Deep learning for Quranic reciter recognition and audio content identification

Mohamet TALL[1,2], Thierno Ibrahima DIOP[1], Ndeye Fatou NGOM[2], and El hadj Abdoulaye Thiam[1]

[1] Baamtu, Sénégal
[2] LTISI laboratory, Ecole Polytechnique de Thiès, tmohamet@ept.sn

Abstract. This paper presents a novel approach for identifying the reciter, sura, and verse of a given Quranic passage using pre-trained embedding models and transfer learning. Our approach involves training a deep learning model on a large Quranic recitation audio recordings dataset and using the resulting embeddings to compare and classify different reciters. We also present a workflow for identifying the specific sura and verse of a Quranic passage using a speech-to-text model and elastic-search for the query. We evaluate our approach using a variety of metrics and demonstrate its effectiveness in accurately identifying the reciter, sura, and verse of a given passage. We discuss the potential applications of this approach in the fields of Quranic studies and Islamic education and outline directions for future work in this area.

Keywords: Speaker identification · Speaker recognition · Deep learning · Transfer learning · Audio embedding

1 Introduction

The automatic identification and verification of speakers through representative audio continue to gain the attention of many researchers with diverse domains of applications [11]. Speaker recognition is categorized into speaker identification and speaker verification. Verification is the task of automatically determining if a person is a person. Identification is the mapping of a speech signal from an unknown speaker to a database of known speakers. With existing audio datasets, there are a lot of research on speech and speaker recognition as well as the recognition and identification of imitation sound clips of a speaker. Speech processing is more complicated than other pattern recognition tasks such as text classification and image recognition. Speech recognition is a multidisciplinary domain that includes statistics, signal processing, phonetics, linguistics, and deep learning. Although considerable research has been devoted to English speech recognition and emotion recognition system based on speech signals [4], [5], [16], less attention has been paid to the Arabic speech recognition [1].
This study presents a novel approach for accurately identifying the reciter, sura, and verse of Quranic passages using deep learning techniques. One key contribution of this study is the use of pre-trained embedding models to classify different

reciters, which demonstrates high accuracy in identifying the reciter. The proposed workflow involving speech-to-text and elastic-search to identify the sura and verse of a given passage offers a number of benefits. The presented approach has potential applications in Quranic studies and Islamic education, such as the creation of digital Quranic libraries or educational tools. The study also suggests directions for future work, including the exploration of different audio embedding models and the use of larger or more diverse data-sets, as well as the potential to extend the approach to other languages or text-based data-sets. Overall, these results demonstrate the feasibility of using deep learning techniques to accurately identify the reciter, sura, and verse of Quranic passages. This also have the potential to make a significant contribution to the fields of Quranic studies and Islamic education.
The rest of the paper is organized as follows. Section 2 discuss previous research and identify gaps in the literature. Section 3 describes the research design and data collection methods, as well as the techniques used for reciter classification and sura and verse identification. Section 4 describes the process of identifying the sura, verse and challenges encountered. Section 5 presents the results of the reciter classification experiments and discusses notable trends or patterns, summarize the main findings, and discusses the implications and applications of the results. Section 6 summarizes the key contributions and suggests directions for future research.

2 Related work

The Holy Quran is considered the primary reference to approximately 1.6 billion Muslims around the world [13]. Recitation and listening to the Holy Quran are essential activities of a Muslim. There are many known Quranic recitations or reading methods. Tajweed is a set of rules to read the Quran in a correct Pronunciation of the letters with all its Qualities while Reciting the Quran [2]. Two well-known narrations (Rewayah) exist in each recitation, which has been authenticated and passed down by qualified and experienced scholars (Sheikhs) to their students. The most popular recitation is that of Hafs Bin Suleiman, on the authority of Asim Al-Koufi, which is being recited in Arabia, Egypt, India, Pakistan, a Turkey [11]. The sacred Arabic text is the Holy Quran and 78,000 words of the Quran form 114 chapters (Sura). A word denotes many concepts (polysemy) and a concept can be denoted by many words (synonyms). Challenging points regarding the structure of the Holy Quran exist when searching for certain information on Quran reciters or retrieves verses.
A system for factoid questions specialized in Islamic sciences such as prophetic tradition (Hadith), Hadith narrator, and Quran interpretation (Tafsir) is proposed in [12]. The proposed QA System follows a symbolic approach composed of query formulation, information search, and answer extraction and treatment. In [17], Continuous Hidden Markov Models (CHMMs) were used to identify Arabic speakers automatically from their voices and the Mel-Frequency Cepstral Coefficients (MFCCs) were selected to describe the speech signal. In the

text-independent experiments, the identification rate was found to be 80%. An automated system can assist in identifying the specific voice of a Qari from numerous available offline and online recitations. In [19], a support vector Machine (SVM) was chosen by the author as a classifier model for identifying narrator name entities in Hadith documents. In [8], the authors analyzed the recitation of the Twelve Qari, reciting the last ten Surah of the Quran, thus representing a 12-class problem. The Mel-frequency Cepstral Coefficient (MFCC) and pitch were used as the features for model learning. The features were learned with the naive Bayes, j48, and random forest, being selected due to their overall excellent performance in the state of the art.
In [6], the authors show that speech representations extracted from a specific type of neural model (i.e. Transformers) lead to a better match with human perception than two earlier approaches on the basis of phonetic transcriptions and MFCC-based acoustic features [6]. They furthermore find that features from the models can generally best be extracted from one of the middle hidden layers than from the final layer. In [14], the authors propose a simplified deep-learning approach to accomplish the speaker identification task using as little training data. To represent the feature vectors of over 4,000 speakers using approximately 343 hours of speech signals, the MFCC method was utilized. Bidirectional LSTM neural networks provided up to 76.9% accuracy rate for individual voice segments, and 99.5% when considering the segments of each speaker as a bundle. In [3], the authors introduce an isolated word speaker identification system based on a new feature extractor and using an artificial neural network. The classification of the features, extracted MFCC, is done using Multi-layer perceptron with back-propagation algorithm.
This study presents a novel approach for accurately identifying the reciter, sura, and verse of Quranic passages using deep learning techniques. One key contribution of this study is the use of pre-trained embedding models to classify different reciters, which demonstrates high accuracy in identifying both the Quran reciter and audio content.

3 Methodology

3.1 System design

The proposed workflow (Figure 1) is mainly based on a fine-tuned deep audio embedding model that extracts the acoustic characteristics of the input audio and a part represented by the classification layer whose purpose is to classify the reciter. Fine-tuning a network is a procedure based on the concept of transfer learning.
The fine-tuning consists of implementing first a transfer learning by configuring the pre-trained model layer with the parameter *trainable=False*. We train our model by adding some layers and the classification layer. Once convergence is reached, we change the previous parameter to *True* and then re-train the model with a low learning rate. It is important to note that it is possible to have overfitting. Fine-tuning a network is a huge advantage because it allows initializing

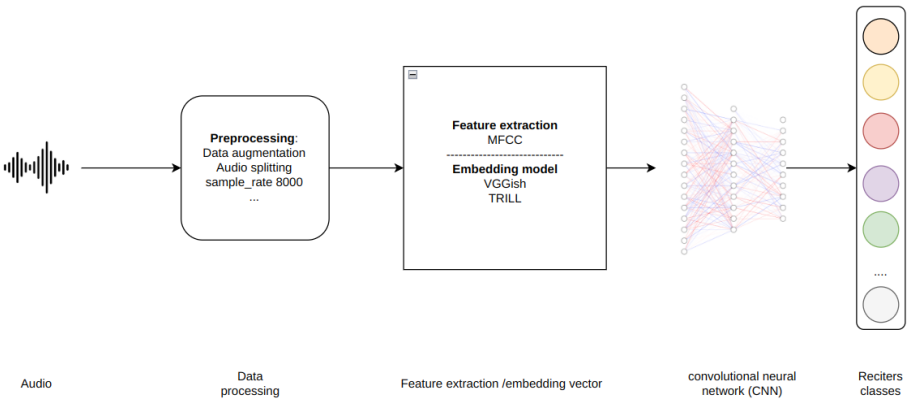


Fig. 1. Classification of Quran reciters using a fine-tuned deep audio embedding model.

the parameters of our neural network with those of a pre-trained model. We can expect a fast convergence and an increase in the learning precision of our classification model for our downstream task, reciter identification. For some types of problems, it can be difficult to find a large volume of data to implement our deep learning model. Transfer learning can be used to train our deep learning model with a low volume of data. Deep audio embedding methods allow us to efficiently capture high-dimensional features into a compact representation. The power of deep audio embedding is to automatically identify predominant aspects in the data at scale [10].

3.2 Dataset

The dataset is obtained by downloading audio from sheiks through several open-source platforms. The raw data used are mainly in mp3 format of variable length, divided into folders for each sheik (a reciter of the Quran). Each folder contains the audio files of the different suras recited by the sheik. The global dataset contains 169 sheiks, i.e. 169 folders representing our model's classes. For each sheik and each of his suras, the sound is cut according to the silence. Indeed, the psalmodist makes momentary pauses to mark the end of a verse or to apply the rules of the stop during the recitation.
We also added the last class composed of random sounds (music, discussion, ambient noise, etc.) to allow our model not to classify a random song as a sheik. So in total, we will have 170 sheiks for our experiment.

3.3 Data preprocessing

For audio samples, we use a sample rate of *8000 Hz*. The audio sample rate is a measurement of the samples per second taken by the system from a continuous digital signal.
There are several ways of reciting the Quran. These different readings are methods of vocalization that arise from the addition of diacritical marks on the consonantal Quranic skeleton. We distinguish several riwayat [18]. In the dataset, we have only two of them. However, there are some sheiks who have two classes with different riwayats. These classes are grouped into one since we do not predict the riwayat but the sheik in person. This increases the number of classes from 169 to 170.
Data augmentation is a common strategy adopted to increase the quantity of training data, avoid over-fitting and improve the robustness of the models [9]. We added different random noises in the audios of our data-set to not allow our model to learn with too perfect data. Indeed, the audios in our data-set were recorded in a studio, and wanting to train our classifier with these data can create a high bias. Adding noise to our training dataset can have a regularizing effect and reduce over-fitting.

3.4 Pre-trained embedding models

In this section, we will introduce the pre-trained embedding models that we used in our experiments. Embedding models are machine learning algorithms that transform input data, such as audio recordings, into a fixed-length vector representation. These vectors, known as embedding, capture the important characteristics of the input data and can be used for a variety of tasks, such as classification and comparison. The embedding models we used have been trained on large data-sets of audio recordings, which allows them to effectively extract the key features of an input sound. As a result, they are an efficient and effective way to represent and analyze audio data.

TRILL is an embedding model trained in a semi-supervised way on audio clips of the audio-set [7, 15] data-set proposed in the article *Towards Learning a Universal Non-Semantic Representation of Speech* [15]. *TRILL* transforms audio into a fixed-size vector with the principle that sounds that are close in time are also close in embedding space. In the original paper [15], the authors show that this definition of the loss function (*triplet loss*) is very efficient to learn a robust representation for several non-semantic tasks. *TRILL* was evaluated on the *Non-Semantic Speech Benchmark (NOSS)* [15], designed by the same authors, by training several small models and comparing their performances. By performing transfer learning for several sub-tasks, the models obtained give good results on the proposed benchmark. With transfer learning, the authors have proposed a new state-of-the-art for several tasks, thus outperforming several previously proposed methods.

VGGish is a pre-trained audio embedding model developed by researchers at Google. It is designed to extract features from short (1-second) segments of audio and has been trained on a data-set of approximately 2 million YouTube audio recordings. The VGGish model converts the raw audio waveform into a spectrogram representation using a short-time Fourier transform (STFT) and then applies a series of convolutional and fully connected layers to the spectrogram to generate embeddings. These embeddings capture the key characteristics of the input audio and can be used for tasks such as music classification, speaker identification, and sound event recognition. The VGGish architecture consists of a series of convolutional layers, which apply filters to the spectrogram and extract relevant features, followed by fully connected (dense) layers, which refine the features and generate the final embeddings. The embeddings are then passed through a layer with a softmax activation function, which converts the embeddings into a probability distribution over a set of classes. Overall, the VGGish model is effective at extracting meaningful features from audio data and generating embeddings that can be used for a variety of tasks.

3.5 Features extraction with Mel Frequency Cepstral Coefficients (MFCC)

The MFCC (Mel Frequency Cepstral Coefficients) is a method of extracting characteristics of the signal developed around the Fast Fourier Transform (FFT) and the Discrete Fourier Transform (DCT) on a Mel scale. The Mel scale is a logarithmic transformation of the frequency of a signal. The fundamental principle of this transformation is that sounds of equal distance on the Mel scale are perceived as being of equal distance to humans. The transformation from the Hertz to Mel scale is given by the following formula:

$$m = 1127.log(1 + \frac{f}{700})$$

The calculation of the MFCC is done through the following steps:

1. Split the signal between several overlapping windows (windowing),
2. To reduce the spectral distortion created by the overlap, apply a Hamming window on the signal,
3. apply the Fast Fourier transform to the window (which gives the spectrum),
4. move on to the previously explained Mel scale,
5. finishes with the conversion of Mel's logarithmic spectrum into time using the Discrete Cosine Transform. This reduces the number of data characterizing the signal and we obtain a limited number of cepstral coefficients (13 for our experiment) per window.

The MFCC is used for automatic speech recognition or for audio denoising. The number of coefficients used is a hyper-parameter of the model. It can vary according to the problem to be solved. To implement the MFCC with speaker identification, we can feed a neural network with the extracted features (MFCC) as input and the number of speakers to identify as the classification layer.

4 Verses and sura identification

To identify the sura and the verse using the audio file as input, we follow different steps of preprocessing, content extraction, and database query (Figure 2). Indeed, first, we apply different layers of pre-processing by adapting the format and the different intrinsic characteristics such as the *sample rate*, etc. Then, we use a *Speech-To-Text model* [3] specifically trained with the Quran data which returns a textual output. However, the Quran is a sacred text so we have to make sure that we have consistent and correct results. The last step is to return the correct part of the Quran extracted from the audio file. This step consists in querying an *elasticsearch* [4] instance which contains all the verses of the Quran and is extracted according to a defined format. For each verse, we have the verse number, the textual content in Arabic, the transcription in Latin, the translation in several languages, and the information of the concerned sura. In summary, the workflow presented in Figure 2 consists of extracting an audio file's textual content with a *speech-to-text* model and then querying an *elasticsearch* instance to obtain the results.

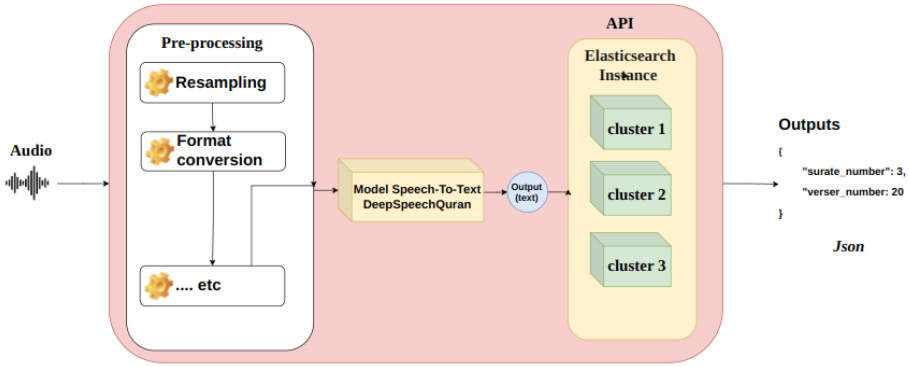


Fig. 2. Sura and verse identification workflow

5 Experimental results

5.1 Experimentation

With the above-mentioned models, we trained a neural network with as output the classification layer i.e. the reciters of the Quran to classify. With the feature

[3] https://github.com/tarekeldeeb/DeepSpeech-Quran
[4] https://www.elastic.co/fr/elasticsearch/

extraction methods, we trained a convolutional neural network and for the embedding models an extraction mode where we will use the model, extract the embedding vector, and then ingest it into our convolutional neural network.
For the experiments, we used a computer with the following characteristics:

- processor: Intel Core i7-7700HQ CPU 2.80GHz x 8
- Graphic cards: GPU NVIDIA Corporation GP107M [GeForce GTX 1050 Mobile] / NVIDIA GeForce 4 Go

	Name	Runtime	accuracy	best_epoch	best_val_loss	epoch	loss	val_accuracy	val_loss
0	VGGish	47097	0.684158	97	0.935010	99	1.124289	0.744445	0.946231
1	MFCC	5178	0.881004	88	0.336364	99	0.445884	0.903174	0.373936
2	TRILL	25821	0.975997	70	0.126861	99	0.115250	0.974353	0.136676
3	TRILL	30864	0.980501	162	0.125846	11	0.096234	0.975837	0.133122
4	VGGish	37980	0.689695	80	0.963028	99	1.101851	0.738093	0.975881

Fig. 3. Model performance

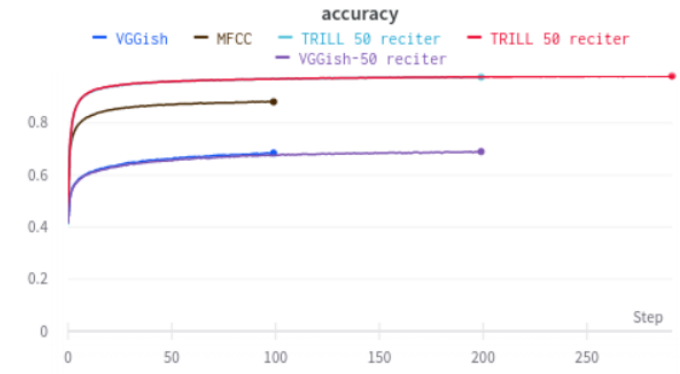
Figure3 gives detailed information on the performance of the models. Figure 4 shows that *Trill* model performs better than the other models in terms of accuracy and loss.

5.2 Model performance analysis

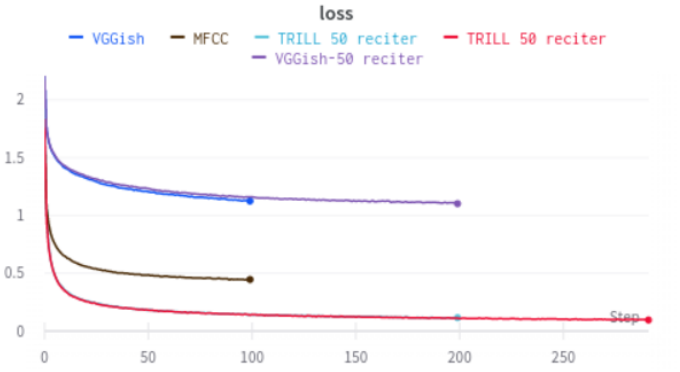
The analysis of the results of the previous experiments allows us to choose the most suitable model to classify the sheiks. Indeed, after convergence, the Trill model has an accuracy of 98% while the *MFCC* and the *VGGish* have respectively 88% and 68%. The validation accuracy is 97.58% for the *TRILL* model, 90% for the *MFCC*, and 73% for the *VGGish*. For the loss *(Sparse Categorical Cross-entropy)*, the Trill model has a value of 0.096, the *MFCC* 0.44, and the *VGGish* model 1.11. For the validation loss, Trill has a value of 0.13, 0.37 for the *MFCC*, and 0.99 for the *VGGish*. On the other hand, Trill has a much higher training time with a time equal to 2h 5m 15s than *MFCC* which is 1h 14m 22s. *VGGish* takes much longer with 4h 6m 57s. The Trill model converges faster followed by *MFCC*. In sum, the results obtained show that the *Trill* model performs better than the other models in terms of accuracy, loss, validation accuracy, and validation loss with sufficient training time and graphics card usage.

5.3 Discussion

One of the key contributions of this study is the use of pre-trained embedding models to classify Quranic reciters. These models have been shown to be effective



(a) Accuracy



(b) Loss

Fig. 4. Analysis of the performance

at extracting meaningful features from audio data and generating embeddings that can be used for a variety of tasks. By training our model on a large dataset of Quranic recitation audio, we were able to achieve high accuracy in identifying the reciter of a given passage, demonstrating the effectiveness of our approach. In addition to the use of embedding models, the proposed workflow for identifying the sura and verse of a Quranic passage using a speech-to-text model and elasticsearch offers a number of benefits. Speech-to-text models are able to accurately transcribe spoken language into written text, making it possible to search for specific passages within the Quran. By combining this transcription with the use of elasticsearch, we were able to quickly and efficiently identify the sura and verse of a given passage.
The proposed method can be used to create digital Quranic libraries, allowing users to easily search and identify specific reciters, suras, and verses. It also can be used to create educational tools that help students and scholars learn more about the Quran and the various styles of recitation.
The experimental results elastic search feasibility of using deep learning techniques to accurately identify the reciter, sura, and verse of Quranic passages. We believe that our approach has the potential to make a significant contribution to the fields of Quranic studies and Islamic education and to provide a foundation for future research in this area.

5.4 Application

The proposed method of Quranic reciter recognition using deep learning can be applied to the development of a platform like *Shazam* [5], but for the Quran, to identify the reciter and specific verse being recited. The platform can improve the understanding of the Quranic text and recitation, provide personalized learning experiences for students to learn from their favorite reciters, and enhance recitation skills with feedback from the algorithm. Ultimately, such a platform has the potential to revolutionize the way people engage with the Quran and its recitation.

6 Conclusion

In this study, we presented a novel approach for accurately identifying the reciter, sura, and verse of a Quranic passage using deep learning techniques. We demonstrated the effectiveness of using pre-trained embedding models to classify different reciters and showed that our approach was able to achieve high accuracy in identifying the reciter of a given passage. Additionally, we described a workflow for using a speech-to-text model in conjunction with elasticsearch to identify the specific sura and verse of a Quranic passage. Our results have the potential to be applied in a variety of settings, including Quranic studies and Islamic education, and provide a foundation for future research in this area.

[5] an application that can identify music

Overall, our approach offers a promising solution for accurately and efficiently identifying the reciter, sura, and verse of Quranic passages.
In future research, it would be valuable to explore the use of different deep learning architectures or pre-trained embedding models in order to potentially improve the accuracy of the reciter classification presented in this study. It could also be interesting to investigate the possibility of using this approach to detect the specific riwayat, or tradition of Quranic recitation, associated with a given passage. This could potentially be achieved through the incorporation of additional data sources or the use of specialized deep learning techniques.

References

1. Al-Anzi, F., Abuzeyna, D.: Synopsis on arabic speech recognition. Ain Shams Engineering journal **13**, 101534 (2022)
2. Alagrami, A., Eljazzar, M.: Automatic recognition of arabic quranic recitation rules
3. Antony, A., Gopikakumari, R.: Speaker identification based on combination of mfcc and umrt based features. Procedia Computer Science **143**, 250–257 (2018)
4. Aouani, H., Ayed, Y.: Speech emotion recognition with deep learning. Procedia Computer Science **178**, 251–260 (2020)
5. Atmaja, B., Sasou, A., Akagi, M.: Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech communication **140**, 11–28 (2022)
6. Bartelds, M., De Vries, W., Sanal, F., Richter, C., Liberman, M., Wieiling, M.: Neural representations for modeling variation in speech. Journal of phoneticc **92**, 101137 (2020)
7. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)
8. Khan, R., Qamar, A., Hadwan, M.: Quranic reciter recognition: a machine learning approach. ASTESJ **4**, 173–176 (2019)
9. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Sixteenth annual conference of the international speech communication association (2015)
10. Koh, E., Dubnov, S.: Comparison and analysis of deep audio embeddings for music emotion recognition. arXiv preprint arXiv:2104.06517 (2021)
11. Lataifeh, M., Elnagar, A.: Arabic diversified audio dataset. Data in Brief **33**, 106503 (2020)
12. Maraoui, H., Hadar, K., Roomary, L.: Arabic factoid questions-answering system for islamic sciences using normalized corpora. Procedia computer science, 25th International Conference on Knowledge based and intelligent information and engineering system **192**, 69–79 (2021)
13. Mohamed, E., Shokry, E.: Qsst: a quranic semantic search tool based on word embedding. Journal of King Saud University Computer and Information Sciences **34**, 934–945 (2022)
14. Nammous, M., Saeed, K., Kobojek, P.: Using a small amount of text-independent speech data for a bilstm large scale speaker identification approach. Journal of King Saud University Computer and Information Sciences pp. 764–770 (2022)

15. Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., Quitry, F.d.C., Tagliasacchi, M., Shavitt, I., Emanuel, D., Haviv, Y.: Towards learning a universal non-semantic representation of speech. arXiv preprint arXiv:2002.12764 (2020)
16. Singkul, S., Woraratpanya, K.: Vector learning representation for generalized speech emotion recognition. Heliyon **8**, e09196 (2022)
17. Tolba, H.: A high-performance text-independent speaker identification of arabic speakers using a chmm-based approach. Alexandria Engineering Journal **50**, 43–47 (2011)
18. wikipedia: Lectures du Coran (2022), https://fr.wikipedia.org/wiki/Lectures$_du_Coran$
19. Yusup, F., Bijaksana, M., Huda, A.: Narratorś name recognition with support vector machine for indexing indonesian hadith translation. Procedia Computer Science, 4th International conference on computer science and computational intelligence **157**, 191–198 (2019)