of this process, a total of 764 word productions have been used, 287 normal pronunciations and 477 containing one type of speech disorders. More details of the data set are shown in table 1 above. Since this amount of data is relatively small for dividing into two subsets; training and testing, a kind of cross-validation strategy was used for training and testing our systems. This means, for the set of words beginning with [r] (second row in table 1), one word production from each class was left for testing and the remaining (245) were used for training. This process was repeated 50 times, resulting in 200 testing segments for each system. The same procedure was used for the other two words (i.e. /faraʃhe/ and /bader/).

4. System overview

4.1. Feature extraction

The first stage in any speech pattern classification process is to convert the speech waveform into a sequence of acoustic feature vectors. 12-dimensional Mel Frequency Cepstral Coefficients (MFCC) features were extracted from 20ms frames, with a frame shift of 10ms. Each feature vector is 36-dimensional, comprising of 12 features plus 12 "delta" and 12 'delta-delta' parameters, giving 36 features per frame at frame rate of 100 frames per second. RASTA filtration is applied to the power spectra and feature mean and variance normalization at spoken word level was applied on the final feature vectors.

4.2. GMM-UBM

Gaussian Mixture Model is widely and successfully used in various speech processing applications such as speaker, language and accent identification [11-13]. It is also used for classifying normal and impaired speech [14] and automatic assessment of language environment [15]. This motivated us to use it for identifying speech disorder types, in addition to normal speech.

A Universal Background Model (UBM) is a GMM trained on acoustic features (36 MFCCs) extracted from all training word productions of the four classes. The K-means clustering algorithm is used for finding initial parameters of UBM GMM (means, diagonal covariance matrices and weights). GMM parameters were soft tuned by 4 iterations of EM algorithm.

A class-dependent GMM is obtained by MAP (Maximum A Posteriori) adaptation (means only) of the UBM using the class specific enrollment features. The result is one UBM model and three class-dependent models, one for each class.

We have tried a different number of Gaussians for GMMs: 16, 32, 64, and 128. UBM with 64 components is found to have the best performance. Consequently, UBM of 64 components is used in all of subsequent GMM-UBM experiments presented in this paper.

4.3. I-vector system

Our second speech disorder identification system is based on I-vectors, a technique introduced in [16] for speaker identification. This technique has also been proven to work well in language and accent identification [11, 17]. An I-vector classifier is based on a configuration determined by the size of the UBM, the number of factor dimensions for the total variability subspace, as well as the various compensation methods to attenuate within-class speaker variability. Feature vectors of each word production in the training and testing data are used for adapting means of the UBM (which is trained on all available training data) in order to estimate a word dependent GMM using the eigenvoice adaptation technique.

The eigenvoice adaptation technique assumes that all the pertinent variability is captured by a low rank rectangular matrix T known as total variability matrix. The GMM supervector M (vector created by concatenating all mean vectors from the word dependent GMM) for a given word production can be modeled as follows:

$$M = m + Tx + \varepsilon \quad (1)$$

Where m is the UBM supervector, the I-vector x is a random vector having a normal distribution N (0, I), and the residual noise term $\varepsilon \sim N(0, \Sigma)$ models the variability not captured by the matrix T. In training the total variability matrix for disorder speech identification, we assumed that every word production for a given disorder class is considered a different class. Additional details on the I-vector extraction procedure are described in [16].

Linear Discriminant Analysis (LDA) is used for reducing the I-vectors dimension. The LDA procedure consists of finding the basis that maximizes the inter-classes variability while minimizing the intra-class variability. Each class is represented by mean of its I-vectors produced form training dataset after LDA dimension reduction.

4.4. Visualization

I-vector system maps a word production into a 100 dimensional vector space for classification. To obtain insight into how I-vector works, this space can be visualized by projecting it onto a suitable 2-dimensional subspace using LDA. Fig. 1 shows 2-dimentional I-vectors representing four classes of phoneme [r] at beginning of Arabic word /rʔas/ (head). It is clear from the figure that substitution speech disorder is close to normal speech, while deletion is far away in the representative space. A possible explanation of this is that pronunciation of [ɣ] and [l] is close to [r], therefore realization of words with substitute [ɣ] and substitute /l] are close to each other and also close to correct pronunciation of [r]. On the other hand, deleting [r] from a word starting with [r] makes its realization very far from the correct pronunciation, hence difficult for understanding..

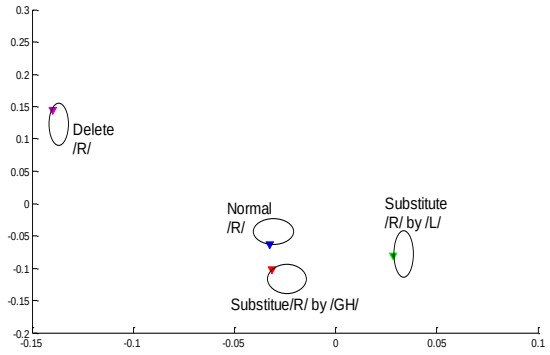

Figure 1: Visualization of average I-vectors of four /r/ classes

5. Experiments and results

5.1. Experimental setup

As we have declared earlier, we are interested in classifying error types of phone [r] at three positions in a word; beginning, middle and end. Therefore, we built three GMM-UBM systems and three I-vector systems, one for each case, i.e. one system for identifying disorders of [r] when it comes in the beginning of a word (column 2 in tables 2 & 3), one for disorders of [r] when it comes in the middle (column 3 in tables 2 & 3) and one when it comes at the end (column 4 in tables 2 & 3). In each case, the systems are trained and tested independently using corresponding word production.

One UBM was represented as a diagonal covariance GMM with different number of components range from 8 to 64. It was trained on all the speech segments of the training set for the three positions of [r], balanced over the four classes. The variance flooring was used in each iteration of EM algorithm during the UBM training. This UBM was used for both GMM-UBM and I-vector disorder speech identification systems.

For each case of [r] position, all class-specific training words were used for MAP adapting UBM means with relevance factor 16. The adapted means with the UBM covariance's and weights form a class-dependent GMM.

In the recognition phase, acoustic features are extracted from testing spoken words and then evaluated against four class-dependent GMMs. The class model which gives the highest log-likelihood of a given test word is the identified class.

For the I-vector classifier of the three positions of [r], various UBM sizes (16, 32, 64, and 128) and different factor dimensions (50 to 200 in steps of 50) were investigated. I-Vectors are derived directly from the 36-dimensional feature vectors over an entire spoken word. The feature vectors from training and testing words are used to construct one I-vector per word production.

During testing, I-vectors of a word production are projected to low-dimensional space by LDA projection matrix. Evaluation scores are calculated by computing dot product of testing I-vectors by representative means of each class, after applying LDA dimensionality reduction. For all experiments, the best LDA dimension reduction is dimension equal to 50.

5.2. Results and discussion

A GMM-UBM system was used for identifying speech disorders of [r] in the three selected words. That is, three GMM-UBM systems were trained and evaluated on the recordings of the three words, i.e. one system for recognizing four classes of [r] in word /rʔas/, one for word /faraʃe/ and one for word /bader/. Results of each GMM-UBM system with a different order of UBM are presented in table 2 below.

Similarly, three I-vector systems have been trained and evaluated on speech of the three words, in the same way for GMM-UBM systems described earlier.

In order to study the effect of number of total variability factors, i.e. eigenvoices, on the performance of our disorder speech identification system, four different numbers of dimensions were tried (50, 100, 150 and 200) using UBM of 16 components and word production with [r] in the beginning. Results of these experiments suggest that total variability with 100 dimensions gives the best performance. Subsequently, 100 is fixed for all of our I-vector experiments.

UBM order	/r/ at the beginning	/r/ in the middle	/r/ at the end
8	59	35	34.3
16	60.8	40.2	39
32	58.2	34	35.7
64	56.6	32.8	33.2

Table 2: *Performance (accuracy %) of GMM-UBM systems with different orders and different positions of /r/ phoneme.*

The same UBMs with different number of Gaussians, which were used in GMM-UBM system described earlier, were also used for training I-vector systems.
Results of I-vector system for each case of [r] position and each order of UBM, are presented in table 3, below..

UBM order	/r/ at the beginning	/r/ in the middle	/r/ at the end
8	65.8	59.2	58.3
16	75	65	62.5
32	73.3	60.8	60.8
64	62.5	51.7	44.2

Table 3: *performance [accuracy %] of I-vector system with different UBM order and different positions of /r/.*

The table shows that UBM with 16 components gives the best performance for two identification systems and for the three positions of [r]. As expected, I-vector system outperformed GMM-UBM system for the three positions of [r]. The best result, 75%, was achieved by I-vector system when detecting disorders of [r] in the beginning of a word, using UBM of 16 components.

It is also interesting to note from these results that recognizing disorder of [r] seems to be easier when it comes at the beginning of a word than when it comes in the middle or in the end. A possible explanation for this, is that the pronunciation of [r] when it comes at the beginning needs more emphasis than when it comes in between or after different sounds. Therefore, the influence of the preceding and following sounds on [r] makes it less clear.

5.3. Normal-disorder two-class system

In this system, we combine the training dataset of the three classes of speech disorders; substitute by [l], substitute by [ɣ] and deletion, into one class, called 'disorder', and keep the 'normal' class. This system is used to classify testing word production as normal speech or disorder speech. Our GMM-UBM and I-vector, with GMM order 16, systems (described above) are re-configured and re-trained for this two-class classification task. The results of these experiments are shown in table 4.

Performance of our two systems are significantly improved when we reconfigure our problem as a two-class task for distinguishing between disorder and normal speech. A possible explanation for this improvement could be stated in the following two ways: The task of classifying disorder speech and normal speech (i.e. two classes) is easier than classifying four classes (three types of disorder and normal). Secondly, the amount of training data is larger for the two-class system.

System	[r] at the beginning	[r] in the middle	[r] at the end
GMM_UBM	83.4	82.6	79
I-vector	92.5	78	77

Table 4: *Performance (accuracy %) of GMM-UBM and I-vector systems with different positions of [r].*

6. Conclusion

The aim of this paper was to evaluate the automated recognition of speech disorders associated with [r] phoneme from Arabic children's speech. We have used and presented the GMM-UBM and I-vector based systems which are used commonly in speaker, language and accent recognitions. Both systems were used for recognizing disorders of [r] in three different places in a word (begin, middle, end). The best performance (75%) was obtained by I-vector system trained and evaluated on words with [r] in the beginning. We also conclude that performance of our two systems is much better when they used only to separate speech into normal and abnormal classes.

The preliminary results are promising, but to prove the usefulness of our method, a larger study with larger dataset is needed.

7. References

[1] L. M. Bedore and E. D. Pena, Assessment of Bilingual Children for Identification of Language Impairment: Current Findings and Implications for Practice, the International Journal of Bilingual Education and Bilingualism, Vol. 11, No. 1, 1-29, 2008.

[2] P. Dejonckere, P. Bradley, P. Clemente, G. Cornut, L. Crevier-Buchman, G. Friedrich, P. Van De Heyning, M. Remacle, V. Woisard, "A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques," Guide lines elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS) 258, 77–82 (2001).

[3] J. Revis, A. Ghio, and A. Giovanni, "Phonetic labeling of dysphonia: a new perspective in perceptual voice analysis," in 7th Int. Conf. Advances in Quantitative Laryngology, Voice and Speech Research, Oct. 2006 [Online]; http://aune.lpl.univaix.fr/~ghio/doc/Bib2006AQLRevis.pdf: (last access 23.02.2015).

[4] N. Ramou, and M. Guerti. "Automatic detection of articulations disorders from children's speech preliminary study." Journal of Communications Technology and Electronics 59.11 (2014): 1274-1279.

[5] G. Georgoulas, V. C. Georgopoulos, and C. D. Stylios. "Speech sound classification and detection of articulation disorders with support vector machines and wavelets." Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE, 2006, pp. 2199-2202.

[6] C. Hernandez Espinosa, M. Fernandez Redondo, P. Gomez Vilda, J. I. GodinoLlorente, and S. Aguil era Navarro, "Diagnosis of vocal and voice disorders by speech signal," in Neural Networks, IEEEINNS ENNS Int. Joint Conf., 2000 (IEEE, New York, 2000),Vol. 4, pp. 253–258.

[7] J. Y. Lee, S. Jeong, and M. Hahn, "Classification of pathological and normal voice based on linear discriminant analysis," Comput. Sci. 4432, 382–390 (2007).

[8] A. Maier, F. Hoenig, T. Bocklet, E. Nöth, F. Stelzle, E. Nkenke, and M. Schuster, "Automatic detection of articulation disorders in children with cleft lip and palate," J. Acoust. Soc. Am. 126, 2589–2602 (2009).

[9] S. A. Farag, M. El Adawy, and A. I. Shahin. "A computer-aided speech disorders correction system for Arabic language." Advances in Biomedical Engineering (ICABME), 2013 2nd International Conference on. IEEE, 2013, pp. 18-21.

[10] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using MFCC and GMM," GCC Conference and Exhibition (GCC), 2013 7th IEEE, Doha, 2013, pp. 292-297.

[11] Hanani, Abualsoud, Martin J. Russell, and Michael J. Carey. "Human and computer recognition of regional accents and ethnic groups from British English speech." Computer Speech & Language 27.1 (2013): 59-74.

[12] Bahari, Mohamad Hasan, Rahim Saeidi, and David Van Leeuwen. "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech." 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 7344-7348.

[13] A. Hanani, M. J. Carey, M. J. Russell. Improved language recognition using mixture components statistics. In INTERSPEECH 2010, 2010, pp. 741-744.

[14] A. Zulfiqar, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam. "Vocal fold disorder detection based on continuous speech by using MFCC and GMM." In GCC Conference and Exhibition (GCC), 7th IEEE 2013, pp. 292-297.

[15] M. Najafian, D. Irvin, Y. Luo, B. S. Rous, and J. H. L. Hansen, "Employing speech and location information for automatic assessment of child language environments," in International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE), 2016, pp. 65–69.

[16] Glembek, O., Burget, L., Matějka, P., Karafiát, M., & Kenny, P. "Simplification and optimization of i-vector extraction." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 4516-4519.

[17] D. M. Gonzlez, O. Plchot, L. Burget, O. Glembek, and P. Matejka,"Language recognition in i-vectors space." in Interspeech. ISCA, 2011, pp. 861–864.

[18] Hanani, A., Basha, H., Sharaf, Y., & Taylor, S."Palestinian Arabic regional accent recognition." Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. IEEE, 2015, pp.1-6.