International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 8 (2018) pp. 5633-5639

Automatic Speech Attribute Detection of Arabic Language

Hager Morsy[1], Mostafa Shahin[2,*], Naif Aljohani[3], Mahmoud Shoman[1] and Sherif Abdou[1]

[1]Information Technology Department, Faculty of Computers and Information, Cairo University Cairo, Egypt.
[2] Department of Electrical and Computer Engineering, Texas A&M University, Doha 23874, Qatar.
[3] Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia.

[*]Corresponding author

Abstract

Recently, the speech attribute features caught the interest of the speech processing society and successfully employed in a wide variety of applications. In this paper we introduce the first intensive study of speech attribute detection in Arabic language. For each speech attribute, namely the manners and places of articulation, a binary Deep Neural Network (DNN) classifier is trained to recognize the existence or absence of the attribute. The DNN consists of multiple fully connected hidden layers and a two-way output softmax layer. The DNN is fed by mel-scale filter bank features extracted from the speech signal. We further adopted the dropout regularization technique to alleviate the classifier overfitting. The system tested on a speech corpus of 90 hours collected from Quranic Arabic reciters. The results show that the speech attribute detectors achieved classification accuracies ranging from 76% to 95%.

Keywords: Speech attributes; deep neural network, Arabic language

INTRODUCTION

The speech attribute features are shown to be very robust against the speaker variation due to the age, gender or even the dialect. These kind of features are proved to outperform the traditional speech features such as the Mel Frequency Cepstral Coefficients (MFCC) in different speech processing problems.

In [1] Lee et al. proposed the Automatic Speech Attribute Transcription system (ASAT) where a bank of speech attribute detectors were trained to measure the existence or absence of each attribute and the output features is then merged and used for performing Automatic Speech Recognition (ASR). This bottom-up approach is known as knowledge-based speech recognition [2].

The powerful of the speech attribute features is that they are shared among languages and therefore speech corpora from multiple languages can be used in modeling a universal speech attribute detectors [3].

In addition to the ASR system, the speech attributes features were used for other speech processing problems. Zhang et al. [4] investigated the effectiveness of such features in achieving speaker verification and the results showed that the proposed system outperform all other speaker verification methods. Furthermore, the attribute features were successfully utilized for foreign accented characterization [3] and spoken language recognition [5].

The speech attribute features are very helpful in the pronunciation verification problem. In fact, the phoneme is considered mispronounced when one or more of its attributes are changed. Several attempts for using the speech attribute features in tackling pronunciation verification problem in the literature. The speech attribute features were utilized in [6] to improve the mispronunciation detection and provide diagnostic feedback for Mandarin learners. In [7] the author introduced the so called articulatory Goodness Of Pronunciation (aGOP) score where the articulation features were used for estimating the phoneme posterior probability.

However, the Arabic speech attribute features received very little attention in the literature. Hammady et al. proposed a hidden Markov model (HMM) for the detection of Arabic speech attributes [8]. While in [9] Ziedan et al. used the speech attribute features to discriminate among different Arabic dialect and accent.

In this paper we investigated the speech attribute features of standard Arabic language. We trained a bank of binary Deep Neural Network (DNN) classifiers to classify each speech frame as belongs to a specific attribute or not. The DNN is a feed forward neural network with multiple fully connected hidden layers and softmax output layer consists of 2 neurons, one of which is fired if the sample is positive, the attribute is exist, while the other one is fired when the sample is negative, the attribute is absent. Moreover, the dropout was employed to cope with the overfitting behavior of the DNN. We trained and evaluate the classifier using Quranic Arabic speech corpus of around 90 hours.

The rest of the paper is organized as follows. Detailed description of the speech corpus is demonstrated in section 2. In section 3, we explain the details of the proposed system. The results are presented in section 4. Finally, the conclusion is drawn in section 5.

SPEECH CORPORA

The advantage of the Quran is that it is a standard Arabic closed vocabulary text (around 14716 unique words) with massive amount of speech data available from hundreds of different reciters. The Quran text consists of 114 chapters vary

in their size from 12316 words to 25 words and each chapter consists of multiple verses. The duration of recording full Quran text from one speaker of average recitation speed is around 30 hours. Most of the available speech recordings are available on chapter-level, where each complete chapter saved as one continuous speech file. While few recordings are manually segmented into verse-level, where each verse saved as one continuous speech file.

The speech corpus used in this work is a Quran corpus segmented in verse-level (VER) and consists of 30 speakers reciting the last 56 chapters of Quran with total duration of around 90 hours. This corpus segmented from chapter-level to verse-level manually by EveryAyah project [10] .The data released in mp3 format with different bit rate. This dataset divided into 3 subsets training, validation and testing which contains 22, 4 and 4 speakers respectively as shown in Table 1.

Table 1. Quran speech corpora

Corpus	N# Speakers	Duration
VER-train	22	~66 hours
VER-valid	4	~12 hours
VER-test	4	~12 hours

METHOD

System Description

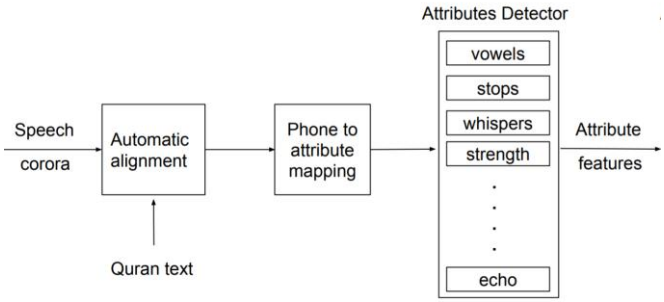

Figure 1. system flow diagram

Figure 1 shows the system flow diagram. First, the VER Quran speech corpus along with the Quran text pass through a segmentation and alignment module. This module consists of an intensity-based Voice Activity Detection (VAD) method and ASR method based on HMM acoustic models and n-gram Language Model (LM). The VAD used for segmenting the long speech files into short segments according to the silence position. The phoneme alignment is then performed using the ASR method.

Each phoneme is then mapped to its corresponding attribute according to a predefined Quran mapping rules. For each attribute we trained a binary DNN classifier to classify each speech frame as positive, when the attribute is exist, or negative, when the attribute is missing. The samples from all phonemes belongs to a specific attribute is used as a positive samples while samples from all other phonemes are forming the negative ones.

The speech frame passed through the bank of pre-trained speech attribute detectors to extract the speech attribute feature vector.

Voice Activity Detector (VAD)

Applying this module on each speech file that contains either one verse from one reciter in order to detect the position of pauses. Most of the materials used in this work were recorded in a noise clean environment such as studios, hence a simple intensity-based algorithm is used. The adopted method is controlled by three parameters the silence threshold, the minimum speech duration and the minimum silence duration. The minimum silence duration is used to eliminate short silence segments that occurs during production of some phonemes, e.g. closure duration in plosive phonemes such as (kalkala قلقلة). The minimum speech duration used to cope with the short noise burst during silence intervals (e.g. microphone noise). Finally, the discrimination between the speech and silence segments performed based on the value of the silence threshold. Because of the variations in the voice level of the reciters and the recording environment, we computed the silence threshold for each speech file based on the values of the 5th and 95th percentile of the intensity within the current speech file.

Silence threshold value (ST) calculated as following:

$$ST = P(05) + 0.2\,[P(95) - P(05)] \quad (1)$$

The 5th percentile intensity value $P(05)$ and the 95th percentile intensity value $P(95)$ were used as alternative to the minimum and maximum intensity values in order to reduce sensitivity to outliers.

Automatic alignment of speech corpora

In this section we describe in details the automatic alignment method which used to obtain the time boundary of each phoneme in the speech corpus VER. Since each speaker has around 3 hours of speech data, we built a speaker dependent acoustic model that is used for aligning speech data of each speaker. Although each speech file in the VER corpus contains exactly one verse, a common behavior by reciters that part of the verse is repeated once or more. Therefore, simple forced alignment method will lead to an inaccurate phoneme alignment. Therefore, the speech file is first passed through a VAD module which detects the existence of pauses and their positions. If no pauses detected, so most likely this speech file has no repetition and contains exactly the verse phoneme sequence. This process repeated for all speech files for a specific speaker and filtering out all files with pauses and using the rest of the data to build a flat-start speaker-dependent HMM acoustic model.

The speech files that contain pauses is first segmented into short segments and then decoded using these initial speaker-

dependent acoustic model along with a bi-gram language model created for each verse. Furthermore, we used all the speech data of specific speaker and trained a final speaker-dependent HMM acoustic model and then re-aligned its speech files to produce more accurate phoneme time boundaries. This process was repeated for all VER speakers.

The HMM acoustic models is a tied-states context-dependent with 32 mixtures per state. The models trained using 13 MFCC features extracted from 25 msec window sampled every 10 msec. The delta and acceleration are further computed to form an input feature vector of size 39.

Speech attribute detection

The speech attributes of the Arabic language is a controversially issue and some of them are not agreed among all linguistics. In this work we adopted 37 attributes as listed in Table 2 following mainly the study in [11]. The Arabic phoneme symbols used are listed in Table 3.

For each attribute we built a binary DNN-based classifier to discriminate between frames where this specific attribute is exist (positive samples) and other frames where the attribute is absent (negative samples).

Table 2. Speech attributes for Quranic Arabic and the corresponding phonemes

	Feature	Phonemes	Feature	Phonemes
Places of articulation	Oral cavity	a:, u:, i:	Interdental	Z, ~z, t_h
	Pharynx	@, h, ~@, ~h, g_h, x	Alveolar	t, d, s, n, z, T, D, S, r, l
	Deep tongue	q, k	Post-alveolar	s_h, j
	Middle tongue	j, s_h, y	Palatal	Y
	Tongue tip	T, d, t, Z, ~z, t_h, S, z, s, n, r	Velar	x, g_h, k
	Tongue border	D, l	Uvular	q
	Labial	f, m, w, b	Pharyngeal	~h, ~@
	Bilabial	b, m, w	Glottal	@, h
	Labiodental	f		
Manners of articulation	Whisper	f, ~h, t_h, h, s_h, x, S, s, k, t	Deviate	l, r
	Strength	@, j, d, q, T, b, k, t	Hiding	h, a:, u: , i:
	Moderate	l, n, ~@, m, r	Echo	q, T, b, j, d
	Softness	D, f, g_h, h, ~h, s, S, s_h, t_h, w, x, y, z, ~z, Z, a, a:, i, i:, u, u:	Stops	b, t, T, d, D, k, q, @
	Silence	Sil	Fricatives	f, s, S, z, t_h, ~z, Z, s_h, x, g_h, ~h, ~@, h
	Elevation	x, S, D, g_h, T, q, Z	Affricates	j
	Adhesion	T, Z, S, D	Glides	y, w
	Whistle	S, z, s	Lateral	l
	Prolongation	D	Vowels	a:, u:, i:, u , a, i
	Spreading	s_h	Repetition	r

Table 3. Quranic Arabic phoneme set

Phoneme	Description	Phoneme	Description	Phoneme	Description
@	ء	s_h	ش	n	ن
b	ب	S	ص	h	ه
t	ت	D	ض	w	و
t_h	ث	T	ط	y	ي
j	ج	Z	ظ	a	فتحة
~h	ح	~@	ع	u	ضمة
x	خ	g_h	غ	i	كسرة
d	د	f	ف	a:	مد فتحة
~z	ذ	q	ق	u:	مد ضمة
r	ر	k	ك	i:	مد كسرة
z	ز	l	ل		
s	س	m	م		

The DNN classifier is consists of 6 fully connected hidden layers with typically 2048 neuron in each layer. The output layer is a softmax layer consists of 2 neurons, one of which is fired in case of positive sample while the other one is fired in case of negative one. The rectifier linear units (RELU) activation function was adopted for all hidden neurons. The RELU function is proved to speed up the training of the DNN and avoid the vanishing gradient problem. Therefore, the time and resource consuming pre-training step becomes less effective and hence we did not perform it. The binary cross entropy was used as an objective function.

Unlike the HMM acoustic model, the DNN was trained using filter bank features which are used commonly with DNN speech models and achieved better performance over the traditional MFCC features [12]. Here also the speech signal divided into frames of 25 msec and 15 msec overlap. For each frame we extracted 21 filter banks plus the delta and acceleration components. We further concatenated each 11 frames (5 frames preceded and 5 frames succeeded the current frame) to form an input feature vector of size 693 per sample.

The mini-batch Stochastic Gradient Descent (SGD) method was utilized for the fine tuning of the DNN model with batch size of 200 samples. The learning rate was controlled by the newbob method where the learning rate starts with 0.1 and remains constant for the following epochs as long as the improvement of the classification accuracy of the validation set is greater than 0.05. Once the improvement in the classification accuracy of the validation set fell under the 0.05, the learning rate scaled by 0.5 during each of the remaining epochs. The training is terminated when the learning rate reaches a minimum value of 0.00001.

Furthermore, we adopted the dropout regularization technique to alleviate the effect of the overfitting over the training data [13]. The idea is to dropout part of the neurons in each hidden layer in the training phase by removing their connections to the neurons in the next and previous layers and not updating their weights during the dropout epoch. This performed by ignoring each neuron with probability p and keep it with probability $(1 - p)$ in each training epoch. On the other hand, all neurons will be fully connected during test with weights multiplied by p.

The samples from all phonemes belongs to a specific attribute were used as the positive samples in the training of the binary classifier while the negative samples are chosen from the frames of the others phonemes. To imbalance training of the classifier, we choose equal number of positive and negative samples for each attribute. Moreover, both the positive and negative samples are distributed equally over all phonemes.

EXPERIMENTAL RESULTS

We trained one binary DNN classifier for each attribute to discriminate between frames belongs to this attribute and frames where the attribute is absent. We use the VER-train dataset for the training and the VER-valid and VER-test for validation and testing of the attribute detectors respectively. The VER-valid was used only to control the learning rate scheduling and early stopping of the training process while the training set was used for computing the gradients and updating the weights in the back propagation mechanism. The final accuracy reported with the VER-test dataset. As aforementioned, the number of positive samples, where the attribute is exist, and negative samples, where the attribute is absent, in both the training, validation and test datasets is balanced and hence we used the frame level accuracy as our performance measure. Table 4 summarizes the overall accuracy and the number of samples of each attribute in the training, validation and testing datasets.

Overall, the manners of articulation behave better than the places of articulation with average test accuracy of 84% and

stander deviation of 4.7% compared to 83% and stander deviation of 4.4% respectively. The "spreading" attribute achieved the best test performance of 94% followed by "affricates" and "Post-alveolar" of 91% each.

We further adopted the dropout as a regularization technique to cope with the overfitting problem and improve the model generalization. The dropout value is fixed to 0.3 for the input layer and 0.2 for all hidden layers. The effect of using dropout is summarized in

Figure **2**. As shown in the figure, the dropout improved the performance of almost all the attribute detectors by 14% to 1% reduction in the error rate.

Table 4. The speech attribute detectors performance for the training, validation and testing datasets

	Speech Attribute	Number of samples			Overall Accuracy (%)		
		Training	Validation	Testing	Training	Validation	Testing
Places of articulation	Oral cavity	2042210	449286	408442	95.8	85.1	85.1
	Pharynx	927860	204129	185572	88.1	78.4	77.9
	Deep tongue	334610	73614	66922	95.0	87.3	86.8
	Middle tongue	368060	80973	73612	93.9	85.0	84.9
	Tongue tip	1606020	353324	321204	91.1	77.9	77.7
	Tongue border	530960	116811	106192	96.2	82.6	82.2
	Labial	1150060	253013	230012	91.9	76.9	77.4
	Bilabial	1007320	221610	201464	90.2	77.9	78.4
	Labiodental	142740	31403	28548	96.6	84.7	85.2
	Nasal cavity	472690	103992	94538	94.3	88.0	87.7
	Interdental	149720	32938	29944	100.0	79.0	80.3
	Alveolar	1987260	437197	397452	90.3	76.8	76.6
	Post-alveolar	122590	26970	24518	100.0	91.4	91.5
	Palatal	245470	54003	49094	92.0	86.3	87.0
	Velar	246640	54261	49328	97.9	85.2	84.0
	Uvular	165360	36379	33072	94.1	87.9	87.0
	Pharyngeal	233570	51385	46714	100.0	87.7	88.0
	Glottal	616900	135718	123380	92.8	78.6	77.8
Manners of articulation	Whisper	1255000	276100	251000	93.8	86.6	86.1
	Strength	1317780	289912	263556	93.8	83.2	83.7
	Moderate	1999410	439870	399882	93.3	76.3	76.5
	Softness	3317380	729824	663476	95.7	74.7	75.0
	Silence	3502210	770486	700442	96.5	90.6	88.4
	Elevation	410360	90279	82072	96.5	86.8	86.7
	Adhesion	167610	36874	33522	98.9	85.5	88.0
	Whistle	241050	53031	48210	97.8	85.7	89.6
	Prolongation	25250	5555	5050	94.4	87.6	85.8
	Spreading	58810	12938	11762	100.0	95.1	94.8
	Deviate	856430	188415	171286	99.6	81.3	81.1
	Hiding	2366440	520617	473288	93.0	84.1	84.0
	Echo	657100	144562	131420	99.1	85.9	86.2
	Stops	1279250	281435	255850	95.0	82.5	82.8
	Fricatives	1227510	270052	245502	94.8	81.8	81.7
	Affricates	63780	14032	12756	99.8	90.0	91.4
	Glides	509770	112149	101954	92.8	80.8	80.5
	Lateral	505710	111256	101142	95.5	84.2	83.5
	Vowels	4109510	904092	821902	95.3	79.2	79.1
	Repetition	350720	77158	70144	96.8	85.7	85.9

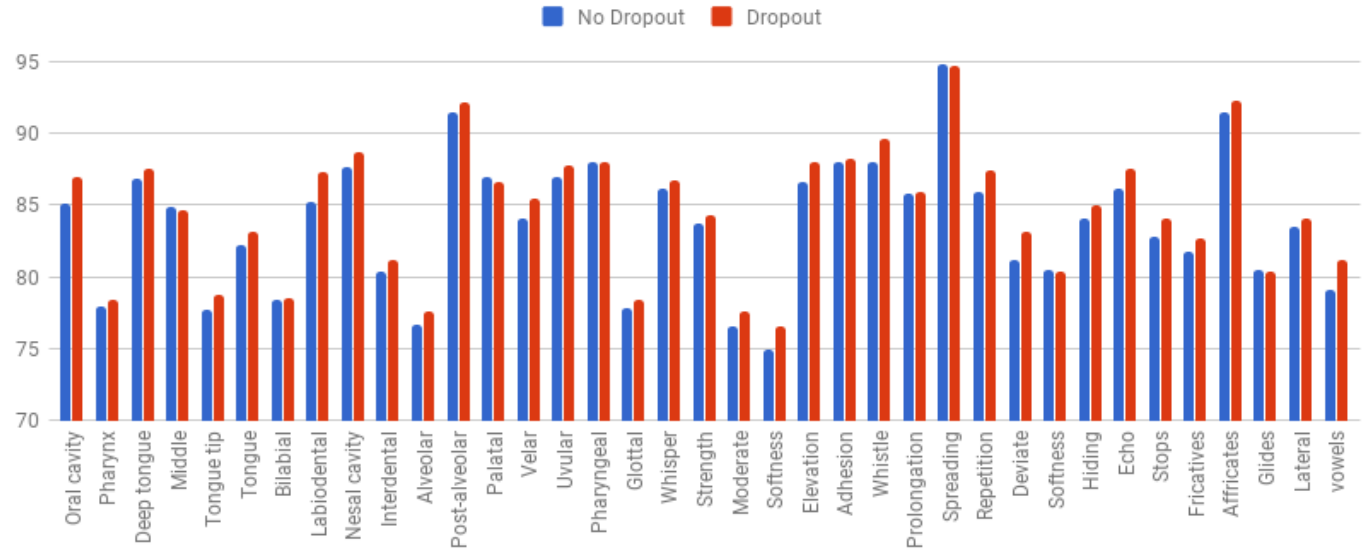

Figure 2. The effect of the dropout regularization method

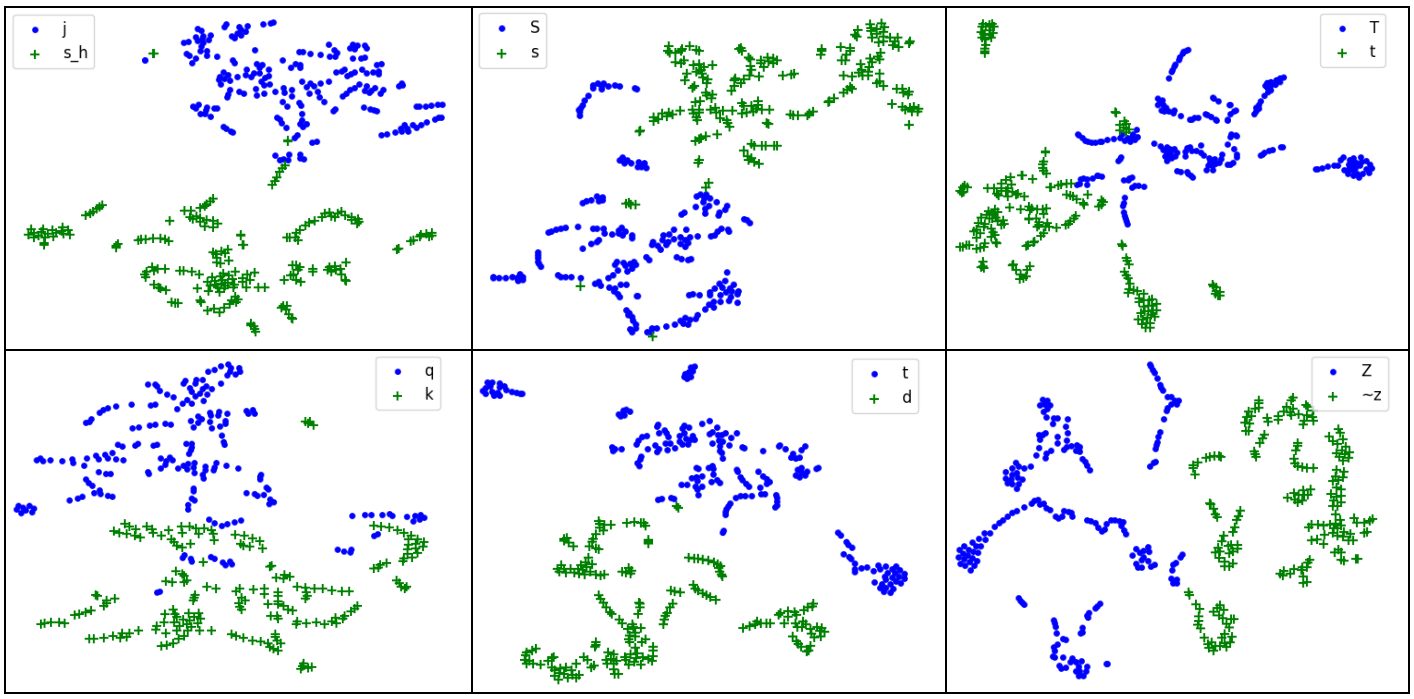

Figure 3. 2D scatter plot of the speech attribute features of 6 pairs of confusable phonemes

In order to demonstrate the powerful of the attribute features in discriminating between phonemes, we draw a scatter plot for the speech attribute features of each pair of phonemes that are considered similar in articulation such as /m/ and /n/, /q/ and /k/, /t/ and /d/, etc. The t-SNE [14] is used to project the speech attribute feature vector from 38 to 2 dimensions. Figure 3 shows the 2D scatter plot of random samples selected from the validation set of 6 confusable phoneme pairs. It is obvious from the figures that each phoneme has clear separate region(s) with some minor overlaps.

CONCLUSION

In this paper we explored the speech attribute features of standard Arabic language. A bank of speech attribute detectors, namely the manners and places of articulation, were built for estimating the existence or absence of each specific attribute. These detectors are based on DNN architecture fed by filter bank features extracted from each speech frame. The attribute detectors achieved average accuracies of 84%±4.7% and 83%±4.4% for the places and manners of articulations respectively.

The paper introduced the first study of the speech attribute detectors of Arabic language and open the door for further research in this area.

REFERENCES

[1] C.-H. Lee, M. A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang*, et al.*, "An overview on automatic speech attribute transcription (ASAT)," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[2] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. ICSLP*, 2004.

[3] V. Hautamäki, S. M. Siniscalchi, H. Behravan, V. M. Salerno, and I. Kukanov, "Boosting universal speech attributes classification with deep neural network for foreign accent characterization," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[4] S. Zhang, W. Guo, and G. Hu, "Exploring universal speech attributes for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 5355-5359.

[5] Y. Wang, J. Du, L. Dai, and C.-H. Lee, "A fusion approach to spoken language identification based on combining multiple phone recognizers and speech attribute detectors," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, 2014, pp. 158-162.

[6] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 6135-6139.

[7] H. Ryu, H. Hong, S. Kim, and M. Chung, "Automatic pronunciation assessment of Korean spoken by L2 learners using best feature set selection," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, 2016, pp. 1-6.

[8] H. Hammady, O. Badawy, S. Abdou, and M. Rashwan, "An HMM system for recognizing articulation features for Arabic phones," in *Computer Engineering & Systems, 2008. ICCES 2008. International Conference on*, 2008, pp. 125-130.

[9] R. Ziedan, M. Micheal, A. Alsammak, M. Mursi, and A. Elmaghraby, "A Unified Approach for Arabic Language Dialect Detection," in *29th International Conference on Computers Applications in Industry and Engineering (CAINE 2016), Denver, USA*, 2016.

[10] *Every Ayah*. Available: http://everyayah.com/

[11] A. Ragheb, "Quran Phonology; Quran reciting rules based on modern acoustics," M.Sc Thesis Cairo University, 2004.

[12] A. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4273-4276.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research,* vol. 15, pp. 1929-1958, 2014.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research,* vol. 9, pp. 2579-2605, 2008.