Aswat: Arabic Audio Dataset For Automatic Speech Recognition Using Speech-Representation Learning

Lamya Alkanhal[1] **Abeer Alessa***‡[2] **Elaf Almahmoud***‡[3] **Rana Alaqil** ‡[4]
[1]Saudi Technology and Security Comprehensive Control Company (Tahakom), Saudi Arabia.
[2]King Saud University, Saudi Arabia [3]Center for Complex Engineering Systems, Saudi Arabia,,Massachusetts Institute of Technology, USA, [4]Intelmatix, Saudi Arabia

lalkanhal@tahakom.com aalessa5.c@ksu.edu.sa elaf@mit.edu ralaqil@intelmatix.ai

Abstract

Recent advancements in self-supervised speech-representation learning for automatic speech recognition (ASR) approaches have significantly improved the results on many benchmarks with low-cost data labeling. In this paper, we train two self-supervised frameworks for ASR, namely wav2vec, and data2vec, in which we conduct multiple experiments and analyze their results. Furthermore, we introduce Aswat dataset, which covers multiple genres and features speakers with vocal variety. Aswat contains 732 hours of clean Arabic speech that can be used in the pretraining task for learning latent speech representations, which results in achieving a lower word error rate (WER) in Arabic ASR. We report the baseline results and achieve state-of-the-art WERs of 11.7% and 10.3% on Common Voice (CV) and the second round of Multi-Genre Broadcast (MGB-2) respectively, as a result of including our dataset Aswat.

Index Terms: Automatic speech recognition, Self-supervised learning, wav2vec, data2vec.

1 Introduction

Automatic speech recognition (ASR) is the task of transcribing speech audio into text. Supervised deep learning has shown a notable improvement in speech recognition, providing significant gains in tasks rich in labeled data. Unfortunately, this reliance on labeled data limits the extent to which deep learning can advance, primarily because of the scarcity of labeled data in some tasks. Recently, self-supervised approaches have overcome this problem and made it possible to reach outstanding results with a limited labeled dataset (Baevski et al., 2020; Hsu et al., 2021; Baevski et al., 2022). Self-supervised learning leverages raw waveforms to learn representation that captures low level features and underlying structure of the data. The learned representations in the pretraining phase are used in downstream tasks in a supervised phase with a minimal amount of labeled data.

Arabic is one of the most spoken languages worldwide, with over 400 million speakers (Graves, and Jaitly, 2014). It is considered challenging to process automatically due to various internal factors, including multiple dialects, ambiguous syntax, syntactical flexibility, and diacritics (Hussein et al., 2022). However, Modern Standard Arabic (MSA) is one formal dialect that is understood by the majority of Arabic speakers. It is the formal spoken and written dialect that is often used in formal speech, news broadcasts, radio, and newspapers. It is also taught in schools and universities (Ryding , 2005).

In our work, we utilized the self-supervised frameworks wav2vec (Baevski et al., 2020) and data2vec (Baevski et al., 2022), and released dataset Aswat (Voices), on which we trained the ASR systems. Aswat is a well-organized, unannotated dataset of Arabic speech, of which 66% is in MSA (Modern Standard Arabic). We carefully curated and manually cleaned it, and it includes speakers from various demographic backgrounds. It has 732 hours of speech constructed from audio files on the internet; thus, it covers a variety of audio files recorded from different speakers and under various recording setup environments. Aswat leads to the learning of useful latent speech representations during the pretraining task in wav2vec (Baevski et al., 2020) and data2vec (Baevski et al., 2022). This results in state-of-the-art performance in Arabic with a word error rate (WER) of 11.7% on Common Voice (CV) and 10.3% on MGB-2, achieved with fewer training instances compared to the second round of Multi Genre Broadcast (MGB-2). The original audio files are crawled

*Equal contribution

‡Work done in Tahakom

Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023), pages 120–127

Dataset	Dialect	Domain	Split	#Hours	#Segments
Common Voice	MSA	Monologues	train	31.5	27,823
			valid	12.7	10,386
			test	12.6	10,388
MGB-2	MSA (70%), DA (30%)	News: Conversation (63%), interview (19%), report (18%)	train	1,128	376,011
			valid	8.5	5,002
			test	9.6	5,365
Aswat	MSA (66%), Saudi (27%), other dialects (7%)	Monologues (45%), Dialogues (55%)	train	724.6	502,391
			valid	7.3	5,065

Table 1: Comparison between CommonVoice, MGB-2 and Aswat.

from YouTube and Soundcloud; therefore, they are subject to copyright. We made the dataset publicly available[1] for non-commercial purposes. This paper's contributions can be summarized as follows:

- Releasing baseline results in Arabic for some of the most prominent self-supervised models in speech, namely wav2vec and data2vec.
- Providing 732 hours of a high-quality diverse Arabic speech dataset.
- Comparing the results obtained from pretraining wav2vec and data2vec on Aswat with two of the most well-known Arabic benchmarks in ASR with extensive analysis, with which we were able to achieve the lowest WER.

2 Background

2.1. Self-supervised speech models

Self-supervised approaches have led to significant advances in the field of speech recognition [1,2,3]. Wa2vec2.0 (Baevski et al., 2020) is the most prominent self-supervised approach in speech, and data2vec (Baevski et al., 2022) is an approach that produced state-of-the-art results on Librispeech.

2.1.1 Wav2vec

The architecture consists of three components: a feature encoder where the audio waves are encoded with a stack of 1-D convolutional layers, a quantization module to map the resulting latent representations into a discretized space, and a contextual network used during the pretraining where a span of the resulting representations are masked and fed into a context network that follows the transformer network. It learns contextualized representations and tries to distinguish them from quantized distractors via a contrastive task. The pretrained model is fine-tuned by projecting a linear head on the top of the context network with connectionist temporal classification (CTC) loss (Baevski et al., 2020).

2.1.2 *Data2vec*

Data2vec is a unified framework that works with three modalities (images, text, and speech) separately. It learns to construct representations that are continuous and contextualized. For speech data, the audio inputs are encoded by 1-D convolution layers. Then, the resulting latent representations are fed into a standard transformer network. The architecture consists of a single model with two modes: student and teacher. In the student mode, the model encodes a masked version of the representation, and in the teacher mode, it encodes the unmasked version of the representation to construct the training targets. The model's training mode is parameterized by an exponential moving average (EMA) of the student's parameters. The student's learning task is to minimize the objective function of the student's prediction of a target that is constructed by the teacher's parameters. Similar to wav2vec, the model is fine-tuned with CTC loss (Baevski et al., 2022).

2.2. Annotated datasets

While the audio datasets in Arabic are still scarce compared to other languages, there is an increase in the recent work to bridge the gap such as: the datasets of Multi-Genre Broadcast challenge, MGB-2 (Ali et al., 2016), MGB-3 (Ali et al., 2017), MGB-5 (Ali et al., 2019), Arabic Mozilla's Common Voice[2], ADI-17 (Shon et al.,

[1] https://github.com/AswatDataset/AswatDataset

[2] https://commonvoice.mozilla.org/ar/datasets

2020), QASR (Mubarak et al., 2021), MASC (Al-Fetyani et al., 2021), and SADA[3]. In our work, we consider the most well-known Arabic labeled datasets in ASR, namely Common Voice and the second round of MGB. Moreover, they are publicly available datasets that focus on MSA speech and are commonly used in literature, we used them for comparison and benchmarking.

2.2.1 Common Voice

Mozilla's CV is a platform that provides a public audio dataset with multiple languages powered by the voices of volunteers around the world, it allows users to record and validate other people's recordings. In this paper, we used Arabic CV version 8.0 that was released on January 19, 2022 and recorded by 1,216 volunteers[2].

2.2.2 MGB-2

MGB-2 uses a multi-dialect dataset with 70% MSA and 30% Dialectal Arabic (DA). It includes programs recorded from 2005 to 2015. The training script is aligned using the QCRI Arabic LVCSR system, and it is manually transcribed but not always verbatim; it includes rephrasing, removal of repetition, and summarization, whereas the validation and test sets are transcribed verbatim. These alterations lead to variation in the transcripts' quality; the WER between the original transcribed text to the verbatim version is about 5% in the validation set (Ali et al., 2016). The dataset includes a large corpus of 130 million words from Al-Jazeera website. We used this corpus for language modeling.

Table 1 depicts the two datasets' information, excluding the overlapping segments from MGB-2 in the validation and test sets.

3 Related Work

In (Ashish et al., 2017), the first transformer-based architecture was introduced to better parallelize self-attention mechanisms. Furthermore, when applied to ASR tasks, Karita et al., (2019) demonstrated that transformer-based models outperformed state-of-the-art recurrent neural networks (RNNs). In the ASR task, self-supervised approaches, such as [1, 3], have recently shown significant improvement. The main difference between them is that wav2vec learns discrete units of speech during pretraining through a quantitation process, and data2vec directly predicts contextualized latent representations without quantization.

Although the literature on E2E models trained on Arabic speech is limited, researchers have done valuable work that is essential to the community. In (Ali et al., 2018), the authors used CTC and RNNs, and the reported results were on the MGB-2 development set, without any further results on the test set. In (Belinkov et al., 2019), the authors analyzed the learned internal representations and compared phonemes and graphemes as well as various articulatory features using DeepSpeech2, an end-to-end ASR model. In Taha Zouhair's work[4], the author used wav2vec model on CV benchmark, achieving a WER of 24 %. Belinkov et al. (2019) utilized the transformer architecture with CTC and attention objectives resulting in a WER of 12.5 % in an MSA task on MGB-2. More recently, Chowdhury et al. (2021) proposed a multilingual strategy for dialectal code switching in Arabic ASR. Using end-to-end transformer models reported in (Belinkov et al., 2019) for Arabic, they achieved state-of-the-art results with a WER of 12.1 % demonstrating the effectiveness of multilingual approaches. In our work, we constructed a high-quality dataset and reached state-of-the-art WERs on two well-known benchmark datasets, by pretraining self-supervised architectures, namely wav2vec2.0 (Baevski et al., 2020) and data2vec (Baevski et al., 2022).

4 Aswat Dataset

4.1. Dataset construction

During the dataset construction phase, we started by selecting Arabic audio data with clear pronunciation, and targeted various speech data recorded under multiple settings, such as audiobooks, news, podcasts, and lectures. It covers multiple genres, including politics, philosophy, history, health, folklore, religions, sports, economy, and science. The data includes clear conversation in an interview-like setting without any overlapping speech. We obtained 1060 audio files from two platforms: YouTube and SoundCloud. The former is a video-sharing

[3] https://www.kaggle.com/datasets/sdaiancai/sada2022

[4] https://www.diva-portal.org/smash/get/diva2:1579121/FULLTEXT01.pdf

service, and the latter is a service for sharing audio and music.

4.2. Dataset cleaning

We cleaned the dataset manually to improve speech intelligibility and find better speech representations. All audio files were reviewed to remove noise such as background music using Audacity tool[5].

4.3. Data preprocessing

We reduced the number of channels from stereo to mono-channel and resampled the wave rate to 16 kHz. Finally, we split the audio into segments ranging in length from 3 to 27 seconds, based on silence regions using Pydub Python Package[6]. Details of Aswat are presented in Figure 1.

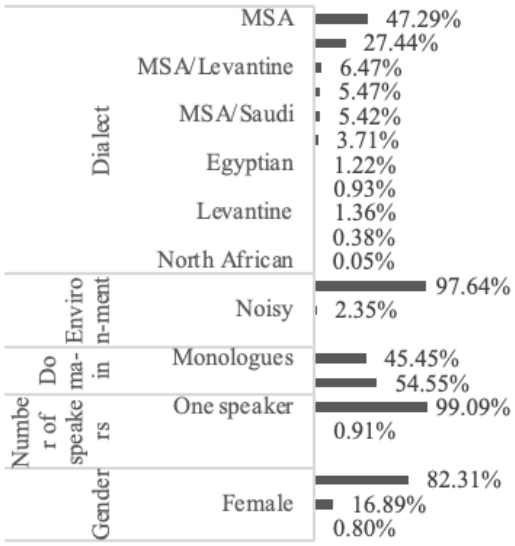

Figure 1: *Aswat Statistics.*

5 Experimental Settings

5.1. Acoustic model

5.1.1 Data preparation

For the acoustic modeling, we segmented the MGB-2 audio files on the timing information provided in the XML files. Then, we converted the audio files of CV and MGB-2 to mono-channel, resampled their rates to 16 kHz, and exported the audio files into FLAC format. We excluded the overlapped speech from MGB-2 validation and test sets.

For the transcription, we preprocessed the transcripts by removing punctuation, diacritics, and any other characters except for the Arabic letters. For the numbers in MGB-2 transcription, we reported the results of two different preprocessing techniques: 1) converting numbers to numerals (words); and 2) removing data entries in the training set that have numbers in their transcription.

5.1.2 Pretraining

We used the implementation of wav2vec and data2vec in fairseq (Ott et al., 2019). We considered only the BASE models and used the same fairseq hyper-parameters (Ott et al., 2019). Moreover, we initialized the models with the fairseq pretrained weights of Librispeech and started the training without resting the optimizer. For pretraining, we ran three experiments: 1) we trained the models on Aswat; 2) we trained the models on MGB-2; and 3) we trained the models on a combined dataset (C.Dataset) of Aswat and MGB-2. The purpose of these experiments is to compare Aswat to MGB-2 and determine which model provides better speech representations for Arabic when they are fine-tuned on the same task. We did not train a model on CV because it is relatively small and pretraining requires a large dataset. For the validation task in the first experiment, we randomly sampled 1% of Aswat dataset and set it as the validation set, and we used the rest for training because self-supervised approaches need substantial data for the pretraining task.

In pretraining data2vec models, the training crashed in the early epochs of the model that was trained on MGB-2, and it crashed in the later epochs of the two other models with the following message: “Minimum loss scale reached (0.0001).” which is caused by the loss overflow. We were able to delay the crashing to later epochs by setting the fp16 scale tolerance to 0.25. We used 16 Tesla V100 (32GB) GPUs for each experiment and chose the training checkpoints with the lowest loss on the validation set.

5.1.3 Fine-tuning

In this stage, we fine-tuned the pretrained models on the labeled data CV and MGB-2 separately. For hyper-parameter selection, we used the configurations of Librispeech-100h for CV and Librispeech-960h for MGB-2 as they resulted in

[5] https://www.audacityteam.org
[6] https://github.com/jiaaro/pydub

the best WERs compared to other Librispeech configurations, we used the same settings except for the max update where we increase it to 640000.

However, we encountered the same issue in fine-tuning that appeared while pretraining our model: the training crashed at early epochs. To train the model for longer epochs, we reduced the batch size and switched from fp16 to fp32. We conducted each experiment on 8 Tesla V100 (32GB) GPUs and chose the models with the lowest WER on the validation set.

5.2. Language Model

We considered a transformer-based language model (LM) provided in fairseq (Ott et al., 2019) to decode the results of the speech recognition models. We used MGB-2 corpus for this task and cleaned the text by removing extra new lines and any non-Arabic characters. Then, we split the text into sequences with a maximum length of 300 words and an overlap of 50 words.

The model was trained on 8 Tesla V100 (32GB) GPUs with the same data splitting approach and hyper-parameters in fairseq (Ott et al., 2019). We tuned the hyper-parameters lm_weigh and word_score and obtained the best results from the values 0.2 and −0.2, respectively.

6 Results and Discussion

6.1 Fine-tuning on Common Voice

We used Arabic CV version 8.0 in training the speech models. Table 2 shows the results of evaluating the models on CV test set, and we decoded the results using LM with beams 5 and 20.

Model	Unlabeled Data	No LM	LM, beam=5	LM, beam=20
wav2vec	Aswat	**16.4%**	**16.1%**	**15.9%**
	MGB-2	18%	17.3%	17.2%
	C.Dataset	16.5%	16.3%	16.1%
data2vec	Aswat	12.1%	13.1%	13%
	MGB-2	15.5%	15.5%	15.3%
	C.Dataset	**11.7%**	**12.6%**	**12.5%**

Table 2: WER on the CV test when training on the CV training set. The best results in each framework are in bold, and the second best results are underlined.

For fine-tuning on CV, we achieved the best results for data2vec models from pretraining on the combined dataset, followed by Aswat, and then MGB-2. For wav2vec, pretraining on Aswat yielded a lower WER than the combined dataset, as Table 2 shows. Additionally, the significantly lower WER achieved by pretraining on Aswat compared to MGB-2 could be attributed to one of two factors: (1) the similarity between Aswat and CV, as they both contain monologue speech, or (2) Aswat has better speech representation, and better generalization. We were able to achieve a state-of-the-art WER of 11.7% on Arabic CV benchmark with the ASR model alone. Decoding with LM resulted in improving the WER of the wav2vec models, but it increased the WER for data2vec, except for the model trained on MGB-2.

Our explanation for the LM performance is that the LM is trained on news data (MGB-2) which has a different domain from common voice (i.e. blog posts, books, movies). In data2vec, the acoustic model (AM) has good results, but LM tends to replace unseen or infrequent words generated by AM with words from its dictionary, which results in increasing WER score. In wav2vec, the AM generates texts that contain non-real words, which are subsequently corrected by the LM. While it's true that the LM occasionally replaces correct words with incorrect ones, the frequency of such cases is significantly lower than the instances where it makes correct predictions. As a result, this contributes to an improvement in WER.

Analysis of the best model errors in Table 2 shows that most errors are substitution errors. Such errors occur due to the similarity in pronunciation of some Arabic sounds between MSA and DA. For instance, the model has many substitutions between سِين (sīn) and صَاد (ṣād), ضَاد (ḍād) and دَال (dāl), and ضَاد (ḍād) and ظَاء (ẓāʾ). Also, some errors occur due to the features that cannot be automatically captured by the model, such as the rules of writing the variations of the هَمْزة (hamzah) which indicates a glottal stop. In Arabic, there are two types of هَمْزة (hamzah) or glottal stops: Hamzat Al-Wasl and Hamzat Al-Qata'a. Hamzat Al-Wasl is written as an أَلِف (ʾalif) without the هَمْزة (hamzah) marker, and it is only pronounced if it is in the beginning of an utterance. In contrast, Hamzat Al-Qata'a is written as an أَلِف (ʾalif) with the هَمْزة (hamzah) marker, and it is always pronounced.

6.2. Fine-tuning on MGB-2

For MGB-2, we used PyArabic Python package[7] to transform numbers to their verbatim form. The WERs of MGB-2 results are reported using the evaluation script provided in the MGB challenge Github repository[8]. The table below depicts the results of testing the model with LM decoded with beams 5 and 20.

Model	Unlabeled Data	No LM	LM, beam=5	LM, beam=20
wav2vec	Aswat	14.7%	13.1%	12.9%
	MGB-2	14.2%	**12.8%**	12.6%
	C.Dataset	**14.1%**	12.9%	**12.5%**
data2vec	Aswat	13%	12.3%	12.1%
	MGB-2	12.6%	11.9%	11.8%
	C.Dataset	**12.1%**	**11.6%**	**11.4%**

Table 3: WER of the first experiment on the MGB-2 test. The best results in each framework are in bold, and the second best results are underlined.

Table 3 shows that the best obtained models in wav2vec and data2vec were those pretrained on the combined dataset, followed by MGB-2, and then Aswat. The addition of Aswat to the pretraining improved the WER from 14.2% to 14.1% in wav2vec and 12.6% to 12.1% in data2vec. The model pretrained only on MGB-2 has an advantage over the model that was pretrained only on Aswat because it has seen all of the data used for fine-tuning, so it has learned better speech representations for MGB-2 and therefore yields a better WER.

We noticed from analyzing the errors of the best model in Table 3 that most errors are substitutions in numeral words. The model substitutes the word for "fifteen" in DA"خمستاشر" (xmsta:ʃr) with its equivalent in MSA " خمس عشرة" (xms ʕʃrt), the word for "sixteen," "ستاشر" (sta:ʃr) with "ست عشرة" (st ʕʃrt), the word for "seventy," "وسبعين" (wsbʃjn) with "وسبعون" (wsbʃwn), and "two thousand" "ألفين" (ʔfjn) with "ألفان" (ʔfa:n). These errors come from using PyArabic tool in preprocessing; it converts every number to its MSA form and uses one grammatical case: Al-Raf'á case (the nominative case). We tackled this issue in the second experiment by dropping examples with numbers from the training set, resulting in removing 11.4% of the training data and reducing the WER by 9.6%.

Model	Unlabeled Data	No LM	LM, beam=5	LM, beam=20
wav2vec	Aswat	12.8%	11.8%	11.6%
	MGB-2	12.8%	11.7%	11.6%
	C.Dataset	**12.4%**	**11.4%**	**11.2%**
data2vec	Aswat	11.4%	10.8%	10.7%
	MGB-2	11.3%	10.7%	10.7%
	C.Dataset	**10.9%**	**10.5%**	**10.3%**

Table 4: WER of the second experiment on the MGB-2 test. The best results in each framework are in bold, and the second best results are underlined.

Table 4 depicts the result of the second experiment. The ASR model shows the best results yielded from fine-tuning the data2vec model that was pretrained on the combined dataset. In addition, the decoded output shows that the model predicts the numerical words correctly. Evaluating the models with the LM reduced the WERs and closed the gap between WERs of wav2vec models. We reached a state-of-the-art (SOTA) WER of 10.3% on the MGB-2 benchmark and outperformed the previous result of 12.1% (Chowdhury et al., 2021).

Analyzing the errors shows that most of them are substitution errors between different Hamza variations and between تاء مربوطة (tāʾ marbūṭah) and هَاء (hāʾ). In addition, some substitutions come from removing the Arabic definite article "ال" (Al) and the connected prepositions and conjunctions from the beginning of the word, such as removing فَاء (fāʾ), بَاء (bā)ʾ, and وَاو (wāw).

Additionally, we observed that the model removes words that are pronounced with an American English accent, even if they are Arabic words. This behavior could be attributed to removing Latin letters from the training script, although the presence of these letters was very small in the dataset.

Finally, Tables 2, 3, and 4 show that data2vec produced better results in all of the experiments, as (Baevski et al., 2022) claimed that discrete units are not required with the use of rich contextualized targets and that learning contextualized targets during the pretraining phase leads to better performance. Our empirical research shows that this claim holds true for Arabic speech data.

[7] https://pypi.python.org/pypi/pyarabic
[8] https://github.com/qcri/ArabicASRChallenge2016

Limitations

While our work achieved state-of-the-art performance, it has three main limitations. First, although our dataset was carefully curated and meticulously cleaned to meet our research objectives; it is important to note a limitation in speaker diversity. This imbalance in gender representation within our dataset can potentially affect our model's performance indicating the need for future experiments with more diverse set of speakers and conducting experiments on the effect of gender bias in our model's performance. Second, while our research used a self-supervised approach, we confined our experimentation with fine-tuning on ASR only, which limited our exploration of other downstream tasks that may benefit from our dataset. The focus on ASR was an intentional choice given its prominence and frequent usage among speech tasks. Nevertheless, we acknowledge that the broader applicability of our dataset across different tasks remains an open question. Third, we did not use the larger version of wav2vec and data2vec models. Although the larger model may potentially yield better performance, the primary goal of this paper was to improve Arabic ASR results and reach SOTA results with our current model configuration. Our findings have successfully demonstrated the benefits of our dataset.

7 Conclusion

In this work, we provide the community with 732 hours of a clean and organized Arabic speech dataset. We report state-of-the-art results for ASR with data2vec architecture, and by combining Aswat with MGB-2 in the pretraining stage, we achieved a WER of 11.7% on CV and 10.3% on MGB-2. In the future work, we plan to improve our methods by using automatic audio cleaning tools[9] and tool in (David at al., 2018) to collect bigger data and include more dialects. In addition, we plan to use the LARGE data2vec and adjust the hyper-parameters based on the training data to enhance the results.

References

Ahmed Abdelrahman, Yasser Hifny, Khaled Shaalan, and Sergio Toral. 2018. *"End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks."* Computational Linguistics, Speech and Image Processing for Arabic Language: 231–48.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. "*The MGB-2 challenge: Arabic multi-dialect broadcast media recognition"*. In Proc IEEE SLT.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. *"Speech recognition challenge in the wild: Arabic MGB-3.* IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. "*The MGB-5 Challenge: Recognition and Dialect Identification of Dialectal Arabic Speech"*. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. *"Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations."* Advances in Neural Information Processing Systems 2020-December: 1–12.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. *"data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language,"*. In Proceedings of the 39th International Conference on Machine Learning. PMLR.

Alex Graves, and Navdeep Jaitly. 2014. *"Towards End-to-End Speech Recognition with Recurrent Neural Networks."* 31st International Conference on Machine Learning, ICML.

Amir Hussein, Shinji Watanabe, and Ahmed Ali. 2022. *"Arabic Speech Recognition by End-to-End, Modular Systems and Human."* Computer Speech and Language 71: 1–39.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *"Attention is all you need."* In

[9] https://github.com/wiseman/py-webrtcvad

Advances in Neural Information Processing Systems, pages 6000–6010.

David Doukhan, Eliott Lechapt, Marc Evrard, and Jean Carrive. 2018. *"Ina's mirex 2018 music and speech detection system."* Music Information Retrieval Evaluation eXchange (MIREX 2018).

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. *"QASR: QCRI Aljazeera Speech Resource A Large Scale Annotated Arabic Speech Corpus."* In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2274–2285, Online. Association for Computational Linguistics.

Karin C. Ryding. 2005. *"A Reference Grammar of Modern Standard Arabic"*. Cambridge: Cambridge University Press, 2005.

Mohammad Al-Fetyani, Muhammad Al-Barham, Gheith Abandah, Adham Alsharkawi, and Maha Dawas. 2021. *"MASC: Massive Arabic Speech Corpus."* IEEE Spoken Language Technology Workshop (SLT). doi:10.21227/e1qb-jv46

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. *"FAIRSEQ: A fast, extensible toolkit for sequence modeling."* In North American Association for Computational Linguistics (NAACL): System Demonstrations.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. *"Towards one model to rule all: Multilingual strategy for dialectal codeswitching Arabic."* Interspeech 2021. pages. 2466–2470. [Online]. Available: https://www.isca-speech.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, Wangyou Zhang. 2019. *"A Comparative Study on Transformer vs RNN in Speech Applications."* 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019 - Proceedings 9(4): 449–56.

Suwon Shon, Ahmed Ali, Younes Samih, Hamdy Mubarak, and James Glass. 2020. *"ADI17: A Fine-Grained Arabic Dialect Identification Dataset"*. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. *"HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units."* IEEE/ACM Transactions on Audio Speech and Language Processing 29(Cv): 3451–60.

Yonatan Belinkov, Ahmed Ali, and James Glass. 2019. *"Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition."* Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.