Transactions on Machine Learning and Artificial Intelligence - Vol. 9, No. 6
Publication Date: December, 25, 2021
DOI:10.14738/tmlai.96.11039.
Mohamed, O., & Aly, S. A. (2021). ASER: Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset. Transactions on Machine Learning and Artificial Intelligence, 9(6). 01-08.

ASER: Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset

Omar Mohamed
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt

Salah A. Aly
Computer Science Section, Faculty of Science, Fayoum University, Egypt

ABSTRACT

Recently, there have been tremendous research outcomes in the fields of speech recognition and natural language processing. This is due to the well-developed multilayers deep learning paradigms such as wav2vec2.0, Wav2vecU, WavBERT, and HuBERT that provide better representation learning and high information capturing. Such paradigms run on hundreds of unlabeled data, then fine-tuned on a small dataset for specific tasks. This paper introduces a deep learning constructed emotional recognition model for Arabic speech dialogues. The developed model employs the state of the art audio representations include wav2vec2.0 and HuBERT. The experiment and performance outcomes of our model overcome the previous known results.

Keywords: Deep Learning; Machine Learning; Emotion Recognition; Arabic Speech; automated speech recognition (ASR).

INTRODUCTION

Researchers and scientists have used deep learning as Feature extraction in recent decades, utilizing the power of deep learning to capture important features that have resulted in significant improvements in literature and real-world applications.

There are several hitches when pursuing research work in emotion recognition:

- emotion detection of humans does not reply on the text or on their speeches, but also in the way they talk to other persons,
- the leak of the available dataset sets particularly in Arabic dialogues,
- emotion detection does not depend only on one word but also in several words in the contexts,
- Some words can be used in different styles, in which they express the speakers' attitude and emotion.

The prosodic properties of human speeches are represented by acoustic features such as pitch, intensity, duration, and voice quality.

Despite the enormous success contributions in emotion recognition in English datasets, there is still a gab in Arabic dataset and emotion recognition systems utilizes these Arabic datasets.

Various Arabic speeches emotion datasets have been proposed in the literature, whether audio or visual, see [1]-[4].

There are well-known several datasets for English, Basic Arabic Vocal Emotions Dataset (BAVED) is a dataset that contains Arabic words spelled in different levels of emotions recorded in an audio/wav format[5].

The problem of emotion recognition analysis on written text or audio speeches has an immense impact and can affect many sectors in the society and relations between persons. A Multi-Task Learning Emotion recognition system has been proposed to detect hate speeches and offensive languages, see[6] and the references therein. This work can also be extended to detect wav audio hate and offensive speeches.

The paper structure is described as follows. In Sections 2 and 3, we introduce the related work and note about the BAVED dataset, respectively. In Section 4 we describe the proposed mispronunciation error detection model. In Section 5 we present simulation studies for the proposed model, and finally, the paper is concluded in Section 6.

RELATED WORK

Recent tremendous results in speech emotion recognition (SER) have been focused on the utilizations of deep learning and convolutional networks [7],[9],[10],[12]. The task is also investigated in Arabic speech emotion recognition (ASER) in several recent results [13]-[14].

The problem has a business side effect in the case of customer's satisfactions and given services. For example, the model can measure if the customers are satisfied about certain products in the market. The system can also be used to happiness and sadness of persons by listening many of their conversations.

Attention-based deep neural networks (DNNs) are employed to give better results than classical neural networks. Klaylat etc. proposed Arabic emotion recognition system based on a data TV news for three labeled emotions: happy, angry or surprised [15]. In their work, classification models are proposed that gave approximately 90% Accuracy.

Recent progress in emotion recognition for certain Arabic dialect has been also investigated, see for example [16],[17].

KSU Emotions was developed by King Saud University (KSU) and contains approximately five hours of emotional Modern Standard Arabic (MSA) speech from 23 subjects. Speakers were from three countries: Yemen, Saudi Arabia and Syria[3].

Klaylat in [18] described an Arabic dataset that consists of eight videos of live calls between an anchor and a human outside the studio were downloaded from online Arabic talk shows. Each video was then divided into turns: callers and receivers. To label each video, 18 listeners were asked to listen to each video and select whether they perceive a happy, angry or surprised emotion. Silence, laughs and noisy chunks were removed. Every chunk was then automatically divided into 1 sec speech units forming our final corpus composed of 1384 records.

Arabic BAVED Dataset

Despite the enormous success contributions in emotion recognition in English datasets, there is still gab in Arabic dataset and emotion recognition systems utilizing these Arabic datasets. Some Arabic speeches emotion datasets have been proposed in the literature, see [1],[3],[5], [19]. Each dataset has a different set of classes or labels, for example, the Arabic audio acted dataset proposed in [20] has five labels (Happiness, Sadness, Neutral, Anger, Fear), and the dataset proposed in [15] has three classes (Happy, Surprised, and Angry), while the dataset proposed in [19] has labels (Happy, Sad, Neutral, Angry, Surprise, Disgust).

0- اعجبني =like
1- لم يعجبني=dislike
2- هذا = this
3- الفيلم = film
4- رائع = good
5- مقبول = neutral
6- سيئ = bad

Figure 1. BAVED defined Words, see[5]

BAVED dataset is a collection of audio/wav recorded Arabic words spoken in various expressed emotions [5]. The BAVED dataset includes 7 words given as 0-like, 1-unlike, 2-this, 3-file, 4-good, 5-neutral, and 6-bad. The BAVED dataset word is pronounced in three levels corresponding to the person's emotions: 0 for low emotion (tired or exhausted), 1 for neutral emotion, and 2 for high emotion positive or negative emotions (happiness, joy, sadness, anger). The dataset contains 1935 recordings that are recorded by 61 speakers (45 males and 16 females). This is a drawback in the dataset that we will investigate in a future work.

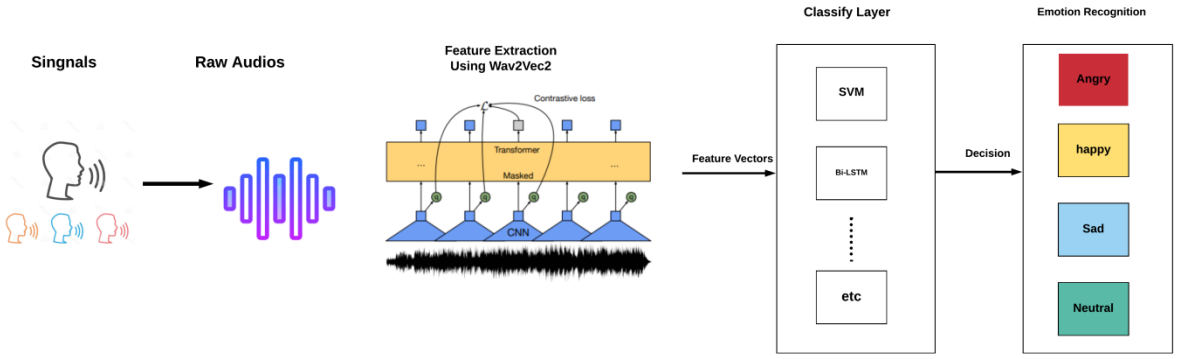


Figure 2. Architecture of the system for emotion recognition of Arabic speeches

METHODS

Feature Extraction

It's relatively simple for a human to understand what's in an image—finding an object, such as a car or a face; classifying a structure as damaged or undamaged; or visually identifying different land cover kinds are all basic tasks. The task is far more complex for machines. To tackle real-world problems, however, it's vital to be able to leverage and automate machine-based feature extraction.

Deep learning is a machine learning technique for detecting features in images. It makes use of a multi-layer neural network, which is a computer system that mimics the functions of the human brain.

Researchers and scientists have used deep learning as Feature extraction in recent decades, utilizing the power of deep learning to capture important features that have resulted in significant improvements in literature and real-world applications.

In the machine learning community, representation learning has evolved into its own area, commonly referred to as Deep Learning or Feature Learning. Although depth is an important aspect of the story, there are many other priors that are intriguing and may be easily captured when the challenge is framed as learning a representation. a more accurate representation allows the model to comprehend the data.

When it comes to representation learning, two recent advanced algorithms produce the state of the art in the field of speech recognition: Wav2vec2.0 [21] and HuBERT [22].

Wav2vec2.0: Wav2vec2.0 is a self-supervised speech representation model that pursues to capture the crucial properties of raw audios by using the power of transformers and Contrastive learning. The wav2vec2.0 training procedure is divided into two phases: i) the model is trained on hundreds of unlabeled data in the first phase, ii) fine-tuned on a small dataset for specific tasks.

The Wav2vec2.0 model consists of:

- convolutional layers that process the raw waveform input to get latent representation – Z,
- transformer layers, creating contextualized representation – C, linear projection to output – Y.

We used the pre-trained model Elgeish [23], which is Fine-tuned facebook/wav2vec2-large-xlsr-53 on Arabic using the train splits of Common Voice and Arabic Speech Corpus.

HuBERT: Innovative method for self-supervised speech representation learning HuBERT for speech representation learning matches or outperforms SOTA techniques for speech recognition, generation, and compression.

HuBERT learns the structure of spoken input by predicting the proper cluster for masked audio segments using an offline k-means clustering step. By alternating between clustering and prediction processes, HuBERT improves its learnt discrete representations over time. Furthermore, the high quality of HuBERT's learned presentations allows for simple deployment to a wide range of downstream speech applications.

HuBERT uses continuous inputs to train both acoustic and linguistic models. The model must first encode unmasked audio inputs into meaningful continuous latent representations, which correspond to the traditional acoustic modeling problem. Second, the model must capture the long-term temporal relationships between learned representations in order to reduce prediction error. One key finding driving this research is the importance of consistency, not simply correctness, of the k-means mapping from auditory inputs to discrete targets, which allows the model to focus on modeling the sequential structure of input data.

If an early clustering iteration can't tell the difference between the /k/ and /g/ sounds, the prediction loss will learn representations that explain how additional consonant and vowel sounds operate together with this super-cluster to generate words, resulting in a single super-cluster comprising both of them. As a result, the next clustering iteration uses the newly learned representation to produce superior clusters. Our results show that by alternating clustering and prediction phases, representations improve with time.

MLP and Bi-LSTM Classifiers

After extracting features with wav2vec2.0 and HuBERT, we feed the output into a classifier head: We utilized MLP Classifier stands for Multi-layer Perception Classifier, which is linked to a Neural Network by its name, and a Bi-LSTM Layer with 50 hidden units, Both classifiers produced results that were close to each other.

RESULTS AND PERFORMANCE EVALUATIONS

To test our models, we use three different measurements to evaluate the performance: F 1 score, validation loss, and confusion matrix.

F 1 Score

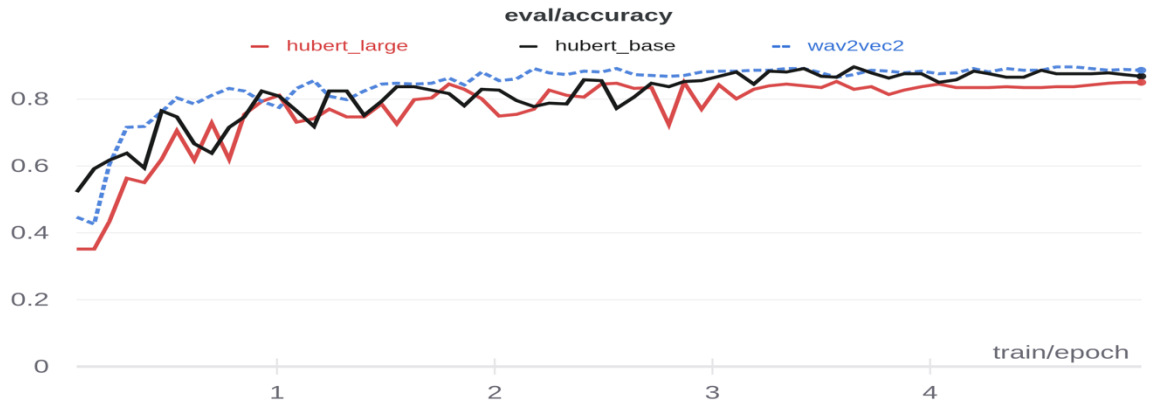


Figure 3. Comprise of the results and accuracy of the proposed three models

Figure 3, illustrates that Wav2vec2.0 achieves best accuracy and converge faster than the two other model. HuBERT is unstable during the process of training.

Validation Loss and Training Loss:

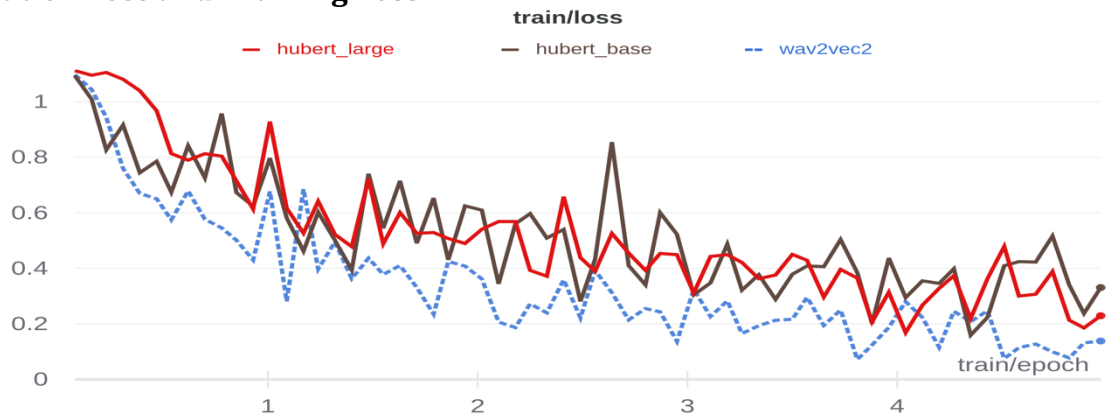


Figure 4. The training loss of the three models

The models were run for a total of 5 epochs. The Wav2vec2.0 and HuBERT models' outputs are represented in the performance measures used to calculate the categorization problem's performance. Wav2vec2.0 converges faster than HuBERT model and is more stable during the training phase.As shown in **Figure 4 ,5**, the train loss of wave2vec2 has the lowest training loss in comparison to the other three models. Evaluation loss of wav2vec2 has the lowest among the three models , and more stable in the converging process

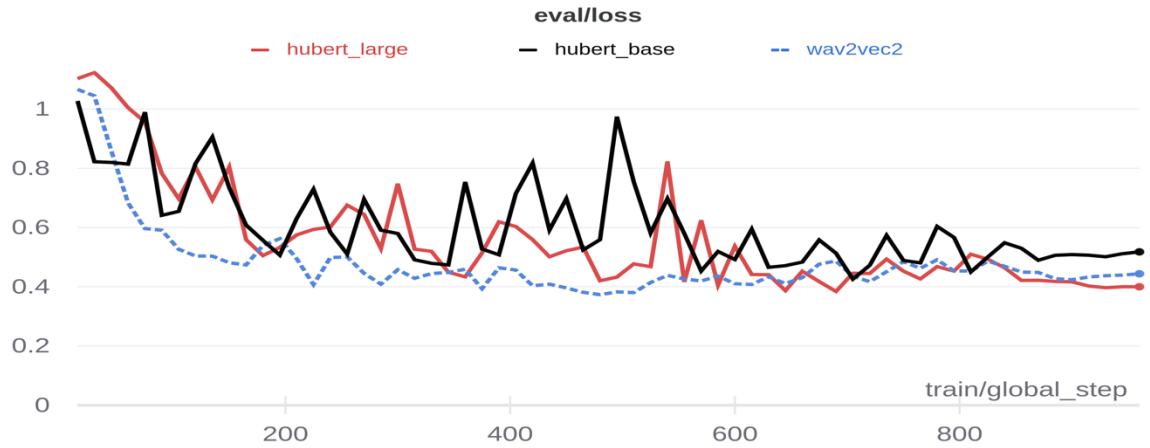
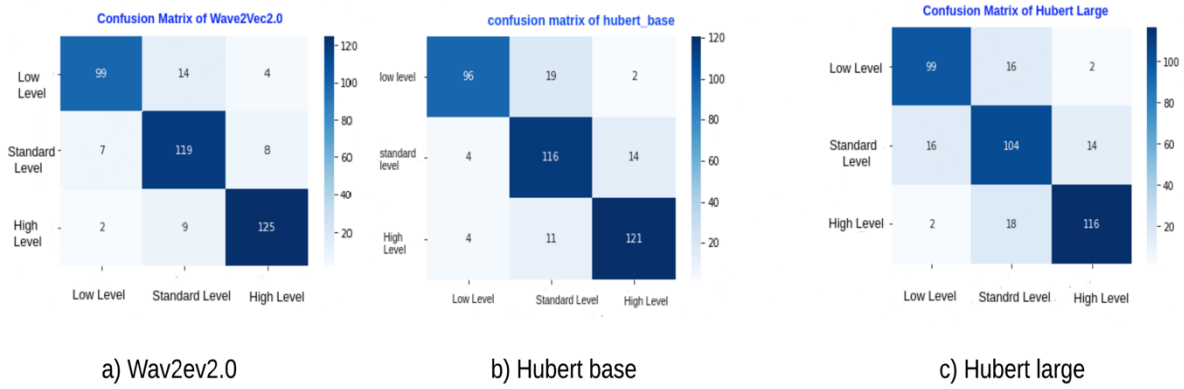


Figure 5. The evaluation loss of the three models.

Confusion Matrix

Figure 6 displays that the prediction matrix of the start of the art models.



a) Wav2ev2.0 b) Hubert base c) Hubert large

Figure 6. Figures: (a) Results of Confusion Matrix wav2vec2.0, (b) Results of Confusion

Matrix HuBERT base, (c) Results of Confusion Matrix HuBERT

Wav2vec2 achieves the best accuracy among the three models. Despite HuBERT achieves the best results on the downstream tasks, and captures important feature representation, the wav2vec outperforms for some reasons: i) Wav2vec2.0 has been trained on Arabic dataset especially Elgeish Pre-trained model, it has been trained on Common Voice[24] and Arabic Speech Corpus[25], ii) Both Hubert models (base , large) have been trained on Multi-language Models. On other tasks HuBERT could outperform the wav2vec2 model.

As shown in table I, Wav2vec2.0 outperforms HuBERT base and large because Wave2vec2.0's pre-trained model "Elgeish" was trained on Arabic datasets (common voice and Arabic speech corpus), whereas both HuBERT models were trained on multi-language tasks. Despite HuBERT's robustness against noise and ability to capture more information than Wav2vec2.0, it failed this task.

Table I. THE RESULTS FOR THE PROPOSED THREE MODELS

Model	Length	no. records	Accuracy
wav2vec2.0	19 Min	1935	89
Hubert Base	19 Min	1935	87
Hubert Large	19 Min	1935	84

CONCLUSION

We developed the state of the art deep learning models to detect emotions of Arabic speeches. We implemented these learning models and demonstrated their results on the Arabic BAVED audio dataset. Several experiments are performed using wav2vec2 and HuBERT different validation techniques. The model is successfully performed by using wav2vec2 and yielded an accuracy of 89%. As future work, we plan to extend the proposed method to incorporate more feature sets and increase the size of the dataset for words, sentence, and paragraph recognition.

ACKNOWLEDGMENT

This research is partially funded by a grant from the academy of scientific research and technology (ASRT), 2020, research grant number 6547.

References

[1]. K. Noh, C. Jeong, J. Lim, S. Chung, G. Kim, J. Lim, and H. Jeong, “Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets,” Sensors, vol. 21, 1579, 2021

[2]. A. Almahdawi and W. Teahan, “A new arabic dataset for emotion recognition. in: Arai k., bhatia r., kapoor s. (eds),” Intelligent Computing. CompCom 2019, Advances in Intelligent Systems and Computing, vol998. Springer, Cham., 2019.

[3]. A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, “Ksuemotions ldc2017s12.” Web Download. Philadelphia: Linguistic Data Consortium,2017 https://catalog.ldc.upenn.edu/LDC2017S12, 2017

[4]. I. Shahin, A. B. Nassif, N. Nemmour, A. Elnagar, A. Alhudhaif, and K. Polat, “Novel hybrid dnn approaches for speaker verification in emotional and stressful talking environments,” Neural Computing and Applications, June 2021.

[5]. A. Aouf, “Basic arabic vocal emotions dataset (baved) – github,” https://github.com/40uf411/Basic-Arabic-Vocal-Emotions-Dataset, 21 September, 2019

[6]. F. M. P. del Arco1, S. Halat, S. Padó, and R. Klinger, “Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language,” Forum for Information Retrieval Evaluation, Virtual Event, December 13–17, 2021

[7]. A. Satt, S. Rozenberg, and R. Hoory, “Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms,” in Proc. Inter speech 2017, pp. 1089–1093, 2017.

[8]. E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, “A review on speech emotion recognition using deep learning attenion mechanism,”electronics,vol.10,1163,2021. https://doi.org/10.3390/electronics10101163.

[9]. P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, “Speech emotion recognition using spectrogram & phoneme embedding,” In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018.

[10]. S. Zhang, S. Zhang, T. Huang, and W. Gao, “Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching,” IEEE Trans. Multimed, vol. 20, p. 1576–1590, 2018.

[11]. R. Khalil, E. Jones, M. Babar, T. Jan, M. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” 2019, vol. 7, p. 117327–117345, IEEE Access.

[12]. W. Z. Zheng and Y. Zong, “Multi-scale discrepancy adversarial network for crosscorpus speech emotion recognition,” Virtual Real. Intell. Hardw., vol. 3, 65–75, 2021.

[13]. Y. Hifny and A. Ali, “Efficient arabic emotion recognition using deep neural networks,” in IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP): https://github.com/qcri/deepemotion, 2019.

[14]. Y. Hifny and A. Ali, “Efficient arabic emotion recognition using deep neural networks,” IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 6710–6714, 2019 doi: 10.1109/ICASSP.2019.8683632.

[15]. S. Klaylat, Z. Osman, L. Hamandi, and R. Zantout, “Emotion recognition in arabic speech,” Analog Integr Circ Sig Process, vol. 96, 337–351,2018.

[16]. R. Y. Cherif, A. Moussaoui, N. Frahta, and M. Berrimi, “Effective speech emotion recognition using deep learning approaches for algerian dialect,”Intern. Conf. of Women in Data Science at Taif University (WiDSTaif),2021.

[17]. L. Abdel-Hamid, “Egyptian arabic speech emotion recognition using prosodic, spectral and wavelet features,” Speech Commun., vol. 122, pp. 19–30, 2020.

[18]. S. Klaylat, “Arabic natural audio dataset,” 2019.

[19]. F. A. Shaqra, R. Duwairi, and M. Al-Ayyoub, “The audio-visual arabic dataset for natural emotions,” 7th International Conference on Future Internet of Things and Cloud (FiCloud), pp. 324–329, 2019.

[20]. M. Meddeb, H. Karray, and A. Alimi, “Speech emotion recognition based on arabic features,” 2015 15th IEEE International Conference In Intelligent Systems Design and Applications (ISDA), pp. 46–51, December 2015

[21]. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec2.0: A framework for self-supervised learning of speech representations,” CoRR, vol. abs/2006.11477, 22 Oct 2020. Facebook Wav2Vec2.0: https://ai.facebook.com/blog/wav2vec-20-learning- the-structure-of-speech-from-raw-audio/.

[22]. W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021.

[23]. Elgeish, “https://huggingface.co/elgeish/wav2vec2-large-xlsr-53-arabic,”2020.

[24]. CommonVoice “https://commonvoice.mozilla.org/en/datasets,”ar-137h-2021-07-21,2021.

[25]. N. Halabi, “Arabic speech corpus,” http://ar.arabicspeechcorpus.com/,2021.