Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

A framework for pronunciation error detection and correction for non-native Arab speakers of English language

Bandar Ali Al-Rami[a] and Yousef Houssni Zrekat[b*]

[a]*Faculty of Computer Studies, Arab Open University, Saudi Arabia*
[b]*Faculty of Language Studies, Arab Open University, Saudi Arabia*

C H R O N I C L E

Article history:
Received: February 23, 2023
Received in revised format: April 2, 2023
Accepted: May 8, 2023
Available online: May 8, 2023

Keywords:
Machine Learning
Pronunciation Errors
Speech Recognition
Phonological Feature

A B S T R A C T

This paper examines speakers' systematic errors while speaking English as a foreign language (EFL) among students in Arab countries with the purpose of automatically recognizing and correcting mispronunciations using speech recognition, phonological features, and machine learning. Accordingly, three main steps are implemented towards this purpose: identifying the most frequently wrongly pronounced phonemes by Arab students, analyzing the systematic errors these students make in doing so, and developing a framework that can aid the detection and correction of these pronunciation errors. The proposed automatic detection and correction framework used the collected and labeled data to construct a customized acoustic model to identify and correct incorrect phonemes. Based on the trained data, the language model is then used to recognize the words. The final step includes construction samples of both correct and incorrect pronunciation in the phonemes model and then using machine learning to identify and correct the errors. The results showed that one of the main causes of such errors was the confusion that leads to wrongly utilizing a given sound in place of another. The automatic framework identified and corrected 98.2% of the errors committed by the students using a decision tree classifier. The decision tree classifier achieved the best recognition results compared to the five classifiers used for this purpose.

1. Introduction

Automatic speech recognition (ASR) is an artificial intelligence technique that allows humans to create a speech dialogue with the machine, which interprets the speech using a pre-trained model. Generally, ASR is implemented by processing the speech received by the machine as a wave file. The speech is cleaned to remove the noise and extract the spectrograms and other features. The acoustic model is then used to recognize the underlying phonemes sounds in the speech. The chain of phonemes is processed stochastically using Hidden Markov Model (HMM) to construct the most likely word using the language model. ASR requires clear pronunciation to complete the recognition process correctly. Thus, the mispronounced words challenged the ASR and led to poor results. Similarly, the variations in word pronunciations among individuals of different mother tongues challenge the ability to recognize an identical word with varied pronunciations. Moreover, as ASR is used for teaching individuals learning English as a Foreign Language (EFL), using an automatic technique to overcome their errors in pronunciation and improve their speech fluency through ASR is demanded to reduce human labor in doing so. Accordingly, automatic error detection and correction are required to improve the accuracy of the ASR and reduce the efforts and preserve time in teaching EFL (Lai & Chen, 2022).

* Corresponding author.
E-mail address: y.zrekat@arabou.edu.sa (Y. H. Zrekat)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.5.004

Speech-related applications depend either on using a knowledge-driven approach, which requires experts to record their speech in canonical form, or a data-driven approach, which depends on recorded speech of non-standard English. Although, the knowledge-driven approach is more efficient for speech-generation applications. The data-driven approach can effectively recognize speech, detect pronunciation errors and distinguish the variation in word pronunciation as it involves more pronunciation variations. Generally, in natural language processing, the data-driven approach and supervised machine learning techniques (i.e., classification tasks) are commonly implemented for complicated detection and prediction tasks, such as optical character recognition (OCR), document classification, and sentiment analysis (Nijhawan et al., 2022; Paula et al., 2022).

Accordingly, in the proposed framework, a data-driven approach is used for error detection, which is implemented by training a classification algorithm with sample data annotated with associated labels (correct vs. incorrect). Then, the trained machine will be used to assign class labels for samples with unknown labels. Yet, the trained data should comprise the characteristics expected for the unknown samples to be recognized. Thus, in the proposed model, the errors are manually analyzed before the application is constructed, define the possible errors, and provide sufficient samples to construct the model. In the proposed framework, the variation of the correct samples is addressed using the speech of multiple individuals, both native English speakers, and speakers of the targeted population. As for the incorrect variations, before collecting the corresponding samples, these variations will be identified, quantified, studied, and analyzed. Accordingly, this study will investigate phonological errors and their corrections. Studying phonological errors is required as these errors cannot be generalized from learners of other languages, as the committed errors depend on the learners' mother tongue (Bensalah & Betta, 2022; Syafrizal et al., 2022). The first language's phonological features influence the EFL learners and lead to pronunciation variations and errors. Accordingly, this study fills in the literature gap about the phonological errors committed by Arab EFL learners while speaking English and implements an automatic system for detecting and correcting these errors to help and encourage learners towards correct pronunciations. The proposed automatic detection and correction framework used collected and labeled data to construct a customized acoustic model to identify and correct incorrect phonemes. Based on the trained data, the language model is then used to recognize the words. The final step includes construction samples of both correct and incorrect pronunciation in the phonemes model and then using machine learning to identify and correct the errors. Generally, the classification algorithms can be classified into decision-tree (DT) classifiers, support vector machine (SVM), probability-based classifiers, instance-based classifiers, and artificial neural networks (ANN). These classifiers are used to implement the detection process and provide feedback for the learners.

The remaining part of the study was arranged as the following. Section two discusses the literature review and previous studies on phonological error identification, detection, correction, and automatic speech recognition and error detection. In section three, the dataset is collected with the goal of optimizing and covering different accents in a particular language (English) by non-English native speakers. Besides, the theoretical finding of the study is argued and presented in this section in order to identify the common errors. Finally, the machine-learning model is discussed and clarified. In section four, the results of the machine learning process are presented and discussed accordingly. Finally, the paper concludes in Section five.

2. Background Study

Every language has its rules for combining sounds to make a meaningful set of words. Children start learning the pronunciation of sounds by adopting a trial-and-error approach. Over time, they produce meaningful sounds, thus developing linguistic efficiency. On the other hand, when learners of English as a foreign language (EFL) start to acquire a new language, they face difficulties pronouncing its sounds correctly due to the differences between the phonological systems of the learner’s mother tongue and the target language. Generally speaking, foreign language learners encounter problems while speaking the target language, which comprises the four basic skills of language: listening, speaking, reading, and writing (Ambrozová, 2014; El Zarka, 2013; Hassan, 2014). Likewise, Arab students encounter various problems concerning pronunciation, grammar, and vocabulary, especially when speaking. Arab speakers of English also encounter phonological problems because of the differences in the phonological properties between their mother tongue, Arabic, and the target language, English (Alqarni, 2013). However, little research has focused on this important issue as the main part conducted in Arab-speaking countries related to English in EFL/ESL situations focused on reading and writing skills. In contrast, spoken English in situations was less studied (Zrekat & Al-Sohbani, 2022).

2.1 Phonological Errors Committed by Arab Students

The phonological errors committed by the Arab students are because English and Arabic languages have differences in the underlying phonemes. According to Daana and Khrais (2018), Arabic has twenty-eight consonants, while English has twenty-four. The differences between the constants in these languages are not in terms of numbers; some constants are presented in one and not in the other, and vice versa. In this case, these absent consonants may create a certain problem for Arab learners of English, particularly with the consonants absent in their classical or colloquial Arabic. Daana (2009) affirms that the sequences of vowels in these languages are greatly different, as Arabic consonant clusters consist of only two consonant sounds. In comparison, English may have up to four consonant sounds. Dobrovolsky and Katamba (1996) denoted that "native speakers of a given language realize that some words of the foreign languages are phonetically unfamiliar, they change the sound segment of the given words to fit the systematic pronunciation of their mother tongue". Accordingly, most of the learners of

non-native English speakers found it difficult to pronounce some English words with more than two consonants. Words like "spring, screen, splash, texts and costs" were very problematic for them. The respondents frequently add a vowel after a word's initial, second or third consonants.

Hassan (2014) investigated the problems faced by the Sudanese English language students at the Sudan University of Science and Technology (SUTC). The study relied on a self-monitoring strategy to get enough information about the mistakes students commit. The results showed that a significant number of students' mistakes in the sounds of /s/ as /θ/, /z/ as /ð/, /p/ as /b/, /v/ as /f/, and /ʃ/ is mixed with /tʃ/ mostly, which lead to mispronunciation errors. Alqarni (2013) investigated different voiceless postalveolar affricate /tʃ/variations by Saudi learners. The tool utilized to carry out this study contained 16 words that started with the sound /tʃ/ with various word combinations. Data were recorded and analyzed using a speech analyzer and SPSS software. The study investigated the Suadi learners found that, due to the lack of the sound /tʃ/ in the phonemic system of Arabic, native speakers of Arabic mostly make mispronunciations of /tʃ/ as an independent phoneme. The sound /tʃ/is present in some dialects of Arabic but is an allophone of other phonemes such as /q/ and /k/.

2.2 Error Detection and Correction for ASR

ASR is developed using trained data, consisting of a lexicon of all the words and phonological sounds in a language to interpret the human dialogue conversation afterward. Although training the language model with canonical language phonemes is required, recording all phonemes by experts is of words difficult for the following reasons: 1) recording a huge number in a language is time and resource-consuming. 2) Even with pre-written pronunciation, errors might be made by the recorder or the pronunciation writer. 3) The pronunciation variation, which is presented among experts, cannot be tolerated. Accordingly, the alternatives are either using an automatic pronunciation generator, which is called grapheme to phoneme (G2P) (Bisani & Ney, 2008; McGraw et al., 2012), or using the data-driven approach (Lu et al., 2013; Rutherford et al., 2014). The advantage of the pronunciation generator is saving time and resources. In contrast, the advantage of the data-driven approach is the ability to record errors and the variation of the correct pronunciation. The data-driven approach outperformed the linguistic expert-based for pronunciation name entities, as discussed by Rutherford et al. (2014). Although G2P saves time and effort, the variation in pronunciation cannot be captured using such a model solely. Accordingly, to gain the advantage of capturing vast pronunciation variations and saving time and resources, the data-driven approach can be combined with the pronunciation generation techniques, which will be addressed in the proposed framework.

Tepperman (2006) used a data-driven approach to verify the children's pronunciation to verify and detect pronunciation errors. Phonemes-based features are extracted and classified into correct or incorrect phonemes. Accordingly, the classification output and the acoustic features verify the input pronunciation. Molina et al. (2009) used classification for pronunciation evaluation, similar to using classification for error detection. Besides, various applications for pronunciation validation were developed by classifying the extracted phonological features into correct and incorrect classes. The Computer Assisted Pronunciation Training (CAPT) tool is used for pronunciation validation and provides feedback on pronunciation errors using posterior probability and HMM recognition process (Gambari et al., 2014). These applications are post-ASR, which depends on the quality of the saved ASR data. For example, Fluency (Eskenazi, 2009) was developed to validate the pronunciation of non-native speakers and detect errors based on ASR and statistical classification algorithms. The Computer-Aided Language Learning (CALL) technique (Peabody, 2011) detects the mispronunciation of naïve speakers based on the distance between the phoneme features and the gold–standard of the correct phoneme. Similarly, Ai (2015) developed an error detection technique using phoneme recognition. Overall, these techniques depend on extracting features from phonemes and using these features to evaluate the goodness of the evaluated phoneme (Li et al., 2017).

Various deep-learning-based classification techniques were used for error detection because they could process huge and complicated data. A convolution neural network was used for feature extraction and data classification for pronunciation error correction (Lee, 2016). The deep neural network was also used for speech recognition with variation (Cai & Liu, 2018). These models improve speech recognition with variations of non-native speakers based on the phoneme features. The correctness of pronunciation is assisted using a corpus of trained data rated on a scale of [1-5]. A machine learning classification is trained with the scaled corpus, and the trained model is then used to assist the pronunciation of non-English speakers (Kotani & Yoshimi, 2018). Generally, although phoneme-based features are robust tools for pronunciation error detection, these post-ASR models depend on the ASR output, which is presented as a complete word. The disadvantage of such an approach is not the detecting abilities but the inability to provide accurate feedback on each phoneme. Accordingly, the focus should be on the acoustic model's output, which addresses the phoneme part.

3. Methodology

The proposed work identifies the common mispronunciation errors committed by Arab learners while learning English. The significance of the proposed work is inspired by the fact that learning English pushed globalization forward, as there is a critical mutual relationship between globalization and English (Poggensee, 2016). More and more people are trying to learn English today, and the number is increasing rapidly; people have realized how important this language is, at least the communicative part of it (Zrekat, 2021). Besides, the proposed work aims to develop a framework for mispronunciation detection

and correction based on ASR. The significance of ASR is realized as its facilities and applications are growing rapidly, including smartphone interfaces, smart homes, robotics, call centers, etc.

3.1 Data Collection Methodology

Data was collected through oral interviews; such a methodology was found to be the best for data collection. Accordingly, in the oral interview, the phonemes which need to be pronounced are recognized and understood. Accordingly, the active participation of the interviewers and the interviewees is essential for collecting the phonological dataset.

3.2 Population and Sample of the Study

The datasets are collected from participants of Arab students from different Arab countries who study at the Arab Open University in Saudi Arabia (AOU/KSA). The respondents speak Arabic as a Native language and English as a foreign language. The study sample examined 30 Arab students (15 males and 15 females) specializing in Education, Business Administration, and Information Technology. The qualities of the students' samples are displayed in Table 1. The student's English pronunciation was recorded to identify the systematic mistakes in pronouncing the consonant sounds. Besides the recorded speech, general background information, such as gender, age, marital status, and the place they had learned English, were collected, as given in Appendix A.

Table 1
Students' Demographic Background

Accent	Sample Size	Speech Duration (Minutes)	# Students
Saudi	20	40	10
Egyptian	10	20	5
Moroccan	6	12	3
Syrian	10	20	5
Sudanese	4	8	2
Yemeni	10	20	5
Total	60	120	30

3.3 Data Collection

Data was collected using a qualitative method to explore various phonological and pronunciation errors committed by the interviewees while speaking English. Two rounds of interviews were conducted to identify the mispronunciation errors. The first is a semi-structured interview conducted with the respondents by the researchers. The respondents were asked to share their views on the difficulties they encountered when speaking English. The researcher followed a focus group interview where the sample was divided into six groups, five respondents in each. Open-ended questions were directed at the respondent's about the difficulties students undergo when speaking English in and outside classrooms. Using open-ended questions will not restrict the respondents, and at the same time, they will feel free to say whatever they like. Accordingly, the researcher will listen to the problems according to the view of the respondents themselves while at the same time recording and listening to the phonological errors they commit while speaking.

Next, the records were given to three experts who specialized in phonology to listen to the phonological errors committed by the students while speaking in English for the experts to analyze and identify the common problems. In the second round, an oral interview is conducted as it is best understood and efficiently analyzes the respondent's pronunciation. A list of words on a paper was given to the students to read (Information is provided in Appendix-A). The participants read the words slowly and carefully while their speech was being recorded. The choice is random; however, the researchers sought equal balance for students of the genders. To ensure the results' accuracy, the recordings were examined carefully, and the data was revised repeatedly. After recordings, the researcher and the three experts analyzed all the sounds. Recordings were analyzed and coded based on the errors committed by the respondents. The errors were coded based on the groups as student one from focus group 1 was coded as (S1, FG1). Overall, the experts found that five sounds were identified as frequently mispronounced: /p/, /tʃ/, /dʒ/, /ɹ/, and /ŋ/.

3.4 Statistical Analysis

The data is analyzed, and the summary of the captured mispronunciation is discussed based on the data recorded by the students as they answered the questions during the interviews. Table 2 shows different realizations of the phoneme sound /p/ in English in all word positions in the data analysis. As noticed, some respondents made errors by replacing this sound with the voiced /b/ (males and females). For instance, in the word "pain", the initial sound was pronounced /b/ by 60% of the respondents instead of /p/. When asked to pronounce the word "Paper", 70% of the respondents mispronounced the initial or the medial consonant or both of them. In a word that ends with /p/, like "sharp", 70% mispronounced it. Approximately, students made 66.6% incorrect pronunciation for the sound /p/ in all positions of the word.

Table 2
Realizations of /p/ Sound

	Realization	Male		Female		Overall	
		Number	Percentage	Number	Percentage	Number	Percentage
Initial	/p/	5	16.6%	7	23.3%	12	40%
	/b/	10	33.3	8	26.6%	18	60%
Medial	/p/	4	13.3%	5	16.6%	9	30%
	/b/	11	36.6%	10	33.3%	21	70%
Final	/p/	4	13.3%	5	16.6%	9	30%
	/b/	11	36.6%	10	33.3%	21	70%

As listed in Table 3, the respondents were asked to pronounce "champion, launching, and church", including the sound /tʃ/ in different word positions. The respondent had two variations; the correctly pronounced one /tʃ/ and the incorrectly pronounced one /ʃ/. When the sound came initially, the percentage of it being mispronounced was 56%. Regarding the middle of the word, 50% was the percent of the students who gave incorrect realization of the sound /tʃ/. When it finally, 63% of the respondents mispronounced the sound. The overall percentage of mispronunciations was 56.6%.

Table 3
Realizations of /tʃ/ Sound

	Realization	Male		Female		Overall	
		Number	Percentage	Number	Percentage	Number	Percentage
Initial	/tʃ/	6	20%	7	23%	13	43%
	[ʃ]	9	30%	8	26%	17	56%
Medial	/tʃ/	7	23%	8	26%	15	50%
	[ʃ]	8	26%	7	23%	15	50%
Final	/tʃ/	5	16%	6	20%	11	36%
	[ʃ]	10	33%	9	30%	19	63%

As listed in Table 4, when the respondents were asked to pronounce the sound /dʒ/ in different word positions like "Germany, Plagiarism, Judge", we had two variations, a correct pronunciation as [dʒ] and incorrect pronunciation as [ʒ]. When it is positioned initially, 63% of the students pronounced the incorrect allophone. On the other hand, when it comes to the middle, 66% of it was pronounced incorrectly. When it was finally positioned, only 23% used the incorrect variation. The total mis-pronunciation in all positions reached 50%.

Table 4
Realizations of /dʒ/ Sound

	Realization	Male		Female		Overall	
		Number	Percentage	Number	Percentage	Number	Percentage
Initial	/dʒ/	5	33%	6	40%	11	36%
	/ʒ/	10	66%	9	60%	19	63%
Medial	/dʒ/	4	26%	6	40%	10	33%
	/ʒ/	11	73%	9	60%	20	66%
Final	/dʒ/	10	66%	13	86%	23	76%
	/ʒ/	5	33%	2	13%	7	23%

As given in Table 5, for the sound /ɹ/, students have the correct variations used in English, whether British or American and the wrong variation, which is the sound [ɾ] used in Arabic. When the students were asked to pronounce the word “rabbit”, the initial consonant was 76% wrongly pronounced by the students. 60% of the respondents had incorrectly pronounced the middle consonant in "carry". Yet, when it came to an end, as in "hour", it was wrongly pronounced by 73% of the respondents. The total wrong pronunciations of the sound /ɹ/ were 69.6%.

Table 5
Realizations of /ɹ/ Sound

	Realization	Male		Female		Overall	
		Number	Percentage	Number	Percentage	Number	Percentage
Initial	/ɹ/	3	20%	4	26%	7	23%
	[ɾ]	12	80%	11	73%	23	76%
Medial	/ɹ/	5	33%	7	46%	12	40%
	[ɾ]	10	66%	8	53%	18	60%
Final	/ɹ/	3	20%	5	33%	8	26%
	[ɾ]	12	80%	10	66%	22	73%

Table 6 summarizes the pronunciation errors committed by Arab students, as conducted in this survey, and the difficulties are summarized 7.

Table 6
The problematic sounds

Correct Pronunciation	Wrong variation	Mispronounced Percentage	Words enlisted in the study
/p/	[b]	66%	"Sharp", "paper", "pain".
/tʃ/	[ʃ]	56.3%	"Church", "launching", "Champion".
/dʒ/	[ʒ]	50%	"plagiarism", "Germany", "judge".
/ɹ/	[ɾ]	69.6%	"Rabbit", "carry", "hour".

Table 7
Difficulties EFL students encountered at AOU

No.	*Sound*	*Mispronounced sound*	*Reason*
1-	/p/	/b/	confused
2-	/tʃ/	/ʃ/	***mispronounced***
3-	/dʒ/	either use /dʒ/ for /dʒ/ and /ʒ/,	confuse them like saying "strange" and "garage" as /streɪndʒ/ and /gəræʤ/ or as /streɪnʒ/ and /gəræʒ/.
4-	/ɹ/	/ɾ/.	Pronounced based on the Arabic /ɾ/.
5-	/ŋ/	/ng/ , /ŋg/	1) /ng/ or as /ŋg/ when it comes at the end. 2) /g/ following and when does not need it following in the middle of the word position.
6-	The major factors for the mispronunciation are mainly linked to the fact that the Problematic sounds are not found in the phonological system of Arabic; if they are found in Arabic, they are not phonetically realized similarly; and that English pronunciation is not practiced sufficiently by Arab learners of English.		

3.5 The Automatic Framework

An automatic framework for pronunciation error detection and correction is proposed. The audios recorded for the Arab EFL learners are used as the inputs, with the experts' labels of the frames corresponding to the underlying phonemes. Moreover, the proposed framework is supplied with a set of words of correct pronunciations from the Speech Accent Archive (Weinberger & Kunath, 2011), which consists of the pronunciation of standardized English paragraphs that is formed to contain all phonemes of English. Volunteers of various native English speakers recorded this archive. The archive is annotated with the corresponding words and paragraphs. Moreover, in the developed framework, the phoneme model generates a variation of words based on the correct and incorrect phonemes, similar to the G2P model, as illustrated in Fig. 1. The model modifies the correct words systematically, such as the correct pronunciation is replaced with the confused utterances of the incorrect utterances, based on the theoretical analysis results (Table 2-Table 7). The correct and incorrect pronunciation and the modified words' pronunciation are saved. The framework involves correct and incorrect words of 41 canonical and ten non-English phonemes. Given that the data contains non-English phonemes, for the incorrect pronunciation, due to the inability to be aligned automatically with the correct sounds, these data are annotated manually by the experts.

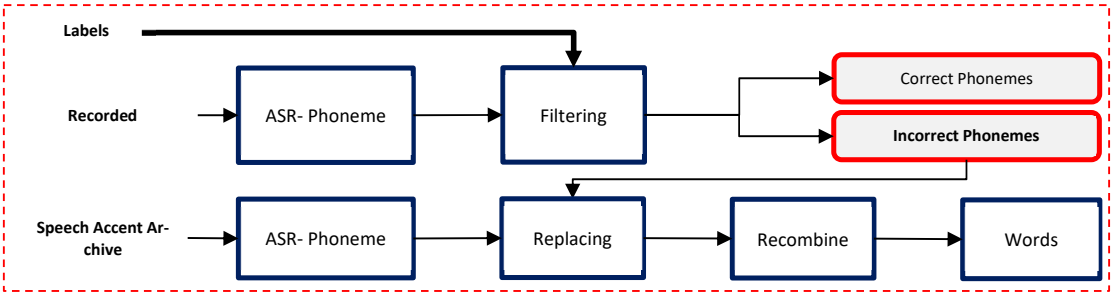

Fig. 1. Speech Generation Implementation Steps

The audio is first cleaned using spectral gating, which uses an estimated noise threshold. Accordingly, the waves are converted into frequencies, which are used with the spectrogram's threshold to remove the noise below some frequency-varying threshold. Next, features are extracted from the spectrograms. Mel-Frequency Cepstral feature (MFCC) is used, which is a spectral feature extracted in the input signal's frequency domain. The continuous signal is divided into 25 ms frames with 10 ms overlapping. The time frames are then combined with the hamming window. The results are processed using Fast Fourier Transform (FFT). The log energy is finally calculated as an output for each frame. To capture the phonemes of the speech, dynamic features are used besides static ones. A static feature represents each frame, the dynamic delta feature represents the difference between the static features and the successive frames, and the dynamic delta-delta is the difference between the delta features of the frame and the delta of the successive frames. Accordingly, the feature vector is formed of a thirty-nine-length feature vector out of the thirteen static features. Based on the constructed acoustic model, the extracted features corresponding to the phonemes are recognized, and the language model is used to construct the output word as the typical ASR. Yet, the language model received inputs of the same word from correct and incorrect pronunciation. Accordingly, the model is trained to receive, besides the correct words, incorrect words that result from incorrect pronunciation. Similarly, correct and incorrect pronunciation is used as input to the classifier, preceded by the acoustic model for recognition. As such, the ASR

model recognizes the words regardless of the limited errors within the word, as they have been trained to do so. The classification model, developed for all the phonemes, is used to recognize and distinguish incorrect phonemes from the correct ones. The proposed framework is implemented, illustrated in Figure 2. Besides pre-processing, three main models are used within the framework: the ASR model, the phoneme model, and the classification model.

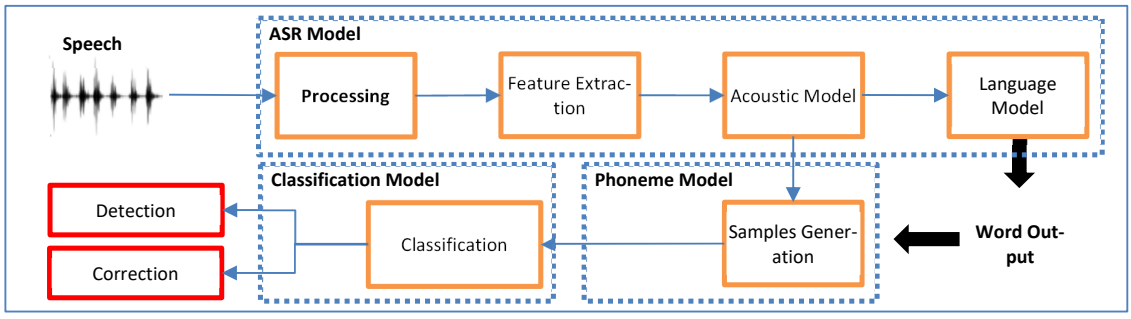

Fig. 2. Mispronunciation Detection and Correction Framework

The ASR model is an engine established as a dynamic HMMs and DNN model with the words' pronunciation processed, segmented, and represented as a set of utterances. The ASR converts the input sequence into an output sequence with the highest posterior probability among the possible sequences. The probability of the word output is the conditional probability of the output word, given the input word, subject to the chain rules generated with the HMM. This model is trained using both correct and incorrect pronunciation of each word to recognize the incorrect pronunciation of the word, given that the samples provided for that word are both correct and incorrect. The ability to recognize correct and incorrect pronunciation is guaranteed as the ASR model used post-probability maximization to recognize the word together, with a limited variation of the underlying consistent phonemes. The acoustic model within the ASR is used to align the speech segment and the phonemes. Accordingly, it is used to find the phoneme segment in the input signal. Then, each segment's features are saved and sent to the phoneme model. Three states represent each phoneme in the HMM, and the DNN is used to estimate the posterior of the states in the HMM model. The acoustic parameters, represented as MFCC extracted from the recorded voice, are used as input. The phoneme model saves the features of the utterances to create a set of feature samples for each phoneme, along with the audio segment corresponding to each sample. The phonemes identification depends on the experts' data labeled in advance. As such, it is used to identify and label the sample and create a variation of the words similar to the G2P model. The classification model classifies the samples based on the trained model into correct or incorrect pronunciation. A single classifier is trained for each phoneme of the 41 canonical phonemes, with both correct and incorrect output, based on the features extracted from the correct and incorrect pronunciation of the recorded, generated by the G2P-similar mechanism and acquired data from the speech archive. Given that the model is designed with phoneme level, mispronunciation can be detected by training a classifier with each phoneme transcription. The output of each classifier is either 0/1, which refers to the quality of the pronunciation of that specific phoneme. In the training phase, a set of feature vectors for each phoneme's correct and incorrect pronunciation is used with their true labels (correct/ incorrect). The trained model is then used to identify whether the phoneme is pronounced correctly or not. The feedback is represented as highlighting the mispronounced phoneme compared to the correct one, besides using the specific classifier to decide whether it is in its correct form. The input vector is used as input for the rest of the classifiers. Accordingly, the output of the classifiers with the 1 output is said to be the pronounced phoneme and is used as feedback to the user as the incorrect substitute phoneme. Yet, there is a possibility that more than a single classifier produces 1 as an output. In such a case, the decision maker makes the final decision. 1) If one of the classifiers produces 1 as an output and that classifier is the one that corresponds to the correct phoneme in the word with the highest probability generated by the ASR, then the pronunciation is correct. 2) If more than the classifier produces 1 as an output, and one of these classifiers corresponds to the correct phoneme in the word, then the pronunciation is correct. 3) If one or more classifiers produce 1 as an output and no one corresponds to the correct phoneme in the correct word, then the pronunciation is incorrect, and the feedback is generated by presenting both the correct and incorrect pronunciation. This classification framework is illustrated in Fig. 3.

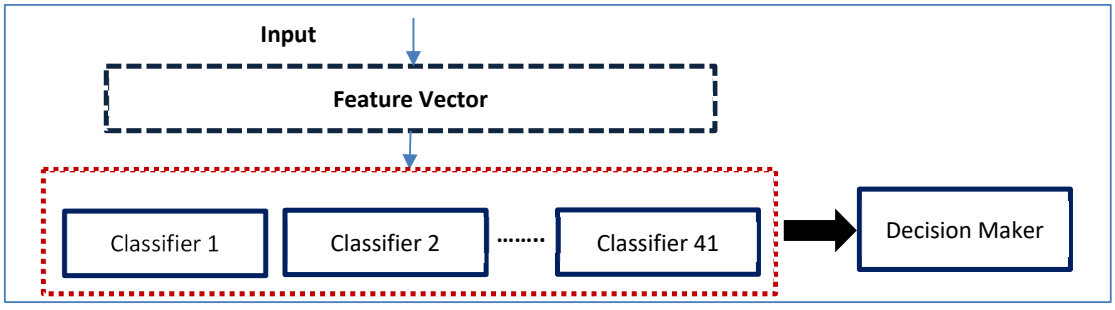

Fig. 3. The Classification Model

4. Experimental Results

The dataset used for the experiments includes correct and incorrect pronunciation of English words, as discussed earlier, which are recorded for the Arab EFL learners, generated by the phoneme model, and collected from the Speech Accent Archive—a summary of the utilized dataset is given in Table 8. After the data is generated using the phoneme model to create a variety of incorrect words and increase the volume of the incorrect subset, the final statistics of the dataset in terms of phonemes are presented in Table 9. The implementation of the detecting framework is illustrated in Fig. 4.

Table 8
Summary of the Collected Dataset

Item	Number of Words
Correct Students Records	450
Incorrect Students Records	480
Speech Accent Archive	1000 (selected words)
Total Phonemes	**1930**

Table 9
Summary of the Experimental Dataset

Item	Number of Items
Unique Phonemes (Canonical & Non-English Phonemes)	49 (41 & 8)
Unique Words (Correct and Incorrect)	450 (200 & 250)
Total Words (Repeated Correct and Incorrect)	2400 (1200 & 1200)
Non-English Phonemes in Each incorrect Word	1-2
Total Phonemes	**9580**

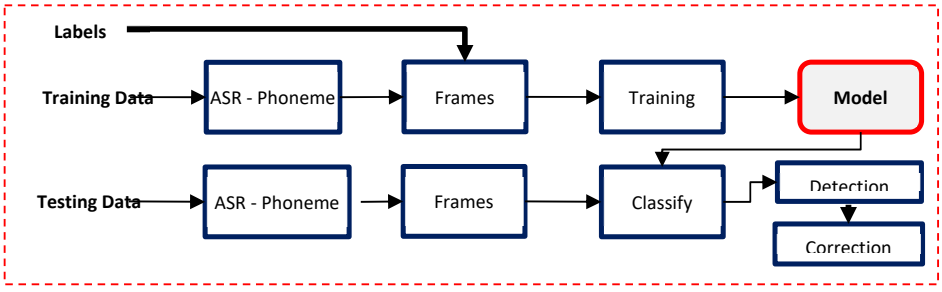

Fig. 4. Detection and Correction Implementation Steps

In the implementation of the speech generation, both the ASR and the phoneme models are implemented to identify the frame(s) in which the incorrect pronunciation has occurred. Because the data is labeled, identifying the incorrect phonemes' frames is applied directly. Librosa library and Python programming are used to convert the audio files and segments and filter the frames. The frames of incorrect pronunciations are then used to replace correct pronunciation in canonical English. As illustrated in Fig. 4, the detecting framework implementation is conducted using Python with Librosa for audio processing and feature extraction, with Scikit, learn for classification techniques. The raw audio is converted to Mel Spectrograms, and the threshold is applied, then the MFCC features are extracted for each frame. The individual frames are re-extracted using the ASR and the phoneme models (See Fig. 2) together with the true labels in the training process. Then, in the testing phase, the frames are extracted and classified accordingly based on the trained classifier for each phoneme.

The training and testing datasets are divided into 80%-20% for training and testing and 90%-10% for other experiments. Also, the results are reported for experiments in a 10-fold manner, in which the data is divided into ten folds, and ten experiments are conducted; in each, nine folds are used for training, and the other fold is used for testing. The overall accuracy, precision, recall, and f-measure of the whole data in the testing phase are reported for the five classifiers used in the experiments. The results of the classification techniques are presented in Table 10, Table 11, and Table 12 and illustrated in Fig. 5, Fig. 6, Fig. 7, and Fig. 8 (for 10-Fold Only).

Table 10
Results of the Detection Proposed Framework in 80%-20% Splitting

80-20	KNN	Bayes	DT	NN	SVM
Accuracy	97.3	96.3	97.6	97.3	94.3
Precision	0.98	0.973	0.986	0.974	0.948
Recall	0.961	0.962	0.975	0.97	0.938
F-Measure	0.975	0.97	0.981	0.971	0.95

Table 11
Results of the Detection Proposed Framework in 90%-10% Splitting

90-10	KNN	Bayes	DT	NN	SVM
Accuracy	97.5	96.2	97.8	97.2	94.4
Precision	0.981	0.974	0.987	0.974	0.952
Recall	0.965	0.97	0.971	0.971	0.94
F-Measure	0.977	0.971	0.983	0.971	0.95

Table 12
Results of the Detection Proposed Framework in 10-Fold Splitting

10-Fold	KNN	Bayes	DT	NN	SVM
Accuracy	97.3	97	98.2	97.6	94.8
Precision	0.981	0.98	0.988	0.976	0.97
Recall	0.969	0.972	0.98	0.972	0.96
F-Measure	0.976	0.969	0.982	0.972	0.95

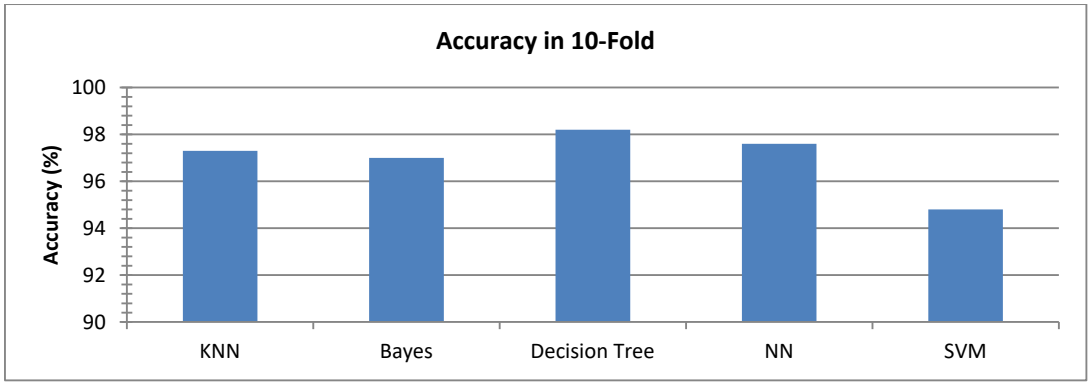

Fig. 5. The Accuracy of the Proposed Framework with 10-Fold Splitting

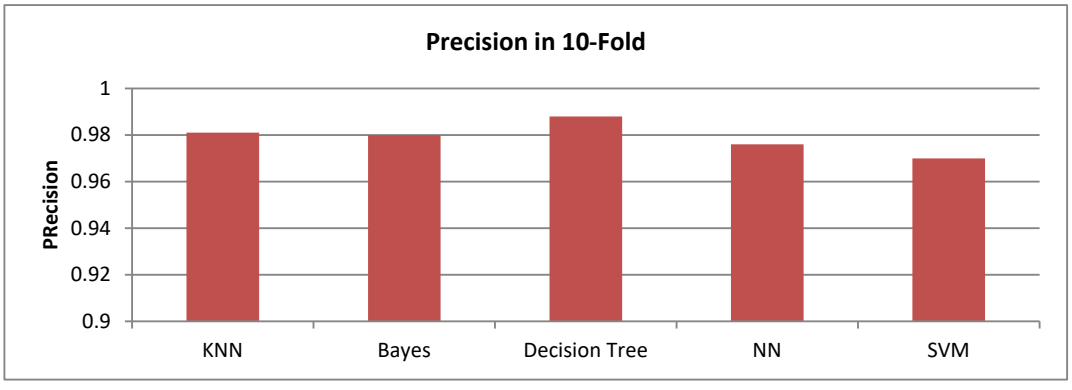

Fig. 6. The Precision of the Proposed Framework with 10-Fold Splitting

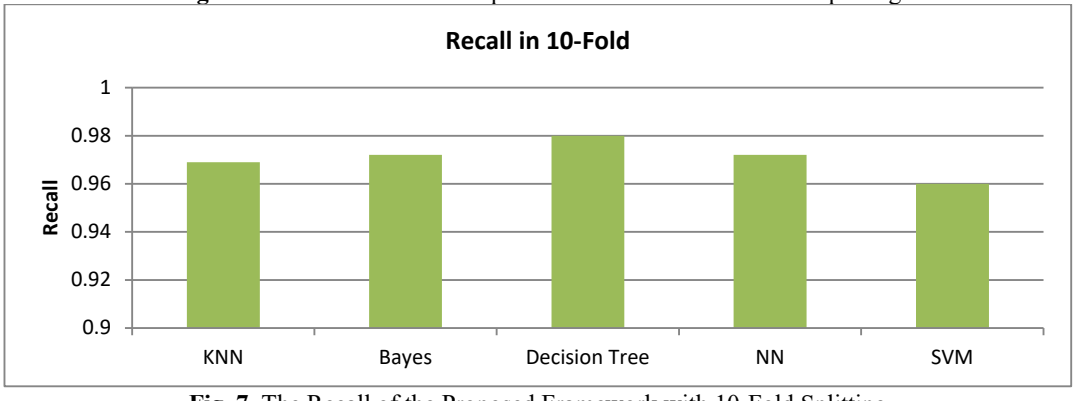

Fig. 7. The Recall of the Proposed Framework with 10-Fold Splitting

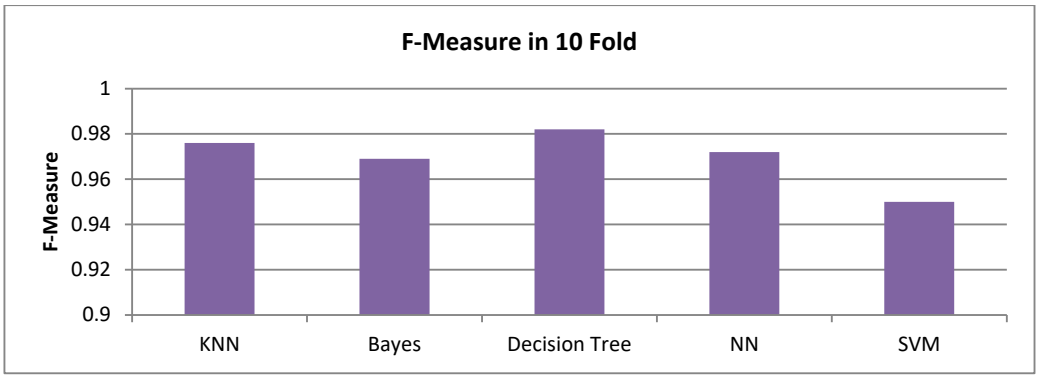

Fig. 8. The F-Measure of the Proposed Framework with 10-Fold Splitting

As noted, the decision tree performed the best among the tested classifiers with an accuracy of 98.2%, followed by the ANN, KNN, Bayesian, and finally, the SVM. The results showed that the precision of identifying the incorrect pronunciation is always higher, while the recall is low. The high precision-low-recall results are due to the data resulting from the G2P, which makes some of the wrong words similar to some correct variations in terms of features. Overall, the results showed the high ability of the proposed framework, with four out of five classifiers in detecting pronunciation errors of Arab EFL learners.

5. Conclusion

This paper proposed an automatic framework for pronunciation detection and correction. A set of contributions have been presented toward this goal. First, a qualitative methodology has been adopted to investigate the phonological and pronunciation errors for EFL/ESL students at AOU/KSA. The proposed solution has optimized and covered different accents in a particular language (English) by non-English native speakers. Second, the data were enriched using an outsourced archive and G2P model, which generates incorrect words from correct words modified by the wrong phoneme. Third, the automatic framework was developed with three models and five different machine learning methods to decide the right pronounced sounds using a prediction model. The result showed that the decision tree performed best, followed by the ANN, KNN, Bayesian, and the SVM. The results also showed that the precision of identifying the incorrect pronunciation is always higher than the recall.

Acknowledgements

The authors extend their appreciation to Arab Open University for funding this research.

References

Ai, R. (2015). Automatic pronunciation error detection and feedback generation for call applications. International Conference on Learning and Collaboration Technologies,

Alqarni, A. A. (2013). *The realization for the English voiceless postalveolar affricate/t [esh]/in Najdi Saudi ESL learners production*. Southern Illinois University at Carbondale.

Ambrozová, M. (2014). English Pronunciation difficulties among Czech students: Causes and compensation strategies. *Unpublished Thesis, Tomas Bata University, Zlin*.

Bensalah, R. F., & Betta, A. (2022). *The Impact of the Mother Tongue on the Phonetic Realization of Foreign Language Allophones. Algerian Arabic VS Received Pronunciation English* Université Ibn Khaldoun-Tiaret-].

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, *50*(5), 434-451.

Cai, J., & Liu, Y. (2018). Research on English pronunciation training based on intelligent speech recognition. *International Journal of Speech Technology*, *21*(3), 633-640.

Daana, H., & Khrais, S. (2018). The Acquisition of English and Arabic Onset Clusters: A Case Study. *English Linguistics Research*, *7*(1), 13-33.

Daana, H. A. (2009). *The development of consonant clusters, stress and plural nouns in Jordanian Arabic child language* The University of Essex].

Dobrovolsky, M., & Katamba, F. (1996). Phonology: the function and patterning of sounds. *Contemporary Linguistics: An Introduction. Essex: Addison Wesley Longman Limited*.

El Zarka, A. M. E. S. (2013). *The pronunciation errors of L1 Arabic learners of L2 English: The role of Modern Standard Arabic and vernacular dialects transfer* The British University in Dubai (BUiD)].

Eskenazi, M. (2009). An overview of spoken language technology for education. *Speech communication*, *51*(10), 832-844.

Gambari, A. I., Kutigi, A. U., & Fagbemi, P. O. (2014). Effectiveness of Computer-Assisted Pronunciation Teaching and Verbal Ability on the Achievement of Senior Secondary School Students in Oral English. *Gist Education and Learning Research Journal*, *8*, 11-28.

Hassan, E. M. I. (2014). Pronunciation problems: A case study of English language students at Sudan University of Science and Technology. *English Language and Literature Studies*, *4*(4), 31.

Kotani, K., & Yoshimi, T. (2018). Machine learning classification of pronunciation difficulty for learners of English as a Foreign Language. *Research in Corpus Linguistics*, 1-8.

Lai, K.-W. K., & Chen, H.-J. H. (2022). An exploratory study on the accuracy of three speech recognition software programs for young Taiwanese EFL learners. *Interactive Learning Environments*, 1-15.

Lee, A. (2016). *Language-independent methods for computer-assisted pronunciation training* Massachusetts Institute of Technology].

Li, W., Chen, N. F., Siniscalchi, S. M., & Lee, C.-H. (2017, August 20-24). *Improving Mispronunciation Detection for Non-Native Learners with Multisource Information and LSTM-Based Deep Models* Interspeech, Stockholm, Sweden.

Lu, L., Ghoshal, A., & Renals, S. (2013, December 8-12). *Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition* 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic.

McGraw, I., Badr, I., & Glass, J. R. (2012). Learning lexicons from speech using a pronunciation mixture model. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(2), 357-366.

Molina, C., Yoma, N. B., Wuth, J., & Vivanco, H. (2009). ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. *Speech communication*, *51*(6), 485-498.

Nijhawan, T., Attigeri, G., & Ananthakrishna, T. (2022). Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data*, *9*(1), 1-24.

Paula, A. J., Ferreira, O. P., Souza Filho, A. G., Filho, F. N., Andrade, C. E., & Faria, A. F. (2022). Machine Learning and Natural Language Processing Enable a Data-Oriented Experimental Design Approach for Producing Biochar and Hydrochar from Biomass. *Chemistry of Materials*, *34*(3), 979-990.

Peabody, M. A. (2011). *Methods for pronunciation assessment in computer aided language learning* Massachusetts Institute of Technology].

Poggensee, A. (2016). *The effects of globalization on English language learning: perspectives from Senegal and the United States* Western Michigan University, Kalamazoo, Michigan].

Rutherford, A., Peng, F., & Beaufays, F. (2014, September 14-18). *Pronunciation learning for named-entities through crowd-sourcing* Interspeech, Singapore.

Syafrizal, S., Wahyuni, S., & Syamsun, T. R. (2022). Pronunciation Errors of the Silent Consonants of Pariskian Junior High School Students. *ELTICS: Journal of English Language Teaching and English Linguistics*, *7*(2), 155-165.

Tepperman, J., Silva, J. F., Kazemzadeh, A., You, H., Lee, S., Alwan, A., & Narayanan, S. S. (2006, September). Pronunciation verification of children²s speech for automatic literacy assessment. In *INTERSPEECH*.

Weinberger, S. H., & Kunath, S. A. (2011). The Speech Accent Archive: towards a typology of English accents. In *Corpus-based studies in language use, language learning, and language documentation* (pp. 265-281). Brill.

Zrekat, Y. (2021). The Effectiveness of Blended Learning in Efl Context: An Experimental Study at Arab Open University-Ksa. *Turkish Online Journal of Qualitative Inquiry*, *12*(7).

Zrekat, Y., & Al-Sohbani, Y. (2022). Arab EFL University learners' perceptions of the factors hindering them to speak English fluently. *Journal of Language and Linguistic Studies*, *18*(1), 775-790.

Appendix A. Data Collection Information

Table A.1
General Information about the Interviewees

S/N	Code	Native Language	Age	Nationalities	Student Level (Semester)	English Usage
1	**RS-01**	Arabic	22	Saudi	2nd	Always
2	**RS-02**	Arabic	27	Saudi	2nd	Usually
3	**RS-03**	Arabic	23	Saudi	1st	Sometimes
4	**RS-04**	Arabic	25	Saudi	2nd	Sometimes
5	**RS-05**	Arabic	29	Saudi	3rd	Occasionally
6	**RS-06**	Arabic	30	Saudi	5th	Usually
7	**RS-07**	Arabic	21	Saudi	1st	Sometimes
8	**RS-08**	Arabic	22	Saudi	1st	Sometimes
9	**RS-09**	Arabic	21	Saudi	2nd	Rarely
10	**RS-10**	Arabic	28	Saudi	2nd	Sometimes
11	**RS-11**	Arabic	25	Egyptian	5th	Always
12	**RS-12**	Arabic	33	Egyptian	8th	Usually
13	**RS-13**	Arabic	22	Egyptian	2nd	Sometimes
14	**RS-14**	Arabic	27	Egyptian	7th	Sometimes
15	**RS-15**	Arabic	31	Egyptian	7th	Usually
16	**RS-16**	Arabic	35	Moroccan	3rd	Rarely
17	**RS-17**	Arabic	28	Moroccan	4th	Sometimes
18	**RS-18**	Arabic	28	Moroccan	5th	Rarely
19	**RS-19**	Arabic	21	Syrian	2nd	Always
20	**RS-20**	Arabic	26	Syrian	6th	Occasionally
21	**RS-21**	Arabic	20	Syrian	1st	Occasionally
22	**RS-22**	Arabic	20	Syrian	1st	Rarely
23	**RS-23**	Arabic	29	Syrian	5th	Sometimes
24	**RS-24**	Arabic	23	Sudanese	4th	Occasionally
25	**RS-25**	Arabic	20	Sudanese	1st	Rarely
26	**RS-26**	Arabic	19	Yemeni	1st	Rarely
27	**RS-27**	Arabic	32	Yemeni	7th	Always
28	**RS-28**	Arabic	27	Yemeni	7th	Usually
29	**RS-29**	Arabic	24	Yemeni	3rd	Sometimes
30	**RS-30**	Arabic	22	Yemeni	4th	Sometimes

Table A.2
Recorded Speech

ROUND	TEXT
1ST ROUND	I face difficulties in ……, also, I cannot pronounce ….. (*Complete sentences*).
2ND ROUND	“Sharp”, “paper”, “pain”, “Church”, “launching”, “Champion”, “plagiarism”, “Germany”, “judge”, “Rabbit”, “carry”, “hour”, "strange" and "garage" (*List of Words*).