Arabic Mispronunciation Recognition System Using LSTM Network

Subjects: Computer Science, Artificial Intelligence

Contributor: Abdelfatah Ahmed , Mohamed Bader , Ismail Shahin , Ali Bou Nassif , Naoufel Werghi , Mohammad Basel

The widespread use of CALL (computer-assisted language learning) systems attests to their success in helping people improve their language and speech skills. CALL is predominantly concerned with addressing pronunciation errors in non-native speakers' speech. Accurate mispronunciation detection, voice recognition, and accurate pronunciation evaluation are all activities that may be accomplished with CALL.

artificial intelligence deep learning long short-term memory Mel-frequency cepstral coefficients

1. Introduction

The widespread use of CALL (computer-assisted language learning) systems attests to their success in helping people improve their language and speech skills. CALL is predominantly concerned with addressing pronunciation errors in non-native speakers' speech. Accurate mispronunciation detection, voice recognition, and accurate pronunciation evaluation are all activities that may be accomplished with CALL. Similarly, there are a plethora of studies on speech processing that have been implemented in numerous languages with the aim of facilitating language learning. Breakthroughs in AI and other areas of computer science have permitted extensive study of CALL. Due to the inability of their mouth muscles to articulate the intricacies of a particular language, speakers of different languages are prone to committing pronunciation problems while speaking a particular language. For this reason, academics often explore mispronunciation in English, Dutch, and French, while Arabic literary studies are scarce. However, Arabic studies have increased in recent years. Arabic, the most widely spoken language with approximately 290 million native speakers and 132 million non-native speakers, and one of the six official languages of the United Nations (UN), has two major dialects, Classical Arabic (CA) and Modern Standard Arabic (MSA). Classical Arabic is the language of the Quran, whereas Modern Standard Arabic is a modified form of the Quran used in daily conversation. In order to retain the right meaning of the phrases, the rules for pronouncing the Quranic language are quite well-defined.

Table 1 illustrates the most mispronounced Arabic letters in the field of pronunciation. Therefore, the effect of employing long short-term memory as a classifier blended with Mel-frequency cepstral coefficients as the feature extractor is observed. The LSTM network is well suited for speech recognition due to its ability to model the

complex temporal relationships in speech signals, adapt to variations in the input data, and handle sequences of variable lengths.

Table 1. Most common disordered Arabic letters.

No.	Arabic Letter	Phonetic Symbol
1	س	/s/
2	ر	/r/
3	ق	/q/
4	ج	/ʒ/
5	ك	/k/
6	خ	/x/
7	غ	/ɣ/
8	ض	/d/
9	ح	/ḥ/
10	ص	/Ṣ/
11	ط	/ŧ/
12	ظ	/ð/
13	ذ	/ð/

2. Arabic Mispronunciation Recognition System Using LSTM Network

Numerous mispronunciation detection and diagnosis (MD&D) research methods attempt to utilize both auditory and linguistic input elements. However, the absence of a substantial quantity of annotated training data at the phoneme level constrains the improvement of performance. Recent advancements in speech recognition, such as the LAS-Transformer, which is an enhanced Transformer based on the Local Attention Mechanism, suggest potential applications in improving mispronunciation detection by utilizing more advanced attention mechanisms in processing auditory inputs [1]. To construct a more robust MD&D system, authors combined the embedding properties of acoustic, phonetic, and linguistic data, abbreviated as APL. In [2], the suggested method has a detection accuracy of 9.93% higher than the baseline, a diagnosis error rate of 10.13% lower, and an F-measure of 6.17% higher than the baseline. These results were acquired through experimental work performed on the L2-ARCTIC database. The authors introduced a phoneme-level MD&D system that utilizes acoustic embedding

References

1. Pengbin Fu; Daxing Liu; Huirong Yang; LAS-Transformer: An Enhanced Transformer Based on the Local Attention Mechanism for Speech Recognition. *Inf.* **2022**, *13*, 250.
2. Wenxuan Ye; Shaoguang Mao; Frank Soong; Wenshan Wu; Yan Xia; Jonathan Tien; Zhiyong Wu. An Approach to Mispronunciation Detection and Diagnosis with Acoustic, Phonetic and Linguistic (APL) Embeddings; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, United States, 2022; pp. 6827-6831.
3. Kun Li; Xiaojun Qian; Helen Meng; Mispronunciation Detection and Diagnosis in L2 English Speech Using Multidistribution Deep Neural Networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **2016**, *25*, 193-207.

4. Mostafa Shahin; Beena Ahmed; Anomaly detection-based pronunciation verification approach using speech attribute features. *Speech Commun.* **2019**, *111*, 29-43.
5. Moner N. M. Arafa; Reda Elbarougy; A. A. Ewees; G. M. Behery; A Dataset for Speech Recognition to Support Arabic Phoneme Pronunciation. *Int. J. Image, Graph. Signal Process.* **2018**, *10*, 31-38.
6. Sura Ramzi Shareef; Yousra Faisal Muhammad Al-Irhayim; Comparison Between Features Extraction Techniques for Impairments Arabic Speech. *AL-Rafidain Eng. J. (AREJ)* **2022**, *27*, 190-197.

Computer-aided pronunciation training systems necessitate reliable automated pronunciation error detection techniques to recognize human faults. Yet, the overall number of mispronounced speech data utilized to train these algorithms and their manual annotation reliability greatly affect their performance [4]. To resolve this issue, the authors in [4] employed anomaly detection methods to identify mispronunciation. Their anomaly detection model was the One-Class SVM, using phoneme-specific models. A bank of binary DNN speech attribute detectors retrieved manners and locations of articulation for each model. Multi-task learning and dropout were implemented to reduce DNN speech attribute detector overfitting. The model was trained using the WSJ0 and TIMIT standard datasets, which contain solely native English speech data, and then assessed it using three datasets: a native English speaker corpus with fake mistakes, a foreign-accented speech corpus, and a children's disordered speech corpus. Lastly, the proposed approach was compared to the usual goodness-of-pronunciation (GOP) algorithm to prove its efficacy. The technique lowered false-acceptance and false-rejection rates by 26% and 39% compared to the GOP technique. Furthermore, in [5], the authors presented a speech recognition system capable of detecting mispronunciations. The dataset contains 89 students, 46 of whom are female. Ten times 28 Arabic phonemes are voiced. MFCCs are retrieved from 890 utterances for modeling with five different machine-learning models. K nearest neighbor (KNN), support vector machine (SVM), naive Bayes, multi-layer perceptron (MLP), and random forest (RF) are some of them. The experimental findings show that the random forest approach achieves an accuracy rate of 85.02%, which is higher than that of other machine learning models. Shareef et al. [6] emphasize the extensive comparison of feature extraction algorithms for the purpose of identifying impaired Arabic speech. The feature extraction approach is based on several wavelet transformation variants. LSTM and CNN-LSTM models are built to identify the impairment in Arabic speech. The combination of MFCCs and LSTM achieves the highest classification accuracy (93%), followed by CNN-LSTM (91%).

Table 2 summarizes the various approaches used in these studies for mispronunciation detection and diagnosis. However, these studies have largely overlooked the influence of gender on pronunciation, an aspect that may hold key insights for personalized language learning.

Table 2. Comparative analysis of different mispronunciation detection and diagnosis methods.

Work	Classification Algorithm	Data Utilized	Performance Metrics	Results
Ye et al. [2]	Acoustic, Phonetic, and Linguistic Data Embedding	L2-ARCTIC database	Detection Accuracy, Diagnosis Error Rate, F-Measure	Accuracy: 9.93% DER: 10.13% F-measure: 6.17%
Li et al. [3]	Acoustic-Graphemic Phonemic Model (AGPM) Using Multi-Distribution Deep Neural Networks (MD-DNNs)	Not specified	Phone Error Rate (PER), False Rejection Rate (FRR), False Acceptance Rate (FAR), Diagnostic Error Rate (DER)	PER: 11.1%, FRR: 4.6%, FAR: 30.5%, DER: 13.5%
Shahin and Ahmed [4]	One-Class SVM, DNN Speech Attribute Detectors	WSJ0 and TIMIT standard datasets	False-Acceptance Rate, False-Rejection Rate	Lowered FAR and FRR by 26% and 39% compared to the GOP technique
Arafa et al. [5]	Random Forest (RF)	89 students' Arabic phoneme utterances	Accuracy	85.02%
Shareef and Al-Irhayim [6]	LSTM and CNN-LSTM	Not specified	Classification Accuracy	LSTM: 93%, CNN-LSTM: 91%

Researchers emphasize the prominence of "speech signal processing" in diagnosing Arabic mispronunciation using the "Mel-Frequency Cepstral Coefficients" (MFCCs) as the optimum extracted features in their proposed system. In addition, Long Short-Term Memory (LSTM) has also been utilized for the classification process. Furthermore, the analytical framework has been incorporated with a gender recognition model to perform two-level classification. The results show that the LSTM network significantly enhances mispronunciation detection along with gender recognition. The LSTM models have attained an average accuracy of 81.52% in the proposed system, reflecting a high performance compared to previous miss pronunciation detection systems.