Computer-Aided Pronunciation Training (CAPT) System for Non-native Learners of the Arabic Language

1- Introduction

In the following I will present the details of my research work in the field of Computer-Aided Pronunciation training (CAPT) for Arabic language. My work in the area started by supervising a master thesis titled “Automatic Pronunciation Error Detection of Nonnative Arabic Speech”. I continued the work by supervising a PhD thesis as co-supervisor with the title “Deep Learning based Pronunciation Error Detection for Non-native Learners of the Arabic Language”. I did these two investigations as the head of the speech processing group at CCIS. Due to the group expertize in the area we were able to secure a funded project with the title “Computer-Aided Pronunciation Training System for Non-native Learners of the Arabic Language”, where I was the PI of the project.

In the following I will first present the abstracts of Master thesis, PhD thesis, followed by a list of the papers in non-native Arabic pronunciation error detection and their abstract. Then I present the letter of NPST about funding the project, and I will end this report by a detailed report of our work in the funded project.

2- Master Thesis

- Title: Automatic Pronunciation Error Detection of Non-native Arabic Speech
- Date of defense: 5/6/2014
- My Role: Supervisor, with Prof. Ghulam Muhammed and Prof. Saad Alqahtani as co-supervisors.

Abstract:

Computer assisted language learning (CALL) and, more specifically, computer assisted pronunciation training (CAPT) have received considerable attention in recent years. CAPT systems can provide many potential benefits to both the language learner and the teacher. They allow continuous feedback to the learner without requiring the sole attention of the teacher; they facilitate self-study and encourage interactive use of the language in preference to rote-learning. One of the important processes in CAPT system is error detection, which locates the errors in the utterance. Although Arabic is currently one of the most widely spoken languages in the world, there has been relatively little research about detection of the pronunciation error by nonnative speakers compared to the other languages. This research is concerned with detecting pronunciation errors of nonnative Arabic speakers from Pakistan and India. The sounds in this study were taken from KSU database. By analyzing the speech of the Pakistani and Indian speakers in KSU database we found that five speech sounds (Tha'a ث, Ha'a ح, Sad ص , Dad ض , Tha'a ظ) were often mispronounced by non-native speakers, hence this study will concentrate on these five pronunciation errors. The speech recognition techniques used was Hidden Markov Model (HMM). The system was built with native and non-native speakers, and tested with nonnatives only. Goodness of Pronunciation (GOP) was calculated to detect if the phoneme was pronounced correctly or not. Comparison between the CAPT system judgment and the human judgment was performed. The result showed that GOP gave high accuracy, where the scoring accuracy were very good to excellent from 87% to 100% and the false rejection was from 0% to 10%.

3- PhD Thesis

- Title: Deep Learning-based Pronunciation Error Detection for Non-native Learners of the Arabic Language
- Date of defense: 26/1/2022
- My Role: Co-Supervisor, with Prof. Hassan mathkour as supervisor

Abstract:

In the recent decade, there has been great interest in computer-assisted pronunciation training (CAPT) systems. Many CAPT systems have been created for second language (L2) learners of various languages. Although Arabic is one of the most commonly spoken languages in the world, with the fifth highest number of speakers, little attention has been dedicated to computerized systems for the detection of pronunciation errors of non-Arabs. The Kingdom of Saudi Arabia is taking charge of serving the Arabic language, hence Arabic CAPT is important for the kingdom. Moreover, the CAPT system will enable the kingdom to help the large number of Muslims in the world to learn Arabic to read the Holy Quran. Mispronunciation detection and diagnosis (MDD) module is a vital component of CAPT systems, because it will detect the mispronounced phonemes and provide different types of feedback to the learner. Compared with other languages, Arabic MDD system needs more investigating due to the scarcity of research in Arabic MDD in

general and in using deep learning techniques for Arabic MDD in particular, and the lack of fully annotated non-native Arabic CAPT corpora. In this thesis, we aim to investigate different cutting edge deep learning techniques to build a high performance MDD system with feedback generation. We tackled the research problem by several folds. In the first fold, the phoneme recognition task of MDD was formulated as an object detection task, where phonemes were considered as objects in spectral images. In the second fold, we designed a system for articulatory feature (AFs) detection by formulating the AFs detection as a multi-label detection problem. The performance of the proposed models was evaluated using Arabic corpus and benchmark English corpus. The system had excellent performance and was also light.

In the third fold, we leveraged the excellent finding of the first and second folds to develop an MDD system for non-native Arabic speech. The proposed system has the ability to detect mispronounced phonemes from the speech at the utterance level, as well as detect the AFs of each phoneme, simultaneously. Through detecting the AFs in addition to the phonemes, our proposed system can provide beneficial feedbacks to the learners, at articulatory level. Moreover, we proposed using genetic algorithm to find the best hyper-parameters of the deep neural network of the proposed models. We compared the performance of the proposed system with the state-of-the art end-to-end MDD systems and our system had better result. In addition, we proposed using fusion between the proposed system and the end-to-end system and got better performance.
To tackle the problem of scarcity of non-native Arabic speech corpora, we investigated solving this by the use of different transfer learning techniques. We also developed a nonnative Arabic speech corpus (Arabic-CAPT). Finally, we investigated using the recent neural Text to Speech (TTS) technique to develop a new synthesized non-native Arabic speech corpus.

4- List of papers in Computer-Aided Pronunciation Training (CAPT) and computerization of Arabic pronunciation training

ISI journals:

1- Algabri, Mohammed, Hassan Mathkour, Mohamed Abdelkader Bencherif, Mansour Alsulaiman, and Mohamed Amine Mekhtiche. "Towards deep object detection techniques for phoneme recognition." *IEEE Access* 8 (2020): 54663-54680.

- **Abstract:** The use of cutting edge object detection techniques to build an accurate phoneme sequence recognition system for English and Arabic languages is investigated in this study. Recently, numerous techniques have been proposed for object detection in daily life applications using deep learning. In this paper, we propose the use of object detection techniques in speech processing tasks. We selected two state-of-the-art object detectors, namely YOLO and CenterNet, based on a trade-off between detection accuracy and speed. We tackled the problem of phoneme sequence recognition using three systems: the domain transfer learning system (DTS) from image to speech, intra-language transfer leaning system (IaTS) between speech corpora within the same language (English to English), and inter-language transfer learning system (IeTS) between speech corpora from dissimilar languages (English to Arabic). For English phoneme recognition, the Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus is used to evaluate the performance of the proposed systems. Our IaTS based on the CenterNet detector achieves the best results using the test core set of TIMIT with 15.89% phone error rate (PER). For Arabic phoneme recognition, the best performance, with 7.58% PER, was achieved using the CenterNet. These results show the effectiveness of using object detection techniques in phoneme recognition tasks.

Furthermore, based on the findings of this study, speech processing tasks may be treated as object detection tasks.

2- Algabri, Mohammed, Hassan Mathkour, Mansour M. Alsulaiman, and Mohamed A. Bencherif. "Deep learning-based detection of articulatory features in arabic and english speech." Sensors 21, no. 4 (2021): 1205.

- **Abstract**: This study proposes using object detection techniques to recognize sequences of articulatory features (AFs) from speech utterances by treating AFs of phonemes as multi-label objects in speech spectrogram. The proposed system, called AFD-Obj, recognizes sequence of multi-label AFs in speech signal and localizes them. AFD-Obj consists of two main stages: firstly, we formulate the problem of AFs detection as an object detection problem and prepare the data to fulfill requirement of object detectors by generating a spectral three-channel image from the speech signal and creating the corresponding annotation for each utterance. Secondly, we use annotated images to train the proposed system to detect sequences of AFs and their boundaries. We test the system by feeding spectrogram images to the system, which will recognize and localize multi-label AFs. We investigated using these AFs to detect the utterance phonemes. YOLOv3-tiny detector is selected because of its real-time property and its support for multi-label detection. We test our AFD-Obj system on Arabic and English languages using KAPD and TIMIT corpora, respectively. Additionally, we propose using YOLOv3-tiny as an Arabic phoneme detection system (i.e., PD-Obj) to recognize and localize a sequence of Arabic phonemes from whole speech utterances. The proposed AFD-Obj and PD-Obj systems achieve excellent results for Arabic corpus and comparable to the state-of-the-art method for English corpus. Moreover, we showed that using only one-scale detection is suitable for AFs detection or phoneme recognition.

3- Algabri, M., Mathkour, H., Alsulaiman, M., & Bencherif, M. A. (2022). Mispronunciation Detection and Diagnosis with Articulatory-Level Feedback Generation for Non-Native Arabic Speech. Mathematics, 10(15), 2727.

- **Abstract:** A high-performance versatile computer-assisted pronunciation training (CAPT) system that provides the learner immediate feedback as to whether their pronunciation is correct is very helpful in learning correct pronunciation and allows learners to practice this at any time and with unlimited repetitions, without the presence of an instructor. In this paper, we propose deep learning-based techniques to build a high-performance versatile CAPT system for mispronunciation detection and diagnosis (MDD) and articulatory feedback generation for non-native Arabic learners. The proposed system can locate the error in pronunciation, recognize the mispronounced phonemes, and detect the corresponding articulatory features (AFs), not only in words but even in sentences. We formulate the recognition of phonemes and corresponding AFs as a multi-label object recognition problem, where the objects are the phonemes and their AFs in a spectral image. Moreover, we investigate the use of cutting-edge neural text-to-speech (TTS) technology to generate a new corpus of high-quality speech from predefined text that has the most common substitution errors among Arabic learners. The proposed model and its various enhanced versions achieved excellent results. We compared the performance of the different proposed models with the state-of-the-art end-to-end technique of MDD, and our system had a better performance. In addition, we proposed using fusion between the proposed model and the end-to-end model and obtained a better performance. Our best model achieved a 3.83%

phoneme error rate (PER) in the phoneme recognition task, a 70.53% F1-score in the MDD task, and a detection error rate (DER) of 2.6% for the AF detection task.

Conferences:

1- Al Hindi, Afnan, Mansour Alsulaiman, Ghulam Muhammad, and Saad Al-Kahtani. "Automatic pronunciation error detection of nonnative Arabic Speech." In 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA), pp. 190-197. IEEE, 2014.

- **Abstract:** Computer assisted language learning (CALL) and, more specifically, computer assisted pronunciation training (CAPT) have received considerable attention in recent years. CAPT allows continuous feedback to the learner without requiring the sole attention of the teacher; it facilitates self study and encourages interactive use of the language in preference to rote learning. One of the important processes in CAPT system is error detection, which locates the errors in the utterance. Although Arabic is currently one of the most widely spoken languages in the world, there has been relatively little research about detection of the pronunciation error by nonnative speakers compared to the other languages. This research is concerned with detecting pronunciation errors of nonnative Arabic speakers from Pakistan and India. All the sounds in this study were taken from King Saud University (KSU) Arabic Speech Database. By analyzing the speech of the Pakistani and Indian speakers in KSU database we found that five phonemes were often mispronounced by nonnative speakers, hence this research will concentrate on pronunciation errors in these five phonemes. The system was built with native and nonnative speakers, and tested with nonnative only. For each phoneme, the Goodness of Pronunciation (GOP) was calculated and compared with a threshold to decide if the phoneme was pronounced correctly or not. The result showed that GOP gave high accuracy, where the scoring accuracy was very good to excellent from 87% to 100%, and the false rejection was zero to less than 10%. This machine judgment is compared with human judgment and the comparison shows excellent agreement between them.

2- Alsulaiman, Mansour, Zulfiqar Ali, Ghulam Muhammad, Afnan Al Hindi, Taha Alfakih, Hussein Obeidat, and Saad Al-Kahtani. "Pronunciation errors of non-Arab learners of Arabic language." In 2014 International Conference on Computer, Communications, and Control Technology (I4CT), pp. 277-282. IEEE, 2014.

- **Abstract:** Arabic is one of the most widely spoken languages in the world, but little attention has been paid to detect pronunciation errors of non-Arabs from different nationalities. In this paper, the speech of four nationalities of Asian non-Arabs speaking different mother languages is analyzed to identify their pronunciation errors while learning Arabic as a foreign language. Two human experts have evaluated the speech of all speakers for mispronunciation, and evaluation show that the nature of errors is almost the same for all nationalities under investigation with the fact that pharyngeal, alveodental, and interdental sounds are difficult to pronounce by learners of Arabic. Some of the errors are due to sounds that are not present in mother languages or the learner is unable to pronounce the phoneme correctly due to similar place of production or/and manner of articulation. An interesting observation is that pronunciation errors of some learners are due to switching pairs of phonemes.

3- Mansour Alsulaiman, Mohammed Algabri, Hassan Mathkour, Mohamed A. Bencherif, Ghulam Muhammad, Saad Al-Gahtani, Mohammed Faisal, Mohamed Amine Mekhtiche, “Development and Analysis of a Versatile Dataset of Speech, Real and Synthesized of Arabic Learners,” 3rd International Conference on Computing and Information Technology (ICCIT), 2023.

- Abstract: Computer-Aided Pronunciation Training (CAPT) systems are gaining popularity recently due to the advancements in deep neural networks (DNN) and machine learning and the availability of databases of speech of language learners. Unfortunately, research in Arabic CAPT systems suffer from the lack of CAPT datasets compared to other languages. In this paper, we present the details of and the ideas used in the development of a versatile dataset of speech, real and syn-

thesized, of Arabic Learners. To develop the dataset, we utilized an existing Arabic speech corpus, King Saud University Speech Database (KSU-SD). KSU-SD's main application was a speaker recognition system, but it was designed to be useful in other applications such as CAPT systems. KSU-SD includes a large number of speakers from diverse nationalities (Saudis, Arabs, and Non-Arabs). KSU-SD contains the recording of about 60 non-native speakers from more than 20 nationalities, hence we selected it to build the non-native Arabic-CAPT corpus. The developed corpus consists of transcribed, segmented, and annotated speech, which makes it suitable for building Arabic CAPT systems. In addition to presenting the details of developing the dataset, we also present an analysis of the text and speech errors of the dataset. The dataset was verified by many CAPT systems.

جامعة الملك سعود
King Saud University

إفادة

تفيد الخطة الوطنية للعلوم والتقنية والابتكار بجامعة الملك سعود بأن سعادة الدكتور / منصور بن محمد السليمان الأستاذ بكلية علوم الحاسب والمعلومات، رقم وظيفي (١١٩٦٢٣)، هو الباحث الرئيس للمشاريع المبينة بالجدول ادناه، وهذه المشاريع ممولة من برنامج التقنيات الاستراتيجية بالخطة الوطنية للعلوم والتقنية والابتكار:

م	رقم المشروع	اسم المشروع	المدة
١	08-INF167-02	التعرف على المتحدث العربي ARABIC SPEAKER RECOGNITION	٢٠١٠-٢٠١٢
٢	١٢ MFD2474-02	تقييم الامراض الصوتية بالحاسب Automatic Voice Pathology Assessment	٢٠١٣ - ٢٠١٥
٣	3-17-09-001-0003	نظام حاسوبي لتعليم اللغة العربية لغير الناطقين بها Computer-Aided Pronunciation Training System for Non-native Learners of the Arabic Language	٢٠٢٠-٢٠٢٢
٤	5-18-02-001-0003	نظام ترجمة محمول للغة الإشارة السعودية Saudi Sign Language Translation Companion System	٢٠٢٠-٢٠٢٢

وقد اعطيت له هذه الإفادة لسعدته بناء على طلبه لتقديمها الى من يهمه الامر ودون أدنى مسؤولية على الوحده.

مدير وحدة العلوم والتقنية والابتكار

أ. د. أحمد بن عبد الله الخازم

Transmittal Letter

Date: 11th/May/2023

Researcher name: Mansour Alsulaiman
College: Computer and Information Sciences
Department: Computer Engineering
Address: P.O. Box 51178, Riyadh 11543
E-mail: msuliman@ksu.edu.sa

Dear Prof. Ahmed Alkhazim

We are submitting to you the final report of our project. The report is entitled technical report for second year (final year) of the project Computer-Aided Pronunciation Training (CAPT) System for Non-native Learners of the Arabic Language. The purpose of the report is to inform you of our work in the project. The report presents the methods that we proposed, the results that we got and the research that we published for the CAPT system. This report also present and discuss the databases that we developed for speech of Non-Arabs. If you should have any questions concerning our project, please feel free to contact Mansour Alsulaiman at 0503255927 or msuliman@ksu.edu.sa.

Sincerely,
Professor
Mansour Alsulaiman (PI)

Affiliation: College of Computer and Information Sciences, King Saud University.

Title Page

Submitted for

National Science, Technology and Innovation Plan (NSTIP)

King Saud University

Project title

Computer-Aided Pronunciation Training (CAPT) System for Non-native Learners of the Arabic Language

Project number

3-17-09-001-0003

Project Investigator

Mansour Alsulaiman

Year

2023

Abstract

Computer-Aided Pronunciation Training (CAPT) System for Non-native Learners of the Arabic Language is an important topic for the Kingdom of Saudi Arabia (KSA), the site of the two holy mosques. This CAPT project is a joint work between two institutes of King Saud University; College of Computer and Information Sciences and the Arabic Linguistics Institute. Both teams used their expertise in machine learning and Arabic Language to conduct research and develop an effective CAPT system. The system can be used offline or online to help learners of Arabic language correct their Arabic pronunciation.

This project has two main goals, one is to design and develop a comprehensive non-native Arabic speech corpus to solve the lack in the non-native Arabic corpora. The second goal is to design and develop an accurate and efficient Arabic CAPT system.

To achieve the first goal, we designed and developed the non-native Arabic speech database, and called it KSU-CAPT-Non-Arabs database. The database was developed in two sessions. Firstly, we designed the text for the first session and recorded the speech of 220 non-native Arabic speakers. Based on our experience in recording session 1, we designed the text for session 2 and recorded the speech of 230 non-native Arabic speakers. Then, the speech of the two sessions was verified and cleaned from extra sounds then sent to experts in Arabic language to annotate it. The experts annotated the speech of 60 speakers from each of the recorded two sessions.

To achieve the second goal, we proposed a new technique of using deep learning for detection and recognition of phoneme and articulatory features (AF). In this proposed technique, we treat the phonemes and AFs as objects in 3 channels spectral images of the speech. By this proposed technique we were able to recognize the sequence of phoneme from the whole utterance of the non-native Arabic speakers, and not only from words. We used the detected phonemes for mispronunciation detection and diagnosis task and the detected AFs for feedback of error in pronunciation. This achievement was published in a Q2 ISI journal. Then, we designed and built a high performance Arabic CAPT system that can detect the error in pronunciation and detect the corresponding AFs, not only in words but even in sentences. The proposed system obtained an

excellent result compared with the state-of-the-art methods. This achievement was published in a Q1 ISI journal. The investigation in the two papers used existing databases, because the project database was still in development at the first stages of the project.

Next, we built a new CAPT system based on the first session of the labeled database and got excellent results. We continued the investigation and developed the CAPT system based on the second session of the database and also got excellent results.

Finally, to let users, test our system, we developed an interface for our Arabic CAPT system.

Acknowledgments

This work is supported by National Science, Technology and Innovation Plan (NSTIP) in King Saud University under grant number 3-17-09-001-0003. The authors are grateful for this support.

Table of Contents

List of Figures

List of Tables

Report Body

1. Introduction

Computer-Aided Pronunciation Training (CAPT) System for Non-native learners of the Arabic Language is an important topic for the Kingdom of Saudi Arabia, as the Kingdom is taking charge of spreading knowledge about Islam and helping Muslims learn Arabic the language of the Quran. In this context, we have been working on this CAPT project which is a joint work between two institutes of King Saud University; the College of Computer and Information Sciences team (CCIS-T) and the Arabic Linguistics Institute team (ALI-T). Both teams used their respective expertise in Machine Learning and Arabic Language to conduct research and develop an automatic solution that can detect errors of pronunciation and detect the location of the error, its type and give a feedback to correct the pronunciation. Output of the project is a pilot CAPT system that can be used in offline or online ways, that can help learners of Arabic language, all over the world, correct their Arabic pronunciation.

Building an Arabic CAPT system requires a Non-Native Arabic Speech database that contains diverse speech and pronunciation errors. Hence this was the first objective of the project. In the proposal we aimed to record two sessions of speech of Arabic learners. The recorded database should stress on pronunciation errors, and have enough speakers with detailed phoneme annotations. ALI-T team had years of expertise in teaching the Arabic language for Non-Arabs, and conducted many research studies to enforce this expertise. Based on this expertise they were able to propose a methodology to construct the texts most suitable for the project, which took a considerable time and efforts. They proposed a text based on this methodology. The text went into many refinements from the whole project team until a set of 25 long sentences and some 61 very special pairs of words were finally selected. After recording and developing session 1 and based on the experience in the development and analysis of session 1, ALI-T refined and enhanced the methodology then selected new text for recording in session 2. The new text contains Arabic words only rather than sentences or paragraphs as was the case in session 1.

Due to the COVID restrictions the recording could not be done in a controlled face to

face set up, and instead the recording was done using an app on a mobile. Details of the database and building it are presented in section 3.

We conducted detailed search in the literature for CAPT systems in general and CAPT for Arabic language in particular. This was the second objective of the project. Details of this search are presented in section 4. From the search we found that some researchers were using long established techniques such as hidden Markov models (HMM), while other researchers were investigating using deep neural networks. We proposed a new technique based on deep neural networks for recognizing phonemes and AFs and built a CAPT system using the proposed technique, hence we accomplished third objective of the project. The proposed method treated phonemes and AFs as objects in 3 channels spectral images of the speech. We published a paper on the proposed new technique. We investigated new improved models of this new technique with one network recognizing the phonemes and AFs in and MDD system and got excellent results that are comparable or better than state-of-the-art published research. We published our second paper with these findings and accomplishments. Details of these CAPT and MDD systems and their results are presented in sections 5 and 6, respectively.

2. Objectives

As we briefly presented in the introduction, the project has three technical objectives, namely:

- building a database of speech of Arabic learners
- conducting a literature review of CAPT systems in general and Arabic CAPT in particular
- Building a CAPT system using the developed speech database.

Building the database and conducting the literature review are two necessary steps in order to accomplish the main objective of the project which is building the CAPT system. In addition to these technical objectives, the project has two objectives namely: establishment of a multidisciplinary research group through the collaboration of the CCIS-team and the ALI-team, and dissemination of the results and conclusions at journals and conferences. In the following we will briefly present our accomplishments in each of these objectives.

2.1. Building the database

To build the database we had do design a text best suitable for building a CAPT system. The text had to include the main or majority of pronunciation errors by Arabic learners and at the same time should be of reasonable length in order to be easily pronounced and easily recorded by non-Arabic speaking volunteers. In the initial proposal, we were hoping to conduct the recording in face to face controlled sessions. Unfortunately, Covid-19 restrictions did not allow this and we had to do the recording using a mobile app that we developed for the project. This has advantages and disadvantages as will be presented in section 3.2. Detail of the recording steps, protocols, cleaning, speech labeling are presented in sections 3.2 and 3.3.

2.2. Literature Review

We conducted a comprehensive review. This review is presented in section 4. From the review we were able to propose a new technique that used object detection techniques to recognize the phonemes and AF and hence build an effective CAPT system.

2.3. Building the CAPT system

The recording and labeling of the developed database took long time, hence we initially used part of KSU database that is owned by the CCIS-team to build the AFs detection module of the CAPT system and got excellent results that we published in an ISI paper. The proposed method and the results of the investigation are presented in section 5. Next, we investigated other improved models of the proposed technique to build an MDD system that recognize phoneme and detect AFs in one model and got excellent results that we published in another ISI paper. The proposed method and the results of the investigation are presented in section 6. The proposed technique and the improved models were used to build a CAPT system using the recording of session 1 of the project database. The performance of the system was evaluated and compared to human judges and gave excellent results. We used the model based on session 1 as the pre-training model to build a CAPT system based on session 2. The performance of the second CAPT system was evaluated and compared to human judges and gave excellent results. The results and analysis of the performance of the first and second CAPT systems will be presented in section 6. Finally, we designed a user interface application that allows the non-native speakers to use the developed system in real time.

2.4. Establishment of a multidisciplinary research group through the collaboration of the CCIS-team and the ALI-team Institute

The CCIS-team and ALI –team worked together and established a CAPT research group that have accomplished: Literature review of Arabic CAPT in the Arabic literature and CAPT and Arabic CAPT in the English literature, designed a text for research on CAPT, recorded speech of Arabic learners, annotate the speech and analyze the errors, and published papers on novel methods for building a CAPT system.

2.5. Dissemination of the results and conclusions at conferences and in journals

We published two ISI papers on novel methods for building a CAPT system. We have a paper accepted at an IEEE conference on describing the KSU-CAPT database. And, we are finalizing a new paper on describing the two sessions of the developed speech database. We are also working in a paper describing the results of the developed CAPT system using sessions 1 and 2.

3. Literature review of CAPT systems

3.1. Arabic CAPT in Arabic literature

(Done in Arabic by ALI-team from the Arabic literature)

تعد المشكلات الصوتية التي تواجه دارسي العربية من الناطقين بغيرها من أولى المشكلات اللغوية التي يعاني منها هؤلاء الدارسون، وأبرزها على الإطلاق؛ إذ يواجهون مشكلة حقيقية في ضبط النطق السليم لعدد من الأصوات العربية وتمييزها عن غيرها مما يشبهها في الخصائص أو يتقارب معها في المخارج، ويلجأ بعضهم إلى أن يستبدلوا بها غيرها من الأصوات الأخرى التي يسهل عليهم نطقها، ويهدف هذا المشروع إلى مساعدة دارسي العربية من الناطقين بغيرها في التغلب على هذه المشكلات. ويجمع هذا المشروع بين مجالين من مجالات التخصص العلمي: مجال دراسة الأصوات العربية وتعليمها، ومجال تصميم النظم الحاسوبية وبنائها، وذلك لإنتاج نظام لتعليم نطق اللغة العربية الصحيحة يسهم في تعليم الطلبة مهارة إتقان اللغة العربية من الجانب النطقي اعتمادًا على تحليل الأخطاء الصوتية التي تم جمعها من خلال قراءة الطلاب لعدد من النصوص تم اختيارها وفقا لقواعد محددة.

وفي هذه الصفحات نقدم موجزًا نعرض من خلاله لاتجاهات البحث في المشكلات الصوتية التي تواجه دارسي العربية من الناطقين بغيرها، والمسارات التي تحكمه، وننتهي إلى تقرير واقع المنجز البحثي في هذا المجال، من حيث مركز الاهتمام، وطبيعة الطرح، وواقع تعليم الأصوات العربية للناطقين بغيرها.

ومن خلال مطالعة المصنفات في هذا المجال تبين أنها اتخذت المسارات الآتية:

- المشكلات اللغوية الشائعة
- المشكلات الصوتية
- تعليم الأصوات
- الأصوات والحاسوب

وفيما يأتي عرض لهذه المسارات وأبرز المصنفات التي تمثلها.

1- الأخطاء اللغوية الشائعة

ونقصد بذلك تلك الدراسات التي تناولت الأخطاء اللغوية الشائعة التي يقع فيها دارسو العربية من الناطقين بغيرها بصفة عامة ومن بينها الأخطاء الصوتية ومن هذه الدراسات دراسة

" لبي2015" حيث هدفت في الفصل الأول منها إلى التعرف على المستوى الصوتي ومعرفة الأخطاء اللغوية الشائعة فيه

لدى طلاب المرحلة الثانوية في المالديف، وطلاب المستويين الرابع والخامس بكلية الدراسات الإسلامية بالماليديف من خلال الاطلاع على قراءات الطلاب وكتاباتهم وذلك عبر زيارتهم في فصولهم وإجراء اختبارات لكشف أخطاء الطلاب في نطق الأصوات الحلقية والمطبقة والأسنانية، وقد خلصت الدراسة إلى أن 74% من طلاب المستوى الأول ينطقون هذه الأصوات نطقًا صحيحًا، و26% يجدون صعوبة في نطقها، وأن 73% من طلاب المستوى الثاني ينطقون هذه الأصوات نطقًا صحيحًا، و27% يجدون صعوبة في نطقها، وأن 67% من طلاب المستوى الثالث ينطقون هذه الأصوات نطقًا صحيحًا، و33% يجدون صعوبة في نطقها، وأن 67% من طلاب المستوى الرابع ينطقون هذه الأصوات نطقًا صحيحًا، و33% يجدون صعوبة في نطقها، وأن 91% من طلاب المستوى الخامس ينطقون هذه الأصوات نطقًا صحيحًا، و9% يجدون صعوبة في نطقها. وقد أرجعت الدراسة ارتفاع نسبة النطق الصحيح عند طلاب الماليديف إلى كون الشعب الماليديفي مسلمًا، وارتباطه بالقرآن قويًا، واهتمام المراكز القرآنية والمدارس الحكومية بتدريب الطلاب على نطق الأصوات العربية نطقًا صحيحًا، بالإضافة إلى وجود تشابه كبير بين أصوات العربية وأصوات اللغة الماليديفية.

ومنها دراسة: (الحديبي 2017م) وهدفت الدراسة إلى تقديم مقترح للتغلب على المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخرى. واعتمدت على المنهج الوصفي. وتكونت مجموعة البحث من 115 مختصا من العاملين في ميدان تعليم اللغة العربية للناطقين بلغات أخرى. وتمثلت أدوات البحث في استبانة المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخرى، وتصور مقترح للتغلب على المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخرى. وتوصلت نتائج البحث إلى تحديد المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخري، وتضمنت مائة واثنتين وعشرين مشكلة، منها ضعف معالجة الجوانب الصوتية للغة العربية في المقرر، وعدم وجود معايير محددة لقبول المعلمين للعمل في جهات تعليم اللغة العربية للناطقين بلغات أخرى. كما توصلت النتائج إلى وضع تصور مقترح للتغلب على المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخري، واستند التصور على النقاط التالية، رؤية التصور، رسالة التصور، أهداف التصور، مصادر أعداد التصور، مسلمات التصور، متطلبات التصور، ضبط التصور، مكونات التصور. وأوصي البحث بضرورة الاهتمام بتصميم المقررات التعليمية المقدمة في برامج تعليم اللغة العربية للناطقين بلغات أخرى، بحيث تكون متوافقة مع المعايير العالمية لإعداد مقررات تعليم اللغة الأجنبية، وضرورة إقامة دورات تدريبية لمعلمي اللغة العربية للناطقين بلغات أخرى مبنية على احتياجاتهم الفعلية، وبحيث تكون مقدمة لحلول فعلية للمشكلات التي يواجهها المعلمون في أثناء الخدمة.

ومنها دراسة (الدياب 2012م) وقد قام بتقسيم هذه الرسالة إلى ثلاثة أقسام رئيسية، تناول الباحث في القسم الأول المشكلات اللغوية وشملت على الصعوبات الصوتية والنحوية والكتابية والدلالية وصعوبات القراءة وقد اندرج تحت كل عنوان الصعوبات التي يواجهها الطلاب الأتراك وختمت كل عنوان بالمقترحات المناسبة. وقد تناول الصعوبات التي تواجهها العملية التعليمية للغة العربية على المستوى الصوتي من خلال المحاور الآتية: النظام الصوتي العربي وأثره في عملية التعليم، واختلاف النظامين الصوتيين العربي والترك، والمعلم وخبرته في نطق الأصوات. ولاحظ أن جميع الأصوات التي يخطئ فيها الأتراك غير موجودة في اللغة التركية، هذا أولاً، وثانياً فإن هذه الأصوات في الأساس مخرجها صعب، أو الصفات التي تمتاز بها تجعل مخرجها صعباً، ويقترح الباحث أن يكون هناك كتاب صوتي يشمل على أهم المشكلات الصوتية التي تعاني منها عملية التعليم

ويكون مرفقاً بقرص للسماع يكون بصوت عربي واضح.

وهذا بيان بأبرز ما ألف من المصنفات في هذا المسار:

- الأخطاء اللغوية الشائعة لدى طلاب المرحلة الثانوية في المالديف، دراسة تحليلية، محمد فارس عثمان لبي، بحث مقدم لنيل درجة الماجستير في الآداب، قسم اللغة العربية، كلية اللغات، جامعة المدينة العالميّة ماليزيا، أغسطس 2015 م
- المشاكل التي تواجه الأتراك في تعليم اللغة العربية والمقترحات، أحمد الدياب، أطروحة ماجستير، قسم اللغة العربية، معهد العلوم التربوية، جامعة غازي، جمهورية تركيا، 2012م.
- أخطاء القراءة الجهرية باللغة العربية للطلبة الناطقين بالملايوية: دراسة وصفية تحليلية، علي، عاصم شحادة، مجلة الدراسات اللغوية والأدبية، الجامعة الإسلامية العالمية، ماليزيا
- تحليل الأخطاء كمدخل لعلاج الصعوبات والأخطاء اللغوية الشائعة في تعليم اللغات الأجنبية، حسين، أحمد علي محمد، مجلة القراءة والمعرفة، الجمعية المصرية للقراءة والمعرفة، كلية التربية، جامعة عين شمس.
- أثر برنامج تدريبي للبرمجة اللغوية العصبية على صعوبات تعلم اللغة العربية للناطقين بغيرها، على، أحمد رمضان محمد، مجلة التربية كلية التربية، ع 155، ج 2، جامعة الأزهر.
- تصور مقترح للتغلب على المشكلات التي تواجه الجهات المعنية بتعليم اللغة العربية للناطقين بلغات أخري، الحديبي، علي بن عبدالمحسن بن عبدالتواب، مجلة كلية التربية، جامعة أسيوط -كلية التربية، مج33, ع1، 2017م
- الصعوبات والأخطاء اللغوية لدي متعلمي اللغة العربية من الصينيين، السيد مصطفي محمد عبيد، مجلة كلية الآداب والعلوم الإنسانية، جامعة قناة السويس -كلية الآداب والعلوم الإنسانية، ع3، 2011م

2- تعليم الأصوات

ونقصد بذلك تلك الدراسات التي تناولت الأصوات العربية وتعليمها للناطقين بغيرها، منها دراسة "جاسم 1994م" ويركز الباحث في دراسته على طبيعة أصوات الحلق، واللهاة، والتفخيم، ويربطها بطرق تعليم الأصوات العربية، ومن خلال دراسته، يكتشف الباحث أن الطلاب الناطقين بغير اللغة العربية في الجامعة الإسلامية العالمية بماليزيا يواجهون صعوبات في نطق الحروف العربية بسبب التداخل مع اللغة الأم. وقد أقام الباحث دراسة مقارنة بين اللغتين للكشف عن مدى اختلاف الظواهر اللغوية بينهما، أي بين اللغة العربية واللغة الأم للمتعلمين من خلال امتحان شفوي وكتابي لدى الطلبة الناطقين بغير العربية. ويهتم الباحث في دراسته بعدد من الدراسات التقابلية بين اللغتين لأجل توضيح العوائق التي يواجهها الطلبة في تعلم أصوات اللغة العربية.

ودراسة "داؤود2018م" وهدفت الدراسة إلى: التعرف على الأصوات اللغوية، وأسس اختيارها في برامج تعليم اللغة العربية للناطقين بغيرها، والطرق العلمية المتبعة في تقديم الأصوات في برامج تعليم اللغة العربية للناطقين بغيرها متبعًا في ذلك المنهج الوصفي. وتوصل البحث الى نتائج منها: أن هناك أسسًا علمية يتم في ضوئها اختيار الأصوات اللغوية في برنامج تعليم اللغة العربية للناطقين بغيرها يجب اتباعها لاختيار الأصوات اللغوية في برامج تعليم اللغة العربية للناطقين بغيرها، وأن هناك

أسسًا علمية لتقديم الأصوات اللغوية في برامج تعليم اللغة العربية للناطقين بغيرها، وهناك أساليب لتقويم الأصوات اللغوية. ويوصي البحث: بمراعاة الأسس العلمية واللغوية في اختيار الأصوات اللغوية وتقديمها في برامج تعليم اللغة العربية للناطقين بغيرها، والتركيز على أساليب التقويم المتنوعة أثناء تدريس الأصوات وبعده.

وهذا بيان بأبرز ما ألف من المصنفات في هذا المسار:

- عبد اللطيف، محمد شاكر القاطوع. (1999م) الأصوات العربية وتعليمها لغير الناطقين بها، بحث مقدم استكمالا لمتطلبات نيل درجة الماجستير في اللغة العربية بكلية الدراسات العليا في الجامعة الأردنية 1999م،
- جاسم، علي جاسم. (1994م) تعليم الأصوات العربية الساكنة الخلفية والمفخمة لغير الناطقين بالعربية. (1994م) رسالة ماجستير منشورة، ماليزيا: الجامعة الإسلامية العالمية بماليزيا.
- محمد، أيوب محمد داؤد (2018م) الأصوات اللغوية وأسس تقديمها في برامج تعليم اللغة العربية للناطقين بغيرها بحث تكميلي لنيل درجة الماجستير في علم اللغة التطبيقي، كلية للغة العربية، جامعة إفريقيا العالميّة 2018م.
- القاويش، حنان رفيق محمد، (2012م) دراسة حول تعليم الصوتيات العربية للناطقين باليابانية، صحيفة الألسن: سلسلة في الدراسات الأدبية واللغوية، جامعة عين شمس -كلية الألسن.
- المسند، حمزة كريم محمد، الدجانى، بسمة أحمد صدقى، (2016م) منهاج تعليم العربية للناطقين بغيرها: تعليم الأصوات أنموذجا، مجلة دراسات وأبحاث، جامعة الجلفة، سبتمبر، ع 24، الجزائر.
- الفاعوري، عوني صبحي، أبو عمشة، خالد، (2005م) تعليم العربية للناطقين بغيرها: مشكلات وحلول الجامعة الأردنية نموذجاً، دراسات -العلوم الإنسانية والاجتماعية، مج 32 , ع 3، الجامعة الأردنية - عمادة البحث العلمي.
- عبد الحميد، عبد الغنى اكوريدى، (2015م) تحديات تدريس الاصوات العربية للناطقين بغيرها وطرق علاجها، أبحاث المؤتمر السنوي العاشر: تعليم اللغة العربية للناطقين بغيرها في الجامعات والمعاهد العالمية، معهد ابن سينا للعلوم الانسانية ومركز الملك عبد الله بن عبد العزيز الدولي لخدمة اللغة العربية.
- المجدوب، عز الدين بن محمد، (1988م) مساهمة في إصلاح نطق العربية لغير الناطقين بها من الفرنسيين، حوليات الجامعة التونسية، ع29.
- أرمبرستر، بوني، (2011م) علم الأصوات أو التعلم القائم على البحث العلمي في القراءة، ترجمة فتحي يونس، مجلة القراءة والمعرفة، جامعة عين شمس -كلية التربية -الجمعية المصرية للقراءة والمعرفة، ع 117.
- خريوش، عبدالرؤوف (2002) كيفية تدريس الأصوات الفصيحة المفقودة في اللهجات العربية الحديثة لمتعلمي اللغة العربية كلفة ثانية، مجلة كلية التربية بأسوان – مصر، ع 17،
- ماهر عيسى حبيب، تعليم الصوتيات العربية وتعلمها بالحاسوب، مجلة جامعة تشرين للبحوث والدراسات العلمية –

سلسلة الآداب والعلوم الإنسانية المجلد 23، العدد 3

- مشكلات تعليم الأصوات العربية للطلبة الصينيين ومعالجتها المرحلة الجامعية المبدئية نموذجا، هو يو شيانغ.
- محجوب، حسن محمد حسن، (2010م) دليل المعلم لتدريس مادة الصوتيات للمبتدئين من الناطقين بغير العربية، العربية للناطقين بغيرها، جامعة أفريقيا العالمية -معهد اللغة العربية، ع10
- ادروا، يوسف، (2014م) دور الصوتيات النطقية في تعليم اللغة العربية للناطقين بغيرها، أبحاث معرفية، ع 2، جامعة سيدي محمد بن عبد الله، كلية الآداب والعلوم الإنسانية مختبر العلوم المعرفية.
- محجوب، حسن محمد حسن، (2008م) مصطلح الثنائيات الصغرى أم الجناس في مجال تعليم الأصوات لغير العرب: دراسة ناقدة، العربية للناطقين بغيرها، ع6، جامعة أفريقيا العالمية، معهد اللغة العربية.

3- المشكلات الصوتية

ونقصد بذلك تلك الدراسات التي تناولت المشكلات الصوتية، ومنها دراسة: " جميل 2010م"

وتعالج هذه الدراسة موضوع تعليم الصوامت الحلقية والمحلقة لمتعلمي العربية الفصيحة باعتبارها لغة ثانية، والمقصود بالصوامت الحلقية في الدراسة هي كل من: العين والحاء، أما الصوامت المحلقة – وهي ما عرفت بالمفخمة – فهي كل من: الصاد والضاد والطاء والظاء. وركزت الدراسة جل عنايتها في محورين، أولهما: بيان الحقائق الصوتية التي جعلت هذه المجموعة من الأصوات تتسم بدرجة غير بسيطة من الصعوبة من جانب، وتعقب أسباب استبدالها بصوامت محددة في العربية فوناتيكيا وفونولوجيا من جانب آخر. أما ثانيهما، فيسلط الضوء على عدد من التقنيات التعليمية العملية التي تعين متعلم اللغة العربية على تجاوز مشكلة ضياع الحدود الفاصلة بين الصوامت على المستويين: العضوي-النطقي، والذهني-الإدراكي. وذكرت الدراسة بأنه يتوجب على المعلم عند البدء بتعليم هذه الأصوات أن يوليها خصوصية معينة، وألا يتعامل معها كما الصوامت الأخرى، وأن مساعدة المتعلم ليتجاوز هذه المشكلة النطقية والإدراكية منذ المراحل التعليمية الأولى يعينه على أن يرفع وعيه تجاه هذه الصوامت وأن يدرك أثر التبدل الصوتي الذي يحدثه أثناء النطق أو يميزه أثناء الإدراك في تشويش الرسائل اللغوية.

ومنها دراسة: " جميل 2016م" وتهدف هذه الدراسة إلى معرفة أخطاء نطق التّفخيم والترقيق لدى الطلاب النّاطقين بغير العربية، وذلك عن طريق جمع عينة من الأصوات المنطوقة من لدن عينات الدّراسة أثناء قراءتهم لأحد النصوص العربية، وقد حاولت الباحثة من خلالها الكشف عن الأخطاء الشائعة المتعلقة بالأصوات المفخمة، حيث إنّ الطلاب النّاطقين بغير العربية يواجهون مشكلة في نطق أصوات التفخيم والترقيق، بسبب عدم وجود مثيل لها في لغتهم الأم، وقد اقتصرت عينة الدراسة على الطلاب الناطقين بغير العربية، وتم تقسيمهم إلى ثلاث مجموعات من الدارسين للغة العربية في الجامعة الإسلامية العالمية بماليزيا، وبلغ عدد المشاركين في التقويم تسعة من الذكور والإناث، وخلصت الباحثة إلى أنَّ الطلاب غير الناطقين بالعربية يواجهون صعوبات في نطق ثلاثة من الأصوات المفخمة، وهي أصوات الصاد، والضاد، والظاء، في حين ينجحون في نطق بقية الأصوات المفخمة بشكل سليم.

ومنها كذلك دراسة (ماسيري ، دكوري و الأمين، سميه دفع الله أحمد 2012م) تهدف هذه الدراسة إلى التعرف على

المشكلات الصوتية التي تواجه الدارس الناطق بغير العربية عند تعلّمه للعربية، والتعرف على التحديات التي تواجه المعلم القائم بتدريس هذه اللغة لهؤلاء الطلاب وترى الدراسة أن المشكلات المتعلقة بالنظام الصوتي من أعمق وأكبر المشكلات في تعلّم اللغة العربية للناطقين بغيرها ؛ وعليه ظهرت الحاجة إلى تحليل هذا المستوى الصوتي لدى دارسي العربية الناطقين من غير أبنائها؛ من أجل الوقوف على أهمِ مشكلات النظام الصوتي، ثم اقتراح بعض الحلول والآليات المناسبة لها ، ومن هنا استخدمت هذه الدراسة المنهج المسحي الكمي، وذلك بالتطبيق على عينة من عشرين طالبًا من مجموع الدارسين بمركز اللغات قسم اللغة العربية، والبالغ عددهم ١٠٠ طالب وطالبة بجامعة المدينة العالمية بماليزيا التي تعنى بتدريس اللغة العربية وامتازت دراسة العينات بتنوع المستهدفين، وقد اعتمدت هذه الدراسة على الاستبانة كأداة لحصر العينات وتحليلها؛ فتوصلت إلى أن نسبة ٩٠ % من أفراد العينة يعانون من صعوبة نطق الأصوات الحلقية (العين، والحاء)، و ٨٠ % منهم ي يعانون من نطق الأصوات الحنجرية (الهاء والهمزة) ومن خلالها توصلت الدراسة إلى توصيات أهمّها: ضرورة بناء تدريس الأصوات اللغوية العربية للناطقين بغيرها على نظام التدرج من السهل إلى الأصعب؛ يبدأ بتعليمِ الأصوات الصامتة :الباء، التاء، الجيم، الثاء، الدال، الراء، الزاي، الذال، السين، الشين، الكاف، اللام، الميم، النون، مع وضعها في كلمات سهلة النطق ذات معان محسوسة، ومِنْ ثَم تعليم الأصوات المطبقة :الصاد، الضاد، الطاء، الظاء، مع وضعها في كلمات سهلة النطق ذات معان محسوسة، ثُم ينتقل بعد ذلك إلى تعليم الأصوات الحلقية :الهمزة، الهاء، العين، الغين، الحاء، الخاء، القاف، ووضعها في كلمات سهلة النطق، ومن ثمّ تقدم الأصوات الصائتة مع التركيز على توضيح بأن الفرق بين الحركات الطويلة والقصيرة هو مدّة النطق.

وهذا بيان بأبرز ما ألف من المصنفات في هذا المسار:

- المشكلات الصوتية في تعلّم اللغة العربية للناطقين بغيرها، جامعة المدينة العالمية أنموذجا، دكوري ماسيري وسميه دفع الله أحمد الأمين، جامعة المدينة العالمية، مجلة المجمع٢٠١٢م، وزارة التعليم العالي الماليزية، دولة ماليزيا.
- أخطاء نطق التّفخيم والترقيق عند النّاطقين بغير العربية، دراسة صوتية حاسوبية، نور مستورة بنت جميل، بحث متطلب مقدم لنيل درجة الماجستير في الآداب، قسم اللغة العربية وآدابها، كلية معارف الوحي والعلوم الإنسانية، الجامعة الإسلاميّة العالميّة ماليزيا، أغسطس 2016 م
- الأصوات الصعبة في نطقها وإدراكها لمتعلمي العربية من الناطقين بغيرها مفاهيم صوتية وتقنيات تعليمية لتدريس الأصوات الحلقية والمحلقة، ابتسام حسين جميل، بحث منشور بمجلة الجامعة الإسلامية للبحوث الإنسانية -شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية، غزة – فلسطين
- المشكلات الصوتية التي تواجه متعلمي العربية من الناطقين بغيرها وطرق علاجها: دراسة تحليلية، المسند، حمزة كريم محمد، رسالة ماجستير، كلية الدراسات العليا، الجامعة الأردنية، الأردن، 2015م.
- صعوبات تعلّم أصوات اللغة العربيّة لدى الماليزيّين في المستوى الثاني: طلبة معهد تعليم اللغة العربيّة لغير الناطقين بها بدمشق أنموذجاً، العليوى، هيام محمود، مجلة جامعة البعث للعلوم الإنسانية، جامعة البعث، سوريا، مج41، ع57، 2019م

- صعوبات النطق عند طلاب الصين في المستوى الأول في معهد تعليم اللغة العربية لغير الناطقين بها في الجامعة الإسلامية بالمدينة المنورة: دراسة مسحية، العنزي، عبد الله بن محمد بن سالم الطويلعى، رسالة ماجستير، الجامعة الإسلامية بالمدينة المنورة، معهد تعليم اللغة العربية لغير الناطقين بها، السعودية، 2015م
- تحليل الاخطاء السمعية في مستوى الاصوات اللغوية المفردة - الفونيمات - عند متعلمي اللغة العربية من الناطقين بغيرها، صفا، فيصل إبراهيم، ومحمد أبو عيد، اللسان العربي، ع 59، 2005م، المنظمة العربية للتربية والثقافة والعلوم
- تحليل الأخطاء النطقية في مستوى الفونيمات المفردة عند متعلمي اللغة العربية من الناطقين بالإنجليزية، أبو عيد، محمد أحمد، إربد للبحوث والدراسات، جامعة إربد الأهلية، م12ع1، 2008م
- الأصوات اللغوية في العربية والفارسية) دراسة تقابلية (عبدالرزاق رحماني، و عبدالله دربان، بحث منشور بمجلة العربية للناطقين بغيرها العدد العشرون يناير 2016 م.

4- الأصوات والحاسوب

ونقصد بذلك تلك الدراسات التي تناولت الأصوات العربية موظفة إمكانيات الحاسوب في تعلمها وتعليمها ، ومنها دراسة: (حبيب 2010م) وتتلخص فكرة البحث في كيفية استخدام إمكانيات الحاسوب في تعلم اللغة العربية وتعليمها، وأهم الطرق التعليمية، وأنواع البرامج الحاسوبية المتبعة في ذلك، ثم يتعرض لأسباب تأخر وضع برامج حاسوبية لتعلم اللغة العربية، والصعوبات التي تعترض ذلك، ويتناول البحث فكرة برنامج الصوتيات العربية الحاسوبي، ومنهجه، وطريقة بنائه، فيذكر أن برنامج الصوتيات العربية يعتمد على الطريقة التحليلية التي تبدأ من الجملة البسيطة، فالمفردة، فالمقطع الصوتي، فالصوت اللغوي، مستخدما أسلوب الاستثارة - الاستجابة، عن طريق استعمال الصور والأسئلة، ومن خلال تنمية مهارات الاستماع / الاستظهار، و يهدف هذا البرنامج إلى إرشاد المتعلمين من غير العرب إلى مخارج الأصوات الصامتة دون صفاتها، مستخدما التجربة الذاتية المصحوبة بصريا بالتمثيل التشريحي المتحرك، كما يحوي هذا البرنامج عددا كبيرا من الألفاظ ذات الدلالات الحسية، بهدف تعلم نطق الأصوات العربية داخل النسج الصوتية العربية، ثم تعلم معانيها عن طريق استخدام الصور، من حيث هي وسيلة تعليمية، بهدف الربط بين الأصوات والدلالة، ولذلك يعد البرنامج معجما لغويا بصريا صوتيا مبسطا، ليصل البحث إلى نتائج المقارنة بين منهجه وطريقة بنائه والطرق التعليمية المتبعة في تعلم اللغات الأجنبية وأنواع البرامج الحاسوبية التعليمية المعروفة.

ومنها دراسة: (مراد 2011م) وتقوم فكرتها على إنجاز نظام لتعليم حروف الأبجدية العربية، موجه خصيصا للناطقين بلغة أخرى غير لغة الضاد وكذلك التلاميذ الجدد. ويسمح هذا النظام بتعلم القواعد المبدئية للغة العربية وتعليم القراءة باستعمال الهجاء الآلي. ولقد تطلب إنجاز هذا النظام التعليمي استعمال نظام آخر لا يقل أهمية عنه وهو يشكل النواة الأساسية له. إنه نظام التعرف الآلي على الكلام باللغة العربية والذي أنجز من طرف الباحث.

وهذا بيان بأبرز ما ألف من المصنفات في هذا المسار:

- تعليم الصوتيات العربية وتعلمها بالحاسوب، ماهر عيسى حبيب، مجلة جامعة تشرين للبحوث والدراسات العلمية –

سلسلة الآداب والعلوم الإنسانية المجلد 23، العدد 3، 2010م.

- اللسانيات الحاسوبية وإشكالات المنهج والأنظمة في ميزان البحث: معالجة تحليلية لرؤى علمية عربية متميزة، بوفلاقة، محمد سيف الإسلام، مجلة الممارسات اللغوية، جامعة مولود معمري تيزي وزو -مخبر الممارسات اللغوية، مج11، ع2الجزائر 2020
- إنجاز برمجية لتعليم اللغة العربية بالاعتماد على نظام للتعرف الآلي على الكلام، عباس مراد، أعمال اليوم الدراسي: المحتوى الرقمي بالعربية في نظام الإدارة الإلكترونية، الجزائر، 2011م
- برامج النطق الآلي أو ما يعرف بمركبات الكلام وعلاقتها باللغة العربية، محمد، جمانة خالد، الأستاذ، جامعة بغداد -كلية التربية ابن رشد، 2012م.

3.2. Arabic CAPT

Several Arabic pronunciation systems that were developed with a limited vocabulary and a small number of speakers are available in the literature. No database is available for Arabic CAPT systems containing words or daily conversation, except those recorded using Quranic verses in [1] and [2]. In [3], a CAPT system was developed for the Arabic language, where the system detects mispronunciation and assesses the pronunciation quality of learners. The ASR-based system contains forced alignment, scoring, normalization of scoring, and quantitative measurement phases. Six speakers of both genders recorded the speech database of three isolated Arabic words. One of the speakers with good pronunciation was used as a reference model, and the remaining five speakers were considered for the evaluation of the system. The reference models for the native speaker were generated by a 19-phoneme HMM. Each phoneme model contained three states and eight Gaussians per state. Thirteen MFCCs, including log energy, were extracted from each frame, and the first- and second-order derivatives of the MFCCs were calculated. Four different measurements, GLL, LLL, ROS, and ROA, were used to evaluate the quality of the phoneme-level pronunciation. A higher measurement score indicated that the quality of pronunciation was good. The GLL score provided 86.66% accuracy in the detection of mispronunciation, which is better than the accuracy of other measures. To determine the value of GLL for a larger corpus, four additional speakers with good pronunciation were added to the database. For the correct acceptance rate, LLL metric results were better than the results of other systems.

Dahan et al. designed, implemented, and evaluated an Arabic speech pronunciation scoring system in [4] for the training of Malaysian teachers of Arabic. The ASR-based system provided feedback on the pronunciation of an L2 learner of Arabic and detected mispronunciation errors. The system extracted features from a signal and fed them to the pattern recognizer. The extracted feature was the MFCC, and HMM was used as the pattern recognizer.

A computer aided language learning (CALL) system capable of providing feedback to L2 students to improve their pronunciation and evaluate learning levels was developed in [5]. The system was based on a new robust speech recognition method that was proposed in the study and implemented using Sphinx3. The method used a three-state HMM with eight Gaussians per state and 13 MFCCs with their first and second derivatives. The method provided output in a form of

phonetic structure that distinguishes the proposed CALL system from other works, as presented in [6], which suppose that learner's speech is already labeled.

In [7], a system to detect mispronunciation in Quranic recitation was implemented by Abdo et al. The system was divided into five parts: (1) segmented features were extracted by a primary feature extractor from the input speech; (2) the boundaries of the targeted segments were determined by a segment analyzer; (3) references for the correct pronunciation and errors were fed to the system through an acoustical database; (4) after the detection of the segment, discriminative features were extracted by a secondary features extractor; and (5) the distance between the test segment and the database was evaluated by a verification module. The accuracy for segmentation detection was 73%, and verification of the samples yielded a 100% accuracy. MFCC performed well as a discriminative feature and provided 95% accuracy, among other features, such as a zero crossing rate, formants, energy, local maxima of spectra, log area ratio, and linear prediction coefficients [8]], [9], and [10].

Deep learning based method for pronunciation error detection for non-native Arabic speakers was proposed in [11]. The authors used a non-native Arabic database, which consisted of recording of 400 Pakistani speakers. Pronunciation error detection for non-native Arabic speakers at word level was proposed in [12]. The authors used recording of Pakistani speakers who were learning Arabic language. Deep CNN features and transfer learning parading were investigated.

3.3. Speech features for CAPT.

3.3.1. Speech features for CAPT.

Before deep learning era, hand-crafted features were playing an important role in designing the recognition systems. For example, in speech recognition systems, many features had been investigated in literature, such as MFCC, which is one of the most famous features, Linear Predictive Coding (LPC), LPCC, PLP, RASTA-PLP, etc. More information of these features and others can be found in [13][14]. With the huge improvement of deep learning technique and computation power, researchers proposed to feed the raw audio to deep learning networks in order to recognize the words/phonemes, such as in [15], [16]. On the other hand, a lot of researchers proposed to convert the audio signal to image representation such as 2 channels spectrogram / 3 channels spectrogram , and dealt with it using image based deep learning

techniques, such as in [17], [18]. As will be shown latter in section 5, we investigated using deep learning techniques with 3 channels spectral images.

3.3.2. Feature reduction for CAPT (e.g., LDA, PCA)

As a middle step between feature extraction and classification steps, feature reduction techniques were used to select the discriminative features in order to optimize the recognition accuracy. In [19], authors used linear discriminative analysis (LDA) to reduce the MFCC dimension for speech recognition system. Authors in [20] proposed using principle component analysis (PCA) as a dimension reduction technique, with MFCC, as a feature extraction technique, to improve the recognition of Indonesian speech system. Details of feature reduction techniques can be found in this survey [21]. As will be shown latter in section 5, we investigated using deep learning techniques and hence did not need to use feature reduction techniques because the network structure takes care of this.

3.4. Classification methods for CAPT

A study of the speaker-independent speech recognition of non-native speakers was conducted by Alotaibi in [22]. An Arabic database from the LDC (language data consortium) was used to observe the effect of a large vocabulary for MSA (modern speech recognition). The database contained speech from 75 native speakers and 35 non-native speakers. The purpose of the study was to determine the phoneme-level differences between native and non-native speakers as well as which type of phonemes contributed to recognition among native and non-native speakers. An HMM-based system provided good results when used by non-native speakers in the training and evaluation phases, and female non-native speakers produced better results than male non-native speakers.

An automatic dialect identification system was developed in Trigui et al. [23] using GMM [24]. Nine different dialect regions (Algeria, Iraq, Morocco, Syria, Gulf countries, Tunisia, Yemen, and Lebanon) were considered in this study. Dialect varies from region to region, with gradual transitions rather than clear boundaries between them.

Another study was conducted by Trigui et al. [25] to classify the Arabic accent of non-native speakers based on a statistical HMM method. The database was recorded by male and female Arabic learners from different countries who spoke different native languages. English, German, and French accents were considered in the study. The system was divided into four

components: data collection, language model construction, acoustic-phonetic decoding, and confusion. The recognition rates were 56% for French accents, 57% for English accents, and 69% for German accents. Speech recognition is an important part of CALL systems. Speaker-independent speech recognition systems can be affected by several factors, such as accent and gender. Many studies have attempted to normalize the regional accents of speakers [26].

An ASR-based CAPT system for L2 learning was implemented in [27] by using prosodic information and the hidden Markov model. For this system, the target language was Indonesian, and the native languages of the learners were Japanese, Peruvian, and Vietnamese. Six graduate students of both genders recorded eighteen target words with all Indonesian phonemes. The pronunciation errors made by the learners depended on their native and target languages. Seven types of pronunciation errors were identified in the study for the Indonesian language learners and were considered during the system development.

End-to-End (E2E) systems based deep learning techniques have obtained promising results in an automatic speech recognition systems, such as: Deep speech [28], Deep speech 2 [29], wav2letter [30], EESEN [31] and end-to-end attention model [32]. Likewise, E2E pronunciation error detection for CAPT systems have been proposed in the last few years and outperform the convention methods. Authors in [33] proposed E2E mispronunciation detection system based on CTC-Attention model. Another E2E system based on CTC-Attention model, for mandarin annotated spoken corpus, were proposed in [34]. Authors in [35] proposed E2E mispronunciation detection and diagnosis system for non-native English speakers. They used TIMIT corpus for native speech and L2-ARCTIC corpus for non-native speech. In our work we also using deep neural to make E2E system but in a new novel technique where we treat the phonemes or the AFs as objects in 3 channels spectral images.

3.5. Scoring in CAPT

ASR-based pronunciation training systems using different speech features and scoring measurement techniques have been developed for pronunciation training in other languages, such as Dutch, Mandarin, and Indonesian. A system to detect errors among Dutch language learners was developed by Doremalen et al. in [36]. Eleven different short and long vowels that are commonly mispronounced by Dutch learners were highlighted in this study. The experiments were conducted using the Spoken Dutch Corpus (CGN). This database contains nine million

words and many speakers belonging to different age, sex, and regional groups [37].

An ASR-based system was implemented in [38] by extracting various phonetic features: the mean pitch and intensity of the segment, and three formants with F2 – F1 calculated at three different locations of the sample. To reduce the variability of the measured parameters between speakers, normalization was performed using Lobanov's Z-score. After normalization, the normalized features are called spectral. In addition to the eight duration features, a zero-coefficient plus twelve MFCCs with their first- and second-order temporal derivatives were measured. The SVM with linear kernel was used for the classification, and the results were provided using different performance parameters, such as EER (equal error rate) and a 95% confidence interval. To obtain the baseline results, segment- and state-based confidence measures (CM) were calculated, and HMM models were trained by SPRAAK for phone alignment. The best EER of 15% was obtained with MFCC, and a result of 12.3% was achieved when MFCC was used with CM and duration features. The MFCC approach yielded better performance than the phonetic features, CM, and the duration features.

In the study [39] by Troung, acoustic-phonetic-based classifiers were developed by implementing linear discriminant analysis [40] and decision trees [41] to discriminate between the correct sounds of native speakers and the incorrect sounds of non-native speakers. The classifiers were designed to detect mispronunciation errors of three phonemes, /a/, /Y/, and /X/, which are frequently made by L2 learners of Dutch. A CAPT system using a confidence measure score, as presented in [42] , to predict pronunciation errors did not have a high correlation between the human score and the machine-calculated score, perhaps because of the use of the same type of features for all sounds without considering acoustic characteristics. Therefore, each classifier was developed for specific errors by analyzing the acoustic differences of each sound in [39]. Moreover, gender-dependent models were used in the classifiers to optimize the performance of the developed pronunciation error detection system. Two experiments were conducted with 20 native Dutch speakers and 60 non-native speakers for both classifiers. In the first type of experiment, the classifiers were tested and trained in two ways: training and testing by native speakers and training and testing by non-native speakers. In the second type of experiment, the classifiers were trained by native speakers and tested by non-native speakers. The accuracy of the experiments for the decision tree and LDA ranged from 75% to 91.7% and from 87% to 95%, respectively.

An automatic pronunciation error detection system for L2 learners of Mandarin was proposed by Xu et al. in [43]. Prior linguistic information was used to improve the performance of the developed CALL system. Methods based on LPP (log posterior probability) and RLPP (revised log-posterior probability) were implemented. The second method used linguistic knowledge, whereas the first did not. The results show that the performance of the method using linguistic knowledge was better than the other method. In another approach, RLPP was used to construct the restricted pronunciation space (RPS) for each phone to observe its pronunciation variation. A database containing 1,585 words pronounced by non-native Mandarin speakers was used to evaluate the system performance, and training was performed by native Mandarin speakers. The accuracy for the detection of pronunciation errors showed that the RPS-based system produced the best performance. Another CAPT system for the Mandarin language was proposed by Liang et al. [44]. The system was divided into two tasks: sentence verification and syllable identification. For sentence verification, acoustic models for the tri-phone, garbage model, pronunciation manner cluster (PMC), and anti-PMC were developed by using HMM. Forty-eight MFCCs with four energy coefficients were extracted from the Formosa Speech Database (ForSDAT) [45] to train the models. Five sentences containing all Mandarin phonemes were recorded for the testing of the system. Syllables were divided into two categories (out-of-task and confusion), and 160 sentences were developed by combining them. The sentence verification task obtained the highest F-measure [46] with the tri-phone vs. anti-PCM acoustic model, at 91%. The output of this model was fed to the syllable identification task. To extend the pronunciation lexicon, pronunciation variation rules were used. The best F-measure for syllable identification was 77.2%.

3.6. Databases for CAPT

Several databases were used in the literature for CAPT system. Some of them are publicly available and some of them are private. Table 1 shows a summary of some CAPT corpora for Arabic and non-Arabic languages.

Table 1: CAPT Databases' Survey.

Database Name	Year	Language	No. of recorded speakers	Total duration (hours)	Recording environment
L2-ARCTIC [47]	2017	English	24	11.2	quiet room
CU-CHLOE Corpus [48]	2015	English from native speakers of Chinese that learn English	211		sound-dampened room
iCALL Corpus [49]	2016	Mandarin	305	142	quiet office rooms,
CSLU [50]	2005 and 2007 update	English	90	30	digital telephone lines
West Point Arabic Speech [51]	2002	Arabic	110	11.42	
The CrossTowns Corpus [52]	2006	German, English, French, Italian, Netherlands	161	16	noise-controlled cabin and small room
Speech accent archive [53]	2016	English	646	N/A	Online
IDEA. The International Dialects of English Archive [54]	1998 -2020	English	N/A	170	Online

4. Development of the databases of Non-Arabs speech

Building an Arabic CAPT system requires an Arabic Speech database that contains diverse speech and pronunciation errors. At the beginning of the project, no database for Arabic L2 speakers with emphasis on pronunciation errors, and enough speakers with detailed phoneme annotations, was available. Hence, we opted as per the proposal objectives to record a new speech dataset, having in mind the quality of the text selection and the coverage of most errors that L2 Arabic speakers might make. ALI-T proposed the text for recording and had many meetings and discussions with CCIS-T, until a set of 25 long length sentences and some 61 very special pairs of words were finally selected for session 1. For the text of session 2 a new methodology for text selection was proposed by the ALI-T that utilized new ideas and our experience in session 1. Session 2 contains about of 582 words. Details of text selection, database specifications, the recording system, registration of the speakers, speech recording are presented in sections 4.2 and 4.3. Details of speech annotation and error analysis will also be presented in sections 4.2 and 4.3. We called these databases KSU-CAPT Non-Arabs Database-Session 1 and KSU-CAPT Non-Arabs Database-Session 2.

We anticipated that the process of building a new speech database for CAPT applications, which is one of the main objectives of this project takes a long time, hence we started our investigation in the project by using part of the KSU speech database. KSU speech database was not developed for CAPT applications and contains the speech of 60 non-native speakers from different nationalities, hence we had to adapt the selected part of the KSU speech database to fulfill the requirement of a database for CAPT applications, and we named it Arabic-CAPT. To have enough speech to train the CAPT models, we used Arabic-CAPT to generate a new database of synthesized speech, and we named it Arabic-CAPT-S (‘S’ stands for synthesized). Details of building the Arabic-CAPT and Arabic-CAPT-S databases are presented in this section.

4.1. Arabic-CAPT and Arabic-CAPT-S

We selected the KSU Speech database to build the Arabic-CAPT for the following reasons:

- The database was developed for Arabic speaker recognition as a main target, but it was also designed to be useful for other applications, mainly speech recognition and speech processing of non-native Arabic speech.
- The corpus was recorded over three different sessions in three different environments and using four different channels.
- The corpus consisted of recording male and female speakers, where the male speakers included Saudi, Arab, and non-Arab speakers.
- The corpus has speakers from 20 nationalities from Asia, Africa and Europe.

4.1.1. Text selection and audio transcript

Each speaker in KSU speech database uttered 16 lists, most of them are common among all speakers and some are distinct. To develop the Arabic-CAPT database, we used only the common lists whose text is presented in Table 2. The KSU speech database provides audio recordings and the associated text (canonical). Native Arabic experts transcribed the audio as pronounced by speakers. Table 3 shows an example of the transcription of the speech of two sentences containing errors in pronunciation.

TABLE 2: THE CANONICAL TEXT OF THE SELECTED LISTS.

List Name	Canonical Text
Numbers	صِفْرْ وَاحِدْ إِثنانْ ثَلاثَهْ أَرْبِعَهْ خَمْسَهْ سِتَّهْ سَبعهْ ثَمَانِيَهْ تِسْعَهْ
Fixed_Sentences	لِمَاذَا سَافَرْتَ إِلَى الخَارِجِ فِي العِيدْ ، جَاءَ الضُّيُوفُ الثَّلاثَةُ بِالذَّهَبِ قَبْلَ الظُّهر
Common_words_**1**	السَّلامُ عَلَيكُمْ لَا نَعَمْ لَا أَدْرِي شُكْراً، اللهُ أَعْلَمْ قَرِيبًا أَخْبَارْ زِيَارَهْ لَحْظَةٌ مِنْ فَضْلِكْ
Common_words_2	كَثِيراً صَحِيحْ غَائِبْ هذا مَنْ مَعِي، تَفَضَّلْ أَهْلاً وَسَهْلا إِنْ شَاءَ اللهْ لَوْ سَمَحْتْ آسِفْ
Paragraph_1	مُحَمَّدٌ رَّسُولُ اللَّهِ وَالَّذِينَ مَعَهُ أَشِدَّاءُ عَلَى الْكُفَّارِ رُحَمَاءُ بَيْنَهُمْ، تَرَاهُمْ رُكَّعًا سُجَّدًا يَبْتَغُونَ فَضْلًا مِّنَ اللَّهِ وَرِضْوَانًا ، سِيمَاهُمْ فِي وُجُوهِهِم مِّنْ أَثَرِ السُّجُودِ، ذَٰلِكَ مَثَلُهُمْ فِي التَّوْرَاةِ وَمَثَلُهُمْ فِي الْإِنجِيلِ كَزَرْعٍ أَخْرَجَ شَطْأَهُ فَآزَرَهُ فَاسْتَغْلَظَ فَاسْتَوَىٰ عَلَىٰ سُوقِهِ، يُعْجِبُ الزُّرَّاعَ لِيَغِيظَ بِهِمُ الْكُفَّارَ ، وَعَدَ اللَّهُ الَّذِينَ آمَنُوا وَعَمِلُوا الصَّالِحَاتِ مِنْهُم مَّغْفِرَةً وَأَجْرًا عَظِيمًا
Phonetic Distinctive Words	فَحْصْ فَحْمْ فَسْحْ قَصْمْ مَرْحْ مَغْصْ نِصْفْ نَهْشْ نَفْعْ نَفْسْ ، حِنْثْ سَهْمْ شَمْعْ صَمْغْ عَزْفْ عَفْشْ عَنْزْ عَنْهُ غُصْنْ غُنْمْ
Common_List_1	بِالوَالِدَينِ إِحْسَانًا، اِستَقِمْ كَمَا أُمِرْتَ ، كَانَ الأَكْلُ لَذِيذًا ،هَلْ هَارَ، أَسَرُونَا بِمُنعَطَفٍ ، جَمَعَ المَوزَ وَ خَلَى ، ضَمِنْتُ شَغَفَكُمْ ، وَرِمَا فَلَنْ يُقَاتِلَا ، هِيَ هُنَا لَقَدْ

	آبَتْ ، أَبْصَرَ ثُعبَاناً وَلَم يَظلِمْهُ

TABLE 3: EXAMPLES OF AUDIO TRANSCRIPT FROM THE LIST "COMMON_LIST 1" OF NON-NATIVE SPEAKER "NS239".

Canonical text	اِستَقِمْ كَمَا أُمِرْتَ
Real transcript	اِستَكِم كَمَا أُمِرتَ
Canonical text	وَرِمَا فَلَنْ يُقَاتِلَا
Real transcript	وريمَاً فلا يُكَاتِل قريْمَاً فَلَا يُكَاتِل

4.1.2. Sentence Segmentation

In this phase, we want to segment the long recorded audio files into small files for easier processing. To accomplish this, we segmented the selected lists into short sentences automatically, by the Aeneas segmentation tool [55]. The input of the Aeneas tool is an audio file and the corresponding text segmented into sentences manually, where each sentence is written in a separate line. The output of the Aeneas tool is a textgrid file for each list consisting of the start and end times of each sentence of this list, as shown in **Error! Reference source not found.**. Next, an Arabic expert manually corrected the output of Aeneas using Praat software, as shown in **Error! Reference source not found.**. We used the final time boundaries to generate the audio file of each sentence and the corresponding transcript.

The previous steps were performed for the selected utterances of all 266 male speakers (native and non-native) in session 1 from the KSU speech database. For speech annotation, we focused only on the speech of non-native Arabic speakers, which included 60 speakers from 20 different nationalities. **Error! Reference source not found.** shows the nationalities and the number of speakers from each nationality. We anticipate from the diversity of the speakers' nationalities that the produced corpus will significantly contribute to the research community in the Arabic-CAPT systems.

4.1.3. Word and phoneme level Segmentation

This is the last step before the annotation process, where we segment each utterance into words and phonemes. For this task, we used the Montreal Forced Aligner (MFA) version 1.0.0 [56].

4.1.4. Annotation

First, three native Arabic experts annotated the utterances. Then, we used an edit distance algorithm to align the canonical words and the annotator's words, and from the output of

alignment we identified the types of error of each phoneme. Finally, to combine the results of the three annotators we took majority voting. If all annotators differed in their judgment, we took the annotation of the most experienced annotator. The total percentage of the mismatch between all annotators was 5.82% from the total number of mispronounced phonemes, which was 3006 phonemes. The mismatch was 104, 15, and 56 for substitution, deletion, and insertion errors, respectively. We noticed that the highest mismatch between the annotators was in substitution errors, which can be attributed to the fact that usually the original phoneme and the substituted phoneme were very near and hard to distinguish. We followed the annotation scheme of L2-ARCTIC [47] and annotated the substitution, insertion, and deletion errors in the following forms: (canonical phoneme, substituted phoneme, S), (@, inserted phoneme, I), and (deleted phoneme, @, D), respectively.

Due to the lack of non-native Arabic corpora, in general, and the low number of pronunciation errors in our Arabic-CAPT corpus, we generated synthesized non-native Arabic speech. We trained the recent neural text-to-speech (TTS), FastSpeech2 [57], to produce high-quality non-native Arabic speech that contains some predefined substitution errors. We selected the FastSpeech2 model because it is state-of-the-art, fast, and supports multi-speaker embedding, so we could generate a synthesized speech using the style of all non-native speakers of our Arabic-CAPT corpus. Because the size of the Arabic-CAPT corpus is not efficient in training FastSpeech2 from scratch, we pre-trained FastSpeech2 using an Arabic native corpus. We selected 5 h of recording of Saudi speakers from the KSU speech database to train FastSpeech2. Then, we fine-tuned the model using the Arabic-CAPT corpus. Finally, we modified the canonical text of the Arabic-CAPT corpus by embedding it with the most common pronunciation errors of non-native Arabic speakers to produce text with substitution errors. Note that we focused on embedding substitution errors in the synthesized corpus because they are common errors in learning Arabic. Then, the generated text was fed to the trained TTS to generate high-quality synthesized speech. Finally, we used the generated transcript and synthesized audio as input to MFA to segment the synthesized Arabic-CAPT-S corpus at the word and phoneme levels. **Error! Reference source not found.** shows the three steps of developing the Arabic-CAPT-S corpus.

4.1.5. Qualitative evaluation of the synthesized spectrograms

We present in this section examples of spectrograms of the synthesized speech and spectrograms of the original corresponding speech. **Error! Reference source not found.** shows the spectrograms of the real and synthesized speech for the Arabic text "هِيَ هُنَا لَقَدْ آبَتْ" using the style of speaker NS251, from Senegal, and the Arabic text "كَانَ الأَكْلُ لَذِيذًا" using the style of speaker NS218, from Nepal. By qualitative evaluation of the real and synthesized spectrograms in **Error! Reference source not found.**, we notice the high similarity between them, which indicates the high quality of the synthesized speech in the developed Arabic-CAPT-S.

4.1.6. Analysis of Pronunciation errors in Arabic-CAPT and Arabic-CAPT-S

Table 4 shows statistics of Arabic-CAPT and Arabic-CAPT-S. It shows the total number of phonemes, not including silence, and the total number of errors for the Arabic-CAPT and Arabic-CAPT-S. For the Arabic-CAPT, we show the number of each of the three types of error, while Arabic-CAPT-S has only substitution errors. We noticed that the number of pronunciation errors in the Arabic-CAPT corpus was 2899, which represented 5.1% of the total number of phonemes in the corpus, where substitution and insertion errors were more frequent than deletion errors. The number of substitution errors in the synthesized Arabic-CAPT-S corpus was 17,422, which represents 6.4% of the total number of phonemes, near the percentage of errors in the Arabic-CAPT. In terms of duration in hours, we can see from Table 4 that the Arabic-CAPT-S consisted of 7.11 recording hours and the Arabic-CAPT consisted of only 2.36 h. We expect that training the proposed system using real and synthesized corpora will make it more generalized and able to detect the most common pronunciation errors of non-native Arabic learners, especially for the most important type of pronunciation error, which is the substitution error.

Table 4: Statistics of Arabic-CAPT and Arabic-CAPT-S.

	Arabic-CAPT	**Arabic-CAPT-S**
Type of data	Real	Synthetic
Speakers	62	62
Utterances	1611	7254
Recording hours	2.36	7.11

Correct phonemes		54,171	255,502
Substitution Errors	#	1080	17,422
	%	2	6.4
Insertions Errors	#	1139	-
	%	2.1	-
Deletion Errors	#	690	-
	%	1.3	-
Total Errors	#	2899	17,422
	%	5.1	6.4

4.2. KSU-CAPT Non-Arabs Database-Session 1

4.2.1. Selection of text for recording of the speech corpus

A main issue in CAPT systems is to select the optimal words and sentences that can cover the majority of errors in learning the pronunciation of L2. The selected texts must contain very specific phonemes that are difficult to pronounce by the speakers, and useful in improving the pronunciation. The text should have the following characteristics:

- Varied text containing rich diversity of phones and di-phones.
- Optimal number of sentences/words that can be pronounced by L2 Arabic learners in a minimal amount of time.

ALI-T team is well qualified for this task, as they have years of expertise in teaching the Arabic language for Non-Arabs, and conducted many research studies to enforce this expertise. Based on this expertise they were able to advice for a methodology to construct the text most suitable for the project, which took a considerable time and efforts. The proposed methodology is as below.

Methodology for the CAPT text selection

The selection of the CAPT sentences were subject to many conditions as follows:

a) Sounds

- Many repetitions of the same sound or phoneme are preferable.
- Appearance of the sound at the start, middle and end of the word.

b) Words

- Common: Common words are preferred over specific words.
- Diversity: words used in diverse Arabic countries are preferred.
- Affinity: Usual and daily words are preferred.
- Inclusion: Words used in many domains are preferred over words used in specific domains.
- Importance: words needed by the learner are preferred.
- Purity: Original Arabic words are preferred over Arabized Arabic words.

c) Sentences of the text

- Sentences must be meaningful.
- Must have Arabic cultural aspect
- Must be valuable.
- Short sentences are preferred, to avoid boringness.
- Must be consistent and clear.
- Must be in accordance to and respect the Kingdom's beliefs and constants.

The selection of the CAPT text passed by two main steps. The first step, contributed by ALI-T experts, consisted of applying the above methodology and suggesting sentences and specific words that contain phonemes that learners of Arabic have problems in pronouncing correctly, in addition to simple phonemes that can be found in other languages, such as `m`, `n`, …etc. The second step, conducted by the project team, was refining the text to the mobile app

and testing it. This second step passed by 4 stages, the first three stages were completed before starting the audio recording.

All the selected CAPT-texts were tested and adjusted over five main criteria:

- Reasonable time to read all the content.
- Complexity of the content, phoneme positions and length of words.
- Richness of the phonemic content in every sentence.
- Dual phones words must contain minimal pair diversity in phoneme pronunciations.
- It is well known that the more the sentences become long, the more the speaker starts damping his voice and is prone to more reading latency and stuttering, not in accordance to what the CAPT system aims to correct.

Both CCIS and ALI teams checked all the sentences and agreed to select an optimal number of 16 sentences with a minimum length of 21 words and a maximum length of 42 words. In addition, they selected a set of 61 minimal dual phones pairs, to target the phoneme dualities that can lead to pronunciation errors, these dual words differ by some phonemes but have a similar structure. Arabic experts from the ALI-T team stressed on the fact that these short and long dual phonetic words are very important in assessing and evaluating Non-Arabs pronunciations.

Samples of two meaningful sentences and eight minimal dual phones pairs are presented in Table 5 below.

Table 5, Sample sentences and dual phonetic words

Sentence 1	أَفْضَلُ النَّاْسِ عِنْدَ اللهِ هُمْ أَصْحَابُ الأَعْمَالِ الفَاضِلَةِ، وَالضَمَائِرِ المُضِيْئَةِ، الَّذِينَ يَرْكُضُونَ إِلَى الخَيْرِ رَكْضاً، يُعِينُونَ الضَّعِيْفَ وَالْمَرِيضَ، وَيُنَاهِضُونَ الضَّلَالَ وَالإِضْرارَ أَمَلًا فِي مَرْضَاةِ اللَّهِ.
Sentence 2	أَفْطَرْتُ بِالأَمْسِ عِنْدَ طَارِقٍ عَلَى طَعَامٍ طَيِّبٍ، وَعِنْدَمَا خَرَجْتُ رَأَيْتُ طُيُورًا فَوْقَ مَبْنَى المَطْبَعَةِ تَطُوفُ بِهَا، ثُمَّ تَحطُّ عَلَيْها لِتَلْتَقِطَ الْحَبَّ، وبَعْضَ الأَطْعِمَةِ، فَطَابَتْ نَفْسِي بِرُؤْيَتِهَا.

Pair of minimal dual phone words 1	خَـيْر / غَـيْر ** غَيْر / خَيْل ** خَائِبٌ/ غَائِبٌ
Pair of minimal dual phone words 2	كُلْ/ قُلْ ** رَقَدَ/ رَكَدَ ** قَالَ/ كَالَ ** كَتَمَ/ خَتَمَ ** مَكْتُومٌ/ مَخْتُوم

The intended recording setup in the proposal was designed to record the students in the Arabic Language Institute at King Saud University, in a controlled live session, face to face. If during the recording sessions, the speaker feels tired or bored, a short pause can be made, and the recording can continue after the pause session. Unfortunately, the COVID-19 restrictions imposed that students cannot come to the university, and we had to move to online recording, through a newly developed mobile app, that will be described later in the section. This online recording had some benefits and some drawbacks, as illustrated in Table 6.

Table 6, Benefits and drawbacks of the online recording solution

Benefits	**Drawbacks**
Any screen can be easily recorded again in a new time if any error is detected without the need to for the student to come back to the recording room (once the admin allows re-recording)	Speaker recordings had to be checked after each speaker completion. This induced extra costs, for the listening and checking stage.
Number of students at ALI was much lower than the number at time of submission of the proposal. Online recording allowed us to record students from inside/outside of Riyadh.	Recording solution must be compatible with diverse screens and phones, which increased the time for the testing and tuning of the App.
Diversity of the students, as students are not from one location or university.	Decrease of the total number of recorded phonemes. (see Figure 1)
Long sentences have been shortened, so the speakers could complete the recording in shorter time.	Additional explanations and discussions were necessary to make the students understand the installation and the use of the recording solution.

Due to the mobile recording constraints, the CAPT selected sentences have been additionally shortened, in order to fit into the screens of the students. Selecting the text in the app pages went into four versions. Statistics of the four versions of the texts are shown in *Table 7*. A

detailed comparison, of the first three versions of the CAPT-Texts for each selection phase, is illustrated in Appendix A. Appendix B lists the text for version 4 (Mobile version) after further simplification and shortening so the read texts can be audio-recorded easily in the Mobile App.

Table 7, Statistics of the CAPT-Text selections for session 1

Statistics	V1	V2	V3	V4 (mobile)
Number of sentences	29	28	16	25
Total number of words (sentences)	1100	767	474	463
Maximum number of words / per sentence	72	41	16	27
Minimum number of words / per sentence	20	20	21	11
Pairs of Minimal Dual words.	113	113	113	61
Number of Screens: sentences	29	28	16	25
Number of screens: words	28	28	28	16

The numbers of phonemes, within the sentences and words of all the text selections versions, are presented in Figure 1. The phoneme distribution remained almost the same although we reduced the total number of phonemes to half from V1 to V3. V4 Mobile version is a reduced version of the V3, to fit in the screens of the Mobile, where sentences were shortened, and 61 important dual pairs were kept for recording.

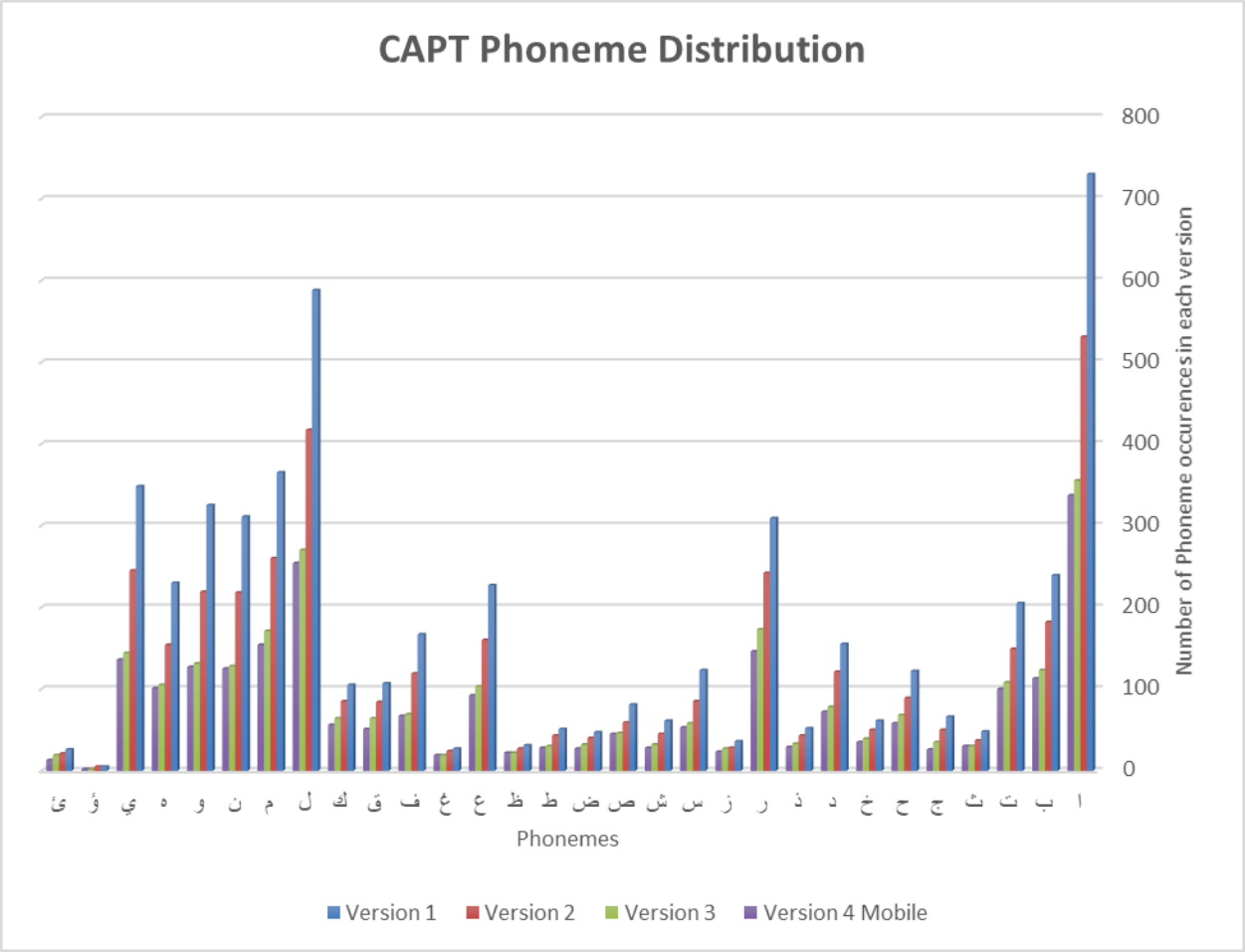


Figure 1, Phoneme distribution in the sentences and dual pairs of the various text selections

A comparative table of the phonemes of each version is detailed in Table 8.

Table 8, Comparative between the text selections of the various versions

Phoneme	Version 1	Version 2	Version 3	Version 4 Mobile
ا	730	531	355	337
ب	239	182	123	113
ت	205	149	108	100
ث	48	37	30	30
ج	66	50	35	26
ح	122	89	68	58
خ	61	50	39	35
د	155	121	78	72
ذ	52	43	33	29
ر	309	242	173	146
ز	36	28	27	23
س	123	85	58	53

ش	61	45	32	28
ص	81	59	46	45
ض	47	40	32	27
ط	51	43	30	28
ظ	31	27	22	22
ع	227	160	103	92
غ	27	24	19	19
ف	167	119	69	67
ق	107	84	64	51
ك	105	85	64	56
ل	588	417	270	254
م	365	260	171	154
ن	311	218	128	125
و	325	219	131	127
ه	230	154	105	101
ي	348	245	144	136
ؤ	5	5	2	2
ئ	26	21	19	13
Total Phonemes	**5248**	**3832**	**2578**	**2369**

4.2.2. Database Specifications

The recording step started by recruiting some sample speakers from the ALI institute, from the fourth level, in order to test the recording time and the quality of the reading. From the initial speakers' samples when recording version 3 of the text, we noticed that the duration of the recording varied between 40 and 45 minutes. The time was still long and we had to reduce the 16 long sentences (in version 3) to 25 short sentences (in version 4) with a maximum of 24 words and a minimum of 11 words per sentence. This was also a good consideration to fulfill the display constraint in reducing the displayed text on the phone screen, as mobile screens do not allow very crowded text and buttons in a convivial application. Sample screens from the Tahadath Mobile App are shown in section 4.2.3.

Once the mobile app was developed and sent to diverse students at different geographical locations, we noticed that Non-Arabs had huge problems in reading texts without diacritics; we then updated the texts with a full diacritization. Screen shots of all the texts in the mobile app are presented in Appendix C.

A sampling rate of 8 KHz was decided upon two considerations:

1. Speech recorded at high sampling rates is in general reduced to 16 kHz or 8 kHz, for easiness of manipulation, and simplicity of use in training large models.

2. Recording from the microphone of the mobile phone allows only 8 kHz. Sample metadata of the recordings are shown in Table 9.

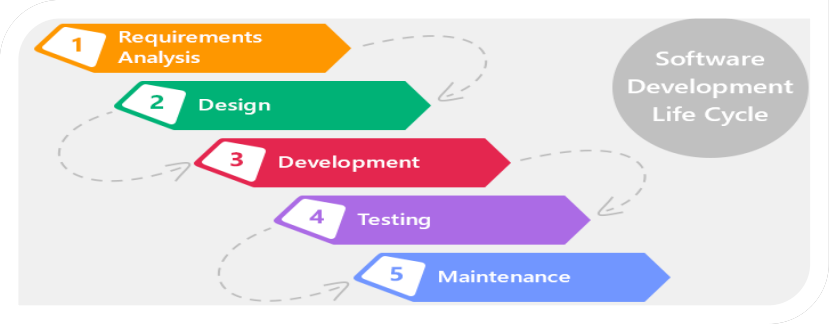
Table 9, Sample Metadata of the CAPT recordings

Input #0, **'mp3'**: □ filetype
com.android.version: 10
Duration: 00:00:06.74, start: 0.000000, bitrate: 20 kb/s
Stream #0:0(eng): Audio: amr_nb (samr / 0x726D6173), **8000 Hz**, mono, flt, 12 kb/s (default)

4.2.3. Establishment of the recording system

As already mentioned, we opted to develop a mobile app instead of a computerized application. We had two options, either use unity to develop the app or android using Java. The development of the recording app was subject to the known software lifecycle, presented in Figure 2.

Figure 2, Lifecycle of the Tahadath Mobile app



We have already developed similar applications while recording speech datasets for a previous KACST funded project [58]. The requirements step was deeply discussed between the members of the team, and the first orientation was the use of a computerized application, but due to the medical restrictions against gathering of students in the institutes, due to COVID-19, we opted to use a mobile app.

The first tentative app was developed by a specialized developer in unity, and has been tested for diverse criteria of screen sizes, colors, etc. We noticed, on the long run, that unity did not support Arabic writing in a very smooth manner, and the app developer had to load and deal with images in the application instead of writing Arabic texts directly in the app. A screenshot of the initial application is presented in Figure 3.

Figure 3, Unity APP screenshot

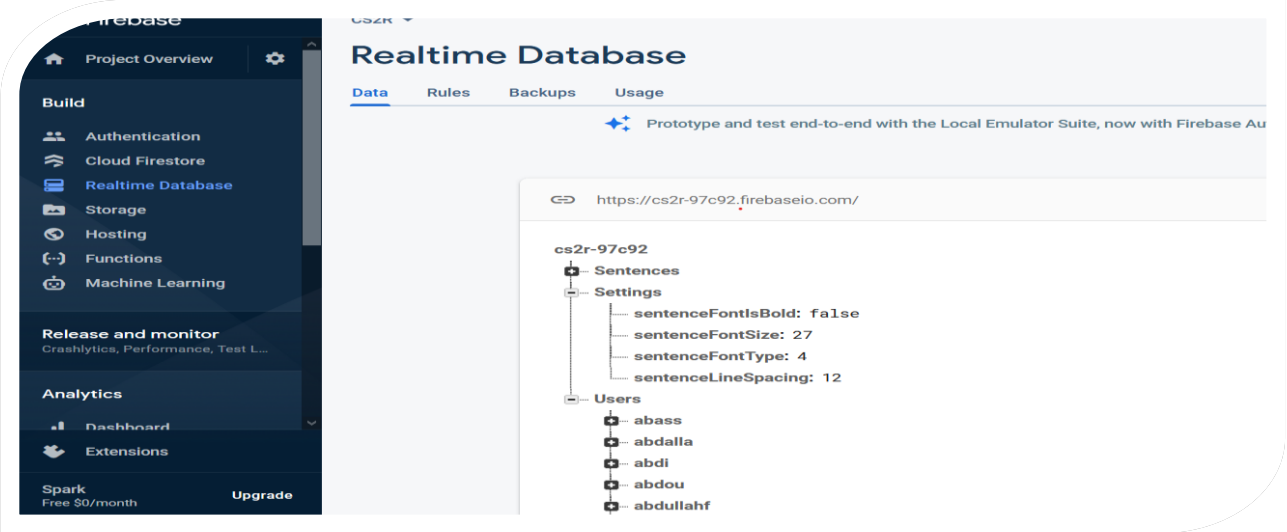
Unfortunately, with the numerous changes of the texts and fonts, we had to move to the development of another application in Java Android. The Java application felt more convivial to Arabic texts and font variations. The Java developer made diverse versions, as per the team requests (design-develop-test). The latest version is the 1.0.10 (10th version), in addition to a second similar application that was also developed for Arabs, as we wanted to split at the database level, Arabs from Non-Arabs recording, for a better management and checking. Some sample screens from the developed Android App "Tahadath" are presented in Table 10, a complete listing is also detailed in Appendix C.

Table 10, Sample screens from the Tahadath Mobile App

The backbone of the Android application is the Google firebase, and the application was subject to very strict access, as speakers are invited by their WhatsApp number and the access to

the application is subject to a fixed name and password, generated by our database manager. Access can be done only when a recording flag is enabled. Once the speaker reads all the lists and approves its recordings, the recording flag is disabled, and no more access to the database is allowed to that speaker, unless it is reactivated by the admin for checking purposes or recording repetitions. A screenshot of the Google Firebase management platform is shown in Figure 4.

Figure 4, cs2r firebase real-time database



We can notice from Figure 4, that the control of the display font and size can be easily controlled from a centralized part of the app, in addition to the possibilities to change the display sizes at the mobile level.

Additional tests have been also made by the database team, in order to test the app in different mobiles and different versions of android. Some problems appeared in fonts and positions and were fixed as per the maintenance step of the software lifetime cycle.

4.2.4. Additional Improvements to the app

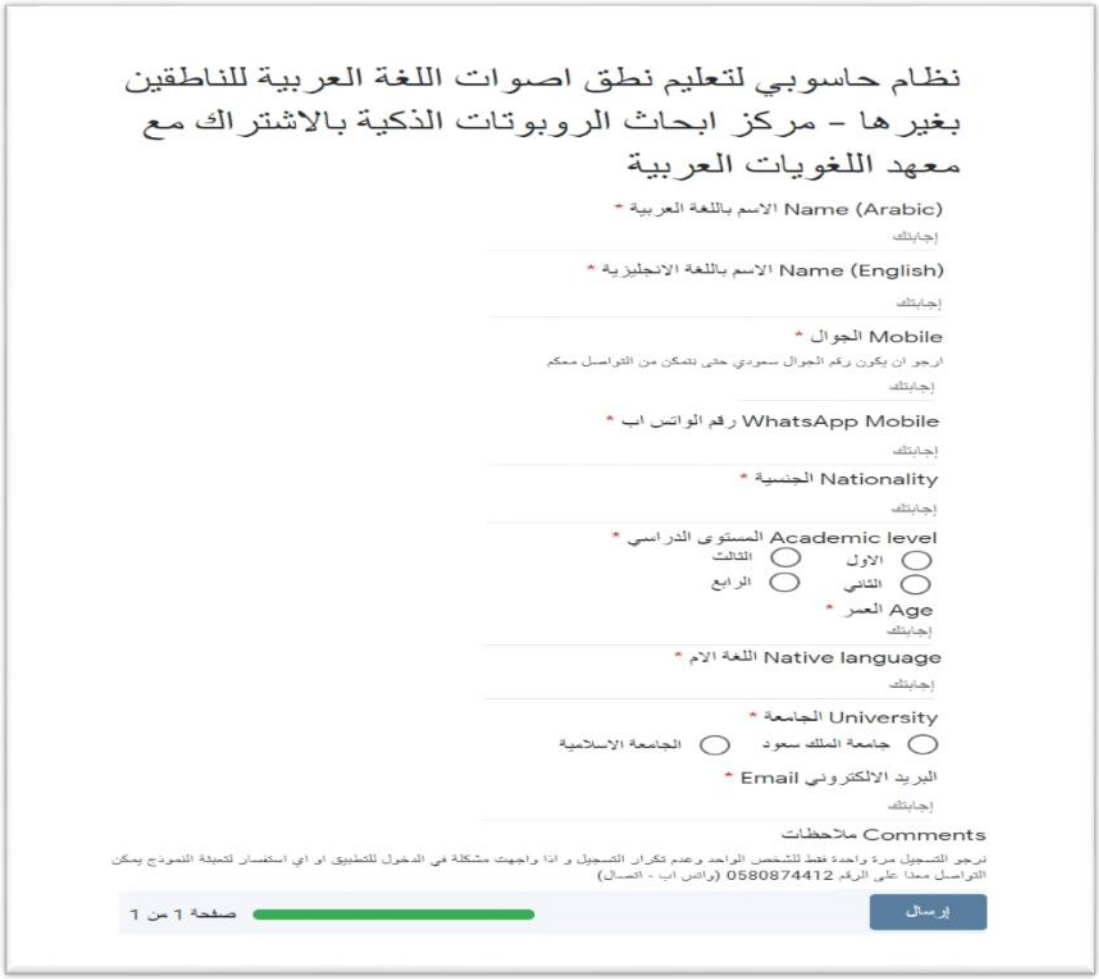
The app that has been developed is a one-way communication, i.e., the speaker records then the recordings are checked. This leads to many problems in terms of quality and recording durations, see Table 6, for benefits and drawbacks of the distant recording. When a speaker records his voice, we had to wait until he finishes his recordings to start checking because different students may be recording at the same time. Hence it was not possible to control every speaker in real time, because each speaker can record at his pace when he feels himself ready.

4.2.5. Speaker Registration

In the project proposal, we intended to record 200 Non-Arabs in the whole project. In order to manage such huge number of students, we followed a sample work methodology, where we start by a small number then increase to the target quota.

To ease the enrollment of the students who were mostly at distant locations, a google form, as shown in *Figure 5*, has been established and sent to the volunteers directly or to a coordinator from each institute who will send to students that he recruits at his institute. Each student needed to fill all the required fields and send it back. Once the forms are collected, the students are contacted by our team for further explanation of the recording steps or to answer any question.

Figure 5, Screen shot of the Google form sent to the students for the CAPT enrollment



In the following part, we will present some statistics of the enrolled students. These statistics include number of students, country of origin, language spoken, level of education, etc.

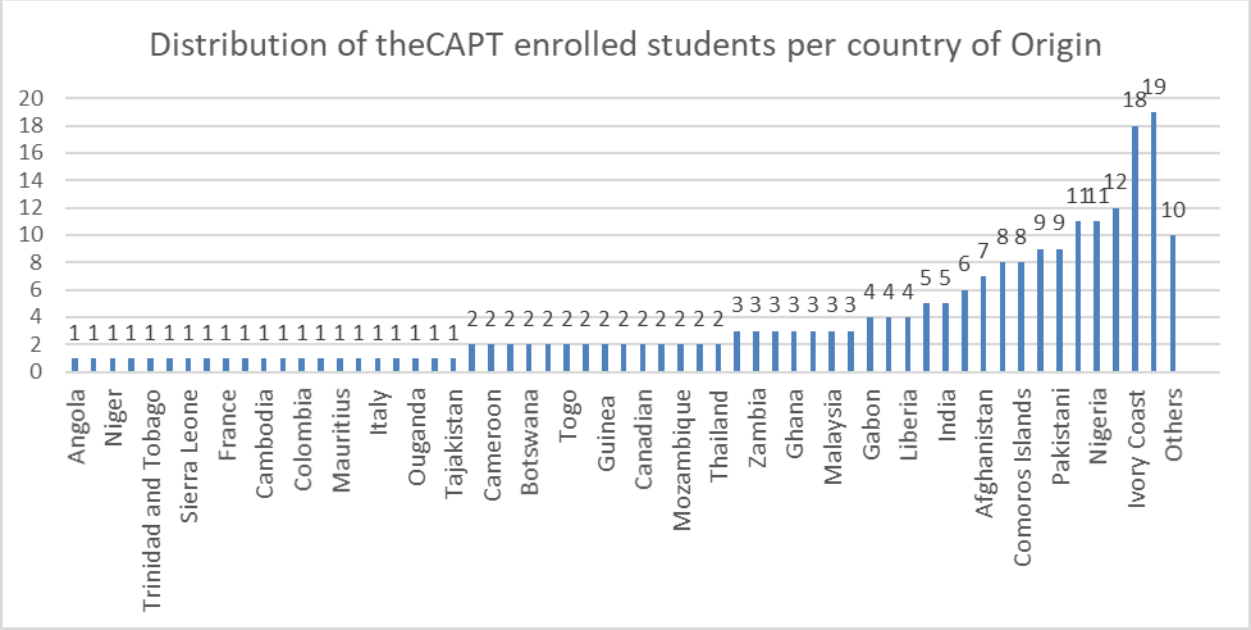
Most of the 371 collected google forms were from the Level 3 and Level 2. After many checking and controls for the validity of the pronunciation of the speakers, and testing their aptitude to pronounce Arabic even with errors, but with a minimal fluency, only 220 students had valid recordings, as shown in Table 11.

Table 11, Number of students that enrolled via the Google Form App

Academic level	Google Form Enrolled Students	Completed Valid Recordings
LEVEL 1	30	13
LEVEL 2	107	82
LEVEL 3	143	77
LEVEL 4	90	48
	370	**220**

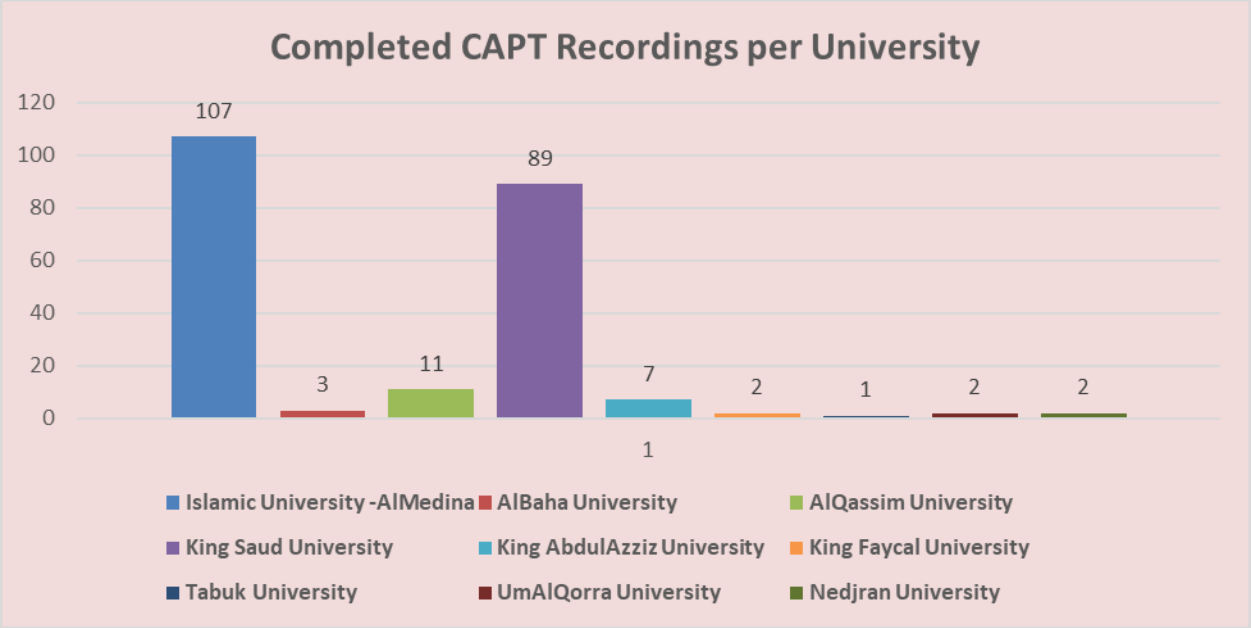
In Figure 6, we present the distribution of the speakers of the 59 nationalities that participated in the recording of session 1.

Figure 6, Statistics of the Nationalities of the CAPT speech recording at session 1



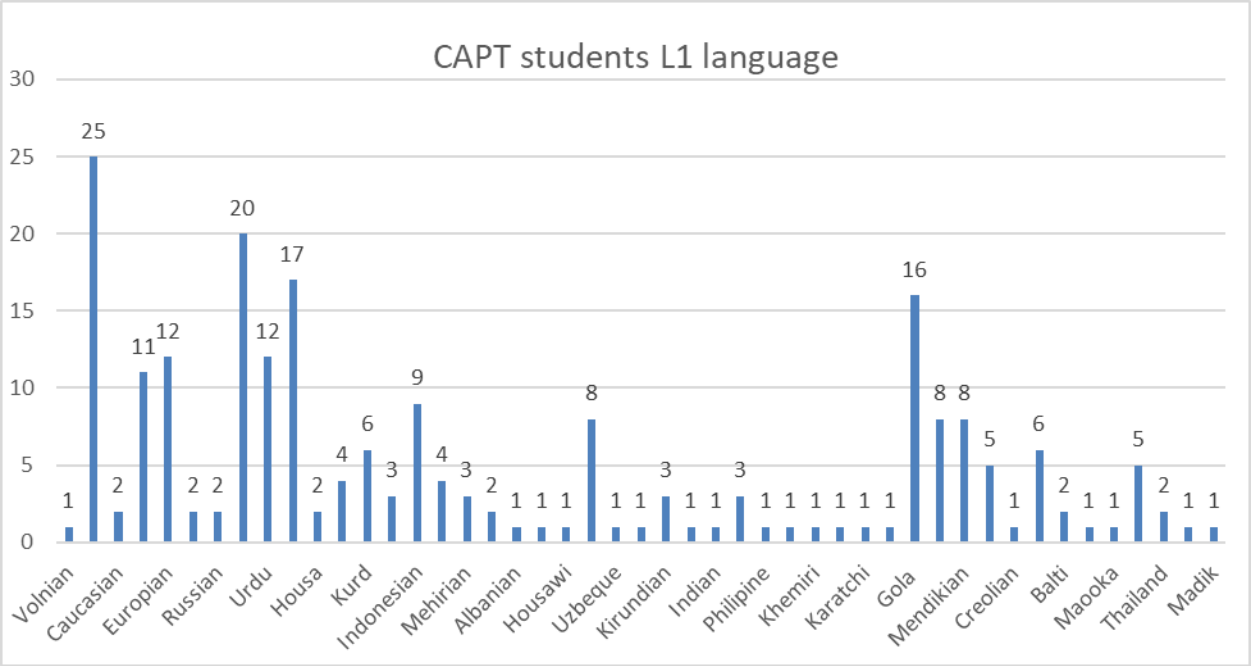
Statistics of the number of completed recordings per university are presented in Figure 7.

Figure 7, CAPT Completed Recordings per University



In Figure 8, we present the number of recorded Non-Arab students from 47 different L1.

Figure 8, Statistics of the L1 of the CAPT speech recording for the session 1 – Non-Arabs



4.2.6. Database Recording

The enrolled students at session 1, were contacted via WhatsApp or by phone in order to officially make them understand the scope of the recording procedure. Each student was provided with a multimedia video tutorial, as a demo of the whole recording made by our database manager, in addition to a manual in Arabic and English, explaining all the steps of the use of the app. Each student received all the items listed in Table 12. A copy of the manual, in both Arabic and English, that was sent to every enrolled speaker is available in Appendix F.

Table 12, Credentials received from /sent to the enrolled speakers (students)

	Item	**Destination**
Student ID:	Ahmed-05555555555	Received within the Google form
Username:	ahmed1	Sent to the student
Password:	123	Sent to the student
Android Application:	Apk format (through WhatsApp)	Sent to the student
Manual:	Tahadath-Manual.pdf	Sent to the student
Video:	App-Demo-Tutorial.avi	Sent to the student
Use of the recorded speech:	Consent screen in the app.	Within the Mobile App

In the manual and video, we tried to be as clear as possible, in order to avoid any inconvenience in the use of the app.

4.2.7. Recording Constraints

- Due to the coronavirus, the decrease in the number of students at the ALI institute forced us to turn to the Islamic University of Al-Madinah, as they have more than 1500 students at their premises from more than 117 nationalities, and this helped us a lot in selecting the quality /quantity required by the project.
- The reason for selecting most of the students from outside of Riyadh, is that ALI student dropped from 300 students at the time of writing the proposal to 70 students at recording time and half of them were not physically present in Riyadh.

- The response from students at Islamic University of Al-Madinah was good at the beginning then stalled, so we recruited students from other universities in KSA

4.2.8. Additional Remarks

- A consent text was written in Arabic in the app, the student needed to approve by clicking a check box, before staring the speech recording session.
- The recording of each speaker was accepted, when it has been double-checked, and passed the quality control criteria defined by the team.
- The student received an honorarium against his participation to the CAPT recordings.
- Many students from the level 1 could not read the texts completely, and were discarded from the recordings.

4.2.9. Recording Arab Speakers

Recording of Arab speakers started after recording of Non-Arabs. The project team tried hard to recruit Arab volunteers by personal invitation and by sending the request to participate in many WhatsApp groups. The response was very slow, hence the number of those who registered in the system database is 58 and among them 32 recorded their speech.

4.2.10. Error detection and analysis by human experts

Before sending the recorded wave files to the human annotators for error annotation, many steps of cleaning and checking have to be realized. These steps range from checking the data consistency to extra sounds removal. This step is a mandatory pre-processing step, and will be described in details throughout the next paragraphs.

4.2.11. Pre-processing of the recorded wave files

Two specific tasks have been assigned to two different per-processing teams that were required to check the content and mention any errors in separate files, and never alter the original wave files by any manner. Details of the tasks of team A & B are described in Table 13.

Table 13, Details of the per-processing teams.

Team ID	Task	Details of Tasks	Number of Checkers per team	Software
A	Content checking:	Check the content of each wave file at the content conformity or consistency level.	2	Audio reader
B	Detailed content analysis: Mark Additional words or sounds	Check for any additional words or sounds that are not part of the CAPT content, using the ELAN software. (Marking temporally the outlier segments)	3	ELAN software

The first **Team A**, had a huge load of daily work, as they were checking each day all the recorded speech and were giving quantitative reports of the correctness and quality of each recorded content. The task was very tedious and took a lot of efforts and time. Detailed samples of the daily reports of team A is presented in Table 14. Where “Pxx” means paragraph number xx (one to three sentences grouped all together), and “SPWyy” means list yy of Sequence of Pairs of Words in the text to be recorded, (each list contains 2 to three pairs of words).

Table 14, Sample reports of the content checking

NAME	DATE	RECORDING STATUS	VERIFICATION	COMMENTS (Coding Used : Pxx : Paragraph xx, SPWyy : List of Sequence of Pairs of Words yy)
Masked Names	17/11/2020	COMPLETE	Done	SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	17/11/2020	COMPLETE	Done	
	17/11/2020	COMPLETE	Done	SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	17/11/2020	COMPLETE	Done	Lot of stutter & repetitions in P1,P2,P3,P4,P5,P10,P17,P18,P23,P25 | SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	17/11/2020	INCOMPLETE		
	17/11/2020	COMPLETE	Done	Small noise in SPW14
	17/11/2020	COMPLETE	Done	Small noise in P7
	17/11/2020	INCOMPLETE		
	17/11/2020	COMPLETE	Done	SPW6,SPW8,SPW9,SPW10,SPW11,SPW12,SPW14,SPW16 | missed the last 2 words in SPW6,SPW8,SPW13,SPW15
	17/11/2020	INCOMPLETE		
	17/11/2020	COMPLETE	Done	
	17/11/2020	INCOMPLETE		
	17/11/2020	COMPLETE	Done	repetitions & noise in P1,P9;/ SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	17/11/2020	INCOMPLETE		
	17/11/2020	INCOMPLETE		
	17/11/2020	INCOMPLETE		
	17/11/2020	COMPLETE	Done	repetition in P5| repetition & noise in P13,P14,P19 | noise in P20 | P24 missed the first word | SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	18/11/2020	INCOMPLETE		
	18/11/2020	INCOMPLETE		
	18/11/2020	COMPLETE	Done	P1, P4 wrong record | stutter & repetitions P7,P8,P19,P21,P23,P24
	18/11/2020	INCOMPLETE		
	18/11/2020	INCOMPLETE		
	18/11/2020	INCOMPLETE		
	18/11/2020	INCOMPLETE		
	18/11/2020	COMPLETE	Done	bad stutter & repetitions P2,P3,P4,P5,P6,P12,P13,P14,P15,P16,P18,P19,P22,P23,P24,P25,SPW1,SPW14,SPW15 | stutter & repetitions & noise P7,P8,P9,P10,P11,P17,P20,P21 | SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	19/11/2020	COMPLETE	Done	stutter & repetitions P3,P20,P22,P23,P14,P16,P25,SPW7 | SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	19/11/2020	INCOMPLETE		
	21/11/2020	COMPLETE	Done	SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	21/11/2020	COMPLETE	Done	stutter, repetitions & noise P1,P2,P3,P4,P5,P6,P7,P10,P11,P13,P14,P16,P17,P20,P23,P24,SPW9,SPW15 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each
	21/11/2020	COMPLETE	Done	stutter repetition P11,P13,P14,P19,P24,SPW1,SPW13
	21/11/2020	COMPLETE	Done	stutter in P4,P10 | SPWrong sentences: P25 | SPW6, SPW8, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	17/11/2020	INCOMPLETE		
	21/11/2020	COMPLETE	Done	Incomplete records P2,P3,P4,P5,P7,P11,P12,P13,P14,P15,SPW5,SPW16 | wrong sentences: P8 | noise in: P21,P24,P25 | empty record: SPW8 | SPW6, SPW7, SPW9, SPW13, and SPW15 are not complete (2 words missed at the end of each record)
	21/11/2020	COMPLETE	Done	stutter & repetition in P3,P10,P11,P21,P22,P25,SPW3 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each
	22/11/2020	COMPLETE	Done	empty record: P1 | stutter & repetition in P2,P4,P5,P6,P7,P8,P9,P10,P11,P14,P15,P16,P17,P18,P19,P20,P21,P22,P23,SPW9,SPW14 | wrong sentence: P24, P25 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each
	21/11/2020	COMPLETE	Done	P1 wrong record | stutter & repetition in P3,P4,P10,P13,P14,P16,P19,P22,SPW7,SPW9,SPW15 | noise in: P5,P11,P23 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each record
	23/11/2020	INCOMPLETE		3 files missed while according to the apps he's finished
	23/11/2020	COMPLETE	Done	stutter repetition P7,P12,P16,P18,P20,P22 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each record
	22/11/2020	COMPLETE	Done	stutter repetition P3 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each record
	23/11/2020	COMPLETE	Done	stutter repetition P2,P3,P9,P14,P15,P17,P20,P25 | noise P4,P6, P19 | SPW6, SPW8, SPW13, and SPW15 are not complete 2 words missed at the end of each record

At this stage, the recordings have been checked for correctness of global content, i.e.: the content of each speaker directory should contain 41 wave files, as described in Appendix B. Once Team A, has completed its task, Team B will continue the verification process by marking each speaker speech for extra sounds that have been introduced during the App recording. Examples of these sounds includes stuttering, door knocking, baby in background, etc…, as many students were recording from home or from the university housing. The main objective of this step in the pre-processing was to be sure that what was written in the App cards, regardless if

it was correctly pronounced or not, is the only speech or sounds within the wave files.

We opted to use the ELAN software due to its capabilities in annotation and export possibilities to text grids, further used by our other software applications. In order for the Team B to work fluently and at fast pace, the technical team concatenated the wave files, to ease the step of listening to the speech and marking the segments of text or special sounds that are not within the original text.

This task was executed by the three checkers in team B and required a lot concentration and repeated listening and required long time. For example, a checker in team B spent one hour to process the speech of one speaker that was recorded in 10 min speech.

Table 15 to Table 17 present some statistics about the minimum, average, and maximum speech duration for the 220 Non-Arabs recorded speakers. A very detailed speaker duration is listed in Appendix D.

Table 15, Statistics about the session 1 speech recording

Item	Details
Minimum time	6.37 min
Average time	12.06 min
Maximum time	37.4 min
Speakers <10 min	66 speakers : (Avg time: 8.98min)
Speaker Rec. >=10 min and <15min	128 speakers (Avg time : 12.13min)
Speaker Rec. >=15 min and <20min	19 speakers (Avg time : 16.96min)
Speaker Rec. >=20 min	7 speakers (Avg time : 26.43min)
Short Paragraphs duration	31.29hours
Sequence of pair of words duration	12.94hours
Total recorded time for all the speakers	44.23Hours
Number of volunteers who registered in the system	371
Number of Speakers who installed the app and started recording	235
Number of Speakers who completed Recording	224
Number of Speakers with valid recordings	220
Number of Speakers with completed transcription checking	132 (60%)

Number of Speakers with completed phoneme level labeling	40 (18%)

Table 16, Additional deep statistics about the session 1 speech recording of Non Arabs: SPW

SPW ID	Avg time (sec)	Min time (sec)	Max time (sec)	Total time for the SPW(Sec)
SPW1	13.95	3.76	27.04	3068
SPW 2	8	3.02	22.02	1761
SPW 3	11.7	5.6	29.2	2574
SPW 4	10.29	5.22	20.86	2265
SPW 5	7.39	3.36	15.94	1626
SPW 6	16.02	6.92	30.56	3524
SPW 7	14.61	6.82	31.08	3214
SPW 8	15.99	6.46	30.16	3518
SPW 9	14.3	5.7	41.6	3146
SPW 10	14.06	1.28	23.68	3092
SPW 11	10.9	5.12	27	2399
SPW 12	10.01	3.7	23.8	2202
SPW 13	15.07	5.92	67.48	3315
SPW 14	15.38	6.5	57.3	3383
SPW 15	20.05	6.16	48.34	4411
SPW 16	14.08	2.7	38.08	3098

Table 17, Additional deep statistics about the session 1 speech recording of Non Arabs: Paragraphs (short paragraph)

Short	Avg. time (sec)	Min time (sec)	Max time (sec)	Total time per type of
P1	21.91	11.08	61.28	4820
P2	25.74	12.66	423.66	5664
P3	27.02	9.86	121.4	5944
P4	22.27	10.38	84.46	4899
P5	18.55	9.24	67	4080
P6	15.82	7.96	111.76	3480
P7	18.68	10.32	74.64	4109
P8	16.52	8.78	66.14	3635
P9	16.64	7.4	71.32	3661
P10	15.73	7.82	78.9	3462
P11	22.89	11.4	78.5	5035
P12	21.94	10.5	71.52	4827
P13	27.36	8.22	78.26	6020
P14	25.61	3.36	72.46	5633
P15	15.88	7.62	89.72	3495
P16	14.76	7.64	50.38	3248
P17	19.8	7.36	89.62	4355
P18	14.23	7.16	61.56	3130
P19	26.8	10.6	206.24	5896
P20	17.8	7.34	76.34	3915
P21	24.8	12.8	77.26	5455
P22	22.43	9.48	89.04	4934
P23	26.79	7.34	87.94	5894

P24	19.26	9.1	100.08	4237
P25	12.81	5.06	82.68	2818

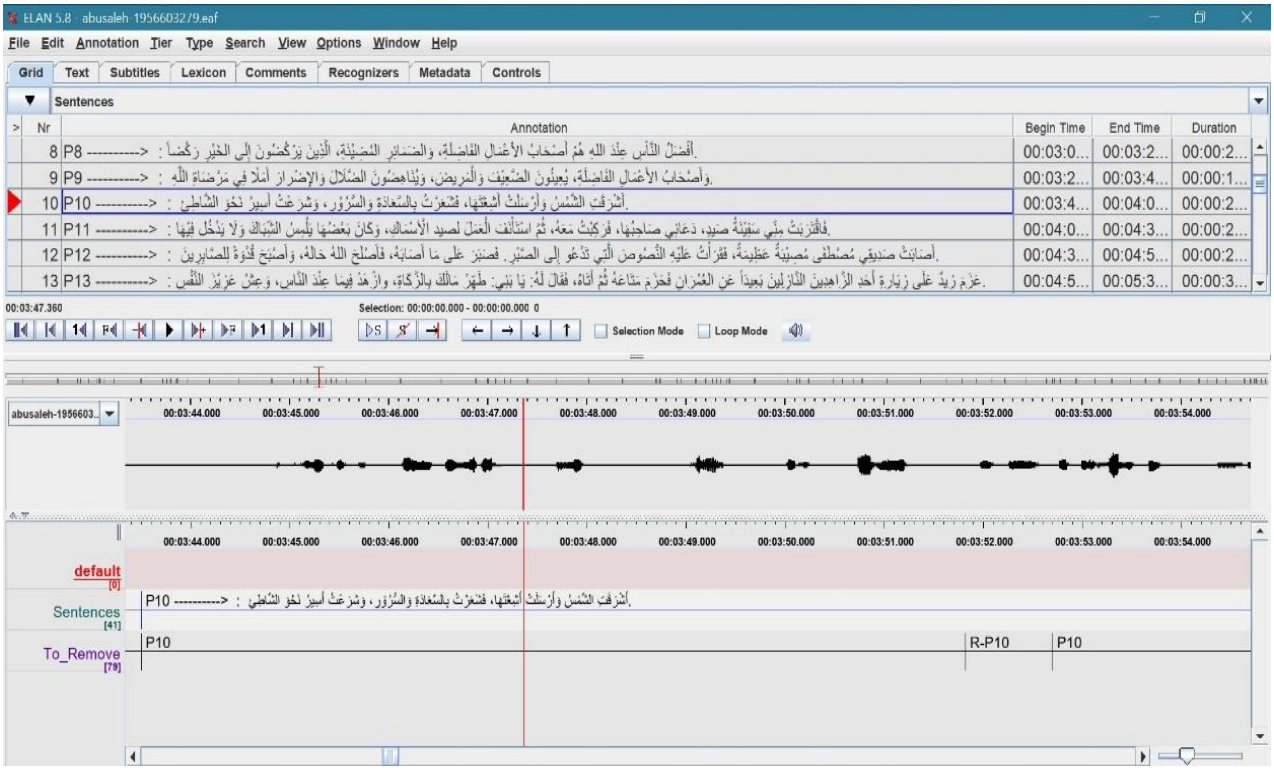
In order to check the conformity of the written transcription (i.e. displayed text in the mobile screens) to the content of the wave file, we opted to use the ELAN annotation tool, which is known for its labeling capability for both audio signals and video sequences, in a single or multi-tiers labeling. Hence we used ELAN software as an aiding tool in speech segmentation. Using ELAN Team B listened carefully to the wave files, and marked any additional noise, repeated word or background sound, by putting time boundaries around them. This process kept the original files as is, and generated a PRAAT TextGrid file, that was used later, by our technical team, for segmenting the wave files, where only the valuable parts were kept.

In order to ensure that, Team B did a job without apparent errors, an additional effort of rechecking 5% of the speakers is performed by the project technical team to recheck the conformity of the content. If no error is detected, no further action is made, else more samples will be rechecked.

All intermediate work and results have a backup copy, in case of need for traceability of errors at the labeling level.

Team B, followed a very strict protocol, from the wave file opening in ELAN to the text-grid generated at the end of the checking session. A copy of the protocol is presented in Appendix E. In Figure 9, we present a screenshot of the ELAN software work for sentence P10.

Figure 9, Screenshot of the ELAN screen work



From the different reports submitted by teams A and B, we collected some remarks that were taken into account while recording the speakers in the next session.

Conclusions from the written feedback of team B.

- The voices of some speakers were not very clear.
- Some speakers had difficulties in reading, hence they read in a very low speed.
- Many stuttering and repetitions.
- Difficulties in reading some sentences or words.

4.2.12. Labeling of speech of session 1

The speech labelling is the process of time labelling the speech, at the sentences, word or phoneme levels, where for each speech wave file a new time label file is produced, and this file contains the timings, generally, the text content, the start time and end time of whatever required modality (sentence, word, phoneme).

The output of Team B, which contains the timed marked errors, is used by our technical team to clean the (concatenated) wave files. The technical team removed all the erroneous segments from the (concatenated) wave files, then they were split back to single wave files as per the initial App cards' content. Next the content was transferred to the phoneme annotators and error analyzers.

At the beginning of the annotation we gave the annotators 41 wave files, and an Excel sheet, having the format shown in Table 18. This table was based on a previous study that we conducted before.

Table 18, Annotator speech phoneme level segmentation sample (Empty form to be filled by annotators)

	Please press here to listen to the speech				
Comments	Deletion	Addition	Substitution	Pronounced text	Reference Text
					كَانَ
					ك
					َ
					ا
					ن
					َ
					لِرَجُلٍ
					ل
					ِ
					ر
					َ
					ج
					ُ
					ل
					ٍ

Speech annotation experts will listen to each wave file and check the phonemes of each word:

- If a word is well pronounced, no remark is written.
- If the word is not well pronounced, by either addition, deletion or substitution of any phoneme, the annotator will mark the location of the error compared to the reference canonical text.

The final output of this step are excel files containing specific information of each wave file, such as the phonemes present in the text (reference text), phonemes correctly pronounced (correct phonemes), phonemes that were not pronounced (deleted), the phonemes badly pronounced (substituted), or added phonemes (added).

After initial work by the annotators we realized there are more in speech labeling and subjective evaluation than the information to put in Table 18. Hence with consultation of the Arabic language experts we came a new format as in the next section.

4.2.13. Subjective evaluation of the pronunciation of L2 learners

As we mentioned in the previous section in addition to the difference between the canonical text and the pronounced speech due to substitution, insertion, and deletion, which are errors, there are other ways of pronouncing that might be correct or wrong. Hence we had to add notation for these differences. To make annotation faster we proposed a new table that will cover speech labeling as well as subjective evaluation. For substitution and addition no symbol is used and the annotators write the canonical and pronounced text as in Table 19, the other two columns are explanation for the report only.

Table 19, Examples of annotating substitution and addition

طريقة تمثيل الاستبدال والإضافة... (بدون رموز)

توضيح **Explanation**	النص المنطوق **Pronounced text**	النص النموذج **Canonical text**	نوع الخطأ **Type of error**
استبدال الياء بالراء	أَنْتَظِيُ	أَنْتَظِرُ	**الاستبدال** **Substitution**
استبدال الذال بالظاء	غَيْذِيَ	غَيْظِيَ	
زيادة همزة القطع، والأصل همزة وصل (التي هي في حكم العدم صوتيا)	إِسْمُهُ	اسْمُهُ	**الإضافة** **Addition**
زيادة الراء بعد الكلمة	أَفْطَرْتُرْ	أَفْطَرْتُ	

For deletion and other operations, we used the symbols as in Table 20. The annotators write the canonical and pronounced text as in in Table 20, the other three columns are explanation for the report only.

Table 20, Examples of annotating deletion and other speech processes

طريقة تمثيل الحذف والظواهر الصوتية الأخرى التي قد يستفاد منها في رصد التباين الصوتي بين مختلف المتحدثين أثناء ملاحظة الأخطاء... (باستخدام الرموز)

Symbol	Process	Example		
		النص النموذج	النص المنطوق	توضيح
#	**التفخيم**	رَضْيَ	رَد#يَ	– استبدال بين الضاد والدال ثم تفخيم الدال.
		طَهِّرْ	طَهِّرْ#	– تفخيم الراء.
@	الترقيق	اللَّهِ	اللَّ@هِ	– ترقيق اللام.
*	الإدغام	وَجْهًا لِوَجْهٍ	وَجْهَنْ* لِوَجْهِ×	إدغام النون الساكنة (التنوين) في اللام، وحذف النون الساكنة (التنوين) في آخر الكلمة الثانية.
&	الإخفاء	حَدِيقَةٌ تُثْمِرُ	حَدِيقَتُنْ& تُثْمِرُ	إخفاء النون الساكنة (التنوين) في التاء.
<	الإقلاب	خَاصَّةٌ بِي	خَاصَّتُنْ< بِيْ	قلب النون الساكنة (التنوين) ميمًا قبل الباء.
×	الحذف	المَظَارِيفَ	الْمَظَارِ×فْ×	هنا حذفان: تقصير الصائت الطويل (الياء الساكنة) إلى صائت قصير (الكسرة) وهو نوع من الحذف، وحذف الصائت القصير (الفتحة) من آخر الكلمة.
^	القلقلة	وَقَدْ تَلَاهُ	وَقَدْ^ تَلَاهُ	قلقلة حرف الدال قبل التاء والأصل يدغم الأول في الثاني.
~	مد في غير موضعه أو استمرار الضغط على الصوت	أَنْقَذَكَ	أَنْ~قَذَكَ	الضغط المستمر على النون
/	الفصل أو التوقف	خَرَجْتُ	خَرَجْ/تُ	التوقف بعد الجيم ثم الانتقال إلى التاء

التوقف بعد الميم ثم الانتقال إلى التاء	الأَطْعِمَ/ةِ	الأَطْعِمَةِ		

4.2.14. Error Analysis

Similar to the work of the researches in L2-Artic and due the budget limitation, we decided to annotate 60 speakers from session 1. The 60 speakers were chosen to cover all nationalities recorded in session 1. Two Arabic experts annotated all the recorded speech of the 60 speakers. Once, the two annotators finished the evaluation process, we check the agreement between them, any utterances with disagreement between the annotators were sent to another Arabic expert to evaluate it. In the following, we present analysis of the annotated pronunciation errors.

Table 21 shows phonemes statistics of the annotated part. We can notice the large numbers of phonemes in the recorded database. We can also notice that the percentage of pronunciation errors is 10.5%, and that the ratios of the different types of pronunciation errors are very near to each other.

Table 21: Phonemes statistics of session 1.

Type	Number	%
Corrections	209825	89.5
Substitutions error	8739	3.7
Insertions error	6626	2.8
Deletions error	9271	4.0
Total	234461	100

Table 22: The top-10 and top-5 substitution errors in KSU-CAPT session 1.

Consonant phonemes				
Canonical phoneme	IPA	Substituted phoneme	IPA	Frequency
ع	ʕ	همزة	ʔ	325
ص	sˤ	س	s	251
ض	dˤ	ظ	ðˤ	215
ن	n	هـ	h	195
ط	tˤ	ت	t	188
ح	ħ	هـ	h	142
ض	dˤ	د	d	139
ظ	ðˤ	ز	z	102
ن	n	ل	l	91
ث	θ	س	s	85
Vowel phonemes				
الف مد	a:	فتحة	a	598
فتحة	a	الف مد	a:	451
كسره	i	فتحة	a	237
فتحة	a	ضمة	u	234
فتحة	a	كسرة	i	230

In Table 22, the top-10 substitution errors are shown. We can notice that most substitution errors were in the vowels phonemes, and we can attribute this to the fact that the non-native Arabic speakers have difficulties in pronouncing the discretized text (النصوص المُشكلة). For the consonant phonemes, the highest number of substitution errors were for substituting the phoneme /ʕ/ with /ʔ/, phoneme /sˤ/ with /s/, and phoneme /dˤ / with / ðˤ /, which agrees with the published research.

The top-10 insertion and deletion errors are depicted in Table 23.

Table 23: The top-10 insertion and deletion errors in the KSU-CAPT session 1.

Insertion errors		Deletion errors	
Inserted phoneme	Frequency	Deleted phoneme	Frequency
فتحة	1721	ضمة	2410
همزة	1201	ن	2121
كسرة	992	كسرة	1258
ضمة	401	فتحة	1140
ن	337	ت	584
ه	231	و	291
الف مد	223	ل	271
ل	213	همزة	206
ي	193	ر	150
و	153	ي	119

4.3. KSU-CAPT Non-Arabs Database-Session 2

4.3.1. Selection of text for recording of the speech corpus

A team member is a linguistic scholar who is also an experienced instructor of Arabic as a second language proposed a new methodology to choose the text to complement the methodology that we used to select the text of session 1. Below is the new methodology in Arabic.

ما زالت دراسة نطق الأصوات العربية تفتقر إلى قاعدة بيانات مناسبة (Data Base) والتي يمكن أن يعتمد عليها في إنشاء برامج حاسوبية تتميز بدقة عالية لتعليم العربية لغير الناطقين بها أو التعرف على أصواتها. ونظرا لصعوبة دراسة نطق جميع الألفاظ العربية في فترة زمنية محدودة، فقد تصل الألفاظ التي تمثل الظواهر الصوتية المختلفة لنطق الصوت الواحد إلى المئات، وهو أمر غير ممكن في ضوء الإمكانات والحدود الزمنية للمشروع الواحد؛ نشأت الحاجة الملحة إلى اختيار عينة مناسبة من الألفاظ بحيث تشمل الظواهر الصوتية المختلفة قدر الإمكان. ومن هنا، قمنا بعمل هذا النموذج ليحقق هذه الغاية، أي اختيار ألفاظ تشمل الظواهر المختلفة لنطق الأصوات العربية لغير الناطقين بها مع مراعاة المدة الزمنية التي تستغرقها قراءة الكلمات بحيث لا تزيد عن عشرين دقيقة تقريبا؛ حيث قمنا باتباع الأسس العلمية والخطوات المنهجية –المبينة أدناه– التي تضمن لنا تحقيق هذا الهدف، وذلك بعد الرجوع إلى أهم الدراسات الصوتية إلى جانب برامج وكتب تعليم اللغة العربية لغير الناطقين بها، فضلا عن خبرة الباحث ومعرفته الطويلة في الظواهر الصوتية العربية ومشكلاتها.

Table 24 shows an example of the selected text for the Arabic phonemes (/خ/،/ق/), which consists of isolated phonemes, words, and minimal pairs. The same procedure has been done for the remaining Arabic phonemes.

Table 24, Sample sentences for session 2

The phoneme /ق/	أَقْ / أقْ إِقْ / قَا قُو/ قِي قَتَادَة / قُنْبُرَةْ قِصَّةْ / بُرْقُوقْ مُبَرْقَعْ / اِقْتَصِدْ تَقَطَّرَ / اِسْتَقْرَأْ مِقْياس / تَقَوقُعْ غاسِقْ / يُشْفِقْ
Minimal pairs of phoneme /ق/	قَسَا /كَسَا يَقِيْل / يَكِيْل تَقْدِيْرْ/ تَكْدِير شَقَّ / شَكَّ مُشْرِق / مُشْرِكْ
Words of phoneme /خ/	أَخْ / أُخْ إِخْ / خَا خو / خِي خَوْلَة / خَدْش خَطَّطَ / خَوْخ خَجْلان / خُضْرَة أخمُص / اِخْشَوْشَن اِسْتِخْراج / بـَخَّاخ مُخ / اِسْتَــرْخِ

Minimal pairs of phoneme Words of phoneme /خ/	خَسْفْ /كَسْفْ خَفَّ /كَفَّ خَاوِيَةْ /كاوِيَةْ تَخَلُّفْ/ تَكَلُّفْ نَخِرَة /نَكِرَة مَسْلُوخْ /مَسْلُوكْ فَخَّ /فَكَّ

Statistics of the selected text of Session-2 is shown in *Table 25*.

Table 25, Statistics of the CAPT-Text selections for session 2

Statistics	
Number of words	579
Total number of phonemes	3199
Maximum number of phoneme / per word	14
Minimum number of phoneme / per word	112
Pairs of Minimal Dual words.	61
Number of Screens : words	31

The numbers of phonemes, within the words and minimal words of all the text of session 2, are presented in Figure 10.

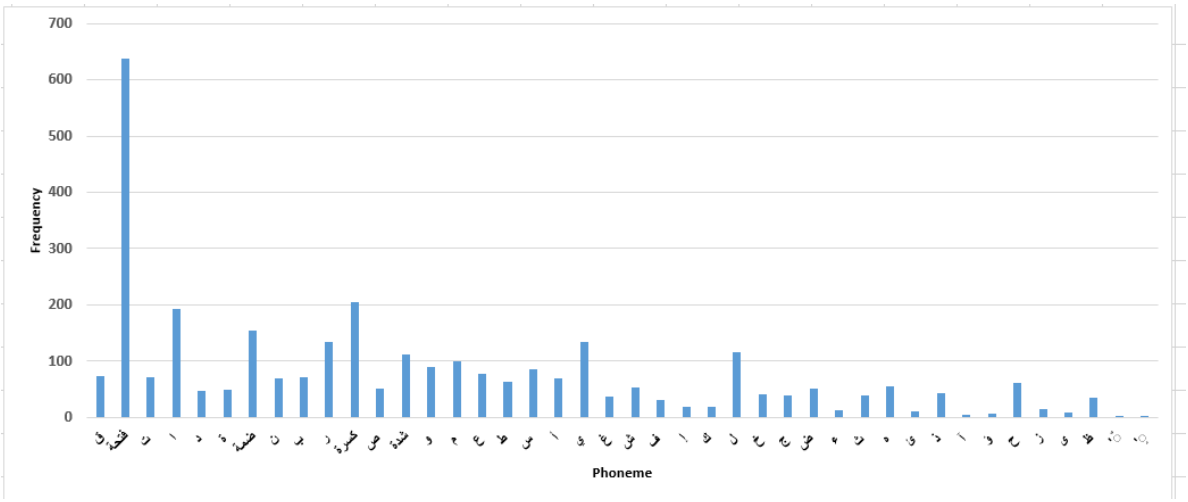


Figure 10: Phoneme distribution in the words and dual pairs of the text of session 2.

After we selected the text of session 2, we updated the mobile recoding app to be suitable for the new text. Screen shots of all the texts of the mobile app are presented in Appendix G.

4.3.2. Speaker Registration

In session 1, we collected the information of 450 non-native volunteers from 10 universities, from these numbers, only 220 speakers completed the speech recording correctly. In session 2, we sent an invitation to all the previous 450 speakers to participate in the new recording, and only 230 speakers accepted and recorded their speech. This time we provided a training sample hosted on a website to allow the speakers to listen to the recordings of an Arab expert, before recording the 42 screens of the mobile app. The average recording duration ranged between 15 and 20 min per speaker.

In the following, we will present some statistics of the recorded speakers. The statistics include number of students, country of origin, language spoken, level of education, etc. Table 26 shows the distribution of the recorded speakers for each academic level. We can clearly see that most of the 230 speakers were from the Level 3, Level 4, and Level 5. This distribution met our objective which is to enroll more students from the middle levels rather than the students from low levels who have some difficulties in pronouncing Arabic language correctly, and students from advanced levels who have fluent Arabic pronunciation.

Table 26, Number of students that recorded in session 2

Academic level	Completed Valid Recordings
LEVEL 1	8
LEVEL 2	22
LEVEL 3	40
LEVEL 4	68
LEVEL 5	74
LEVEL 6	5
LEVEL 7	13
	230

In Figure 11, we present the distribution of the speakers of the 51 nationalities that participated in the recording of session 2.

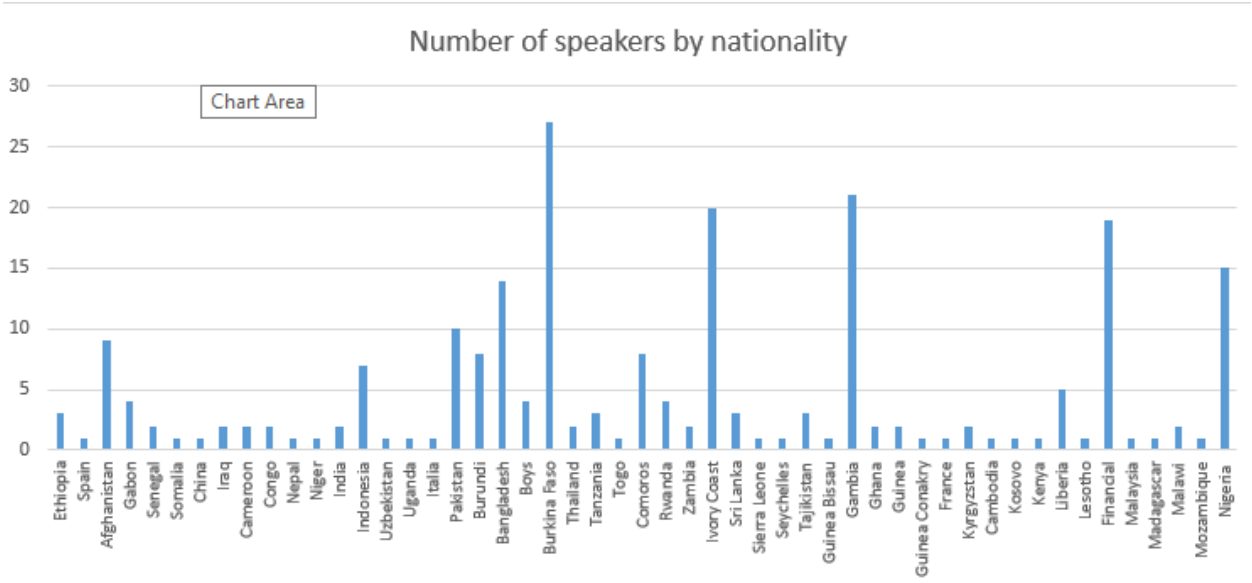


Figure 11: Statistics of the Nationalities of the CAPT speech recording at session 2.

Statistics of the number of completed recordings per university are presented in Figure 12.

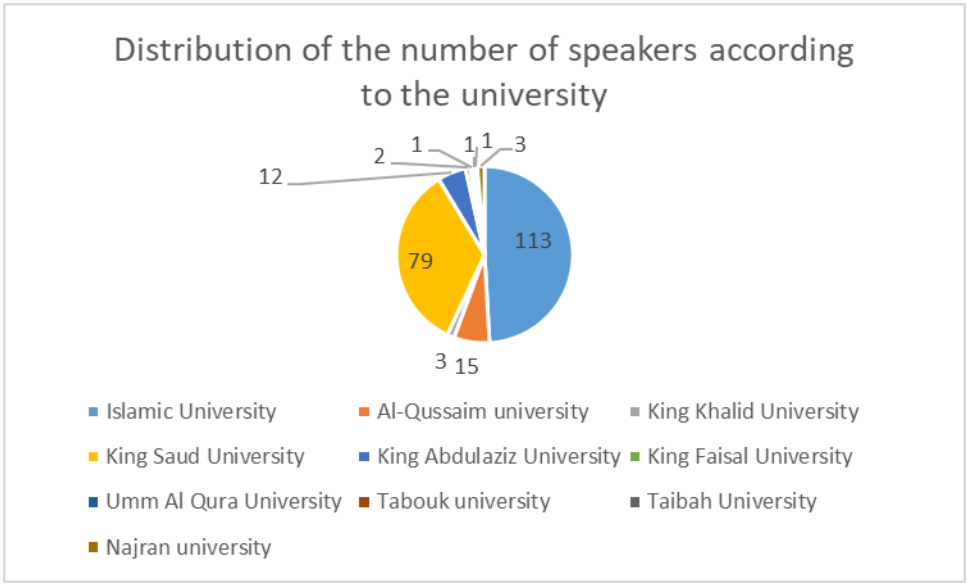


Figure 12: CAPT-Session 2 Completed Recordings per University.

In Figure 13 we present the statistics of the L1 of the speakers of CAPT speech recording for the session 2, where the speakers were from 51 L1.

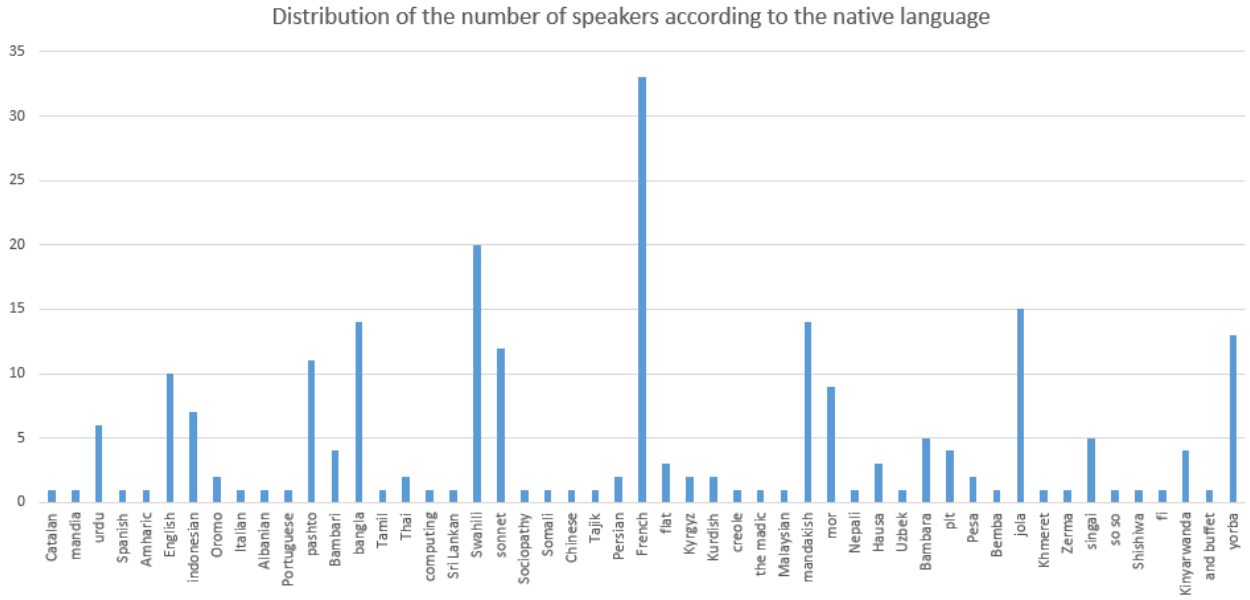


Figure 13: Statistics of the L1 of the CAPT speech recording for the session 2.

4.3.3. Errors annotation and analysis

In session 2, we followed the same process of annotation of session 1, and we annotated 60 speakers as in session 1. Based on the experiments on session 1, we noticed that the 60 speakers were sufficient to train the CAPT system. Number of corrected and mispronounced phonemes in this session is shown in Table 27.

Table 27: Phonemes statistics of session 2.

Type	Number	%
Number of corrections	137475	88.11
Number of substitutions	6268	4.02
Number of insertions	10804	6.92
Number of deletions	1483	0.95

Total	156030	100

Table 28 shows the top 10 substitution errors in consonant and vowel phonemes in the KSU-CAPT session 2, where the occurrences of vowels' substitution decreased dramatically compared with session 1. We can attribute this to the fact that the text of session 2 was words with no diactrics at the end which will produce less number of errors in vowels.

Table 28: The top-10 errors substitution in session 2.

Consonant phonemes				
Canonical phoneme	IPA	Substituted phoneme	IPA	Frequency
t	ʕ	h	ʔ	2043
l	sˤ	HZ	s	563
Z	dˤ	z	ðˤ	298
D	n	Z	h	146
T	tˤ	t	t	130
D		d		86
Z		TH		82
n		m		77
TH		z		74
q		k		72
Vowel phonemes				
a		a 2		99
u		a		98
a 2		u 2		86
a 2		a		83
a		i		78

Insertion errors		Deletion errors	
Inserted phoneme	Frequency	Deleted phoneme	Frequency
HZ	2891	a	267
a	2866	u	115
u	780	r	109
h	424	l	105
i	349	w	91
n	281	n	90
t	231	T	89
y	200	t	81
r	149	HZ	78
m	173	i	71

Table 29: The most frequent insertion and deletion errors in session 2.

5. Design and development of the phoneme recognition and AFs detection for CAPT system.

When we submitted the proposal, we proposed to use the conventional way in speech recognition systems, which consisted of the following steps: feature extraction, feature reduction, and classification. With the great advancement of deep learning and huge improvement in computation power, many of state-of-the-art systems in the current literature are using deep learning to construct end-to-end high performance speech recognition systems. Hence in our work in the first year we investigated using deep learning networks for phoneme recognition system and for mispronunciation detection and diagnosis system (MDD). We proposed a new way of using deep learning for detection and recognition of phoneme and articulatory features (AF). In the new proposed method, we treat the phonemes and AFs as objects in 3 channels spectral images of the speech. By this proposed method we were able to recognize the sequence of phoneme from the whole utterance of the non-native Arabic speakers. Then we used the detected phonemes for mispronunciation detection and diagnosis task.

Providing feedback to non-native learners is very important especially at articulation level. Hence, by the proposed method we detect and recognize the AFs as objects in the 3 channels spectral images, then we use the detected AFs for mispronunciation correction and providing feedback at articulatory level. Figure 16 shows the general architectures of the proposed Arabic CAPT system. In the following sections, we explain the details of the proposed Arabic CAPT system.

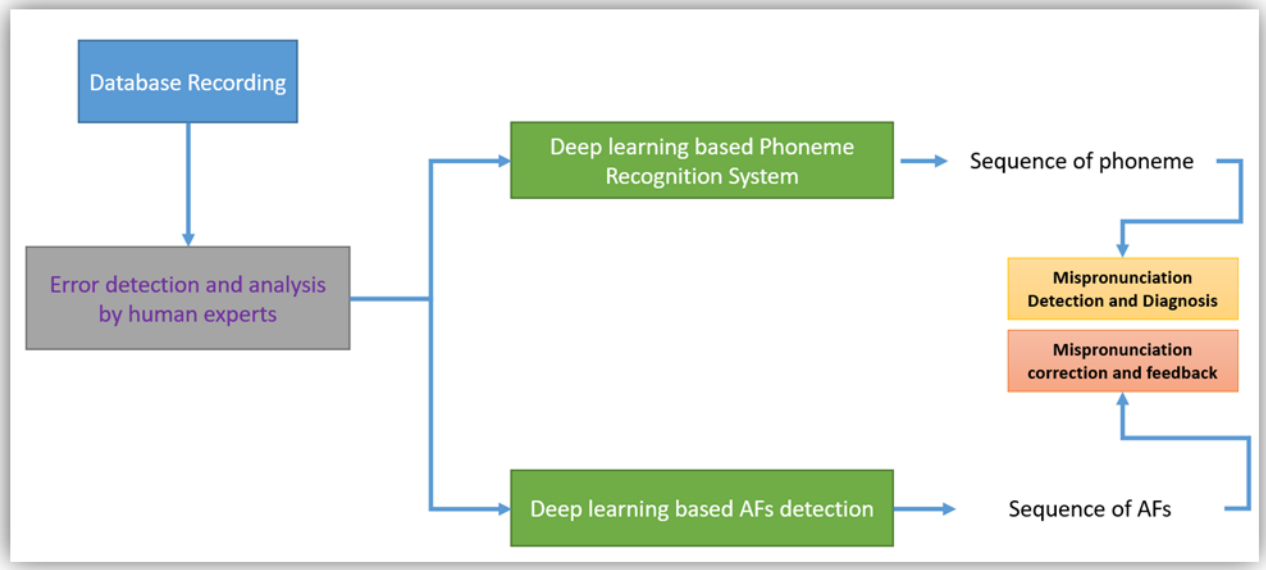


Figure 14: General Architecture of the proposed Arabic CAPT system.

5.1. Proposed method for phoneme recognition in CAPT

The proposal was based on state of the art techniques at time of submitting the proposal, and we mentioned deep neural networks (DNN) as something to investigate. When we started working in the project DNN was widely used in speech recognition, hence we focused on DNN and proposed a new method of using object detection techniques, based on DNN, to detect the phonemes. The proposed method gave excellent results compared to state-of-art techniques and we published our work in [59]. This shift meant that we will not investigate many hand crafted speech features, as we did on some of our previous researches, as the DNN is powerful and has the ability to work in raw data. This also resulted in combining task 3.2.2 of the proposal with task 3.2.3 of the proposal, since the DNN will perform the reduction in its initial layers.

Before presenting the accomplished results and systems of this project and to put our work in the project in prospective we will very briefly summarize our work and results in [59] where we proposed using object detectors to detect the phonemes as objects in an image.

In [59], we proposed the use of object detection techniques for recognizing sequence of phonemes from the whole spectrogram. We converted the utterance of speech to a three channel spectral image then we used a deep object detection models for detecting the phonemes from the spectral image. The novelty of the proposed system is represented by treating of phonemes of utterance as objects in spectral images. We chose YOLO and CenterNet, two cutting-edge object detectors, based on a trade-off between detection accuracy and speed. Our study is the first study in literature, to the best of our knowledge that used object detection for phoneme recognition system. We evaluated the proposed system using native English and non-native Arabic speech corpora. For English phoneme recognition, we used the TIMIT dataset [60] which is a well-

known English speech corpus. For non-native Arabic phoneme recognition, we used a small part of the KSU speech database [61] [58] [62].

Due to the small size of the databases, we investigated using different types of transfer learning techniques as follow:

- Transfer learning from image to speech databases (DTS)
- Transfer learning between speech corpora within the same language (IaTS)
- Transfer learning between speech corpora within the different language (IeTS)

Table 30 shows the result of the two proposed systems DTS and IaTS for the test set of the TIMIT corpus. We can see that the performance of the transfer learning improved the results. We achieved the best PER using the IaTS based on CenterNet detector with DLA backbone network which was 15.89%.

Table 30: PER for the TIMIT test set.

System	***Object Detector***	***Model***	***PER***
DTS	YOLO	YOLOv3-tiny	28.25
DTS	YOLO	YOLOv3	20.2
DTS	CenterNet	ResNet	21.09
DTS	CenterNet	DLA	19.06
IaTS	YOLO	YOLOv3-tiny	25.57
IaTS	YOLO	YOLOv3	16.34
IaTS	CenterNet	ResNet	17.16
IaTS	**CenterNet**	**DLA**	**15.89**

Table 31 presents the performance of the proposed system IeTS for non-native Arabic phoneme recognition. In this experiments, we used speech of 15 non-native speakers and 5 native speakers for training and 11 non-native speakers for testing. The total number of phonemes in training and testing utterances is 14413. The phoneme recognition system based on YOLO detector tiny version achieved 10.15% PER and the one based on CenterNet detector achieved a 7.58% PER.

Table 31: PER for non-native Arabic Speech (Small-Arabic-CAPT).

System	***Object Detector***	***Model***	***PER***
IeTS	YOLO	YOLOv3-tiny	10.15
IeTS	CenterNet	DLA	7.58
HMM [62]			28.8

The results in [59] are comparable or better than state of the art published researches.

5.2. Proposed method for AFs detection in CAPT

In this part we will present the main results of this task out of the project within the first year. The details of the work and the results were published in [63]. In the following we will try to highlight the main work, findings and results in the paper. The paper title is "Deep learning-based detection of articulatory features in Arabic and English speech", nonetheless it also investigated detection of the phonemes in two ways. It detected the phonemes either directly from the spectral images or based on the detected AFs. To detect the AFs, we proposed using object detection techniques to recognize sequence of AFs from speech utterances by treating AFs of phonemes as multi-label objects in spectral images. Note that for phoneme recognition we treated the phonemes as single label objects. We tested the proposed system on English corpus, TIMIT, and on Arabic speech corpus, KAPD [64]. Figure 15 sows the general overview of the proposed systems, where the system to detect the AFs is called AFD-Obj and the system to detect the phonemes is called PD-Obj. By detecting the AFs, we can provide feedback to the non-native Arabic learners at articulatory level. This is a new important and beneficial feature that we included in our Arabic-CAPT system, though it was not in the proposal. Moreover, we studied the effects of the number of detection levels of YOLOv3-tiny detector on AFs detection.

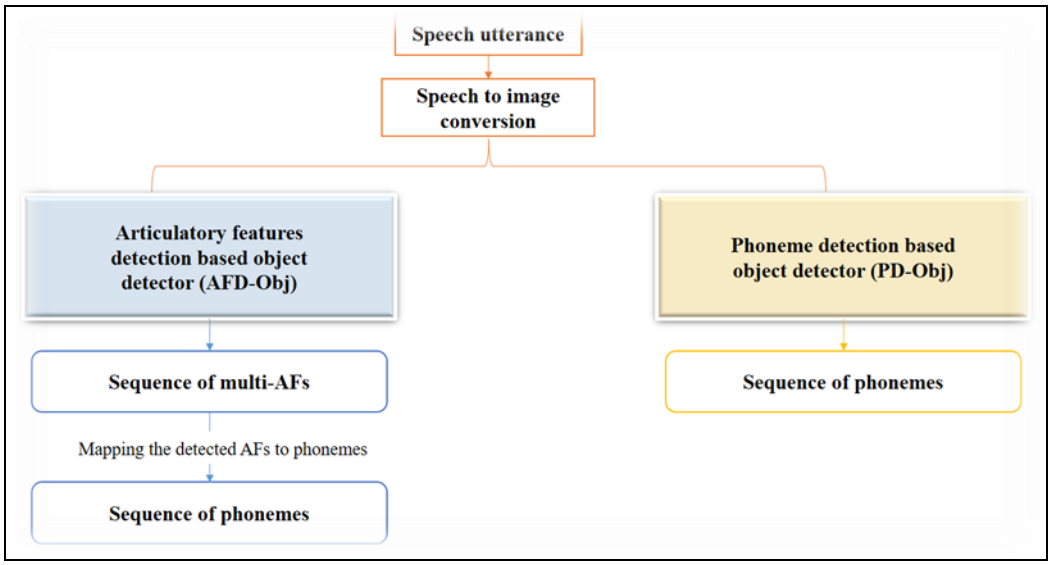
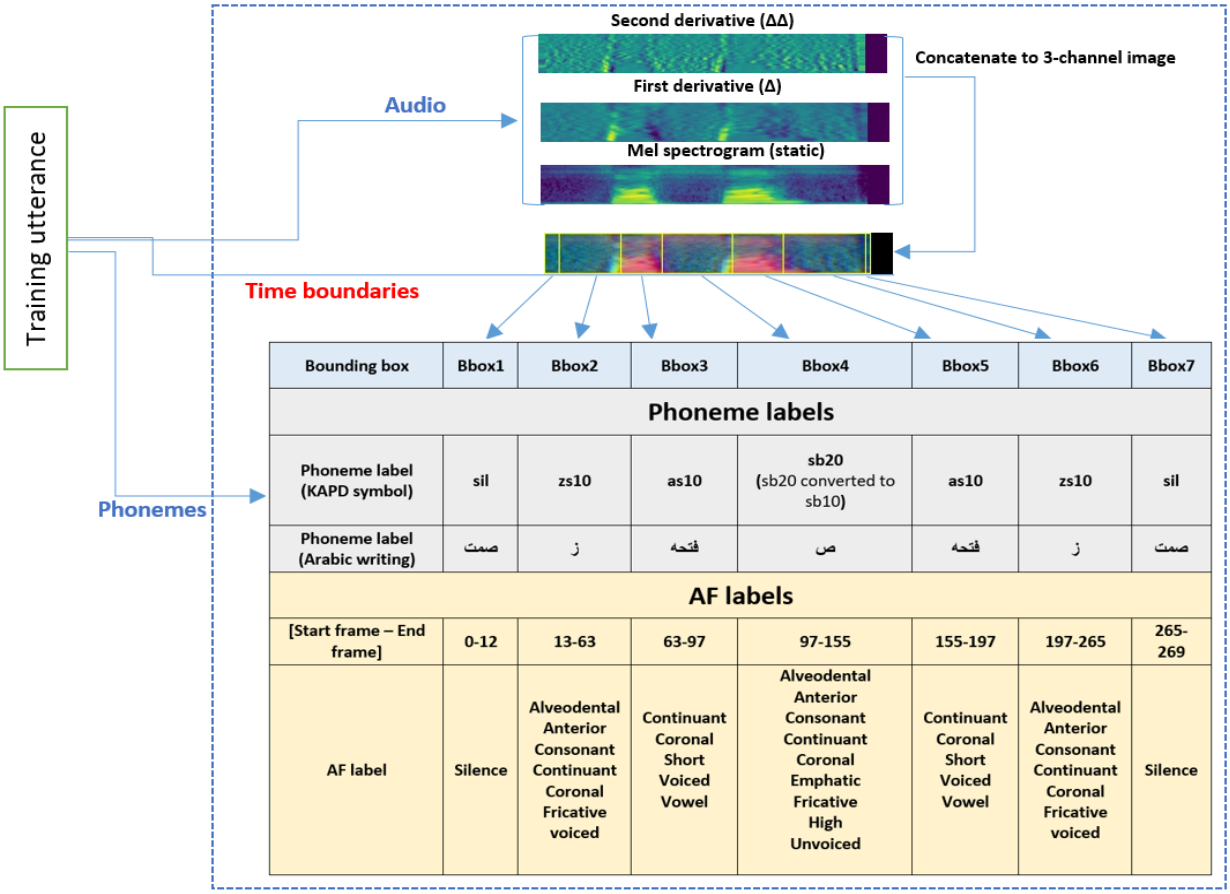


Figure 15, Proposed System of the AFD-Obj and the PD-Obj

5.2.1. Feature extraction

We used the speech-to-image transformation that we presented in detail in [59]. We concatenated the power Mel-spectrogram and the first and second derivatives to generate a three-channel image. Then, using the time boundaries calculated by MFA, we calculated the bounding box of each object. Figure 16 shows a detailed example of the process of creating spectral images with annotations. It shows the process of generating a spectral three-channel image from the speech and creating the associated bounding boxes for the utterance (GHSBGMA) from the KAPD training set [63].



Bounding box	Bbox1	Bbox2	Bbox3	Bbox4	Bbox5	Bbox6	Bbox7
Phoneme labels							
Phoneme label (KAPD symbol)	sil	zs10	as10	sb20 (sb20 converted to sb10)	as10	zs10	sil
Phoneme label (Arabic writing)	صمت	ز	فتحه	ص	فتحه	ز	صمت
AF labels							
[Start frame – End frame]	0-12	13-63	63-97	97-155	155-197	197-265	265-269
AF label	Silence	Alveodental Anterior Consonant Continuant Coronal Fricative voiced	Continuant Coronal Short Voiced Vowel	Alveodental Anterior Consonant Continuant Coronal Emphatic Fricative High Unvoiced	Continuant Coronal Short Voiced Vowel	Alveodental Anterior Consonant Continuant Coronal Fricative voiced	Silence

Figure 16, Example of converting the detected AFs to the corresponding phonemes.

5.2.2. Deep Learning Based AFs Detection System (AFD-Obj)

We selected the yolov3-tiny detector for this investigation because of its real time property and its support of multi-label detection. The real time property will allow our CAPT system to be used online on mobile devices. The detector consists of two main parts: backbone network and the detection layers. We started by training the backbone network of yolov3-tiny detector, which is darknet-reference, for the task of speech command classification using Google speech command corpus (V2) [65]. Then, we used the weights of the backbone network to initialize the weights of yolov3-tiny detector for the task of AFs detection. We trained the proposed system AFD-Obj for AFs detection in the Arabic corpus and for AFs detection in the

English corpus. For each task, we investigated different models of AFD-Obj system by changing the number detection's scale which are: YOLOv3-tiny-1S, YOLOv3-tiny-2S, and YOLOv3-tiny-3S for one scale, two scale, and three scale of detection, respectively. In the following sections, we will present the results of AFD-Obj for detecting AFs in Arabic and English corpora.

5.2.3. Results of AFD-Obj for Detecting AFs in Arabic Corpus

We used the KAPD corpus in this section to detect Arabic AFs and recognize Arabic phonemes from the detected AFs. KAPD was developed by King Abdul-Aziz City for Science and Technology at 2003. In our experiments, we used the latest version of KAPD corpus, which was developed by [64], where the total number of phonemes is 20283 for training and 8138 for testing. For mapping phoneme to AFs, we used the mapping table of [66]. The results of applying YOLOv3-tiny with its three different scales for the AFs detection task in KAPD is presented in Table 32.

Table 32, Performance metrics of the proposed system AFD-Obj for the Arabic AFs.

	YOLOv3-tiny-1S		YOLOv3-tiny-2S		YOLOv3-tiny-3S	
	GM	**F-measure**	**GM**	**F-measure**	**GM**	**F-measure**
affricative	0.929	0.927	0.931	0.929	0.931	0.929
alveodental	0.988	0.982	0.989	0.986	0.992	0.989
alveopalatal	0.938	0.936	0.927	0.925	0.945	0.943
anterior	0.980	0.982	0.985	0.986	0.989	0.990
aspirated	0.988	0.907	0.978	0.918	0.994	0.941
bilabial	0.954	0.876	0.930	0.868	0.940	0.908
consonant	0.998	0.998	0.997	0.997	0.999	0.998
continuant	0.992	0.993	0.994	0.994	0.993	0.994
coronal	0.977	0.975	0.980	0.978	0.984	0.983
emphatic	0.904	0.891	0.912	0.900	0.913	0.904
fricative	0.992	0.990	0.993	0.991	0.990	0.990
glottal	0.968	0.903	0.984	0.915	0.975	0.933
high	0.932	0.918	0.939	0.923	0.927	0.920
interdental	0.856	0.775	0.865	0.811	0.879	0.833
labiodental	0.795	0.721	0.838	0.776	0.803	0.729
labiovelar	1.000	0.967	0.988	0.953	0.988	0.966
lateral	0.960	0.922	0.960	0.897	0.969	0.873
nasal	0.979	0.963	0.951	0.929	0.978	0.973
palatal	0.978	0.967	0.978	0.977	0.967	0.945
pharyngeal	0.984	0.984	0.966	0.960	0.967	0.961
plosive	0.960	0.913	0.961	0.926	0.965	0.936
rounded	0.982	0.940	0.987	0.949	0.990	0.966
semivowel	0.989	0.967	0.994	0.983	0.989	0.972
short	0.997	0.995	0.999	0.998	0.999	0.998

silence	0.999	0.999	0.998	0.998	1.000	0.999
trill	0.955	0.933	0.954	0.932	0.919	0.874
unvoiced	0.985	0.964	0.983	0.972	0.982	0.971
uvular	0.960	0.917	0.953	0.932	0.938	0.926
velar	0.989	0.958	0.968	0.928	0.989	0.948
voiced	0.995	0.996	0.996	0.996	0.996	0.997
vowel	0.999	0.999	1.000	0.999	0.999	0.999
Average	**0.965**	**0.941**	**0.964**	**0.943**	**0.964**	**0.945**

For all AFs, the systems achieved a geometric mean (GM) greater than 80%, except for labiodental. For the F-measure of all AFs, the systems achieved accuracies greater than 80%, except for labiodental, which had an F-measure of 72.1%, 77.6%, and 72.9% using YOLOv3-tiny-1S, YOLOv3-tiny-2S, and YOLOv3-tiny-3S, respectively, and interdental, which had an F-measure of 77.5% using YOLOv3-tiny-1S. In general, we achieved GM and F-measure average accuracies of 96.5% and 94.1% for the YOLOv3-tiny-1S model, 96.4% and 94.3% for the YOLOv3-tiny-2S model, and 96.4% and 94.5% for the YOLOv3-tiny-3S model. These results are better than those of state-of-the-art results [66], where approximately 45% of the AFs obtained less than 80% for GM and approximately 61% obtained less than 80% for the F-measure using their best model (i.e., DBN–DNN). It is to be noted that, we achieved our results using a single network for all AFs, while Ref. [66] used a different network for each AF. Moreover, our testing input is a whole utterance without time boundary information, while their testing input was speech phonemes. We also detected the time boundaries of each AF; therefore, we can calculate the accuracy at the frame level.

5.2.3.1 Extraction of the Arabic Phonemes from the Detected AFs

Each phoneme has a unique vector representing the existences or absences of each AF; thus, we can detect the phonemes and their boundary from the AF vectors of each frame. We used the lookup table provided in [66] to produce the corresponding phoneme from the vectors of the detected AFs. The output of our proposed system is a sequence of AFs; hence, the length of the output sequence is may not be equal the length of the canonical form, hence we calculated the correction rate and the PER by applying sequence alignment between the vector of detected phonemes and the vector of the canonical phonemes. Figure 17 shows an example of recognizing the phonemes from the detected AF vectors and calculating the correction rate.

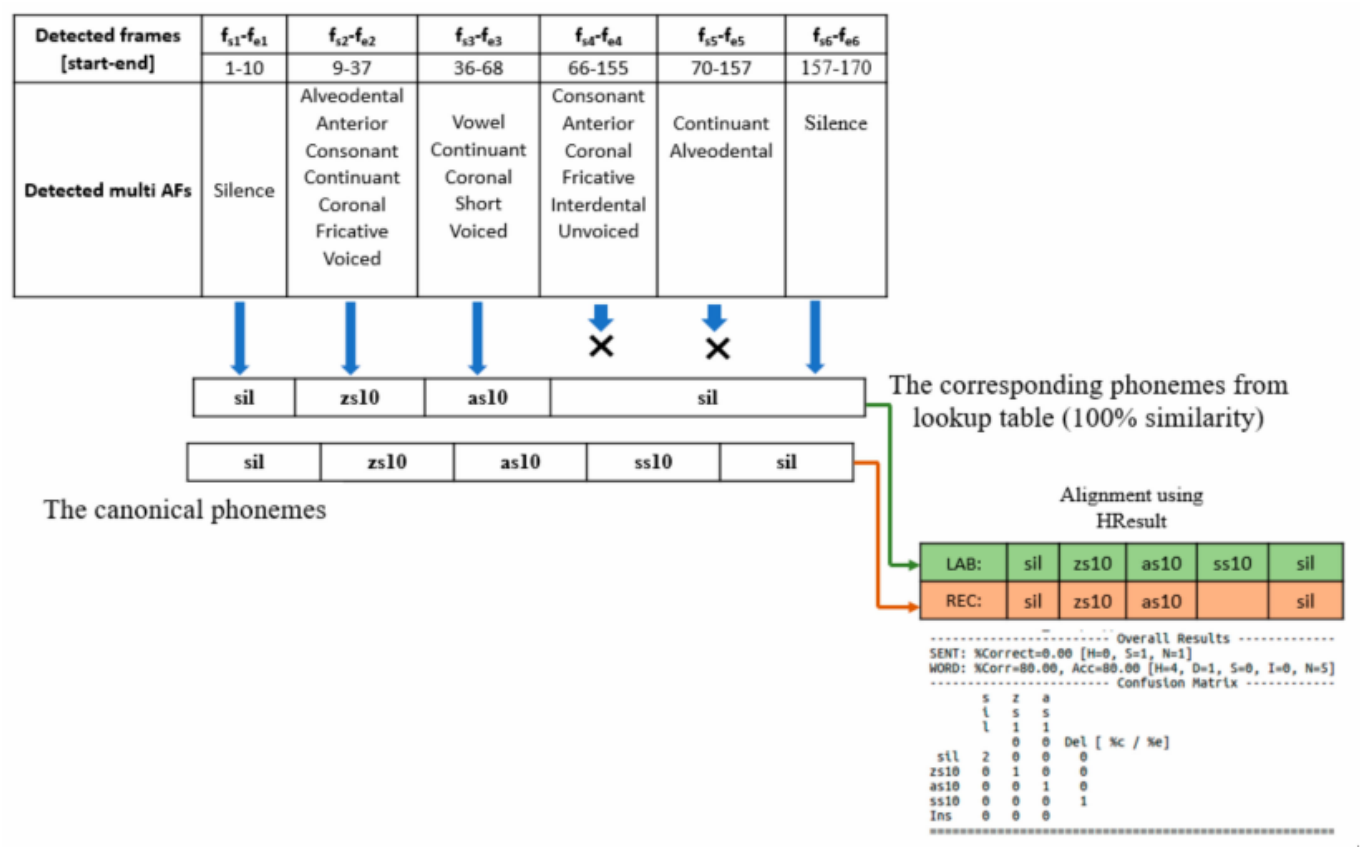


Figure 17, Testing example of converting the detected AFs using the YOLOv3-tiny-1S model to the corresponding phonemes and calculating the percentage of correct phonemes using the HResults tool (file "CMSSSFA") from the KAPD corpus test set. X sign means invalid output, which occurs when the minimum hamming distance is greater than threshold (threshold = zero in case of 100% similarity).

We considered the detected phonemes correct when there is an exact match of 100% similarity between the predicted AF vector and the reference vector for each phoneme. Ref. [66] reported their result for a 3-bit difference between the detected AFs and the lookup table, which amounted to approximately 90% similarity between the predicted and actual vectors. We compared the correction rate of our proposed method and that of [66] using the 100% and 90% similarities. For our best model (i.e., YOLOv3-tiny-1S), we outperformed the matching rate of their best classifier (i.e., DBN–DNN) by almost 40% at 100% similarity and by approximately 4% at 90% similarity [66]. Using 100% similarity, we achieved correction rates of 86.04%, 88.06%, and 89.35% for YOLOv3-tiny-3S, YOLOv3-tiny-2S, and YOLOv3-tiny-1S, respectively, compared to 64% correction rate (which they called matching rate) for the model in Ref. [66]. These values increased to 91.16%, 92.38%, and 92.59%, respectively, when using 90% similarity for all three models compared to 89% for that in Ref. [66]. This increase can be attributed to the fact that the correction rate measure ignored the insertion errors; hence, we ignored many insertion errors when using only 90% similarity.

For 100% similarity, our models obtained PERs of 14.13%, 12.09%, and 10.84%, respectively, which increased to 20.1%, 15.53%, and 12.57%, respectively, for 90% similarity. Ref. [66] did not provide the PER result. Another point to highlight is that these observations confirmed our postulation for not needing the second and third scales of the YOLO detector in the AF detection and phoneme recognition. The PER results also illustrate that using 90% similarity during AF matching to generate the corresponding phonemes is not acceptable because wrong phonemes can be recognized as correct, as shown in Table 33.

Table 33, PER (%) and correction rate (%) for our proposed AFD-Obj system and results of [66].

Matching rate (# bits)	Model	PER (%)	Correction rate (%)
100% (0 bit)	YOLOv3-tiny-3S	14.13	86.04
	YOLOv3-tiny-2S	12.09	88.06
	YOLOv3-tiny-1S	**10.84**	**89.35**
	DBN-DNN [66]	-	64.00 (Exact matching rate)
90% (3 bits)	YOLOv3-tiny-3S	20.1	91.16
	YOLOv3-tiny-2S	15.53	92.38
	YOLOv3-tiny-1S	**12.57**	**92.59**
	DBN-DNN [66]	-	89.00 (Matching rate)

5.2.4. Results of the AFD-Obj System for Detecting AFs in the English Corpus

This section presents the results of applying our proposed system for detecting the English AFs using the TIMIT corpus. To be able to compare the performance of our proposed system with other published research for detecting AFs in TIMIT, we used accuracy at the frame level, where we considered the bounding box coordinates as the start and end frames. Figure 18 shows an example of calculating the frame level accuracy from the detected output.

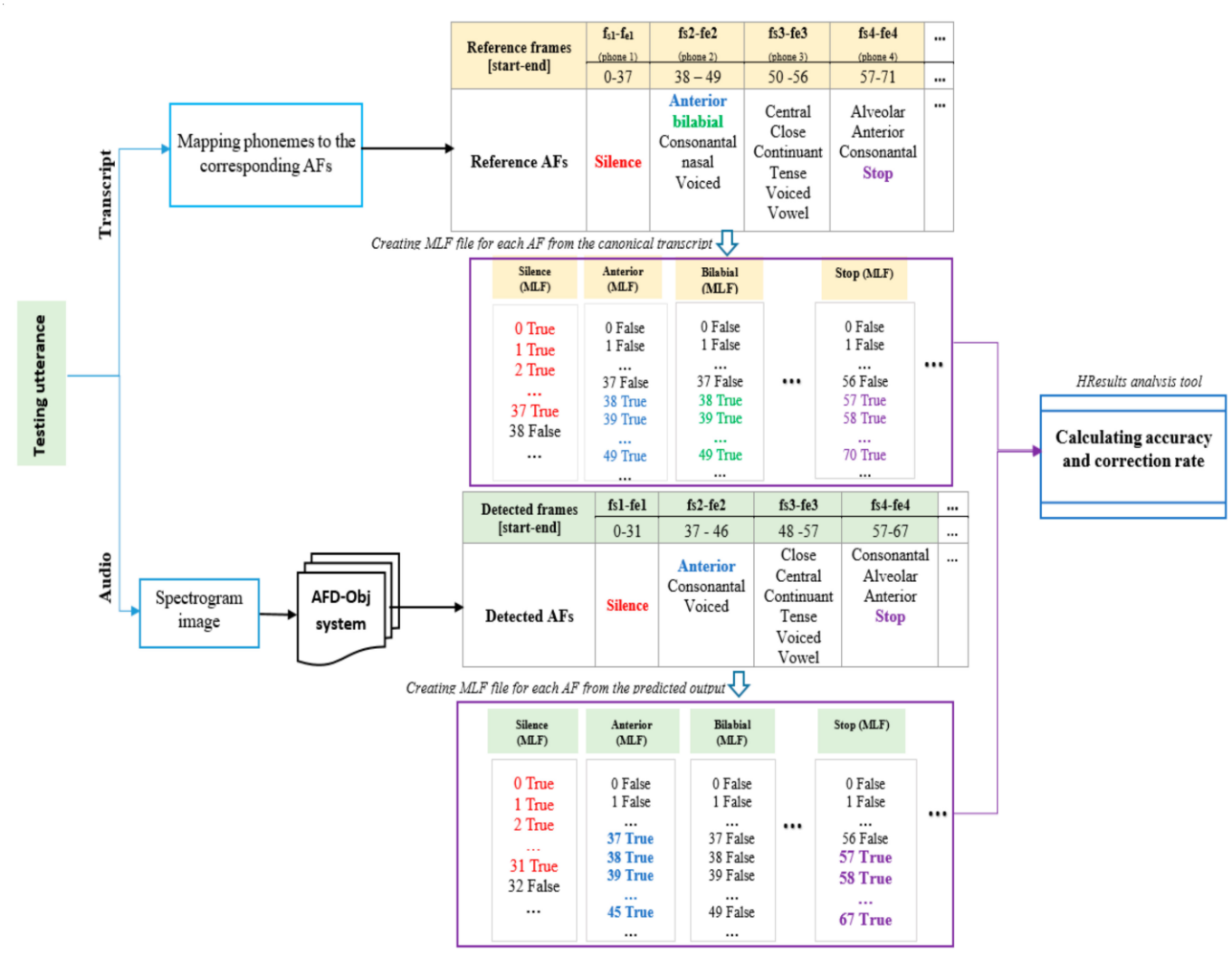


Figure 18, Testing phase of the AFD-Obj system: calculating the frame level accuracy of the detected outputs.

We compared the results of the proposed system with that of the state-of-art published work in AFs detection using TIMIT [67], called LAS-MTL-M. We considered for our comparison the results of LAS-MTL-M which were reported at frame level. Authors of [62] calculated the accuracy in two ways. They used the TIMIT segments markup (time boundaries) to calculate the

accuracies of the column "*markup frames*" and the DTW algorithm to convert soft attention to hard attention to calculate the accuracies of the column "*frames*". In both cases, they dealt with the different number of predicted and target frames by taking the minimum length of target and prediction, as shown in the code they provided. For better comparison, we calculated the accuracy of our proposed system using the coordinates of the detected bounding boxes as markup frames, after taking the minimum length of the predicted and target frames. This result is presented in the column "bounding box coord." of Table 34, which can be compared to the column "markup frames" in [67]. To compare with the column "frames" of [67], we used HResults analysis tool to align the predicted and target frames, then we calculate the accuracy. Moreover, HResults accuracy is more precise because it considers the insertion errors. This accuracy is presented in column "HResult align." of Table 34 . In Table 34, we also show the results of [67], which detected only some of the AFs in TIMIT.

Table 34 presents the result of our proposed system AFD-Obj with the three models. The table shows that our system achieved results comparable to the published result for all AFs. Our models had an average accuracy (with bounding box coord.) of 94.29%, 95.04%, and 95.13%, and average accuracy (using HResult) of 93.23%, 93.47%, and 93.66% for YOLOv3-tiny-3S, YOLOv3-tiny-2S, and YOLOv3-tiny-1S, respectively. In [67], phones-las-frames model had an average detection accuracy of 95.5% using markup-frames. Since the test results of the "markup-frames" of [67] depend on segmenting the speech into markup frames, while our system doesn't depend on any segmentation, hence we think fair comparison should be with the result of the "frames" column of [67].

Table 34, Detection accuracy of all 28 English AFs using the proposed system AFD-Obj and state-of-the-art methods.

Articulatory features	**AFD-Obj system**						**LAS-MTL-M markup-frames** [67]	**LAS-MTL-M frames** [67]	**KT** [68]
	YOLOv3-tiny-1S		**YOLOv3-tiny-2S**		**YOLOv3-tiny-3S**				
	Bounding box coord.	**HResult align.**	**Bounding box coord.**	**HResult align.**	**Bounding box coord.**	**HResult align.**			
Alveolar	91.05	90.22	90.92	90.01	89.31	88.96	95	77	
Anterior	89.69	89.34	89.55	89.02	87.92	88.08	90	69	90
Approximant	97.12	95.39	97.17	95.32	96.87	95.39	98	94	68
Bilabial	97.70	95.89	97.53	95.57	97.30	95.78	98	93	
Central	93.73	92.31	93.73	92.18	93.36	92.27	99	91	
Close	94.13	92.65	94.02	92.46	93.36	92.33	97	88	86
Consonantal	88.97	88.75	88.75	88.42	87.32	87.64	88	64	90

Continuant	91.37	90.46	90.88	90.04	88.60	88.38	89	68	86
Fricative	96.03	94.56	95.73	94.21	95.04	94.06	95	83	88
Front	93.33	91.96	93.42	91.92	92.06	91.12	95	89	84
Glottal	98.67	96.69	98.62	96.48	98.42	96.82	99	98	
labiodental	98.88	96.89	98.80	96.71	98.57	96.94	99	96	
Lateral approximant	98.21	96.34	98.11	96.07	97.88	96.31	99	96	
Mid	90.28	89.09	90.04	88.77	88.87	88.3	97	82	
Nasal	97.59	95.95	97.55	95.72	97.15	95.74	99	93	84
Non sibilant fricative	97.60	95.8	97.50	95.58	97.22	95.74	97	94	
Open	96.09	94.31	95.83	93.98	95.63	94.23	98	91	93
palatal	99.60	97.54	99.63	97.4	99.57	97.81	99	99	
postalveolar	99.18	97.12	99.17	96.94	98.96	97.21	99	97	
Round	94.99	93.36	94.70	92.97	94.30	93.04	98	91	92
Sibilant affricate	99.50	97.41	99.51	97.29	99.38	97.64	99	99	
Sibilant fricative	97.97	96.1	97.81	95.95	97.37	96	98	90	
Silence	96.79	95.21	97.05	95.29	96.68	95.35	80	63	89
Stop	95.03	93.74	95.05	93.61	94.46	93.53	97	85	96
Tense	89.63	88.46	89.92	88.73	88.65	87.94	97	81	87
Velar	98.37	96.47	98.31	96.25	98.01	96.37	99	95	
Voiced	90.86	89.9	90.71	89.69	88.62	88.19	84	72	93
vowel	91.31	90.69	91.19	90.57	89.29	89.26	92	70	92
Average	**95.13**	**93.66**	**95.04**	**93.47**	**94.29**	**93.23**	**95.5**	**86**	

An important observation from Table 34 is that our models detected silence within the utterance with a high accuracy compared to [67], which achieved only 63% and 80% for frames and markup frames, respectively. This high performance in detecting silence in continuous speech is very promising and can be looked at as an important achievement by itself. Our three models, YOLOv3-tiny-1S, YOLOv3-tiny-2S, and YOLOv3-tiny-3S had an average accuracy detection of silence equal to 95.13%, 95.04%, and 94.29%, respectively. These results reinforce our previous assumption for not needing the second and third scales of YOLO detection for our specific application.

5.2.5. Results of PD-Obj for Detecting phonemes in Arabic Corpus

Table 35 presents the results of investigating the proposed PD-Obj system for phoneme recognition using the KAPD corpus. Our proposed system using the YOLOv3-tiny-2S model achieved the lowest PER of 5.63%, while the YOLOv3-tiny-1S and YOLOv3-tiny-3S models achieved 5.79% and 6.29% PER, respectively. These results are remarkable and show that our proposed system has an excellent potential compared to the recent state-of-the-art systems on the whole KAPD corpus [69]. Ref. [69] used the HMM for the Arabic phoneme recognition using the DPF elements. The results also reinforced our previous assumption for not needing the second and third scales of the YOLO detection for our specific application.

Table 35, PER and correction rate of the Arabic phoneme recognition using the proposed models.

Model	PER (%)	Correction rate (%)
PD-Obj (YOLOv3-tiny-3S)	6.29	93.94
PD-Obj (YOLOv3-tiny-2S)	5.63	94.56
PD-Obj (YOLOv3-tiny-1S)	5.79	94.34
AFD-Obj (YOLOv3-tiny-1S)	10.85	89.33
PDF-HMM [69]	39.57	70.68

We observe from Table 35 that the PD-Obj system obtained better results than the system based on the AFD-Obj. However, detecting the AFs is important in building a versatile CAPT system because it allow the CAPT system to provide pronunciation error correction and diagnosis. Moreover, the AFs are universal between many languages. An interesting point for future work is to see how to improve the accuracy of the system that performs phoneme recognition based on the detected AFs.

We also calculated the correction rate of each phoneme for the YOLOv3-tiny-1S model. We found that 79% of the Arabic phonemes had a correction rate greater than 80%, while 44% had a correction rate greater than 90%.

6. Design and development of the CAPT system for L2 learners of the Arabic language.

In this section, we present the complete Arabic CAPT system and its results. We start by presenting proposed models for MDD task in the CAPT system, then the obtained results using the Arabic-CAPT speech database is presented, and then we will present the findings using the developed database KSU-CAPT. Note that the results of using Arabic-CAPT database in this section have been published in ISI journal [70], while the result of using KSU-CAPT databases will be published later.

6.1. Proposed methods for CAPT

Building the project database was anticipated take a long time, and in fact it took even longer than anticipated as explained in chapter 4. Hence to be able to do our investigation in new techniques for building high performance Arabic CAPT system, we built Arabic-CAPT and Arabic-CAPT-S databases as explained in details in section 4.1. We used Arabic-CAPT and Arabic-CAPT-S to continue our investigation of using object detectors for building CAPT systems. In section 5.1 and 5.2, we investigated using object detection for phoneme and AFs detection, while in this section we will investigate using object detection for building high performance MDD system that recognize phonemes and AFs, which is the core of the CAPT system.

We investigated three MDD systems. The first is a proposed MDD system based on object detection and we called it MDD-Object. The second is state-of-the-art end-to-end MDD system that was used in languages other than Arabic but was not used before for Arabic language, and we called it MDD-E2E. The third system is a fusion between MDD-Object and MDD-E2E. In the following, we will present the details of these three systems and we present their results in section 6.2. More details of the proposed systems and results can be found in [70].

6.1.1. MDD-Object

In MDD-Object, we propose object detection technique to build a high-performance versatile CAPT system for MDD and articulatory feedback generation for non-native Arabic

learners. The proposed system can locate the error in pronunciation, recognize the mispronounced phonemes, and detect the corresponding articulatory features (AFs), not only in words but even in sentences.

We formulate the recognition of phonemes and corresponding AFs as a multi-label object recognition problem, where the objects are the phonemes and their AFs in a spectral image. The proposed MDD-Object system consist of the following steps.

Table 36: Mapping phonemes to their corresponding AFs.

		AFs Categories				
		Place	**Manner**	**Manner–Voice**	**Manner–Emphatic**	**Vowel**
Phonemes	/f/	Labio-dental	Fricative	Unvoiced	Non-emphatic	-
	/a/	-	-	-	-	Vowel
	/H/	Pharyngeal	Fricative	Unvoiced	Non-emphatic	-
	/S/	Alveo-dental	Fricative	Unvoiced	Emphatic-fricative	-

6.1.2. Spectral Image Generation and Annotation

We propose addressing the phonemes and the associated AFs of utterances as objects in spectral images and applying object detection techniques to detect them. To accomplish this, we converted the speech utterances into three-channel spectral images to fulfill the requirements of object detectors, as presented in more detail in section 5.2. For each phoneme, we extracted the corresponding AF labels based on the mapping of Table 36. For example, the Arabic word "فَحْصٌ" consists of the following phonemes (/f/, /a/, /H/, /S/), and each phoneme is represented by multi-label AFs, as presented in Table 36. Then, using the time boundary of each phoneme, we calculated the bounding box coordinates of the phoneme and its corresponding AFs to create the annotation file of each utterance.

6.1.3. Multi-Label Object Detector: Selection, Optimization, and Training

The selection of the object detector for our proposed system was based on two conditions: the system should operate in real-time and should be able to detect multi-label objects. Based on our observations in section 5.2, we selected the tiny version of the third YOLO (YOLOv3-tiny) as the detector of our proposed system, since it best satisfied the above two conditions. We trained different models from the YOLOv3-tiny with and without different types of transfer learning, as presented in the following.

We constructed an MDD system based on the YOLOv3-tiny, with default parameters, as a baseline, and we called it the MDD-Object model. We trained the MDD-Object model using only real non-native speech without any initial training. In the second model, we fine-tuned the YOLOv3-tiny parameters using genetic algorithms and we called the system MDD-Object-G. We trained the model from scratch without any initial training using the developed Arabic-CAPT corpus. Then, we trained various versions of it by initially training the detector using native speech, synthesized speech, and a combination of native and synthesized speech, and we called these versions MDD-Object-G/N, MDD-Object-G/S, and MDD-Object-G/NS, respectively, as shown in Table 37. In the MDD-Object-G/N model, we initially trained using native speech, and then we transferred the weights by freezing the first four layers of the backbone network of the detector and fine-tune the remaining layers using the real non-native Arabic-CAPT corpus. In MDD-Object-G/S, we trained the model using synthesized speech from the Arabic-CAPT-S corpus, which is approximately four times the size of the original Arabic-CAPT corpus, as shown in Table 38, then, we transferred the weights of the detector and fine-tuned them using the real Arabic-CAPT corpus. Finally, in MDD-Object-G/NS, we initially trained the model using the combination of native speech and synthesized speech. Then, we also transferred the weights of the detector and fine-tuned them using the real Arabic-CAPT corpus.

Table 37: Speech type for the proposed MDD-Object models. (N: Native, S: Synthesized, and NN: Non-Native)

	Initial Training Phase		Fine Tuning Phase		
	Training Set	**Validation Set**	**Training Set**	**Validation Set**	**Testing Set**
MDD-object (baseline)	-	-	NN	NN	NN
MDD-object-G	-	-			
MDD-object-G/N	N	N			
MDD-object-G/S	S	S			
MDD-object-G/NS	N+S	N+S			
MDD-object-G-Large/NS	N+S	N+S			

Table 38 presents the statistics of non-native, synthesized, and native corpora that were used for the training, validation, and testing of the proposed models. For the Arabic-CAPT corpus, we used 67% of speakers for training, 8% for validation, and the remaining 25% for testing. All speakers in training, validation, and testing were distinct. We wanted to map the diversity of the nationalities in the database in the test set; hence, the 15 speakers of the test set belonged to 12 different nationalities. In Arabic-CAPT-S, we used the synthesized speech of the speakers of the training and validation sets of Arabic-CAPT, as shown in Table 38. In native speech, we used all 146 Saudi speakers in session 1 from the KSU speech database and divided them into 132 for training and 14 for validation.

Table 38: The KSU speech corpora used in the training and testing phases

	Arabic-CAPT			**Arabic-CAPT-S**		**Native Speech**	
Set	**Train**	**Valid**	**Test**	**Train**	**Valid**	**Train**	**Valid**
Speakers	42	5	15	42	5	132	14
Utterances	1091	130	390	4914	585	3426	361
Duration (Hours)	1.65	0.16	0.54	4.79	0.6	4.11	0.43

6.1.4. Decoder phase of the MDD-Object model

The output of the MDD-Object model is a sequence of bounding boxes, where each box has a multi-label (i.e., phonemes and AFs), and some of these boxes overlap and lead to duplicate phonemes and AFs. To deal with this, we propose using decoding techniques that are used in sequence-to-sequence acoustic models, such as CTC. These decoding techniques need a matrix of probabilities for the phonemes of the frames, and so to accomplish this, we started by converting the confidence scores matrix to a matrix of probabilities by applying softmax so that the sum of the probabilities in each frame is equal to one. In this study, we investigated the fast and simplest decoding technique, which is greedy search decoding.

6.1.5. Scoring Metrics Evaluation

In scoring phase, we aligned the detected phonemes of the proposed system, the canonical phoneme, and the annotated phoneme of the human experts to calculate the system performance and evaluation metrics. We calculated the performance metrics of the proposed models for the non-native phoneme recognition task and MDD task as follows:

For the phoneme recognition task, we used a well-known metric in this field, PER, and calculated it as in Equation (1) [71].

$$\mathrm{PER} = 100 - \frac{\mathrm{N} - \mathrm{S} - \mathrm{D} - \mathrm{I}}{\mathrm{N}} \tag{1}$$

where N, S, D, and I are the number of samples, substitution errors, deletion errors, and insertion errors, respectively. Once we detected any mispronounced phonemes, these phonemes were sent to the mispronunciation correction and feedback module to provide corrective feedback.

For the MDD task, we used a hierarchical evaluation that was presented in [72] and used in many MDD studies such as [49] [73], where the phonemes of the sequence annotated by the experts were classified as correct or mispronounced. By comparing the output of our proposed models with the annotations of the experts, we obtained four cases: true acceptance (TA), false acceptance (FA), false rejection (FR), and true rejection (TR). TR was divided into correct diagnosis (CD) and diagnosis error (DE). CD corresponded to the annotator agreeing with system in the recognized mispronounced phoneme, while DE corresponded to the annotator

disagreeing with system in the recognized mispronounced phoneme. We calculated the performance metrics of the MDD task using the following equations:

$$\text{Precision} = \frac{\text{TR}}{\text{TR} + \text{FR}} \tag{2}$$

$$\text{Recall} = \frac{\text{TR}}{\text{TR} + \text{FA}} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{FAR} = 1 - \text{Recall} \tag{5}$$

$$\text{FRR} = \frac{\text{FR}}{\text{TA} + \text{FR}} \tag{6}$$

$$\text{Diagnosis Accuracy (DA)} = \frac{\text{CD}}{\text{CD} + \text{DE}} \tag{7}$$

6.1.6. Mispronunciation Correction and Feedback

The goal of this phase is to correct the mispronunciation of the learner and provide feedback at the articulatory level. The mispronounced phonemes were detected in the previous step and the purpose of the current step is to compare the detected AFs of the mispronounced phoneme with the AFs of the canonical phoneme. From this comparison, we provided corrective feedback to the learner at the articulatory level. As shown in the example in Figure 19, the canonical phoneme /S/ has the following AF classes, fricative, unvoiced, emphatic-fricative, and alveo-dental, and the detected phoneme /s/ has the detected AF classes, fricative, unvoiced, non-emphatic, and alveo-dental. Hence, the learner had a problem with an emphatic class, while he did not have problem with other classes. Therefore, textual feedback can be provided and a 2D/3D animation of the correct pronunciation of phoneme /S/ can be shown to the learners.

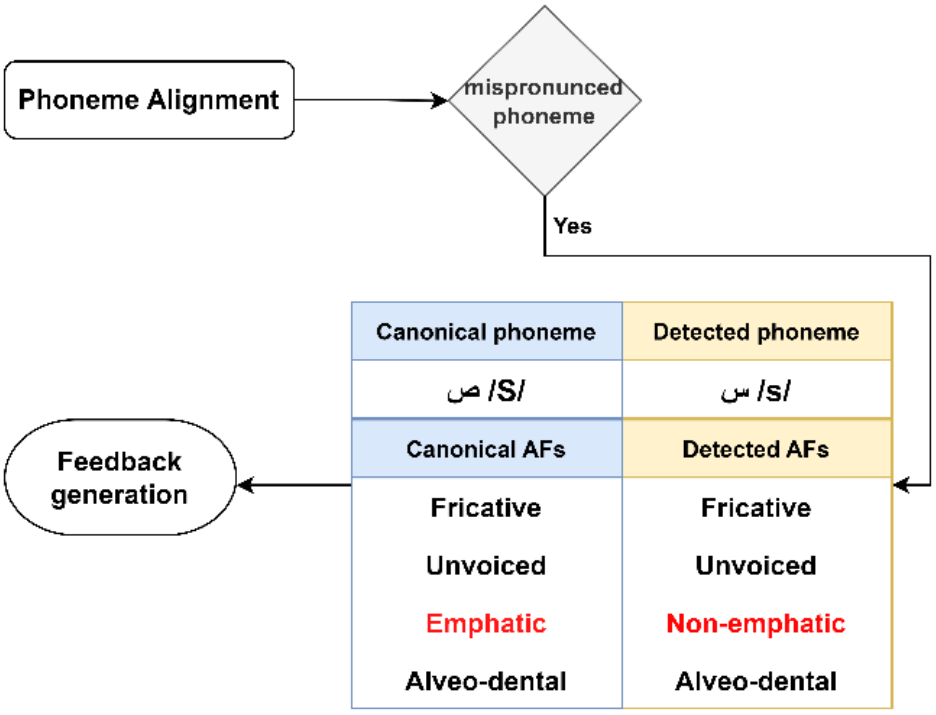


Figure 19: An example of feedback generation.

6.1.7. MDD-E2E

Many of the researchers in end-to-end MDD methods [35] [74]–[76], used CNN-RNN-CTC [77]; hence, we selected it as the end-to-end method to investigate and compare with our MDD-Object. The original CNN-RNN-CTC was proposed for the end-to-end MDD task for Chinese learners of English using the CU-CHLOE corpus. We used the implementation provided by [78] to train and evaluate the CNN-RNN-CTC using our developed corpora. To the best of our knowledge, this is the first investigation of applying end-to-end deep learning techniques in an MDD system for non-native Arabic speech. Furthermore, we enhanced the performance of the CNN-RNN-CTC model by using different combinations of native, synthesized, and non-native speech. We trained the baseline model CNN-RNN-CTC using only the non-native Arabic-CAPT corpus, while we trained other models, namely CNN-RNN-CTC/N, CNN-RNN-CTC/S, and CNN-RNN-CTC/NS, using combinations of the data, as shown in Table 39.

Table 39: Speech type for the proposed MDD-E2E models, N: Native, S: Synthesized, NN: Non Native.

	Training Set	Validation Set	Testing Set
CNN-RNN-CTC	NN	NN	NN
CNN-RNN-CTC/N	N + NN	NN	NN
CNN-RNN-CTC/S	S + NN	NN	NN
CNN-RNN-CTC/NS	N + S + NN	NN	NN

6.1.8. Fusion

Our MDD-Object system and MDD-E2E system have different structures and are based on different principles, and hence to benefit from the performance of the two systems, we propose applying the fusion technique between the MDD-Object system and MDD-E2E system. We used the decision level fusion between the best model of the MDD-Object system, which is MDD-Object-G-Large/NS, and the best model of the E2E-MDD system, which is CNN-RNN-CTC/NS. To achieve this, we calculated the weighted average of the output probabilities of the two models before decoding the final output (i.e., sequence of phonemes).

To make the output of the two models more consistent in the number of frames in the output, we set the frame length and number of Mels of the CNN-RNN-CTC system to the same values used in the MDD-Object system. This change in the CNN-RNN-CTC system made it necessary to change the size of the input RNN layer. These changes improved the performance of the CNN-RNN-CTC system, as will be shown later.

Another point that we addressed to achieve fusion between the two systems is that the CNN-RNN-CTC model reduces the number of frames by a factor of 4. We dealt with this by down sampling the output of the MDD-Object by a factor of 4. Next, we applied a weighted average to the probabilities of the two models as in Equation (8).

$$Probs_{AVG} = \alpha \times Probs_{E2E} + (1 - \alpha) \times Probs_{YOLO} \qquad (8)$$

where $Probs$ is the log softmax array output with the dimensions number of frames × number of phonemes.

6.2. Results using the Arabic-CAPT Database

6.2.1. Results of MDD-Object

For our proposed MDD-Object system, the PER of the baseline MDD-Object model was 7.56%, as shown in Table 40. This value improved to 4.93% when we used the GA to select the optimal parameters of the model in MDD-Object-G. When we used native and synthesized speech for the initial training of the model, the PERs improved to 4.75% and 4.71% using MDD-Object-G/N and MDD-Object-G/S, respectively. The result of the MDD-Object-G/S confirmed our anticipation of the benefit of using synthesized speech in transfer learning to deal with the scarcity of non-native speech. When we used the combination of native and synthesized speech for the initial training of MDD-Object-G/NS, we achieved PERs of 4.54% resulting in relative improvements of 40% compared to the baseline. Though using the native and synthesized speech did not offer a better result than MDD-Object-G/N, it resulted in a better performance in the MDD metrics. The best performance of our proposed system was achieved by MDD-Object-G-Large/NS, which achieved 4.05%, resulting in relative improvements of 46.42% compared to the baseline.

For the MDD metrics, we can see that the results for our baseline model MDD-Object for FAR, FRR, and P were 29.33%, 5.26%, and 40.29%, respectively. These results improved to 24.16%, 2.04%, and 65.06%, respectively, using our best model MDD-Object-G-Large/NS. Our baseline model obtained 76.99% diagnostic accuracy (DA), while our best model MDD-Object-G-Large/NS achieved a DA of 84.97%. Our best model MDD-Object-G-Large/NS obtained a 70.04% F1-score with relative improvement of 36.47% compared to our baseline model which achieved an F1-score of 51.32%.

Table 40: MDD results and PER of the proposed MDD-object models.

System	Model	TA (%)	FAR (%)	FRR (%)	DA (%)	DER (%)	P (%)	R (%)	F1 (%)	PER
MDD-Object	MDD-object (baseline)	94.74	29.33	5.26	76.99	23.01	40.29	70.67	51.32	7.56

	MDD-object-G	97.55	29.79	2.45	82.68	17.32	59.0	70.21	64.12	4.93
	MDD-object-G/N	97.99	32.83	2.01	81.22	18.78	62.70	67.17	64.86	4.75
	MDD-object-G/S	97.63	30.55	2.37	81.62	18.38	59.51	69.45	64.10	4.71
	MDD-object-G/NS	97.84	28.88	2.16	84.19	15.81	62.32	71.12	66.43	4.54
	MDD-object-G-Large/NS	97.96	24.16	2.04	84.97	15.03	65.06	75.84	70.04	4.05

As we mentioned earlier that our proposed system MDD-Object is trained for multi-labels: phoneme and AF detection, so in the following we present the performance of the proposed model for the AF detection task. As a performance metric, we used the detection error rate (DER), which was used in some previous works, such as [78]. Figure 20 shows the DER of the five AF categories using our six proposed models.

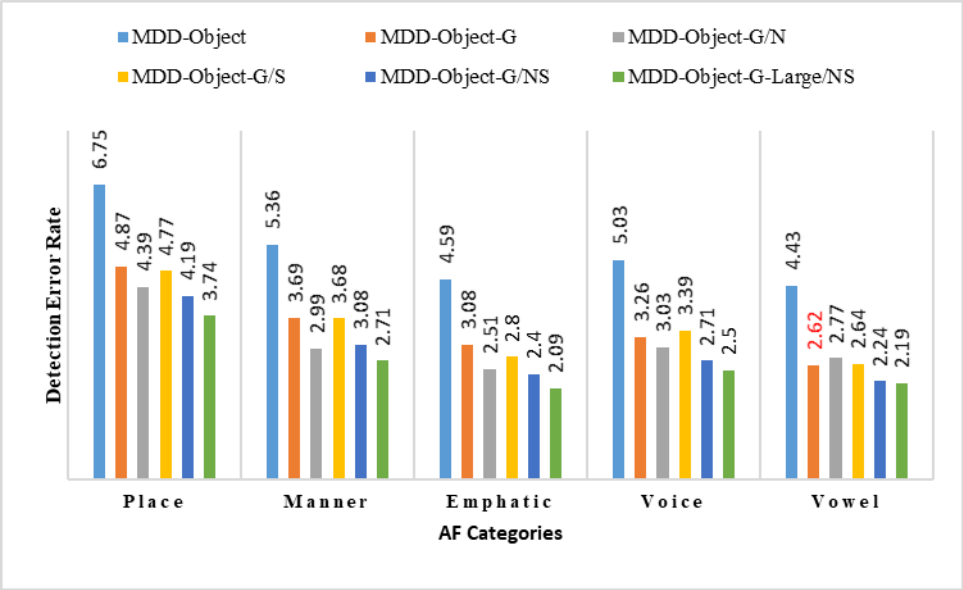


Figure 20:Detection error rate (DER) for each AF category using our proposed models.

We can see from Figure 20 that our best model MDD-object-G-Large/NS achieved DERs of 3.47, 2.71, 2.09, 2.5, and 2.19 for all five categories: place, manner, emphatic, voice, and vowel, respectively. Similar to the phoneme recognition task, we can clearly notice the usefulness of

fine-tuning the hyper-parameters and using synthesized speech to enhance the performance of the proposed system.

Figure 21 presents the confusion matrix of each of the five AF categories using our best model, MDD-object-G-Large/NS. These confusion matrices show the ability of the proposed system to distinguish between the different classes within the categories with little confusion. In the matrices, we can see that the confusion is between near classes. For example, for the place category, the highest confusion occurs in the (glottal vs. pharyngeal) and (alveo-dental vs. interdental) classes, and these two pairs of places are near to each other. Another example is that in the manner category, the highest confusion occurs between stop and fricative classes, which are near to each other, and this is consistent with what was shown on English AF detection in [68]. From these interesting results, we can see that our proposed system gives us the capability to teach non-native Arabic learners the correct pronunciation by providing the ability of suitable feedback at the articulatory level.

Place category	Bilabial	Labio-dental	Inter-dental	Alveo-dental	Palatal	Alveo-palatal	Velar	Uvelar	Pharyngeal	Glottal	Del
Bilabial	**1324**	0	0	4	0	0	0	0	0	2	12
Labio-dental	1	**427**	0	2	0	0	0	1	0	1	2
Inter-dental	0	0	**321**	9	0	0	0	2	1	0	2
Alveo-dental	8	4	16	**3231**	0	4	0	0	0	1	33
Palatal	0	0	0	1	**179**	0	0	0	0	0	3
Alveo-palatal	0	0	0	1	0	**265**	0	0	0	0	0
Velar	0	0	0	0	0	0	**194**	2	0	0	5
Uvular	0	0	0	1	0	1	2	**302**	3	2	1
Pharyngeal	1	0	0	0	0	0	0	4	**555**	10	9
Glottal	3	0	0	3	1	0	0	1	17	**1044**	29
Ins	6	3	11	34	1	2	1	9	8	21	

Manner category	Stop	Fricative	Affricate	Nasal	Laterals	Trill	Del
Stop	**1553**	13	2	0	0	0	19
Fricative	13	**2909**	2	3	0	1	38
Affrictae	0	0	**148**	0	0	0	0
Nasal	1	2	0	**1524**	4	0	26
Laterals	2	1	0	3	**823**	0	7
Trill	0	0	0	1	1	**464**	3
Ins	20	23	0	12	6	2	

Emphatic category	Emphatic	Emphatic-fri	Non-emphati	Del
Emphatic-stop	**103**	0	2	2
Emphatic-fricative	1	**171**	3	6
Non-emphatic	3	14	**7659**	83
Ins	3	8	43	

Voice category	Voiced	Unvoiced	Del
Voiced	**4792**	15	54
Unvoiced	16	**3126**	44
Ins	37	35	

Vowel category	Vowel	Semi-vowel	Del
Vowel	**5472**	3	68
Semi-vowel	2	**470**	12
Ins	45	2	

Figure 21: Confusion matrix of each of the AF categories using our best model MDD-object-G-Large/NS.

6.2.2. MDD-E2E Results

In this section, we will present the results of MDD-E2E different models using Arabic-CAPT database.

Table 41: MDD results and PER of the proposed MDD-E2E models

System	Model	TA (%)	FAR (%)	FRR (%)	DA (%)	DER (%)	P (%)	R (%)	F1 (%)	PER

MDD-E2E	CNN-RNN-CTC (baseline)	94.14	20.67	5.86	75.10	24.9	40.47	79.33	53.59	8.93
	CNN-RNN-CTC/N	97.62	24.32	2.38	82.53	17.47	61.48	75.68	67.85	5.17
	CNN-RNN-CTC/S	96.44	23.40	3.56	82.74	17.26	51.91	76.60	61.88	6.17
	CNN-RNN-CTC/NS	98.09	24.92	1.91	84.01	15.99	66.31	75.08	70.42	4.59

Table 41 shows the results of the MDD-E2E system for phoneme recognition task and MDD task. For phoneme recognition task, the baseline model CNN-RNN-CTC achieved 8.93% PER. This value decreased to 5.17%, resulting in a relative improvement of 42.1% when we added native speech for training the CNN-RNN-CTC/N model. The same observation occurred when we added synthesized speech for training the CNN-RNN-CTC/S model, where we achieved a PER of 6.17%, resulting in a relative improvement of 30.9% compared to the baseline. This improvement confirms our anticipation of the usefulness of synthesized speech for solving the scarcity of non-native speech. When we added native speech and synthesized speech for training the CNN-RNN-CTC/NS model, we achieved a 4.59% PER, which was a relative improvement of 48.6% compared to the baseline.

For MDD metrics, we can see from Table 41 that the CNN-RNN-CTC/NS model achieved the best result among the CNN-RNN-CTC models in FAR, FRR, P, and F1 with values of 24.92%, 1.91%, 66.31%, and 70.42%, respectively. The model also achieved the best result in the mispronunciation diagnosis, where the DA was 84.01%.

6.2.3. Decision Level Fusion between MDD-Object and MDD-E2E

Table 42 shows the results of the fusion model. The system with fusion between the CNN-RNN-CTC/NS and MDD-Object-G-Large/NS models achieved the best PER of 3.83% amounting to 16.5% relative improvement compared to the performance of CNN-RNN-CTC/NS and 5.4% relative improvement compared to the performance of MDD-Object-G-Large/NS. This result demonstrates the benefit of applying the decision-level fusion. Moreover, we noticed that our fusion system achieved the best performance in all MDD metrics, which indicates the usefulness of applying fusion.

Table 42: MDD results and PER of the fusion model.

System	Model	TA (%)	FAR (%)	FRR (%)	DA (%)	DER (%)	P (%)	R (%)	F1 (%)	PER
Fusion	YOLO-CNN-RNN-CTC	98.12	25.08	1.88	85.19	14.81	66.62	74.92	70.53	3.83

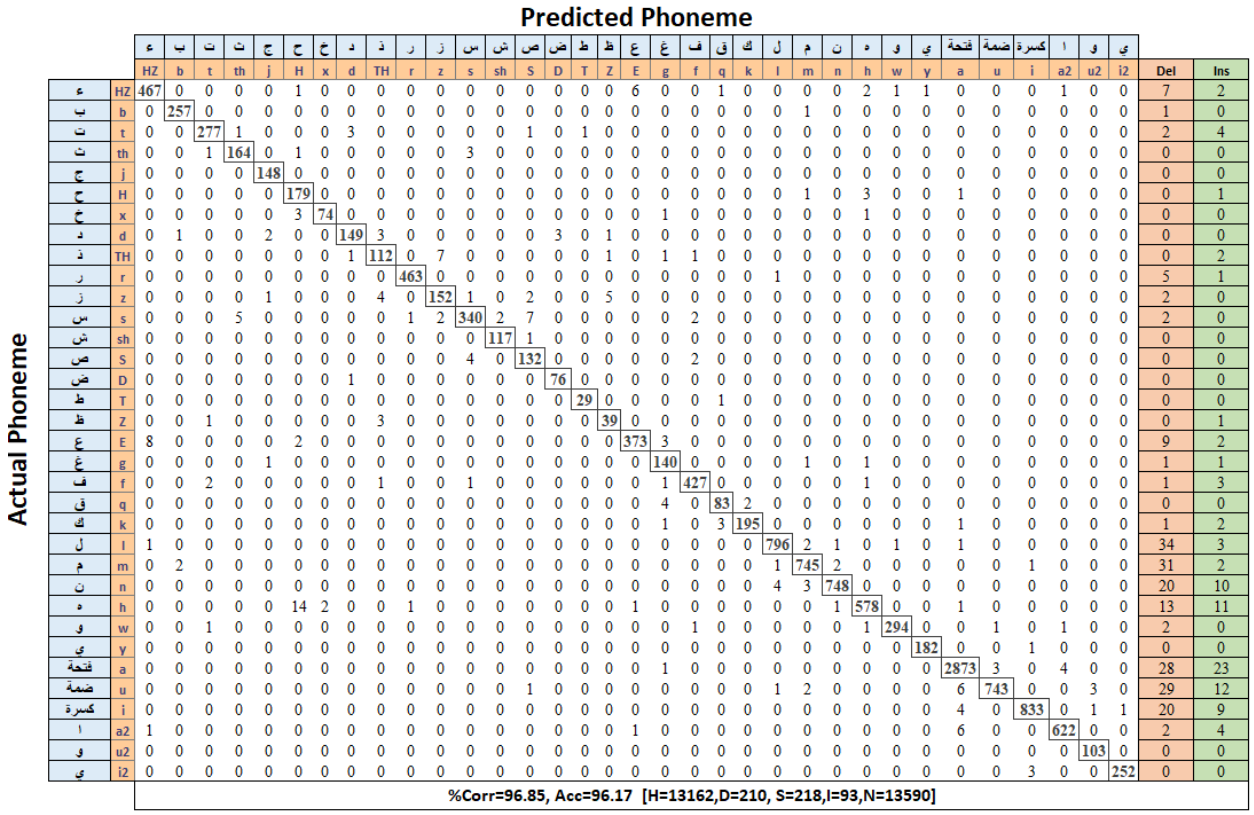
For a better analysis of the performance for each phoneme, we present in Figure 22 and Figure 23 the confusion matrices of our best model MDD-object-G-Large/NS and the fusion model YOLO-CNN-RNN-CTC respectively, using the test set of the Arabic-CAPT corpus. The total number of phonemes in the test set, without leading and trailing silence, is 13,590 phonemes belonging to 390 utterances.

Predicted Phoneme

Actual Phoneme	HZ	b	t	th	j	H	x	d	TH	r	z	s	sh	S	D	T	Z	E	g	f	q	k	l	m	n	h	w	y	a	u	i	a2	u2	i2	Del	Ins	%Corr/phoneme
ء	HZ	465	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	2	1	1	0	0	0	1	0	0	10	3	97.5
ب	b	0	255	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3	0	99.6
ت	t	0	0	277	1	0	0	0	3	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	97.9
ث	th	0	0	1	162	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	97
ج	j	0	0	0	0	148	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
ح	H	0	0	0	0	0	179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0	0	0	0	1	2	97.8
خ	x	0	0	0	0	0	3	74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	94.9
د	d	0	1	0	0	2	0	0	148	4	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	93.7
ذ	TH	0	0	0	0	0	0	0	1	112	0	7	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	91.8
ر	r	0	0	0	0	0	0	0	0	0	463	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	100
ز	z	0	0	0	0	1	0	0	0	3	0	152	1	0	2	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	93.3
س	s	0	0	0	5	0	0	0	0	0	0	2	341	2	6	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	2	4	95
ش	sh	0	0	0	0	0	0	0	0	0	0	0	0	117	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99.2
ص	S	0	0	0	0	0	0	0	0	0	0	0	4	0	133	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	96.4
ض	D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98.7
ط	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	96.7
ظ	Z	0	0	1	0	0	0	0	0	4	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	5	88.1
ع	E	8	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	371	3	0	0	0	0	0	0	0	0	0	0	0	0	0	11	4	96.6
غ	g	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	97.9
ف	f	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	427	0	0	0	0	1	0	0	0	0	0	0	0	0	4	3	99.3
ق	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	83	2	0	0	0	0	0	0	0	0	0	0	0	0	0	93.3
ك	k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	194	0	0	0	0	0	0	0	0	0	0	0	5	1	99
ل	l	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	798	2	1	0	1	0	1	0	0	0	0	29	4	98.9
م	m	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	749	2	0	0	0	0	0	1	0	0	27	4	99.2
ن	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	748	0	0	0	0	0	0	0	0	20	9	99.1
ه	h	0	0	0	0	0	12	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	572	0	0	1	0	0	0	0	22	10	97.1
و	w	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	292	0	0	2	0	0	0	3	0	98
ي	y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	179	0	0	1	0	0	2	0	98.9
فتحة	a	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2872	3	0	4	0	28	22	99.7
ضمة	u	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2	0	1	0	0	8	741	0	0	3	0	28	11	97.9
كسرة	i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	5	0	831	0	0	1	21	10	99.2
ا	a2	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	5	0	0	622	0	0	2	3	98.7
و	u2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	102	0	1	0	100
ي	i2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	253	0	0	99.2

%Corr=96.70, Acc=95.95 [H=13142,D=238, S=210,I=103,N=13590]

Figure 22: Confusion matrix of phoneme detection using the MDD-object-G-Large/NS model.



		HZ	b	t	th	j	H	x	d	TH	r	z	s	sh	S	D	T	Z	E	g	f	q	k	l	m	n	h	w	y	a	u	i	a2	u2	i2	Del	Ins
ء	HZ	467	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	6	0	0	1	0	0	0	0	2	1	1	0	0	0	1	0	0	7	2
ب	b	0	257	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
ت	t	0	0	277	1	0	0	0	3	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4
ث	th	0	0	1	164	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ج	j	0	0	0	0	148	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ح	H	0	0	0	0	0	179	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	1	0	0	0	0	0	0	1
خ	x	0	0	0	0	0	3	74	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
د	d	0	1	0	0	2	0	0	149	3	0	0	0	0	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ذ	TH	0	0	0	0	0	0	0	1	112	0	7	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ر	r	0	0	0	0	0	0	0	0	0	463	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	5	1
ز	z	0	0	0	0	1	0	0	0	4	0	152	1	0	2	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
س	s	0	0	0	5	0	0	0	0	0	1	2	340	2	7	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0
ش	sh	0	0	0	0	0	0	0	0	0	0	0	0	117	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ص	S	0	0	0	0	0	0	0	0	0	0	0	4	0	132	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ض	D	0	0	0	0	0	0	0	1	0	0	0	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ط	T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ظ	Z	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
ع	E	8	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	373	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	2
غ	g	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	140	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1
ف	f	0	0	2	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	427	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	3
ق	q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	83	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ك	k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	195	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2
ل	l	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	796	2	1	0	1	0	1	0	0	0	0	0	0	34	3
م	m	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	745	2	0	0	0	0	1	0	0	0	0	0	31	2
ن	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	3	748	0	0	0	0	0	0	0	0	0	0	20	10
ه	h	0	0	0	0	0	14	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	578	0	0	1	0	0	0	0	0	13	11
و	w	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	294	0	0	1	0	1	0	0	2	0
ي	y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	182	0	0	1	0	0	0	0	0
فتحة	a	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2873	3	0	4	0	0	28	23
ضمة	u	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	0	0	0	0	0	6	743	0	0	3	0	29	12
كسرة	i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	833	0	1	1	20	9
ا	a2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	6	0	0	622	0	0	2	4
و	u2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	103	0	0	0
ي	i2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	252	0	0

Figure 23: Confusion matrix of the phoneme detection task using the YOLO-CNN-RNN-CTC model.

The confusion matrix of Figure 22 asserts the effectiveness of the proposed system, where most of the phonemes were correctly classified. The percentage of confusion was very small and confusion occurred between similar Arabic phonemes. For example, the phoneme /Z/ is wrongly classified 4 times as /TH/, where the two phonemes have the same place, manner, and manner-voice, and they differ only by manner-emphatic. This agrees with the statistics of the errors in the Arabic-CAPT corpus that was presented in chapter 4, where the substitution of phoneme /Z/ by /TH/ is from the top 10 highest substitution errors. The same observation is valid for the phoneme /h/ which was wrongly classified 12 times as /H/.

Figure 23 shows the confusion matrix of the output of the fusion of the two proposed systems. By comparing the matrices in Figure 22 and Figure 23 we notice the effectiveness of the fusion model. For example, the number of deletions reduced from 238 to 210. Nonetheless, this assert the power of our proposed method since the fusion increased the recognition only slightly that our proposed method.

6.3. Results using the KSU-CAPT Non-Arabs Database-Session 1 and Session 2

From the different models of the proposed systems, MDD-Object and MDD-E2E, we selected the best models based on the results of section 6.1. We applied the selected models to the two sessions of the developed database. In this section we present and discuss the result of the two systems.

6.3.1. Results of KSU-CAPT Session 1

The number of annotated speakers of this session is 60, hence we split this number to 45, 5, and 15 speakers for training, validation, and testing respectively. The total numbers of phonemes in the test set is 37943, without silence.

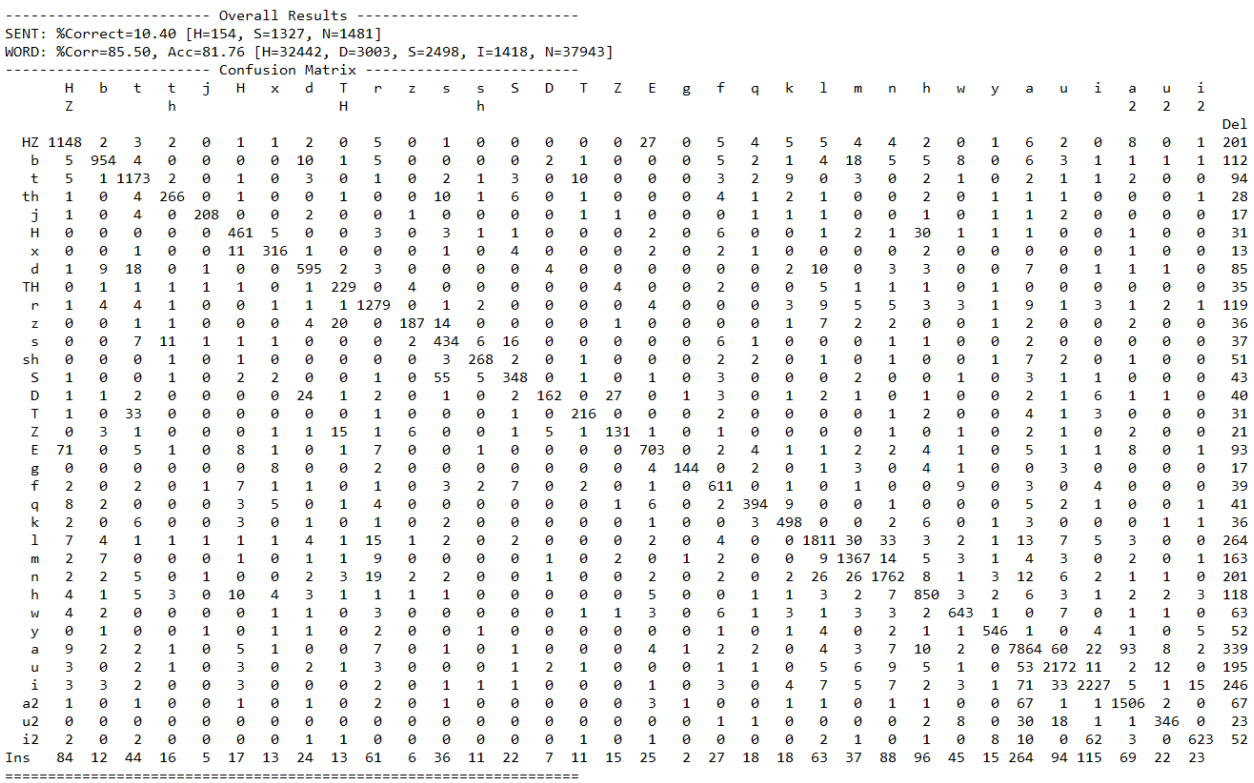
The performance of the proposed systems are shown in Table 43, where we see that the MDD-Object obtained PER of 18.24%, while MDD-E2E obtained 14.65%. The two proposed systems got good results compared to the average PER of the published results of non-Arabic benchmarking databases. We notice that the MDD-E2E obtained better results than MDD-Object, and we attribute this to the fact the MDD-Object used segmented data in training phase, but in our work we did not use segmented data and the segmentation was done automatically using MFA due to budget and time limitation. Moreover, MDD-Object can be used for phoneme and AFs recognition while MDD-E2E can be used only for phoneme recognition.

In terms of MDD metrics, the MDD-E2E achieved a 53.39%, 93.92%, and 68.08% for precision (P), recall (R), and F1, respectively. In contrast, the MDD-Object achieved a 32.74%, 83.77%, and 47.08% for precision (P), recall (R), and F1, respectively. The diagnosis accuracy (DA) of the proposed systems was 74.24% and 77.45% for MDD-Object and MDD-E2E, respectively. The results in Table 43 indicate to the quality of the proposed two Arabic CAPT systems in mispronunciation diagnosis.

Table 43: Performance of phoneme recognition task and MDD using KSU-CAPT session 1.

System	DA (%)	P (%)	R (%)	F1 (%)	PER
MDD-Object	74.24	32.74	83.77	47.08	18.24
MDD-E2E	77.45	53.39	93.92	68.08	14.65

The confusion matrix of the proposed MDD-Object system is presented in Figure 24, where we can see that the number of deletions is higher than number of substitutions and insertions. We can attribute this to the fact that some phonemes cannot be represented in the image of spectrogram due to the speed of the speech of some speakers or the noise in the recording, hence will not be present in the image of spectrogram and cannot be detected which amounts to deleting them from the recognized phonemes. In terms of substitutions, we can notice that most substitutions occur between two similar phonemes, such as the phoneme /H/ is substituted with phoneme /h/ 30 times, and phoneme /T/ substituted 33 times with phoneme /t/.



```
---------------------- Overall Results -------------------------
SENT: %Correct=10.40 [H=154, S=1327, N=1481]
WORD: %Corr=85.50, Acc=81.76 [H=32442, D=3003, S=2498, I=1418, N=37943]
---------------------- Confusion Matrix ------------------------
      H   b   t   t   j   H   x   d   T   r   z   s   s   S   D   T   Z   E   g   f   q   k   l   m   n   h   w   y   a   u   i   a   u   i
      Z           h                   H                   h                                                                   2   2   2
                                                                                                                                              Del
  HZ 1148   2   3   2   0   1   1   2   0   5   0   1   0   0   0   0   0  27   0   5   4   5   5   4   4   2   0   1   6   2   0   8   0   1  201
   b    5 954   4   0   0   0   0  10   1   5   0   0   0   0   2   1   0   0   0   5   2   1   4  18   5   5   8   0   6   3   1   1   1   1  112
   t    5   1 1173   2   0   1   0   3   0   1   0   2   1   3   0  10   0   0   0   3   2   9   0   3   0   2   1   0   2   1   1   2   0   0   94
  th    1   0   4 266   0   1   0   0   1   0   0  10   1   6   0   1   0   0   0   4   1   2   1   0   0   2   0   1   1   1   0   0   0   1   28
   j    1   0   4   0 208   0   0   2   0   0   1   0   0   0   0   0   1   1   0   0   0   1   1   1   0   0   1   0   1   1   2   0   0   0   17
   H    0   0   0   0   0 461   5   0   0   3   0   3   1   1   0   0   0   2   0   6   0   0   1   2   1  30   1   1   1   0   0   1   0   0   31
   x    0   0   1   0   0  11 316   1   0   0   0   1   0   4   0   0   0   2   0   2   1   0   0   0   0   2   0   0   0   0   0   1   0   0   13
   d    1   9  18   0   1   0   0 595   2   3   0   0   0   0   4   0   0   0   0   0   0   2  10   0   3   3   0   0   7   0   1   1   1   0   85
  TH    0   1   1   1   1   1   0   1 229   0   4   0   0   0   0   0   0   4   0   0   2   0   0   5   1   1   1   0   1   0   0   0   0   0   0   35
   r    1   4   4   1   0   0   1   1   1 1279   0   1   2   0   0   0   0   0   4   0   0   0   3   9   5   5   3   3   1   9   1   3   1   2   1  119
   z    0   0   1   1   0   0   0   0   4  20   0 187  14   0   0   0   0   1   0   0   0   0   1   7   2   2   0   0   1   2   0   0   2   0   0   36
   s    0   0   7  11   1   1   1   0   0   0   2 434   6  16   0   0   0   0   0   0   6   1   0   0   0   1   1   0   0   2   0   0   0   0   0   37
  sh    0   0   0   1   0   1   0   0   0   0   0   3 268   2   0   1   0   0   0   2   2   0   1   0   1   0   0   1   7   2   0   1   0   0   0   51
   S    1   0   0   1   0   2   2   0   0   1   0  55   5 348   0   1   0   1   0   3   0   0   0   2   0   0   1   0   3   1   1   0   0   0   0   43
   D    1   1   2   0   0   0   0  24   1   2   0   1   0   2 162   0  27   0   1   3   0   1   2   1   0   1   0   0   2   1   6   1   1   0   0   40
   T    1   0  33   0   0   0   0   0   0   1   0   0   0   1   0 216   0   0   0   2   0   0   0   0   1   2   0   0   4   1   3   0   0   0   0   31
   Z    0   3   1   0   0   0   1   1  15   1   6   0   0   1   5   1 131   1   0   1   0   0   0   0   1   0   1   0   2   1   0   2   0   0   0   21
   E   71   0   5   1   0   8   1   0   1   7   0   0   1   0   0   0   0 703   0   2   4   1   1   2   2   4   1   0   5   1   1   8   0   1   0   93
   g    0   0   0   0   0   0   8   0   0   2   0   0   0   0   0   0   0   4 144   0   2   0   1   3   0   4   1   0   0   3   0   0   0   0   0   17
   f    2   0   2   0   1   7   1   1   0   1   0   3   2   7   0   2   0   1   0 611   0   1   0   1   0   0   9   0   3   0   4   0   0   0   0   39
   q    8   2   0   0   0   3   5   0   1   4   0   0   0   0   0   0   0   1   6   0   2 394   9   0   0   1   0   0   0   5   2   1   0   0   1   41
   k    2   0   6   0   0   3   0   1   0   1   0   2   0   0   0   0   0   0   1   0   0   3 498   0   0   2   6   0   1   3   0   0   0   1   1   36
   l    7   4   1   1   1   1   1   4   1  15   1   2   0   2   0   0   0   2   0   4   0   0 1811  30  33   3   2   1  13   7   5   3   0   0 264
   m    2   7   0   0   0   1   0   1   1   9   0   0   0   0   1   0   2   0   1   2   0   0   9 1367  14   5   3   1   4   3   0   2   0   1 163
   n    2   2   5   0   1   0   0   2   3  19   2   2   0   0   1   0   0   2   0   2   0   2  26  26 1762   8   1   3  12   6   2   1   1   0 201
   h    4   1   5   3   0  10   4   3   1   1   1   1   0   0   0   0   0   5   0   0   1   1   3   2   7 850   3   2   6   3   1   2   2   3 118
   w    4   2   0   0   0   0   1   1   0   3   0   0   0   0   0   1   1   3   0   6   1   3   1   3   3   2 643   1   0   7   0   1   1   0   63
   y    0   1   0   0   1   0   1   1   0   2   0   0   1   0   0   0   0   0   0   1   0   1   4   0   2   1   1 546   1   0   4   1   0   5   52
   a    9   2   2   1   0   5   1   0   0   7   0   1   0   1   0   0   0   4   1   2   2   0   4   3   7  10   2   0 7864  60  22  93   8   2 339
   u    3   0   2   1   0   3   0   2   1   3   0   0   0   1   2   1   0   0   0   1   1   0   5   6   9   5   1   0  53 2172  11   2  12   0 195
   i    3   3   2   0   0   3   0   0   0   2   0   1   1   1   0   0   0   1   0   3   0   4   7   5   7   2   3   1  71  33 2227   5   1  15 246
  a2    1   0   1   0   0   1   0   1   0   2   0   1   0   0   0   0   0   3   1   0   0   1   1   0   1   1   0   0  67   1   1 1506   2   0  67
  u2    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   1   0   0   0   0   2   8   0  30  18   1   1 346   0  23
  i2    2   0   2   0   0   0   0   1   1   0   0   0   0   0   0   1   0   1   0   0   0   0   2   1   0   1   0   8  10   0  62   3   0 623  52
Ins    84  12  44  16   5  17  13  24  13  61   6  36  11  22   7  11  15  25   2  27  18  18  63  37  88  96  45  15 264  94 115  69  22  23
=================================================================
```

Figure 24: Confusion Matrix of the MDD-Object using KSU CAPT Session 1.

In Figure 25, we show the confusion matrix of the MDD-E2E system. By comparing Figure 24 and Figure 25 we can notice that the number of deletions in MDD-Object is higher than in MDD-E2E, which can be justified as in the previous paragraph. We can notice that the most substitution in MDD-E2E were in vowels, such as the phoneme /a/ substituted 103 times for /a2/ and the phoneme /a2/ is substituted 126 times for /a/.

SENT: %Correct=20.12 [H=298, S=1183, N=1481]
WORD: %Corr=88.01, Acc=85.35 [H=33467, D=1998, S=2562, I=1011, N=38027]
---------- Confusion Matrix ----------

	H	a	q	u	i	u	i	t	a	d	m	b	r	h	S	E	T	s	y	w	g	s	f	k	l	x	j	n	D	t	T	H	z	Z	
	Z					2	2		2													h								h	H				
																																			Del
HZ	1175	5	7	0	0	0	0	13	6	1	5	1	6	2	0	42	0	1	3	2	1	0	6	1	6	1	0	3	0	3	3	5	0	1	141
a	3	7968	1	44	23	3	3	2	103	0	0	3	5	5	0	2	1	1	1	0	1	1	1	1	4	0	0	4	0	0	2	3	0	0	304
q	2	0	420	1	0	0	0	3	0	0	0	1	2	2	0	9	1	0	0	0	2	0	4	6	0	6	0	0	1	1	1	3	0	0	21
u	0	46	0	2242	21	10	2	2	4	1	1	0	0	5	1	2	1	0	0	1	1	0	1	2	3	1	0	4	0	0	1	1	0	0	151
i	3	65	0	25	2317	1	23	1	3	1	0	0	3	1	1	0	1	0	3	1	0	0	2	1	4	0	0	10	0	0	0	1	1	1	200
u2	0	32	0	16	1	345	2	0	2	0	0	0	0	0	1	0	0	1	0	5	0	0	0	0	0	1	0	1	0	0	0	1	0	0	23
i2	4	15	0	0	50	1	651	0	1	0	0	0	1	0	0	0	0	0	12	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	32
t	4	2	1	1	3	0	0	1212	0	3	0	1	1	0	1	3	11	1	0	2	0	1	5	6	3	0	1	1	0	1	2	6	2	0	50
a2	1	126	0	1	1	3	0	0	1455	0	0	0	1	0	1	4	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	62
d	2	1	0	2	0	1	0	19	1	627	2	13	3	4	0	1	1	1	0	0	0	0	0	1	9	1	1	3	6	1	2	0	0	0	44
m	2	1	0	4	0	0	0	2	2	1	1452	6	10	2	1	7	0	0	1	8	1	0	0	0	15	0	0	15	1	0	0	1	0	0	69
b	9	3	6	3	1	0	2	5	0	3	16	993	8	3	1	3	2	1	1	2	1	0	5	0	11	1	1	5	2	2	3	1	0	0	61
r	5	10	0	2	2	1	0	2	1	3	3	5	1303	4	1	10	4	4	0	4	0	0	2	0	15	0	1	6	0	1	1	0	0	2	72
h	16	2	2	2	0	2	0	4	0	1	4	1	5	841	1	18	0	0	1	3	0	0	4	0	7	0	0	5	0	2	0	19	1	1	102
S	1	3	1	0	0	0	0	1	0	0	0	0	0	0	374	1	0	47	0	0	1	7	6	1	0	0	0	0	0	2	0	1	1	0	24
E	69	3	5	1	0	0	0	1	1	0	2	0	6	0	0	752	3	1	0	1	1	0	5	1	3	0	0	2	0	2	0	1	0	1	63
T	1	2	0	0	0	0	0	22	0	0	0	1	0	0	5	0	248	0	0	0	0	0	1	2	1	1	0	0	1	2	0	0	0	0	9
s	2	1	0	0	0	0	0	8	1	0	2	1	1	1	18	0	0	453	0	0	0	5	4	0	0	0	0	0	0	8	0	0	0	0	23
y	3	0	0	0	0	0	5	2	0	2	1	0	2	0	0	2	0	0	551	1	0	0	0	0	7	0	1	1	0	0	0	0	0	0	48
w	2	1	2	2	0	5	0	1	0	0	4	3	4	2	0	5	0	0	0	661	0	0	4	1	3	0	1	0	0	0	0	1	0	1	49
g	2	2	1	0	0	0	0	0	0	0	1	1	0	5	1	6	0	0	0	2	153	0	1	0	1	2	0	0	0	0	0	0	0	0	11
sh	0	1	0	0	2	0	0	0	0	0	0	0	0	1	0	1	0	5	0	0	0	311	2	0	3	0	0	1	0	0	0	0	0	0	18
f	4	2	0	1	1	0	0	6	0	1	1	1	1	1	11	4	1	8	0	9	0	0	603	2	2	3	0	2	1	1	0	4	1	0	27
k	6	1	5	1	1	0	0	12	0	4	0	1	0	2	0	1	2	3	0	0	0	0	3	492	2	0	0	0	0	5	0	3	0	0	23
l	6	8	0	1	3	0	1	5	0	1	18	10	15	1	0	2	0	2	2	3	0	1	2	1	1976	0	0	28	0	0	3	0	2	0	128
x	1	0	2	0	0	0	0	0	0	0	0	0	0	2	6	5	0	2	0	1	1	0	4	0	0	306	0	1	0	2	0	13	0	0	9
j	1	0	0	0	2	0	1	1	0	2	0	0	0	0	0	0	0	0	2	8	0	0	0	0	4	0	209	0	0	0	1	0	0	0	12
n	3	8	0	6	5	0	2	3	1	3	28	5	8	3	0	3	0	0	5	0	0	0	0	1	57	1	0	1825	0	0	1	0	2	0	129
D	0	2	1	0	4	0	0	4	0	20	0	8	3	0	1	0	2	0	1	0	0	0	1	1	3	0	1	1	188	0	1	0	0	25	16
th	0	1	0	1	0	0	1	6	0	0	0	0	1	1	5	2	1	10	1	0	0	0	3	1	1	0	0	0	0	280	1	3	0	0	14
TH	0	0	0	0	0	1	0	1	0	1	1	2	2	1	0	2	1	0	2	1	1	0	1	0	6	0	0	3	0	0	249	0	4	1	9
H	5	2	1	0	0	0	0	1	0	0	2	0	1	18	2	4	0	3	0	0	0	0	5	0	2	4	0	1	0	2	0	472	0	0	26
z	0	4	0	1	0	0	0	2	0	2	3	0	1	0	1	0	0	8	1	0	0	1	0	1	6	0	1	0	0	0	19	0	208	3	19
Z	0	1	0	0	0	0	0	2	0	2	0	3	2	0	1	1	2	1	0	1	0	0	1	0	2	0	0	0	1	1	8	0	3	155	9
Ins	51	227	10	65	92	9	8	24	26	16	37	18	45	60	15	24	3	15	18	33	1	9	16	12	50	3	6	75	8	8	6	19	1	1	

==========

Figure 25: Confusion Matrix of the MDD-E2Eusing KSU CAPT Session 1.

6.3.2. Results of KSU-CAPT session 2

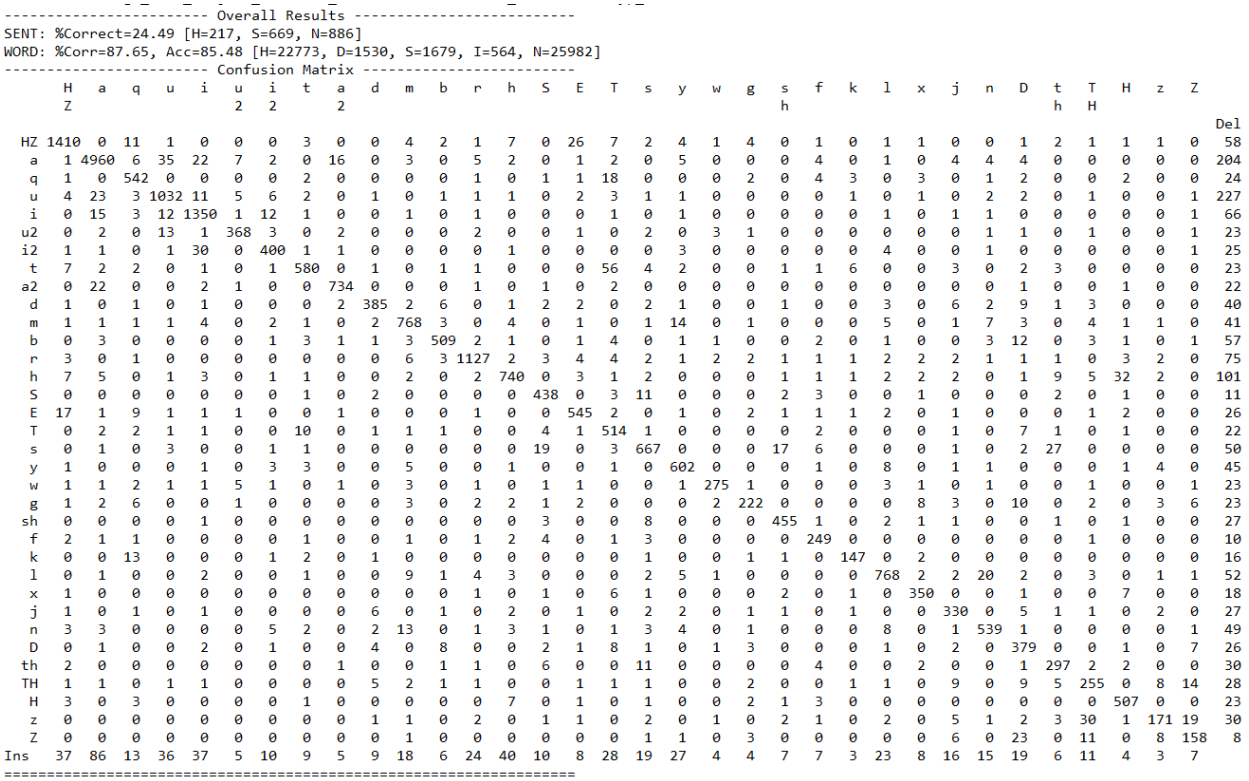
In this section, we repeat the investigation of section 6.2.1 using the data of session 2 of the developed KSU-CAPT database. We can clearly see that the performance of the proposed systems using session 2 is almost identical to the performance of session 1. This performance shows that the proposed method gave excellent results for two different databases with different text and constructions. Figure 26 and Figure 27 show the confusion matrices of the proposed systems. From the two figures we get same remarks that we got for session 1.

Table 44: Performance of phoneme recognition task and MDD using KSU-CAPT session 2.

System	DA (%)	P (%)	R (%)	F1 (%)	PER
MDD-Object	74.91	66.57	92.65	77.47	19.6
MDD-E2E	77.85	76.84	94.58	84.79	14.52

```
------------------------ Overall Results ---------------------------
SENT: %Correct=11.63 [H=103, S=783, N=886]
WORD: %Corr=82.27, Acc=80.40 [H=21295, D=3253, S=1336, I=485, N=25884]
------------------------ Confusion Matrix --------------------------
       H   b   t   t   j   H   x   d   T   r   z   s   s   S   D   T   Z   E   g   f   q   k   l   m   n   h   w   y   a   u   i   a   u   i
       Z           h                   H                   h                                                                   2   2   2
                                                                                                                                                Del
  HZ 1311   0   4   1   0   3   0   2   1   3   0   0   0   2   2   2   0  19   1   1   4   0   0   1   0   6   1   2   4   2   0   1   0   0  177
   b    1 470   2   0   0   0   0   4   3   2   1   1   0   1   4   1   3   0   0   0   1   0   1   4   0   0   0   0   0   1   2   1   0   2  106
   t    2   1 553   2   1   0   0   4   0   4   0   5   0   1   2  22   0   1   0   1   2   3   1   1   0   0   0   2   4   1   2   0   0   0   82
  th    2   1   2 245   0   2   1   0   0   1   0  10   2   1   0   1   0   0   0   2   0   0   0   0   0   2   0   0   2   1   0   0   0   0   85
   j    1   1   1   0 304   1   0   1   0   0   1   0   4   0   1   0   0   0   2   0   2   2   2   0   0   0   0   2   0   0   0   0   0   0   61
   H    1   0   0   2   0 471   1   1   0   1   0   0   1   0   0   0   0   4   0   2   2   0   0   2   0  11   0   0   3   0   1   0   0   0   48
   x    0   0   1   0   1   3 324   0   0   1   0   1   0   0   0   0   0   0   1   1   1   2   2   0   0   0   0   0   0   0   1   0   0   0   50
   d    4   0   3   1   1   0   0 360   1   1   0   0   0   0   5   3   0   0   0   0   0   0   4   0   3   0   0   0   2   1   1   0   1   0   78
  TH    1   1   0   2   6   0   0   1 231   0   9   0   0   0   3   0   9   0   0   0   0   0   0   2   2   1   1   0   1   0   1   1   0   0   0   76
   r    0   0   0   1   1   2   0   0   1 1024   0   0   0   0   0   1   0   1   0   0   0   1   1   0   0   0   0   2   6   1   0   1   0   1  207
   z    2   0   0   1   4   0   0   0  24   2 154   6   1   2   0   0  11   1   0   0   0   0   2   0   1   0   0   2   2   1   0   0   1   0   59
   s    0   0   0  20   0   0   0   0   0   1   0 639  15  23   0   1   0   0   0   2   0   0   0   0   0   0   0   0   0   1   4   2   0   0   89
  sh    0   0   1   0   1   2   0   0   0   0   0   2 446   1   0   0   0   1   0   0   0   0   1   0   0   0   0   0   0   1   0   0   0   0   44
   S    0   0   0   5   0   1   1   0   0   0   0  17  10 376   1   1   0   0   0   0   0   0   0   0   0   1   0   0   3   0   0   0   0   0   59
   D    0  10   1   0   3   0   0   8   1   1   0   0   0   0 296  11   6   1   1   1   0   0   0   1   1   0   0   0   3   2   3   0   0   1   97
   T    2   1  20   0   2   0   0   1   0   2   0   0   0   2   0 412   0   0   0   2   3   2   0   0   0   0   0   0   3   0   0   0   0   0  120
   Z    0   2   1   0   2   0   0   1  10   1   4   0   0   1  14   0 132   1   2   0   0   0   0   1   0   1   0   0   0   2   0   0   1   0   44
   E   16   0   0   0   1   1   1   0   2   2   0   0   0   0   4   1   0 493   1   0   1   0   2   0   0   1   1   0   7   1   0   2   1   0   79
   g    3   0   0   0   1   0  10   1   3   3   1   0   0   0   1   1   4   0 221   0   2   0   0   3   0   1   1   0   0   0   0   0   0   0   43
   f    0   0   1   3   0   3   0   0   0   0   0   1   2   1   0   1   0   0   0 216   0   0   0   0   1   0   0   0   2   0   0   1   0   0   45
   q    4   2   1   0   0   1   5   1   0   1   0   1   0   0   1   9   0   1   0   3 506   9   0   0   1   0   0   0   3   0   0   0   0   0   57
   k    1   0   0   0   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1  13 140   0   0   0   0   0   0   1   1   0   0   0   0   25
   l    2   1   0   0   0   0   1   2   2   4   0   1   0   0   0   0   0   1   0   0   0   1 692   7   7   2   0   0   2   1   3   0   0   0  149
   m    2   3   0   1   0   0   0   1   1   1   0   0   0   0   0   0   0   1   0   0   0   0   6 715   7   1   0   7   1   2   1   0   0   2  116
   n    0   0   1   0   0   1   0   1   0   0   1   0   1   0   1   0   0   0   0   0   0   0   8  11 498   0   2   4   3   4   1   1   2   1   99
   h    6   0   1   3   3  13   1   0   1   4   2   5   1   1   0   2   0   3   1   0   0   1   3   0   0 688   1   2   3   5   1   1   0   0  174
   w    1   0   0   1   0   1   0   0   1   0   0   0   0   0   0   0   0   1   0   0   0   0   1   1   0   1 197   1   2   3   1   0   5   1   44
   y    2   0   2   1   2   0   0   0   1   0   0   0   0   0   0   1   0   0   0   0   1   0   1   1   3   0   0 582   0   1   1   0   0   8   58
   a    8   0   0   1   2   0   0   3   3   3   0   0   1   0   0   2   1   0   0   2   2   0   0   2   1   4   0   3 4784  30  12  11   5   3  403
   u    2   0   1   0   1   1   1   0   2   4   2   0   0   2   1   2   2   2   1   0   3   0   0   1   2   2   0   3  13 1029  13   0   4   1  235
   i    1   1   2   1   2   0   0   1   0   2   1   0   0   0   0   0   0   3   1   0   0   0   0   3   0   2   1   0   3  12   5 1306   1   0   8  112
  a2    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0  13   1   1 728   2   0   41
  u2    3   0   0   0   0   1   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   1   0   2  10   0   0 360   2   42
  i2    1   0   0   0   1   0   0   0   0   2   0   0   1   0   0   0   0   0   0   1   0   0   0   1   0   0   0   0   1   1  18   0   0 392  49
Ins    32   9  13   3   9   6   8  14   6  24   5   9  12  17  12  13   5  14   2   4   6   0   6  13   8  38   6  25  51  53  35   8   7  12
=================================================================
```

Figure 26: Confusion Matrix of the MDD-Object using KSU CAPT Session 2.



```
---------------------- Overall Results -----------------------
SENT: %Correct=24.49 [H=217, S=669, N=886]
WORD: %Corr=87.65, Acc=85.48 [H=22773, D=1530, S=1679, I=564, N=25982]
---------------------- Confusion Matrix ----------------------
      H   a   q   u   i   u   i   t   a   d   m   b   r   h   S   E   T   s   y   w   g   s   f   k   l   x   j   n   D   t   T   H   z   Z
      Z                   2   2       2                                                   h                           h   H
                                                                                                                                  Del
  HZ 1410  0  11   1   0   0   0   3   0   0   4   2   1   7   0  26   7   2   4   1   4   0   1   0   1   1   0   0   1   2   1   1   1   0   58
   a    1 4960   6  35  22   7   2   0  16   0   3   0   5   2   0   1   2   0   5   0   0   0   4   0   1   0   4   4   4   0   0   0   0   0  204
   q    1   0 542   0   0   0   0   2   0   0   0   0   1   0   1   1  18   0   0   0   2   0   4   3   0   3   0   1   2   0   0   2   0   0   24
   u    4  23   3 1032 11   5   6   2   0   1   0   1   1   1   0   2   3   1   1   0   0   0   0   1   0   1   0   2   2   0   1   0   0   1  227
   i    0  15   3  12 1350  1  12   1   0   0   1   0   1   0   0   0   1   0   1   0   0   0   0   0   1   0   1   1   0   0   0   0   0   1   66
  u2    0   2   0  13   1 368   3   0   2   0   0   0   2   0   0   1   0   2   0   3   1   0   0   0   0   0   0   1   1   0   1   0   0   1   23
  i2    1   1   0   1  30   0 400   1   1   0   0   0   0   1   0   0   0   0   3   0   0   0   0   0   4   0   0   1   0   0   0   0   0   1   25
   t    7   2   2   0   1   0   1 580   0   1   0   1   1   0   0   0  56   4   2   0   0   1   1   6   0   0   3   0   2   3   0   0   0   0   23
  a2    0  22   0   0   2   1   0   0 734   0   0   0   1   0   1   0   2   0   0   0   0   0   0   0   0   0   0   0   1   0   0   1   0   0   22
   d    1   0   1   0   1   0   0   0   2 385   2   6   0   1   2   2   0   2   1   0   0   1   0   0   3   0   6   2   9   1   3   0   0   0   40
   m    1   1   1   1   4   0   2   1   0   2 768   3   0   4   0   1   0   1  14   0   1   0   0   0   5   0   1   7   3   0   4   1   1   0   41
   b    0   3   0   0   0   0   1   3   1   1   3 509   2   1   0   1   4   0   1   1   0   0   2   0   1   0   0   3  12   0   3   1   0   1   57
   r    3   0   1   0   0   0   0   0   0   0   6   3 1127   2   3   4   4   2   1   2   2   1   1   1   2   2   2   1   1   1   0   3   2   0   75
   h    7   5   0   1   3   0   1   1   0   0   2   0   2 740   0   3   1   2   0   0   0   1   1   1   2   2   2   0   1   9   5  32   2   0  101
   S    0   0   0   0   0   0   0   0   1   0   2   0   0   0   0 438   0   3  11   0   0   0   2   3   0   0   1   0   0   0   2   0   1   0   0   11
   E   17   1   9   1   1   1   0   0   1   0   0   0   1   0   0 545   2   0   1   0   2   1   1   1   2   0   1   0   0   0   1   2   0   0   26
   T    0   2   2   1   1   0   0  10   0   1   1   1   0   0   4   1 514   1   0   0   0   0   2   0   0   0   1   0   7   1   0   1   0   0   22
   s    0   1   0   3   0   0   1   1   0   0   0   0   0   0  19   0   3 667   0   0   0  17   6   0   0   0   1   0   2  27   0   0   0   0   50
   y    1   0   0   0   1   0   3   3   0   0   5   0   0   1   0   0   1   0 602   0   0   0   1   0   8   0   1   1   0   0   0   1   4   0   45
   w    1   1   2   1   1   5   1   0   1   0   3   0   1   0   1   1   0   0   1 275   1   0   0   0   3   1   0   1   0   0   1   0   0   1   23
   g    1   2   6   0   0   1   0   0   0   0   3   0   2   2   1   2   0   0   0   2 222   0   0   0   0   8   3   0  10   0   2   0   3   6   23
  sh    0   0   0   0   1   0   0   0   0   0   0   0   0   0   3   0   0   8   0   0   0 455   1   0   2   1   1   0   0   1   0   1   0   0   27
   f    2   1   1   0   0   0   0   1   0   0   1   0   1   2   4   0   1   3   0   0   0   0 249   0   0   0   0   0   0   0   1   0   0   0   10
   k    0   0  13   0   0   0   1   2   0   1   0   0   0   0   0   0   0   1   0   0   1   1   0 147   0   2   0   0   0   0   0   0   0   0   16
   l    0   1   0   0   2   0   0   1   0   0   9   1   4   3   0   0   0   2   5   1   0   0   0   0 768   2   2  20   2   0   3   0   1   1   52
   x    1   0   0   0   0   0   0   0   0   0   0   0   0   1   0   1   0   6   1   0   0   0   2   0   1   0 350   0   0   1   0   0   7   0   0   18
   j    1   0   1   0   1   0   0   0   0   6   0   1   0   2   0   1   0   2   2   0   1   1   0   1   0   0 330   0   5   1   1   0   2   0   27
   n    3   3   0   0   0   0   5   2   0   2  13   0   1   3   1   0   1   3   4   0   1   0   0   0   8   0   1 539   1   0   0   0   0   1   49
   D    0   1   0   0   2   0   1   0   0   4   0   8   0   0   2   1   8   1   0   1   3   0   0   0   1   0   2   0 379   0   0   1   0   7   26
  th    2   0   0   0   0   0   0   0   1   0   0   1   1   0   6   0   0  11   0   0   0   0   4   0   0   2   0   0   1 297   2   2   0   0   30
  TH    1   1   0   1   1   0   0   0   0   5   2   1   1   0   0   1   1   1   0   0   2   0   0   1   1   0   9   0   9   5 255   0   8  14   28
   H    3   0   3   0   0   0   0   1   0   0   0   0   0   7   0   1   0   1   0   0   2   1   3   0   0   0   0   0   0   0   0 507   0   0   23
   z    0   0   0   0   0   0   0   0   0   1   1   0   2   0   1   1   0   2   0   1   0   2   1   0   2   0   5   1   2   3  30   1 171  19   30
   Z    0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1   1   0   3   0   0   0   0   0   6   0  23   0  11   0   8 158    8
Ins   37  86  13  36  37   5  10   9   5   9  18   6  24  40  10   8  28  19  27   4   4   7   7   3  23   8  16  15  19   6  11   4   3   7
=============================================================
```

Figure 27: Confusion Matrix of the MDD-E2E using KSU CAPT Session 2.

6.4. Building a Pilot Arabic CAPT system (GUI and Examples)

In this section, we present the Arabic-CAPT system in an application form. We used the proposed MDD-Object because it can detect the phonemes and associated AFs, hence we can detect the pronunciation errors and provide feedback at articulatory level, concurrently. To design the graphic user interface (GUI) for the proposed Arabic CAPT system, we used the python tool called Streamlit. Figure 28 shows the GUI of the proposed system, where the user can select testing his pronunciation either by using already saved wav file or by recording online. Then the user has to select the canonical text that he wants to learn its correct pronunciation. As we can see from Figure 28, the proposed system detected the consonant and vowel phonemes, and the PER of the detected text is 7.14%.

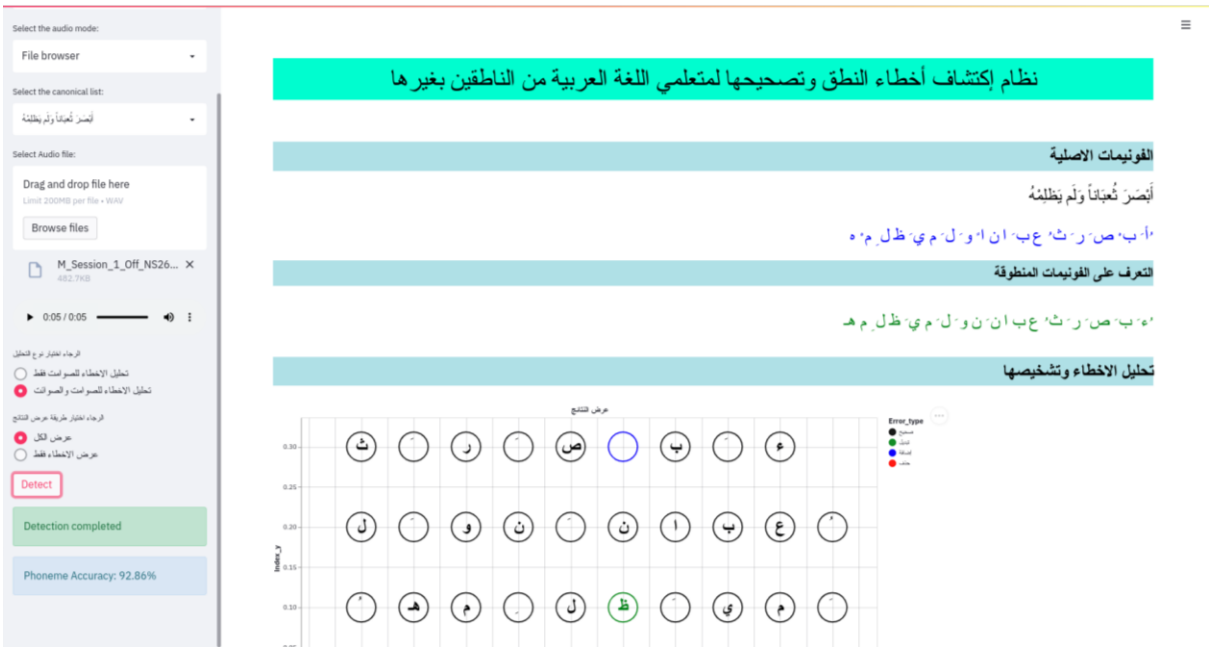


Figure 28: GUI of the proposed Arabic CAPT system, and the first example of using Arabic-CAPT database.

Regarding the mispronunciation analysis and diagnosis, the system shows the detected phonemes in color format: black for correct phoneme, green for substitution error, blue for insertion error, and red for deletion error. For a substitution error, the system provides the details of the error and articulatory feedback to the user once he clicks on the phoneme in error.

Figure 29 shows an example of the system display when pressing on the phoneme with substitution error. The substitution error was in the phoneme /ظ/ of the word "يظلمه". By pressing on the phoneme in error the system displays the box of messages as in the Figure 29. First message is the canonical phoneme. Second message is the detected phoneme which was "ظز" which mean the system cannot assert if the detected phoneme is "ظ" or "ز" because it detected that the phoneme as /ظ/ with score 61% and /ز/ with score 56%. This confusion is because the speaker pronounced this phoneme in a manner that have the characteristics of both "ظ" and "ز". This information is displayed to the user in the diagnosis error in the third message in the box.

As we mentioned before that our proposed system has the ability to detect the phoneme and the AFs simultaneously from the whole utterance, hence by comparing the detected AFs with the canonical AFs, the system has the ability to inform the user of the reason for this error.

This information is passed to the user in the fourth message in the box, hence in this example in the Figure 29, the system tells the user that the main problem to cause the pronunciation error in the phoneme /ظ/ is that the pronounced phoneme has the alveo-dental feature which is not in the canonical phoneme /ظ/.

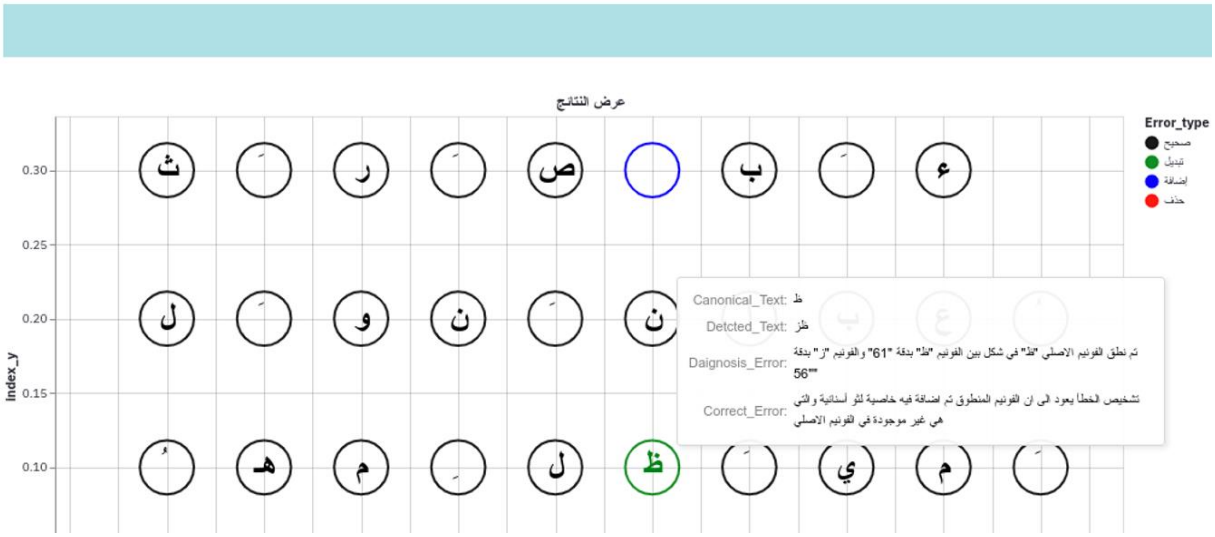


Figure 29: Example of mispronunciation diagnosis and feedback generation in Arabic-CAPT database.

Figure 28 shows an example of mispronunciation diagnosis and feedback generation in MDD-Object system using Arabic-CAPT database. Now, we will show an example from the KSU-CAPT session 1 database using the proposed system MDD-Object. The proposed system detected the pronounced phonemes in the whole utterance as shown in Figure 30. We can notice that the phoneme /ث/ in the word /ثناء/ was pronounced as /س/, hence the system detected it as /س/. The details of the box of messages in this substitution error is shown in Figure 31.

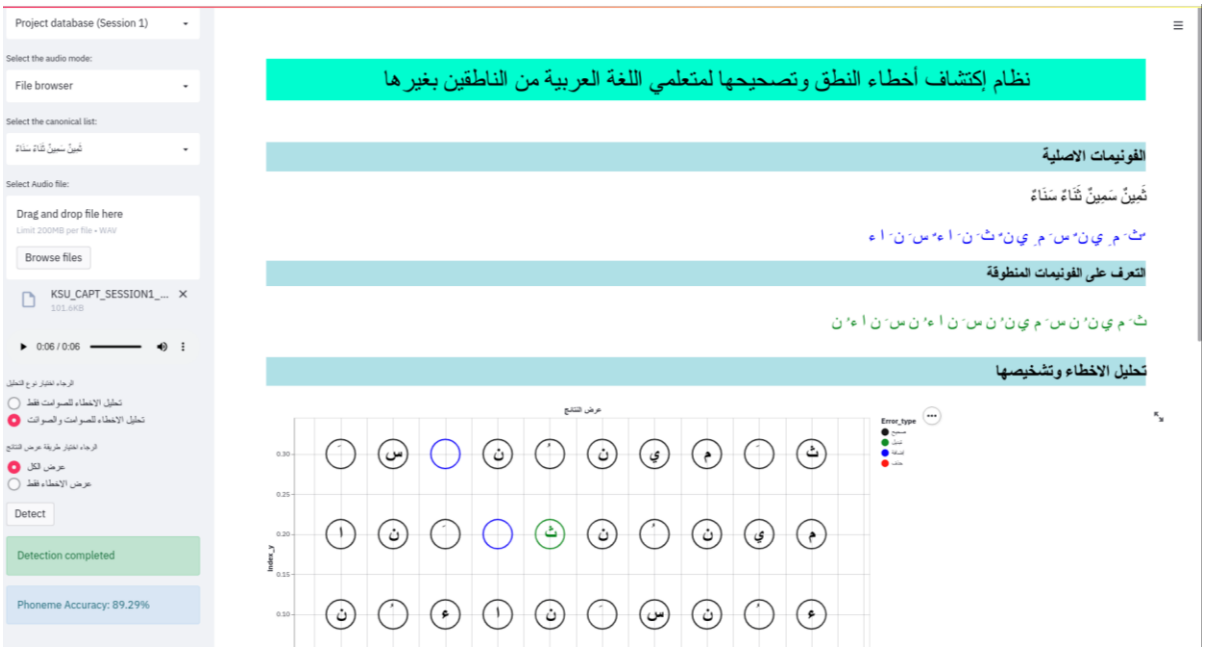


Figure 30: Example of MDD in the proposed Arabic CAPT system using KSU-CAPT session 1 database.

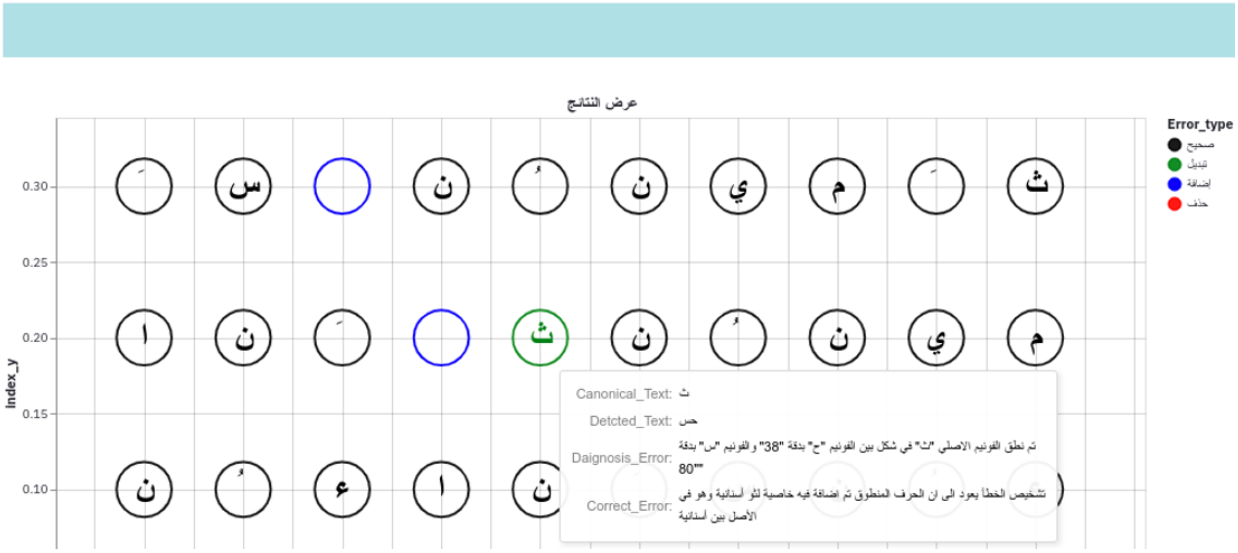


Figure 31:Example of mispronunciation diagnosis and feedback generation in KSU-CAPT session 1.

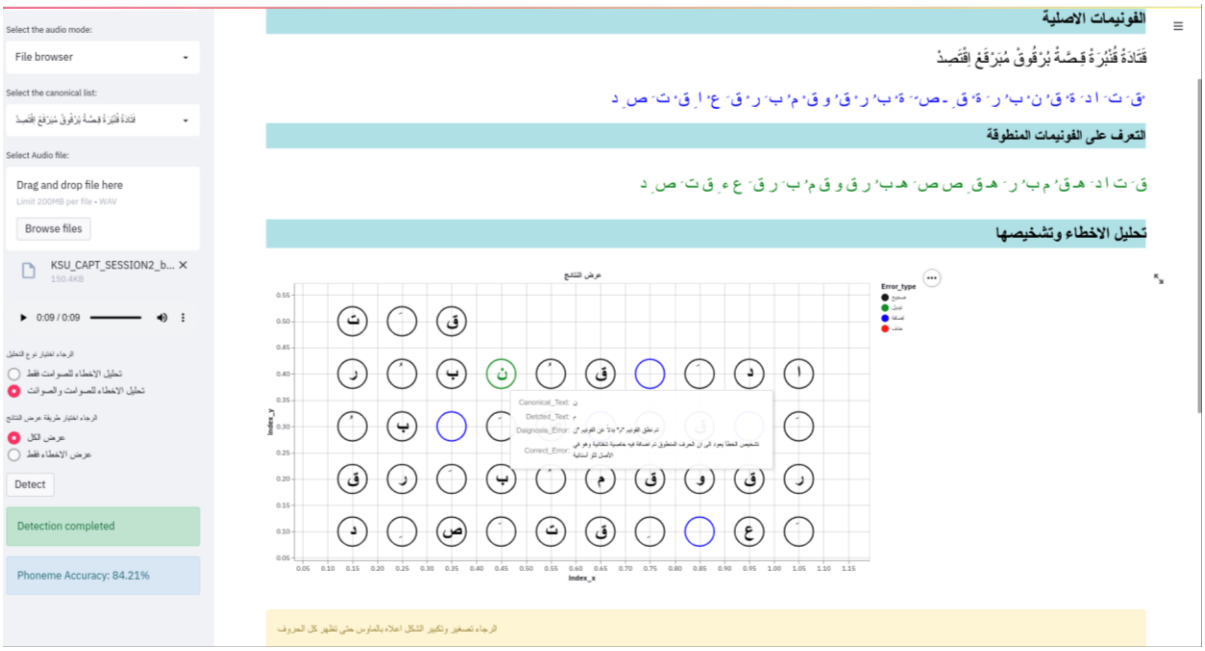


Figure 32:Example of MDD in the proposed Arabic CAPT system using KSU-CAPT session 2 database.

Figure 32 shows an example of testing the proposed MDD-Object system for recognizing a speech sample from KSU-CAPT session 2. We notice an interesting point, which is that when the user pronounces the word "قنبرة" he replaced the phoneme "ن" by "م" following the Tajjweed rules of the Holy Quran and this was detected by the system. The details of the box of messages in this substitution error is shown in Figure 32.

7. Future work

We successfully achieved the project objectives, by developing a comprehensive non-native Arabic speech databases and proposing an effective MDD systems for phonemes and AFs detection. Our aim is to continue in this research area because it has a huge need by non-native Arabic speakers. We can investigate enhancing the performance of proposed MDD systems by annotating the speech of more speakers. We may investigate building a word-based MDD system which is suitable of beginner learners and children. To do that, we have to segment all recorded speech to word level. Moreover, in the direction of converting the project outcomes to a real product, we will upgrade the proposed Arabic CAPT application to work on mobile devices.

We may investigate completing the annotation of the session 1 and session 2 using the proposed system to save time and budget. In terms of the feedback, the propose system can be enhanced by adding the following types of feedbacks: text, 3D, and video animation.

8. References

[1] S. E. Hamid, O. Abdel-Hamid, and M. Rashwan, “Performance Tuning and System Evaluation for Computer Aided Pronunciation Learning,” in *Proceedings of NEMLAR International Conference on Arabic Language Resources and Tools*, 2009, pp. 140–143.

[2] S. M. Abdou *et al.*, “Computer aided pronunciation learning system using speech recognition techniques,” in *Ninth International Conference on Spoken Language Processing*, 2006.

[3] K. Necibi and H. Bahi, “An arabic mispronunciation detection system by means of automatic speech recognition technology,” in *The 13th International Arab Conference on Information Technoloy Proceedings*, 2012, pp. 303–308.

[4] H. Dahan, A. Hussin, Z. Razak, and M. Odelha, “Automatic arabic pronunciation scoring for language instruction,” 2011.

[5] M. Belgacem, A. Maatallaoui, and M. Zrigui, “Arabic language learning assistance based on automatic speech recognition system,” in *Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE)*, 2011, p. 1.

[6] M. S. El-Kasasy, “An Automatic Speech Verification System,” Ph. D. Thesis, Cairo University, Faculty of Engineering, Department of~…, 1992.

[7] M. S. Abdo, A. H. Kandil, A. M. El-Bialy, and S. A. Fawzy, “Automatic detection for some common pronunciation mistakes applied to chosen Quran sounds,” in *2010 5th Cairo International Biomedical Engineering Conference*, 2010, pp. 219–222.

[8] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Tsinghua University Press, 1999.

[9] R. M. Hegde, “Fourier transform phase-based features for speech recognition,” *Indian Inst. Technol. Madras*, 2005.

[10] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[11] F. Nazir, M. N. Majeed, M. A. Ghazanfar, and M. Maqsood, “Mispronunciation detection using deep convolutional neural network features and transfer learning-based model for Arabic phonemes,” *IEEE Access*, vol. 7, pp. 52589–52608, 2019.

[12] S. Akhtar *et al.*, “Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features,” *Electronics*, vol. 9, no. 6, p. 963, 2020.

[13] U. Shrawankar and V. M. Thakare, “Techniques for feature extraction in speech recognition system: A comparative study,” *arXiv Prepr. arXiv1305.1145*, 2013.

[14] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A review on speech recognition technique,” *Int. J. Comput. Appl.*, vol. 10, no. 3, pp. 16–24, 2010.

[15] D. Palaz, R. Collobert, and others, “Analysis of cnn-based speech recognition system using raw speech as input,” 2015.

[16] J. Lee, T. Kim, J. Park, and J. Nam, “Raw waveform-based audio classification using sample-level CNN architectures,” *arXiv Prepr. arXiv1712.00866*, 2017.

[17] G. Kovács, L. Tóth, D. Van Compernolle, and S. Ganapathy, “Increasing the robustness of CNN acoustic models using autoregressive moving average spectrogram features and channel dropout,” *Pattern Recognit. Lett.*, vol. 100, pp. 44–50, 2017.

[18] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3D log-mel spectrograms with deep learning network,” *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[19] N. Souissi and A. Cherif, “Speech recognition system based on short-term cepstral parameters, feature reduction method and artificial neural networks,” in *2016 2nd international conference on advanced technologies for signal and image processing (ATSIP)*, 2016, pp. 667–

671.

[20] A. Winursito, R. Hidayat, and A. Bejo, “Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition,” in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 379–383.

[21] C. O. S. Sorzano, J. Vargas, and A. P. Montano, “A survey of dimensionality reduction techniques,” *arXiv Prepr. arXiv1403.2877*, 2014.

[22] Y. Alotaibi, S.-A. Selouani, and D. O’shaughnessy, “Experiments on automatic recognition of nonnative Arabic speech,” *EURASIP J. Audio, Speech, Music Process.*, vol. 2008, no. 1, p. 679831, 2008.

[23] M. Belgacem and M. Zrigui, “Automatic Identification System of Arabic Dialects,” in *IPCV 2010: proceedings of the 2010 international conference on image processing, computer vision, & pattern recognition (Las Vegas NV, July 12-15, 2010)*, 2010, pp. 740–749.

[24] W. J. J. Roberts and J. P. Willmore, “Automatic speaker recognition using Gaussian mixture models,” in *1999 Information, Decision and Control. Data and Information Fusion Symposium, Signal Processing and Communications Symposium and Decision and Control Symposium. Proceedings (Cat. No. 99EX251)*, 1999, pp. 465–470.

[25] A. Trigui, A. Mars, M. A. Ben Jannet, M. Maraoui, and M. Zrigui, “Foreign accent classification for Arabic speech learning,” in *Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, 2011, p. 1.

[26] W. J. Barry, C. E. Hoequist, and F. J. Nolan, “An approach to the problem of regional accent in automatic speech recognition,” *Comput. Speech Lang.*, vol. 3, no. 4, pp. 355–366, 1989.

[27] L. Indrayanti, T. Usagawa, Y. Chisaki, and T. Dutono, “Evaluation of pronunciation by means of automatic speech recognition system for computer aided Indonesian language learning,” in *2006 7th International Conference on Information Technology Based Higher Education and Training*, 2006, pp. 553–556.

[28] A. Hannun *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv Prepr. arXiv1412.5567*, 2014.

[29] D. Amodei *et al.*, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International conference on machine learning*, 2016, pp. 173–182.

[30] R. Collobert, C. Puhrsch, and G. Synnaeve, “Wav2letter: an end-to-end convnet-based speech recognition system,” *arXiv Prepr. arXiv1609.03193*, 2016.

[31] Y. Miao, M. Gowayyed, and F. Metze, “EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 167–174.

[32] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, “Improved training of end-to-end attention models for speech recognition,” *arXiv Prepr. arXiv1805.03294*, 2018.

[33] L. Zhang *et al.*, “End-to-End Automatic Pronunciation Error Detection Based on Improved Hybrid CTC/Attention Architecture,” *Sensors*, vol. 20, no. 7, p. 1809, 2020.

[34] T. H. Lo, S. Y. Weng, H. J. Chang, and B. Chen, “An effective end-to-end modeling approach for mispronunciation detection,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-Octob, pp. 3027–3031, 2020, doi: 10.21437/Interspeech.2020-1605.

[35] Y. Feng, G. Fu, Q. Chen, and K. Chen, “SED-MDD: Towards Sentence Dependent End-To-End Mispronunciation Detection and Diagnosis,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3492–3496.

[36] J. Van Doremalen, C. Cucchiarini, and H. Strik, “Automatic detection of vowel pronunciation errors using multiple information sources,” in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 580–585.

[37] N. Oostdijk, “The Spoken Dutch Corpus. Overview and First Evaluation.,” in *LREC*, 2000, pp. 887–894.

[38] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, “Combination of machine scores for automatic grading of pronunciation quality,” *Speech Commun.*, vol. 30, no. 2–3, pp. 121–130, 2000.

[39] K. P. Truong, A. Neri, F. de Wet, C. Cucchiarini, and H. Strik, “Automatic detection of frequent pronunciation errors made by L2-learners,” in *Ninth European Conference on Speech Communication and Technology*, 2005.

[40] K. Fukunaga, *Introduction to statistical pattern recognition*. Elsevier, 2013.

[41] L. F. Weigelt, S. J. Sadoff, and J. D. Miller, “Plosive/fricative distinction: The voiceless case,” *J. Acoust. Soc. Am.*, vol. 87, no. 6, pp. 2729–2737, 1990.

[42] Y. Kim, H. Franco, and L. Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction,” in *Fifth European Conference on Speech Communication and Technology*, 1997.

[43] S. Xu, J. Jiang, Z. Chen, and B. Xu, “Automatic pronunciation error detection based on linguistic knowledge and pronunciation space,” in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 4841–4844.

[44] M.-S. Liang, J.-Y. Hung, R.-Y. Lyu, and Y.-C. Chiang, “Pronunciation error detection for computer assisted pronunciation teaching in mandarin,” in *2008 6th International Symposium on Chinese Spoken Language Processing*, 2008, pp. 1–4.

[45] R. Lyu, M. Liang, and Y. Chiang, “Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka, and Mandarin,” in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 2004, pp. 1–12.

[46] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier, “An exact algorithm for F-measure maximization,” *Adv. Neural Inf. Process. Syst.*, vol. 24, pp. 1404–1412, 2011.

[47] G. Zhao *et al.*, “L2-Arctic: A non-native English speech corpus,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, pp. 2783–2787, 2018, doi: 10.21437/Interspeech.2018-1110.

[48] K. Li, X. Qian, and H. Meng, “Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 193–207, 2016.

[49] N. F. Chen, R. Tong, D. Wee, P. Lee, B. Ma, and H. Li, “iCALL corpus: Mandarin Chinese spoken by non-native speakers of European descent,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[50] T. Lander, “CSLU: Foreign Accented English.” Release, 2007.

[51] S. Col, A. LaRocca, and R. Chouairi, “West Point Arabic Speech,” *LDC Cat. LDC2002S02*, 2002.

[52] S. Schaden and U. Jekosch, “‘Casselberveetovallarga’ and other Unpronounceable Places: The CrossTowns Corpus.,” in *LREC*, 2006, pp. 993–998.

[53] S. Weinberger, “Speech accent archive,” *Georg. Mason Univ.*, 2015.

[54] P. Meier, “International dialects of English archive,” *IDEA-The Int. Dialects English Arch.*, 1997.

[55] A. Pettarin, “aeneas is a Python/C library and a set of tools to automagically synchronize audio and text (aka forced alignment),” *GitHub repository*. GitHub, 2017.

[56] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi.,” in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[57] Y. Ren *et al.*, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” in *International Conference on Learning Representations*, 2020.

[58] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, “KSU speech database: text selection, recording and verification,” in

2013 European Modelling Symposium, 2013, pp. 237–242.

[59] M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Towards Deep Object Detection Techniques for Phoneme Recognition," *IEEE Access*, vol. 8, pp. 54663–54680, 2020.

[60] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Tech. Rep. n*, vol. 93, 1993.

[61] M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, and Z. Ali, "KSU rich Arabic speech database," *Inf.*, vol. 16, no. 6 B, pp. 4231–4253, 2013.

[62] A. Al Hindi, M. Alsulaiman, G. Muhammad, and S. Al-Kahtani, "Automatic pronunciation error detection of nonnative Arabic Speech," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 190–197.

[63] M. Algabri, H. Mathkour, M. M. Alsulaiman, and M. A. Bencherif, "Deep learning-based detection of articulatory features in arabic and english speech," *Sensors (Switzerland)*, vol. 21, no. 4, 2021, doi: 10.3390/s21041205.

[64] Y. Seddiq, A. Meftah, M. Alghamdi, and Y. Alotaibi, "Reintroducing KAPD as a Dataset for Machine Learning and Data Mining Applications," in *2016 European Modelling Symposium (EMS)*, 2016, pp. 70–74.

[65] M. Alghmadi, "KACST arabic phonetic database," in *the Fifteenth International Congress of Phonetics Science, Barcelona*, 2003, pp. 3109–3112.

[66] Y. Seddiq, Y. A. Alotaibi, S.-A. Selouani, and A. H. Meftah, "Distinctive Phonetic Features Modeling and Extraction Using Deep Neural Networks," *IEEE Access*, vol. 7, pp. 81382–81396, 2019.

[67] I. Karaulov and D. Tkanov, "Attention model for articulatory features detection," *arXiv Prepr. arXiv1907.01914*, 2019.

[68] S. King and P. Taylor, "Detection of phonological features in

continuous speech using neural networks," 2000.

[69] Y. A. Alotaibi, S.-A. Selouani, M. S. Yakoub, Y. M. Seddiq, and A. Meftah, "A Canonicalization of Distinctive Phonetic Features to Improve Arabic Speech Recognition," *Acta Acust. united with Acust.*, vol. 105, no. 6, pp. 1269–1277, 2019.

[70] M. Algabri, H. Mathkour, M. Alsulaiman, and M. A. Bencherif, "Mispronunciation Detection and Diagnosis with Articulatory-Level Feedback Generation for Non-Native Arabic Speech," *Mathematics*, vol. 10, no. 15, p. 2727, 2022.

[71] S. Young *et al.*, "The HTK book," *Cambridge Univ. Eng. Dep.*, vol. 3, no. 175, p. 12, 2002.

[72] X. Qian, F. K. Soong, and H. Meng, "Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT)," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[73] Y.-B. Wang and L. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 564–579, 2015.

[74] Z. Zhang, Y. Wang, and J. Yang, "Mispronunciation Detection and Correction via Discrete Acoustic Units," *arXiv Prepr. arXiv2108.05517*, 2021.

[75] S. W. F. Jiang, B. C. Yan, T. H. Lo, F. A. Chao, and B. Chen, "Towards Robust Mispronunciation Detection and Diagnosis for L2 English Learners with Accent-Modulating Methods," *2021 IEEE Autom. Speech Recognit. Underst. Work. ASRU 2021 - Proc.*, pp. 1065–1070, 2021, doi: 10.1109/ASRU51503.2021.9688291.

[76] M. Wu, K. Li, W. K. Leung, and H. Meng, "Transformer based end-to-end mispronunciation detection and diagnosis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2, pp. 1471–1475, 2021, doi: 10.21437/Interspeech.2021-1467.

[77] W.-K. Leung, X. Liu, and H. Meng, "CNN-RNN-CTC based end-to-end

mispronunciation detection and diagnosis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8132–8136.

[78] K. Fu, J. Lin, D. Ke, Y. Xie, J. Zhang, and B. Lin, "A Full Text-Dependent End to End Mispronunciation Detection and Diagnosis with Easy Data Augmentation Techniques," *arXiv Prepr. arXiv2104.08428*, 2021.

9. Publications / Presentations

We published the following papers:

1. M. Algabri, H. Mathkour, M. M. Alsulaiman, and M. A. Bencherif, "Deep learning-based detection of articulatory features in arabic and english speech," Sensors (Switzerland), vol. 21, no. 4, 2021, doi: 10.3390/s21041205.

2. Algabri, M., Mathkour, H., Alsulaiman, M., & Bencherif, M. A. (2022). Mispronunciation Detection and Diagnosis with Articulatory-Level Feedback Generation for Non-Native Arabic Speech. Mathematics, 10(15), 2727.

3. Mansour Alsulaiman, et el., Versatile Dataset of Speech Real and Synthesized of Arabic Learners, 3rd International Conference on Computing and Information Technology (ICCIT), TABUK, KINGDOM OF SAUDI ARABIA, 2023**. (Accepted).**

10. Appendices

Appendix A - Text Selection Comparison (V1 to V3)

Appendix B - Selected text for the Arabic CAPT recording system

Appendix C - Tahadath App Screen Cards of Session-1

Appendix D - Durations per Speaker

Paragraphs and SPW for 44 first Speakers

ID	Total(Sec)	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16
1	525.34	19.36	15.74	19.44	13.80	11.72	11.20	17.52	12.44	15.22	12.90	17.42	15.46	19.04	23.84	11.66	10.96	13.50	8.90	16.98	11.58	16.84	15.06	17.88	12.16	10.06	9.22	5.90	7.84	7.92	5.72	12.34	10.52	11.40	9.66	11.00	8.86	6.64	12.68	10.34	14.72	9.90
2	531.72	22.36	21.70	21.08	14.88	12.42	10.28	14.96	14.42	9.62	10.84	17.40	16.98	21.22	17.90	12.20	10.90	15.74	10.32	20.06	14.28	18.90	18.60	16.88	10.14	10.68	8.36	6.56	7.50	8.60	4.92	9.08	9.54	9.54	10.26	11.44	6.56	8.86	12.90	14.56	10.54	
3	779.90	21.56	28.56	24.24	24.16	21.68	17.36	19.02	17.14	22.62	15.28	26.36	24.60	28.74	27.02	15.32	15.72	21.50	13.62	25.48	23.78	16.76	21.38	22.08	27.04	28.86	17.52	15.30	16.68	10.04	15.26	11.64	6.46	13.74	17.36	17.42	17.22	22.80	12.60	13.52	16.04	18.06	20.60	17.52	
4	767.64	22.10	18.00	21.92	18.52	16.60	14.34	16.74	17.34	15.36	15.90	24.22	22.34	26.46	23.08	13.80	14.72	17.86	14.86	23.78	16.76	21.38	22.08	27.04	18.80	10.32	18.40	11.78	13.04	14.74	10.60	22.56	20.90	25.80	18.62	17.38	15.50	11.62	25.34	19.76	26.58	20.70	
5	681.42	21.82	24.90	19.16	19.42	17.10	14.66	15.28	14.58	13.28	13.90	18.60	17.98	21.58	20.14	13.56	12.22	19.48	17.48	23.10	17.92	22.54	25.36	27.36	16.46	9.44	11.78	7.02	10.70	11.36	7.96	15.66	15.68	16.46	14.44	16.80	12.72	9.84	17.42	16.72	23.48	16.02	
6	668.40	16.86	22.04	21.74	19.68	15.56	14.44	18.46	16.68	14.74	12.90	19.00	20.18	27.08	24.32	13.86	13.20	19.06	13.12	21.10	14.26	23.02	19.58	25.98	17.46	9.54	12.40	6.36	9.20	8.60	5.58	16.34	12.68	14.70	13.72	13.56	8.54	9.10	14.86	15.38	19.90	13.62	
7	914.44	25.30	25.06	35.00	25.70	26.76	18.80	24.08	23.16	19.64	24.30	25.36	25.18	33.04	36.58	19.02	21.70	25.90	22.20	32.78	32.72	36.90	24.90	41.06	24.56	24.16	13.18	8.80	12.24	10.76	7.84	21.78	16.74	17.24	11.12	12.70	14.24	7.16	13.08	17.96	34.32	20.62	
8	701.34	20.76	21.50	24.84	20.68	16.92	14.92	17.60	14.56	17.20	14.44	17.50	22.06	19.84	12.56	13.42	16.72	12.00	22.38	17.86	20.52	20.14	25.98	26.30	16.68	8.42	17.48	7.92	14.52	13.56	9.08	23.36	19.54	19.52	17.62	18.26	11.78	11.36	19.58	17.08	25.80	15.86	
9	781.42	21.86	27.16	29.36	27.50	20.80	17.38	24.34	16.06	14.44	17.24	24.12	19.56	32.94	24.68	17.72	15.24	26.58	12.58	31.72	17.24	22.80	25.98	26.30	19.98	9.08	12.26	7.18	13.52	12.14	9.60	18.14	16.80	17.44	18.26	18.14	12.52	12.34	17.44	16.76	19.92	16.20	
10	563.62	17.60	17.42	18.50	13.98	11.82	9.72	11.32	10.38	10.02	9.42	13.60	14.58	17.94	14.86	10.50	9.94	14.38	9.46	18.88	11.28	17.64	15.10	17.88	12.60	7.94	13.68	7.42	10.30	10.24	7.00	17.00	14.02	20.84	15.96	15.18	11.12	10.86	19.32	15.76	22.76	15.40	
11	677.00	20.20	21.00	21.64	17.64	16.22	13.86	14.60	12.88	13.86	12.78	18.98	20.26	26.68	21.32	12.96	11.80	17.30	11.88	23.84	16.96	22.34	19.36	24.50	17.74	8.86	14.38	7.68	11.78	10.40	9.02	16.48	15.96	19.24	15.46	16.86	11.32	10.56	20.14	16.40	25.16	16.70	
12	727.16	20.82	22.54	22.92	20.36	18.80	16.22	20.24	16.58	15.78	16.78	22.18	21.24	29.34	24.62	15.96	15.40	18.50	13.42	27.60	19.10	26.92	23.98	27.38	19.44	15.86	15.34	9.10	12.46	11.46	7.82	15.44	14.36	16.18	13.80	12.92	11.02	9.62	16.88	16.38	19.16	13.24	
13	608.40	15.44	16.68	17.90	16.70	13.82	10.84	12.44	11.86	11.60	10.56	15.98	17.14	21.00	17.14	10.84	11.22	16.12	11.48	20.06	12.70	17.88	15.62	19.24	13.16	11.06	18.06	8.62	12.76	12.44	10.06	15.42	14.96	16.92	16.02	16.40	15.82	13.34	13.48	17.36	17.36	20.42	15.44	
14	567.46	20.20	20.98	18.60	15.66	11.80	10.22	15.38	12.38	13.60	10.56	17.28	15.94	19.12	19.48	10.96	13.32	20.92	15.38	21.90	15.58	22.36	17.38	17.62	13.62	7.42	9.38	4.92	8.96	7.88	5.54	16.58	9.78	12.28	11.82	13.52	9.10	7.26	7.98	9.20	12.00	7.82	
15	562.30	16.40	20.36	22.64	18.92	13.54	12.90	16.72	12.80	14.68	11.96	18.84	22.40	21.88	19.06	11.06	12.30	13.88	10.34	22.72	14.08	18.76	18.42	20.18	12.56	12.48	9.66	6.44	7.80	8.14	5.46	9.28	11.16	9.88	9.40	10.70	10.04	8.34	9.78	12.32	13.38	10.84	
16	801.14	22.12	21.46	20.02	20.84	13.96	13.48	18.00	13.82	13.54	13.10	15.70	20.62	23.44	20.84	13.42	10.32	15.04	11.98	206.24	15.28	24.24	16.54	24.44	15.20	10.26	8.94	6.06	28.12	7.84	5.92	12.06	13.44	13.82	13.76	12.20	7.56	8.68	11.42	10.74	13.98	12.70	
17	700.74	31.70	37.98	42.72	32.36	28.24	26.50	34.86	37.12	38.44	41.86	36.22	10.50	23.88	18.92	16.82	13.78	15.48	12.92	12.72	11.06	16.46	13.38	7.34	10.64	9.68	12.96	7.26	12.68	7.14	5.22	6.92	7.20	11.04	5.70	6.14	6.44	6.60	7.86	6.50	6.16	3.34	
18	746.86	17.56	19.88	23.76	17.04	16.72	17.80	15.22	12.60	18.02	16.38	20.76	17.30	26.72	28.08	19.02	12.36	19.56	14.60	23.82	20.78	28.58	20.56	23.94	18.58	16.72	16.18	8.76	15.14	12.40	8.48	17.32	22.64	16.92	18.14	17.14	13.22	14.58	16.06	18.40	22.78	22.34	
19	551.08	15.46	17.44	16.74	14.00	11.38	10.60	15.90	10.04	9.46	12.40	9.40	18.32	13.42	16.96	15.38	18.12	12.68	7.14	11.30	6.06	8.98	10.56	8.42	17.70	15.96	17.84	13.98	13.72	10.60	10.18	18.94	15.60	22.12	14.38								
20	462.56	16.42	15.72	17.06	15.82	11.50	9.80	11.84	10.08	12.02	10.48	15.74	14.84	16.30	16.72	12.44	10.26	14.84	9.98	18.58	11.72	14.94	16.78	17.36	11.80	11.12	6.20	4.82	8.62	6.52	5.60	7.68	8.20	8.58	8.54	9.30	7.50	6.94	6.44	8.02	7.88	6.76	
21	587.60	19.28	18.10	18.14	17.12	14.56	11.10	12.12	11.22	12.12	9.60	16.82	17.14	21.60	16.90	11.62	11.10	14.90	10.26	20.58	13.58	19.64	19.32	19.92	12.78	6.98	10.86	7.14	10.86	8.98	7.92	18.30	13.20	18.18	15.78	14.64	12.52	10.36	14.78	10.94	21.22	15.62	
22	623.52	17.42	20.64	19.60	16.70	16.12	14.88	17.12	14.56	14.08	14.64	20.36	18.72	23.80	21.34	13.34	11.66	15.30	13.16	19.14	14.48	20.02	17.34	21.44	17.76	14.10	14.30	5.62	9.86	10.52	7.30	14.96	14.02	14.24	14.48	15.18	9.00	9.18	12.90	14.54	17.68	12.02	
23	616.28	21.58	20.10	20.60	18.14	17.10	14.82	14.78	12.88	14.04	14.58	17.26	19.94	23.90	20.70	13.84	13.54	16.38	10.24	22.52	15.54	20.84	19.92	21.48	16.34	10.24	16.12	10.78	10.94	10.82	9.18	19.40	17.62	16.44	13.46	16.08	13.06	12.32	15.66	14.44	21.94	14.48	
24	690.12	21.58	23.64	31.18	21.56	16.56	12.04	18.70	13.74	12.62	19.44	17.82	18.46	22.60	24.30	12.68	16.64	18.90	12.58	21.10	15.92	26.24	20.88	27.86	20.20	12.90	18.48	7.80	12.26	11.06	6.82	16.06	14.32	16.10	12.84	11.90	9.80	9.00	14.58	14.60	19.92	14.44	
25	821.24	24.64	31.84	41.72	28.42	25.36	18.64	24.28	18.56	19.76	21.90	31.04	26.12	34.22	31.52	18.34	20.08	27.68	18.40	32.82	18.24	30.00	24.04	35.90	22.20	19.00	10.78	7.56	12.32	9.70	6.58	11.70	12.36	10.06	10.56	17.16	9.04	8.20	9.72	11.80	15.88	13.10	
26	563.66	17.10	17.70	9.86	16.24	15.90	12.38	17.02	13.92	14.28	11.48	17.08	18.96	20.38	21.86	12.42	12.48	16.78	10.62	21.66	13.58	19.54	17.18	19.24	15.00	8.06	3.76	6.56	9.92	9.28	6.28	13.26	12.54	15.02	11.60	11.46	8.96	7.90	12.38	13.48	17.42	12.92	
27	671.20	23.30	20.74	22.86	16.92	13.88	18.26	14.48	13.94	14.18	13.42	18.08	24.94	21.92	21.26	11.66	10.36	13.94	10.42	20.24	13.62	19.36	20.32	22.90	14.44	8.28	14.08	7.48	11.36	10.40	8.30	18.32	18.24	19.82	15.66	16.02	15.88	12.56	21.56	17.86	26.06	16.56	
28	899.92	25.30	22.58	28.22	22.32	25.42	16.08	28.62	18.68	21.24	19.94	36.26	31.78	29.82	32.98	18.46	16.50	21.98	18.18	32.64	21.96	39.06	27.90	37.70	23.28	12.28	14.50	13.02	13.30	10.82	7.98	22.10	17.32	19.48	17.04	17.50	16.06	11.08	19.30	18.30	33.22	19.62	
29	588.92	22.94	20.16	27.36	19.52	24.94	11.20	15.92	13.42	11.62	14.48	18.26	16.58	21.36	17.88	14.56	11.04	17.64	14.42	19.64	15.30	16.48	18.08	21.66	13.52	16.86	8.10	4.82	6.48	7.32	6.18	8.70	10.70	10.24	13.34	11.02	11.10	8.26	10.32	11.76	15.60	10.14	
30	1000.84	41.00	30.90	43.68	36.62	29.88	21.06	33.22	29.30	21.14	19.78	28.20	32.26	38.04	41.82	22.08	24.72	32.90	16.70	28.64	20.30	36.06	33.18	33.48	24.48	13.02	22.16	10.86	16.12	10.28	8.60	20.14	25.62	20.08	22.40	13.62	16.68	10.10	21.84	15.24	21.54	13.00	
31	569.94	14.46	17.10	16.94	16.32	15.54	10.20	11.26	13.88	13.60	11.54	16.80	16.76	19.72	21.76	11.66	9.96	11.68	8.72	17.76	13.44	21.30	14.92	19.86	12.34	8.70	13.44	7.40	10.12	9.94	6.92	16.16	12.60	15.24	11.34	11.78	11.44	10.66	15.68	15.96	21.94	13.10	
32	932.78	24.18	28.50	45.30	34.56	24.26	17.02	25.96	19.54	25.66	18.94	32.88	28.96	45.68	32.56	18.30	21.94	25.54	16.10	29.20	24.46	32.72	24.92	42.38	21.68	11.54	14.50	15.96	14.58	15.70	8.16	19.50	17.54	20.04	15.70	14.24	11.92	12.06	19.72	18.56	21.82	20.00	
33	856.68	25.12	27.18	34.12	24.08	21.26	17.70	20.76	16.22	17.76	16.32	25.46	26.22	33.88	28.72	18.60	17.12	22.24	18.06	24.82	21.80	27.08	25.88	31.80	19.18	13.04	25.06	9.88	14.92	15.98	10.96	22.26	21.14	23.68	17.44	17.54	10.86	13.70	20.54	20.28	22.72	15.30	
34	700.94	24.02	23.70	28.38	22.80	19.76	13.22	21.56	13.56	13.88	13.18	21.62	21.32	25.70	24.52	15.62	13.42	20.30	12.36	24.20	13.12	23.26	22.60	25.96	14.36	7.90	10.88	10.50	11.22	9.70	8.66	19.64	14.98	20.54	12.04	17.14	11.92	9.84	15.48	13.76	21.14	13.18	
35	691.36	18.14	20.82	18.94	19.72	15.30	14.08	13.22	13.00	13.06	11.90	17.18	19.68	26.30	24.24	16.02	15.06	19.72	15.38	22.32	17.38	24.86	16.64	26.86	21.02	9.52	19.40	9.08	13.44	12.46	8.92	16.82	15.42	19.20	15.26	14.16	12.24	11.58	19.18	15.70	23.04	15.10	
36	655.86	18.20	16.72	21.78	18.24	16.12	13.10	14.78	12.88	14.04	14.58	17.26	19.94	23.90	20.70	13.84	13.54	16.38	10.24	22.52	15.54	20.84	19.92	21.48	16.34	10.24	16.12	10.78	10.94	10.82	9.18	19.40	17.62	16.44	13.46	16.08	13.06	12.32	15.66	14.44	21.94	14.48	
37	670.30	18.64	20.72	21.22	15.96	12.96	11.38	13.32	12.96	12.12	11.98	17.02	17.98	22.22	21.60	11.68	11.88	16.74	11.00	21.84	16.74	19.34	20.88	23.64	16.80	8.86	19.82	8.24	12.52	12.28	8.48	21.66	19.12	21.58	16.66	17.22	11.98	11.30	20.34	17.42	26.02	16.18	
38	1180.36	27.62	30.40	34.68	27.60	32.74	21.52	31.44	33.12	34.56	31.06	36.50	41.22	59.96	54.22	32.02	23.70	44.32	24.74	52.42	25.86	52.60	40.20	57.92	33.82	17.18	19.30	10.42	14.08	13.56	9.56	20.74	18.24	24.66	17.12	22.14	13.88	13.16	20.98	18.20	25.40	17.50	
39	869.70	18.70	26.04	23.86	22.56	25.76	15.20	24.64	15.92	16.40	9.66	78.50	36.94	48.28	40.18	23.78	16.62	30.56	10.62	41.20	18.96	39.72	32.58	39.22	22.98	12.36	11.78	5.54	9.02	8.88	5.22	17.96	13.84	15.28	10.94	12.18	7.18	8.36	14.28	10.54	17.32	10.14	
40	604.14	16.90	17.58	20.06	16.64	14.10	11.78	13.84	13.56	14.16	12.34	17.54	18.14	23.90	22.46	11.92	12.74	15.92	10.82	21.12	13.62	18.74	19.72	21.50	15.92	8.32	14.24	7.60	11.72	10.40	7.04	14.40	12.56	16.78	13.50	11.04	10.56	9.32	16.54	14.34	17.86	12.90	
41	838.84	28.60	25.50	36.06	21.40	19.16	15.64	19.84	17.86	17.06	17.08	23.28	24.26	30.22	27.86	17.26	15.40	20.38	18.86	27.92	18.52	27.08	26.52	28.00	20.26	13.18	16.08	9.54	12.94	12.62	8.98	20.92	16.80	20.02	14.60	22.12	12.68	12.20	17.06	16.68	24.28	17.12	
42	648.50	20.94	18.96	21.70	18.36	16.56	13.16	14.76	13.94	14.58	14.06	22.32	20.14	29.50	25.74	15.24	15.30	18.90	10.50	23.60	18.46	24.88	19.48	23.52	17.66	17.36	13.14	5.76	11.14	9.04	5.90	13.56	12.30	13.62	11.16	12.98	10.78	8.46	12.18	11.06	18.14	9.46	
43	666.46	19.86	29.18	26.44	18.18	18.42	14.74	19.50	15.76	14.72	14.92	21.38	20.00	24.58	25.04	14.32	13.80	15.62	13.84	16.36	16.68	21.16	18.56	26.44	20.76	17.34	12.54	8.50	10.02	8.36	7.50	13.28	12.68	11.46	12.50	11.34	9.98	8.44	11.78	14.18	15.30	11.00	
44	788.08	18.68	18.82	22.32	18.64	16.80	13.32	15.02	14.68	16.10	12.52	20.50	21.14	26.96	20.62	13.10	13.72	17.86	13.38	25.90	18.50	23.60	21.64	25.04	16.76	9.94	25.82	11.88	17.78	17.10	11.94	26.98	22.56	28.44	20.38	22.40	17.08	15.52	27.06	20.08	27.60	19.90	

Appendix E- ELAN Annotation CAPT protocol

In order to unify the checking procedure, Each ELAN-Checker was asked to follow the following protocol: (including an explanatory video in Appendix F)

STEP 1:

1. Open a new eaf file :
2. Include the wavefile of the speakers
3. Import the TextGrid of the same speakers
4. Save the work an speaker EAF file, SAME DIRECTORY

STEP 2:

1. Remove default tier
2. Activate the To_Remove tier

Figure E.1: Elan Tier Selection

1. Listen to the whole segment of the file.
2. The Arabic text is inside the Sentences tier.
3. If any repeated speech is heard:
   a. Try to locate the left and right boundaries of that segment.
   b. Select the tier To_Remove,
   c. Insert the left boundary split
   d. Insert the left boundary split
4. Move to next Speech sentence.

N.B :

Please mark whenever you can:

- Any word not in the text
- Any additional text, at start or end of file.
- Any extra sound (baby, cat, door opening/closing, phone notification,)
- If any speaker is not worth hearing, or understanding, do not continue on its speech, but put a remark in an additional file. (problems.docx).

Appendix F - Tahadath Application User Manual

Tahadth Application User Manual

دليل المستخدم لتطبيق تحدث

نظام حاسوبي لتعليم نطق أصوات اللغة العربية
للناطقين بغيرها

مركز أبحاث الربوتات الذكية
كلية علوم الحاسب والمعلومات

معهد اللغويات العربية

جامعة الملك سعود

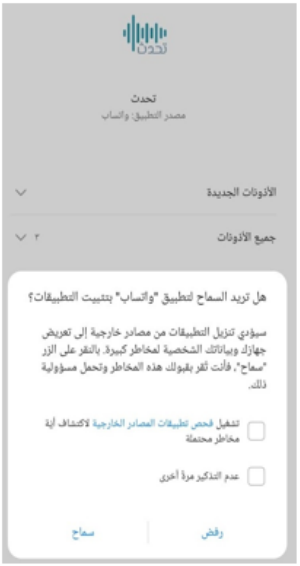
1

إن تطبيق "تحدث Tahadth " يستخدم لتسجيل بعض النصوص العربية.

• لتحميل تطبيق "تحدث Tahadth" إضغط على التطبيق ، ومثل اي تطبيق ستظهر شاشة تحمل المحتوى التالي:

تجاهل محتوى الرسالة ، حيث أن التطبيق آمن طالما تلقيت هذا التطبيق من جهة معروفه.

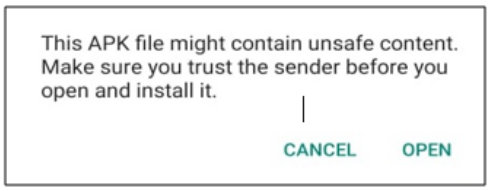
• إضغط على زر "فتح" و بعدها تظهر الشاشة التالية:



• إضغط على زر "سماح" ومن ثم تظهر الشاشة التالية لبدء تثبيت **التطبيق:**

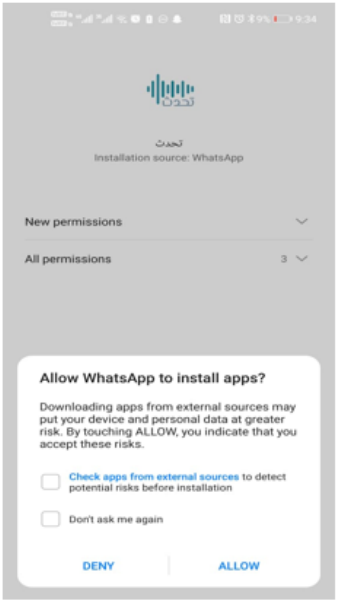
The application "Tahadth" is used to record some Arabic texts.

• To install the "Tahadath" application,
• Click on the application, the following screen will appear:

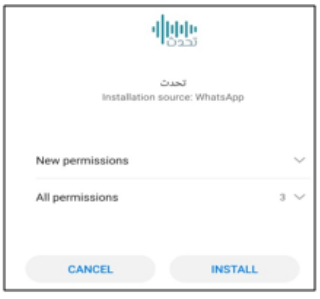


Ignore the message content, this application is safe.

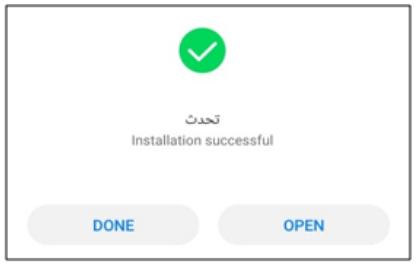
• Click on the "OPEN" button, the following screen will appear:



• Click on the "ALLOW" button, then the following screen will appear:



- Click on the “INSTALL” button to install the application, after that the following screen appears to open the application:



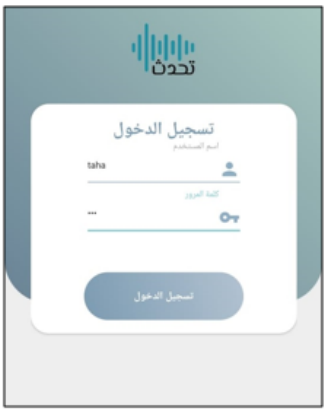
- To open the application, click on the “Open” button, the application has been installed on the device as "Tahadth", the following screen appears:

- Wait until the following login screen appears:

- إضغط على زر "تثبيت" و عند الانتهاء من تثبيت التطبيق تظهر الشاشة التالية لفتح التطبيق:

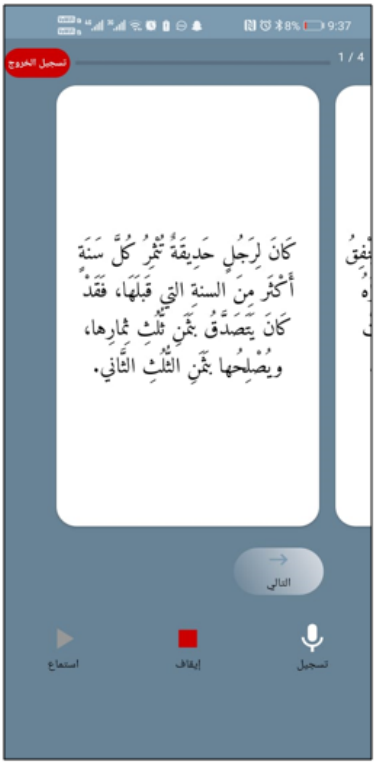
- لفتح التطبيق إضغط على زر "فتح" ، حيث أن التطبيق كذلك تم تثبيته على الجهاز باسم "تحدث" ، ومن ثم تظهر الشاشة التالية:

- إنتظر حتى تظهر شاشة تسجيل الدخول التالية:



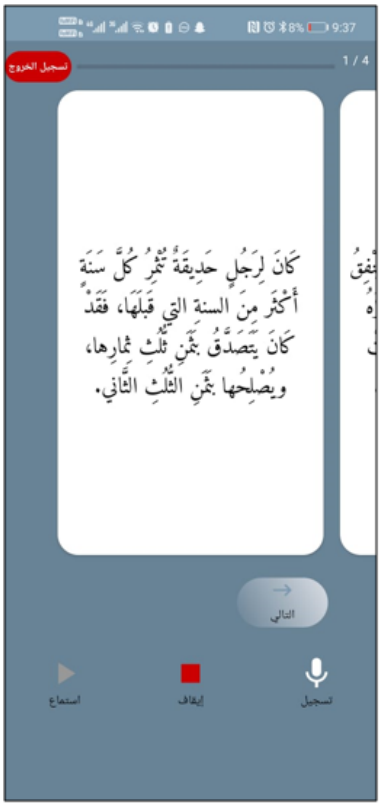
- Enter the user name and password, usually the username is in lowercase letters. (credentials are sent privately to you)
- After the login, the following screen will appear, click on the check box of the phrase "I agree to allow the use of my recordings for research purposes", which confirms your approval to use your recordings for research studies, after that click on "انقر للبدء" button:

- The first screen of the application appears, where the text to be recorded appears in a card:



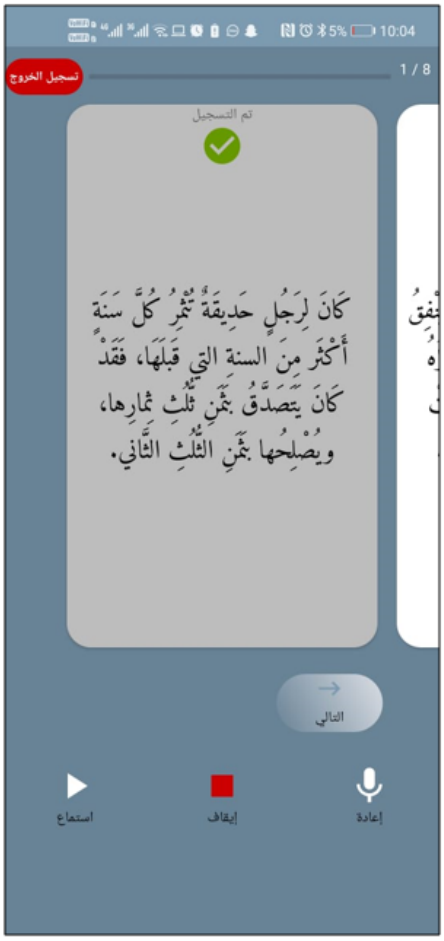
- إدخل اسم المستخدم وكلمة المرور اللذين تم ارسالهما على الخاص وعادةً ما يكون اسم المستخدم بالاحرف الانجليزية الصغيرة.
- بعد عملية تسجيل الدخول ستظهر الشاشة التالية ، قم بالتأشير على عبارة " أقر بالسماح باستخدام التسجيلات لأغراض بحثية" التي تؤكد إقرارك و موافقتك على استخدام تسجيلاتك لأغراض بحثية ، بعدها إضغط على زر "انقر للبدء ":

- تظهر الشاشة الأولى من التطبيق، التي يظهر فيها النص المراد تسجيله:



- Two buttons on the recording screen, a microphone-recording button "تسجيل" and a button to stop recording as a red square "إيقاف". To start the recording process, click on the microphone button and read the text shown on the screen, after you finish click on the stop recording button.

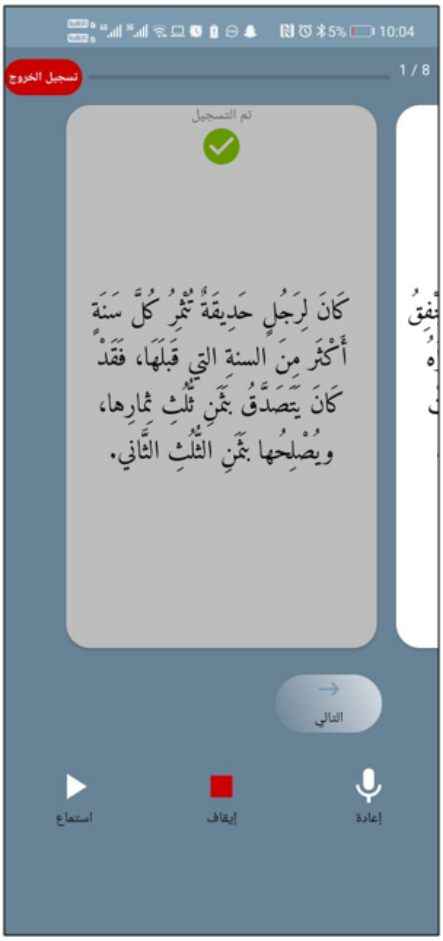
- The recorded sound is automatically approved, and the sign end of recording appears at the top of this screen" تم التسجيل", as shown in the following screen:



- To move to the next text, click on the "التالي" button, or by swiping the screen to the left. You can also listen to your recorded voice by clicking on the "استماع" button.

- في شاشة التسجيل يظهر زر تسجيل بشكل ميكروفون و زر "إيقاف" التسجيل كمربع احمر ، لبدء عملية التسجيل ، إضغط على زر الميكرفون و قراءة النص الظاهر على الشاشة بصوت واضح وعند الانتهاء من قراءة النص إضغط على زر "إيقاف" التسجيل، (يفضل استخدام ميكروفون سماعة خارجية).

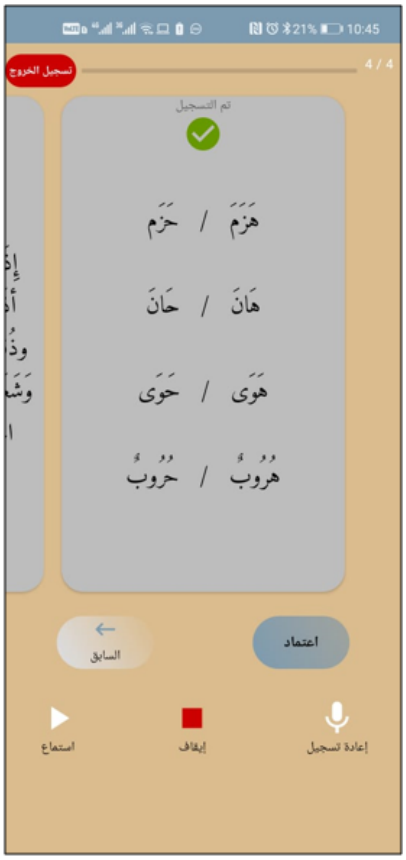
- يتم اعتماد الصوت المسجل آليا و تظهر علامة انتهاء التسجيل لهذه الشاشة و ذلك بظهور رسالة تأكيد إتمام التسجيل في اعلى الشاشة "تم التسجيل" كما هو موضح بالشاشة التالية:



- إنتقل للنص التالي وذلك بالضغط على زر التالي او بسحب الشاشة لليسار. يمكنك قبل الانتقال الاستماع لصوتك المسجل بالضغط على زر استماع.

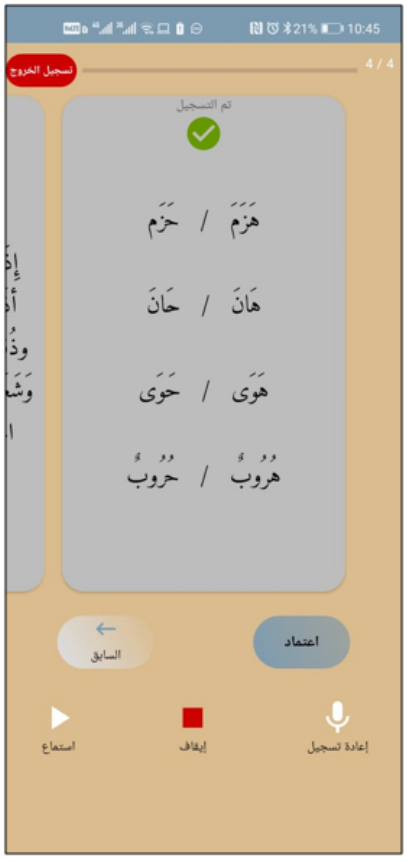
- وهكذا يتم تسجيل نصوص كل الشاشات التالية.
- بعد الانتهاء من تسجيل نص اخر شاشة و في حالة تأكد النظام من تسجيلك لجميع الشاشات ستظهر لك الرسالة المدبقة التالية:

- إضغط على زر "نعم".
- لاعتماد التسجيل إضغط على زر اعتماد الظاهر في اسفل الشاشة والموضح في الصورة التالية:



- Once all the recordings are complete, the following message appears:

- Click on "نعم" button. If you are sure that all screens have been recorded, please click on the approval button "اعتماد" , which appears in the following screen:



- When you click on the approval button "اعتماد", a confirmation screen appears as follows:

- The previous screen represents the recording approval.
- Press the "yes نعم" button, you confirm that you have completed the recording, then you are automatically logged out of the application and back to the login screen.
- Click on the " لا No" option, if you want to listen to the previous recordings. Press the previous buttons to scroll back over the recordings.
- **Note: If you don't press "نعم yes", the recordings are not approved.**
- When you press the "Approval اعتماد" button, and you have some screens that have not been recorded, a pop-up message will appear to inform you of the number of screens that you have not recorded, as in the following example:.

- Press on the "Yes نعم" button, as you will be kept on the last screen. You can return to screens that you did not record by using the " السابق Previous" button to moves between screens
- Once you finish recording all the required texts, go to the last screen and press the approve button "إعتماد" approved to your recordings.

- عند الضغط على زر اعتماد تظهر الشاشة المنبثقة التالية وذلك للتأكد من رغبتك في إعتماد تسجيلاتك .

- إضغط على زر "نعم" لاعتماد تسجيلك و موافقتك على رفع ما قمت بتسجيله ، ويتم بعدها آليا تسجيل خروجك من التطبيق و العودة إلى شاشة الدخول، ولن تستطيع لاحقاً الدخول للنظام .
- اضغط على زر "لا" في حالة ما اذا كنت تريد سماع او إعادة تسجيل اي نص، يمكنك بعد الضغط الرجوع للشاشات السابقة وذلك باستخدام زر "السابق" للتنقل بين النصوص.

ملاحظة : لن يتم اعتماد تسجيلاتك الا بالضغط على زر "نعم" في الرسالة المنبثقة عند ضغطك على زر اعتماد .

- في حالة ما اذا تم الضغط على زر "اعتماد" وكان هنالك شاشات لم يتم تسجيلها ستظهر لك رسالة منبثقة لإبلاغك بارقام الشاشات التي لم تقم بتسجيلها كما في المثال التالي:

- إضغط على زر "نعم" حيث سيتم ابقاءك في الشاشة الاخيرة. يمكنك الرجوع للشاشات التي لم تقم بتسجيلها وذلك باستخدام زر "السابق" للتنقل بين النصوص.
- بعد الانتهاء من تسجيل نصوص الشاشات المطلوبة انتقل إلى الشاشة الاخيرة واضغط على زر "اعتماد" ، لاعتماد تسجيلاتك.

Appendix G - Screen Content of Tahadath App for Session-2