مجلة اللغويات الحاسوبية والمعالجة الآلية للغة العربية

Journal of Computational Linguistics & Arabic Language Processing

مجلة علمية محكمة تصدر عن مجمع الملك سلمان العالمي للغة العربية

- الرياض، المملكة العربية السعودية

Syllable-Based Arabic Speech Recognition Using Wav2Vec

التعرف على الكلام العربي على أساس المقطع باستخدام Wav2Vec

إبراهيم عبد العال[أ] محمد عبد الواحد[ب] مصطفى الشافعي[ج]

[أ] قسم الاتصالات وتكنولوجيا المعلومات, كلية الهندسة, جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر

[ب] قسم الاتصالات وتكنولوجيا المعلومات, كلية الهندسة, جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر

[ج] قسم الاتصالات وتكنولوجيا المعلومات, كلية الهندسة, جامعة العلوم والتكنولوجيا بمدينة زويل، الجيزة، مصر

عبدالعال، ا.، عبد الواحد،م.،والشافعي،م. (2023). التعرف على الكلام العربي على أساس المقطع باستخدام Wav2Vec. *مجلة اللسانيات العربية*، ع*00* (0)، 00-00.

Abdalaal, I., Abdelwahed, M., & Elshafei, M. (2023), Syllable-Based Arabic Speech Recognition Using Wav2Vec. *Journal of Computational Linguistics and Arabic Language Processing*, En(0) 00, 00-00

Submission Date:01/08/2023

Acceptance Date:23/10/2023

تاريخ الإرسال: 01/08/2023

تاريخ القبول:23/10/2023

الملخص

تقدم هذه الورقة طريقة مبتكرة للتعرف على الكلام العربي، بالاستفادة من الخصائص الفريدة للمقاطع الصوتية العربية، ونموذج لغة 5جرام ، بالإضافة إلى نموذج Wav2Vec-2الحديث . تبدأ الطريقة بتقسيم الكلام العربي إلى مقاطع صوتية يتم تكوينها برمجيا باستعمال قواعد نطق اللغة العربية، وتتميز بقدرتها على وصف التفاصيل الصوتية الدقيقة. ولزيادة تحسين الدقة، تم استخدام نموذج لغوي 5جرام للمقاطع الصوتية والذي يأخذ في الاعتبار الظواهر اللغوية المتأصلة في اللغة العربية، بما في ذلك الاختلافات الكبيرة على مستوى الكلمات والتبعيات السياقية. كما تم استعمال نموذج Wav2Vec كنموذج صوتي، وهو معروف بتمثيلاته الصوتية القوية التي تم تعلمها من خلال التدريب المسبق غير الخاضع للإشراف على كميات كبيرة من بيانات الكلام غير المسماة .و تم ضبط النموذج بدقة على مجموعة بيانات أصغر، مما عزز قدرته على التعرف على الكلام العربي. قد حقق هذا المزيج الجديد من التقنيات معدل خطأ في الكلمات بلغ 0.066 مع تشكيل النص، و 0.0596 بدون تشكيل النص، مما يدل على فعالية هذا النهج .

الكلمات المفتاحية :- Wav2Vec- نموذج لغة-المقاطع الصوتية-قواعد النطق

*المؤلف المراسل.:مصطفى الشافعي

البريد الإلكتروني: moelshafei@zewailcity.edu.eg

Abstract

This paper introduces an innovative Arabic speech recognition method that relies on Arabic syllables, a 5-gram language model, and the Wav2Vec-2 architecture. It starts by segmenting Arabic speech into syllables, improving accuracy by handling the language's complexity. The method is rigorously evaluated, showing a significant boost in system performance. To enhance accuracy further, a 5-gram language model is used to address linguistic nuances. The Wav2Vec-2 architecture, known for its robust acoustic representations, is employed as the acoustic model. Fine-tuning on a smaller labelled dataset improves Arabic speech recognition, making it resilient to pronunciation variations and noise. This unique combination yields impressive Word Error Rates (WER) of 0.06624 with Tashkeel and 0.05959 without Tashkeel, demonstrating its effectiveness.

Keywords: Wav2Vec- WER- Language model – Syllables - Pronunciation rules.

1. Introduction

Speech recognition technology has become an integral aspect of our daily lives, playing a critical role in powering interactive education and training, virtual assistants, speech transcription, speech translation, games, and various voice-enabled applications. Meeting the escalating demand for multilingual and dialect-specific speech recognition systems necessitates the development of robust solutions tailored to languages with intricate phonetics and morphology. Arabic, being one of the world's most widely spoken languages, spoken by 300 million, and the official language of 22 countries, presents distinctive challenges due to its complex morphological structure and diverse phonetic variations across different regions and dialects. Conventional methods for Arabic speech recognition predominantly rely on phoneme-based models, which encounter difficulties in capturing the subtle phonetic nuances and contextual dependencies inherent in the language. Consequently, the need for more sophisticated techniques that can effectively handle the intricacies of Arabic pronunciation and morphology has grown substantially. One of the problems in the speech recognition of Modern Standard Arabic (MSA) is the contextual phoneme alteration in cross-word pronunciation and intra-word pronunciation variation. These contextual variations alter the phonetic spelling of words, leading to a high word error rate (WER). In (AbuZeina et al., 2011), the authors presented a knowledge-based approach based on modelling cross-word pronunciation variation by expanding the phonetic dictionary and corpus transcription. They tested the method on a speech corpus of 5.4 hours of Modern Standard Arabic. The WER was reduced by 2.3% below the baseline system.

In data-driven end-to-end speech recognition approaches, the recognition model is provided with the speech utterances and its corresponding text. The problem here is that the phonetic representation of the actual utterance is not produced by a simple grapheme-to-phoneme relationship, but produced by a network of pronunciation rules. For example, a letter is written in the text, but may not be pronounced depending on the context, e.g. the word “Alshams” is pronounced “Ash-shams” without letter L, and introducing a second phoneme “sh”. As a sample of common substitution of letters; the letter N in “man laho” is uttered “mal-laho” with phoneme N substituted by phoneme L. Similarly, the phoneme N in the word “Anbar” is replaced by M and pronounced “Ambar”. However, these variations follow up certain pronunciation rules.

To overcome these problems, the end-to-end approach requires enormous data sets and complex deep learning models to capture the latent pronunciation rules, hopefully by a large number of training examples. Our proposed approach is to apply Arabic pronunciation rules to produce syllable strings reflecting the actual utterances. There are two advantages to this approach. The first is that the syllables represent un-interrupted sound unit. The second advantage is that every uttered phoneme is written and every unuttered phoneme is

removed from the corresponding syllable script, resulting in a one-to-one grapheme to phoneme mapping. This approach could help to use smaller dataset and more efficient models for the same WER.

In the proposed syllable-based approach the text is converted to syllable using the Arabic pronunciation rules, (Alghamdi et al., 2004). If foreign names or words are present in the text, exceptions may occur leading to unknown syllables, or could be forced to follow the Arabic syllable patterns. For example, “street” could be uttered as “es te reet”.

The paper also introduces an innovative method for Arabic speech recognition that utilizes the potential of syllables, a linguistically motivated linguistic unit, in conjunction with a 5-gram language model and the cutting-edge Wav2Vec-2 architecture (Baevski et al., 2020). Embracing a syllable-based approach allows us to overcome the constraints of phoneme-based systems, thereby offering a more resilient and precise solution for Arabic speech recognition.

Syllables, being natural linguistic units in many languages, offer several advantages for Arabic speech recognition. Firstly, syllables offer a concise representation of phonetic data, effectively bridging the divide between phonemes and more extensive linguistic components like words and phrases. This quality enables a more detailed examination of speech, capturing crucial phonetic variations necessary for precise recognition. Secondly, syllables in Arabic are closely linked to morphological units, making them suitable for handling the language's complex word structure and root-based derivational patterns. Leveraging syllables can thus contribute to a more linguistically informed and contextually aware speech recognition system.

To further enhance the accuracy and robustness of the proposed system, a 5-gram language model is integrated. The 5-gram model captures the statistical dependencies between successive syllables, effectively addressing the contextual variations and dependencies present in Arabic speech. By considering sequences of five syllables, the language model can exploit the long-range dependencies that often influence Arabic word formation and pronunciation, leading to improved recognition performance.

As the acoustic model, we adopt the Wav2Vec-2 architecture, which has demonstrated remarkable success in various speech recognition tasks. Wav2Vec-2 employs unsupervised pre-training on vast amounts of unlabelled speech data, allowing it to learn powerful acoustic representations without the need for aligned transcriptions. The pre-trained model is then fine-tuned on a smaller labelled dataset, specifically tailored to the complexities of Arabic speech. The integration of Wav2Vec-2 contributes to the system's robustness against variations in pronunciation, dialects, and ambient noise, making it well-suited for real-world applications.

This paper aims to evaluate the effectiveness of the proposed approach by conducting extensive experiments on a benchmark Arabic speech dataset. The results demonstrate the advantages of employing syllables

alongside the 5-gram language model and Wav2Vec-2, showcasing the system's capability to achieve state-of-the-art performance in Arabic speech recognition.

In summary, this paper introduces a novel approach to Arabic speech recognition that leverages syllables as a fundamental linguistic unit, combined with a 5-gram language model and the powerful Wav2Vec-2 architecture. We anticipate that this combination of linguistic awareness and cutting-edge deep learning techniques will lead to significant advancements in Arabic speech recognition technology, paving the way for more accurate, contextually aware, and robust systems in the future.

The next section provides a brief introduction to the Arabic syllables. Then, we provide a literature review on the state of the art of speech recognition for Arabic language and the use of syllable-based speech recognition. The dataset is described in Section 3.1, followed by the model description in Section 3.2. Section 3.3 covers the Language Model and Decoding followed by the Results and Discussion. Finally, the paper ends with the Conclusion section.

2. Literature review

2.1. **Introduction to Arabic syllables**

Arabic is a Semitic language, and it is one of the oldest languages in the world. It is the 5th widely used language nowadays (Sayed, 2015). Modern Standard Arabic has 28 consonants and 6 vowels, 3 short and three long (Alghamdi et al., 2004). A consonant is a basic speech sound, which is produced when the air flow is partly or fully obstructed at one or more points along the vocal tract. Vowels, on the other hand, are produced when the vocal tract is open and the vocal cords vibrate. Vowels have high energy and distinct formant patterns (Huang et al., 2001). A syllable is a cluster of basic sound units, which is pronounced as a single complete unbroken sound. Words consist of one or more syllables. Arabic is a syllable based language. In Arabic language a syllable must begin with a consonant followed by a vowel, and contains a single vowel, short or long. Modern Standard Arabic MSA Language is based on 5 syllable structures or classes; CV, CVV, CVC. CVVC, and CVCC. Where C stands for a consonant, V stands for a short vowel, and VV stands for a long vowel. The CVCC is a rare syllable and could occur in speech at end of utterance e.g., أمر، شهر، فجر and does not appear in MSA diacritized text. No more than two consecutive consonants could occur in Arabic language (Elshafei, 1991). The Arabic syllable never begins with two consonants. No single phoneme syllable exists in Arabic. All Arabic syllables must contain at least one vowel. Also Arabic vowels cannot be initials and can occur either between two consonants or final in a syllable.

Arabic syllables can be classified as short or long. The CV type is a short one, while all others are long. Syllables can also be classified as open or closed. An open syllable ends with a vowel, while a closed syllable ends with a consonant. For Arabic, a vowel always forms a syllable nucleus, and there are as many syllables in a word as

vowels in it. Syllables play important roles in the articulation, cross articulation, the speech prosodics, and in the speech pronunciation of the allophones of phonemes (Elshafei, 1991). The syllable structure plays a significant role in the Arabic poems and in the music of the language.

A statistical study was conducted using a data set of MSA with full diacritics. The dataset contains 349,981 characters, 39217 words, 4574 lines of text. The text was collected from written Arabic news in the area of Sports and economics. The syllables were generated by a program using the Arabic pronunciation rules in (Alghamdi et al., 2004). The number of all syllables is 101958, from 2498 unique syllables. Table 1 shows each syllable structure and its percentage compared with the whole syllables. Columns 3 and 4 report the most repeated and the least repeated syllables. It is clear that the most frequent syllable is CV, and the least frequent syllable is CVCC. If the last word in a sentence ends with CVC CV, it will be uttered with stop instead of ending with CV, creating a closing CVCC.

According to Table 1, we may infer that on the average, a word with full diacritics contains 8.9 characters. The number of characters per syllable is approximately 3.4, and the average number of syllables per word is 2.66 syllables. However, it was noticed that these statistics were driven from the text only, which include an end case diacritic. If the ending consonant is not vowelized to reflect actual speech utterance more CVCC cases may appear. The above statistics may also differ for text domains, and the size of the text corpus.

Table 1: General Syllables Percentages.

Syllables	%	Most Repeated	Least repeated
CV	42.66957	وَ=3139	1أُ =
CVC	29.76593	تِلْ=1476	* 1=دُضْ
CVV	23.1783	مَا=1025	* 1=ظَا
CVVC	4.385217	لَارْ=202	* 1=وِينْ
CVCC	0.000981	أَمْرْ=1	1= أَمْرْ

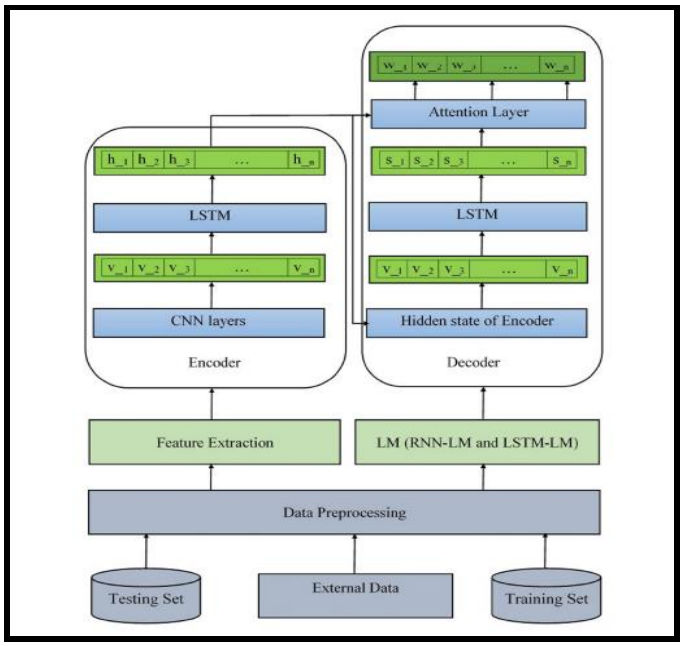
* Indicates that there are other syllables having the same number of occurrences.

Theoretically speaking, the possible number of CV cases is (28 consonants X 3 vowels) = 84 ; for CVC = 28*3*28= 2352, and CVV=84, CVVC=2352, a total of 4872.However, many of them may not appear in the MSA.

2.2. **Related works**

The motivation behind using syllables comes from recent research on syllable-based recognition.
Syllable-based speech recognition was proposed and evaluated for several languages, (Jasmin et al., 2022), (Fantaye et al., 2019), (Aslyan, 2011), (Majewski, 2008), such as Malayalam, Chaha, Turkish, and Polish low resource languages, as well for Mandarin Chinese (Zhou et al., 2018). Syllable role was also investigated in studies of human perception (Kirchhoff, 1996), (Wu et al., 1998), which demonstrate the central role of the syllable played in human perception and generation of speech. One important factor that supports the use of syllables as the acoustic unit for recognition is the relative insulation of syllables from pronunciation variations arising from addition and deletion of phonemes as well as co-articulation.
In the following literature review, we present selected methodologies that have been used for different languages for speech recognition. Developing a speech recognition system with a language like Arabic presents several challenges. The first set of challenges relates to the different dialects of Arabic. While Arabic has millions of native speakers, each area has its dialect of Arabic. Consequently, the number of datasets available for some of these is few while other dialects do not even have datasets available for them (Alsayadi et al., 2022). In addition, some dialects lack a well-defined orthographic system like the Egyptian dialect (Ali et al., 2017). These issues combined led to a lack of performance of the developed Arabic Speech Recognition (ASR) systems compared to what has been achieved in other languages. Furthermore, the absence of the diacritized text (with tashkeel) is a hindering factor in developing high quality speech recognition systems. The studies developing ASR systems using non-diacritized version of Arabic usually leads to a limited accuracy because a non-diactrized Arabic words may have different pronunciation and meaning depending on the context. Finally, the recent work in (Alsayadi et al., 2022) found that focusing on one dialect resulted in better accuracy than developing a multi-dialect system. An approach was demonstrated by students at Zewail City: University of Science and Technology to create an ASR pipeline for Egyptian Arabic dialects by adapting two models and comparing them. These models are bidirectional RNN and unidirectional RNN (Mahrous et al., 2021). They used MGB-3 data set to train their two models. They built their first model by using a gated recurrent unit (GRU), with additive attention. The second model was a bidirectional GRU with attention based layer. They were able to achieve a loss for unidirectional GRU of 0.3473 and a loss of 0.4599 for bidirectional GRU (Mahrous et al., 2021).
Authors of (Alsayadi et al., 2021) propose an end-to-end model for the Arabic ASR system in which they used long short-term memory networks (LSTM). The proposed model makes use of LSTM in learning long term dependencies. In addition, they use CNN to further extract the useful information from the input. As obvious from the proposed model architecture in Figure 1.

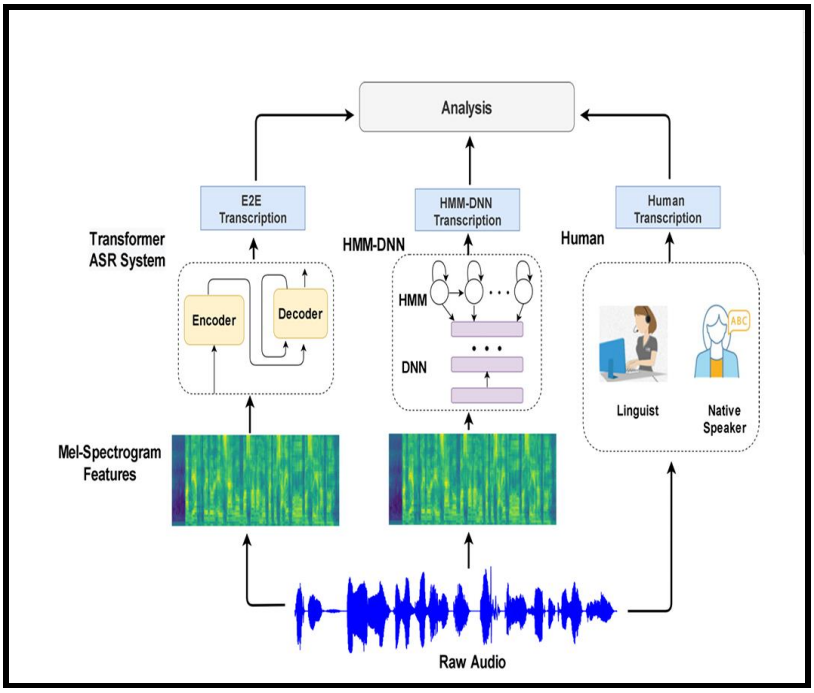
Figure 1: model architecture (Alsayadi et al., 2021)



The first step is pre-processing the data in the corpus, language model and lexicons. Then LSTM and CNN are used for the encoder. The encoder role is to build the acoustic model with CNN to prepare the input as vectors and send them to LSTM layers. For the decoder part, the output from the encoder is received and LSTM with attention decodes the encoder output and obtains the predication. The model was tested in Standard Arabic Single Speaker Corpus (SASSC) and obtained WER 28.48 % . Although the accuracy and WER is not impressive but it gives us more insight on the model and architectures used.

Automatic Speech Recognition (ASR) has made remarkable strides in recent times, achieving levels of accuracy comparable to human transcribers, prompting discussions about whether machines have attained human-level performance. Previous research in this domain has primarily focused on the English language, using modular Hidden Markov Model-deep neural network (HMM-DNN) systems (Hussein et al., 2022). The aimed to fill the gap by conducting an extensive benchmarking study on Arabic speech recognition, encompassing end-to-end Transformer ASR, modular HMM-DNN ASR, and human speech recognition (HSR) across various Arabic dialects. They explore the performance of end-to-end Transformer ASR for Arabic and its dialects, achieving notable Word Error Rates (WER) of 12.5%, 27.5%, and 33.8% for the MGB2, MGB3, and MGB5 challenges, respectively. However, despite the impressive advancements in ASR for Arabic, the study concludes that human performance remains notably superior to machine performance. The authors highlight an absolute WER gap of 3.5% on average, underscoring the continued challenges in achieving human-level recognition accuracy for Arabic speech.

Figure 2: Core Architecture of the paper (Hussein et al., 2022)



The performance of well-trained speech recognition systems, based on high-quality full bandwidth speech data, often suffers when applied in real-world scenarios, especially in telephone speech recognition. The limited bandwidth of transmission channels in telephone communications poses significant challenges for accurate recognition. (Azmi & Tolba, 2008) investigated telephone speech recognition of Egyptian Arabic speech using syllables as acoustic units. The authors designed a speaker-independent Hidden Markov Models (HMMs)-based speech recognition system using the (HTK) tool kit. The database used for both training and testing comprised recordings from forty-four Egyptian speakers. They found that syllable based speech recognition has better accuracy in noisy environment over the conventional phoneme based speech recognition.

One of the advantages of using syllables lies in their longer time frame, typically encompassing two or more phones. This extended duration offers a more parsimonious framework for modelling pronunciation variations in spontaneous speech, contributing to the system's robustness and adaptability to real-world environments.

Additionally, syllable-based recognition requires a relatively smaller number of acoustic units, resulting in faster processing times compared to word-based recognition. This efficiency makes syllable-based ASR systems more suitable for real-time applications and more computationally efficient.

3. Methodology

3.1. **The Arabic News Corpus Dataset**

The primary dataset utilized in this research, referred to as the Arabic News Corpus, is based on TV news transcriptions in diacritized Modern Standard Arabic (MSA). It comprises 15 hours of WAV recordings, annotated with diacritized Arabic syllables. The original corpus was about 5.7 hours developed by a grant from King Abdulaziz City of Science and Technology (KACST) in 2008. Since then, it has been expanded and refined by many grants from public, institutions, and private sources; an additional level of annotation is available, providing fully diacritized word-level transcriptions.

Importantly, the Arabic News Corpus was prepared using an automated software tool that converts the text into syllables. While this process enabled us to generate a substantial volume of data necessary for our task, it is important to note that the automated nature of syllabification may have introduced some errors into the dataset due to errors in the diacritized text or due to errors in the network of pronunciation rules. We estimate these errors to be around 3% of the dataset. Despite this potential source of inaccuracies, our model has been able to deliver significant results, attesting to its robustness and capacity to handle minor inconsistencies in the training data.

The dataset was partitioned into training, validation, and testing sets. The training set was allocated the largest portion, with 12.5 hours of recordings. The validation set was assigned 1 hour, and the remaining 1.5 hours of recordings were reserved for the testing set.

The size of a dataset plays a crucial role in determining the methodology, with smaller datasets potentially necessitating a more specific approach. Our dataset, with a broad range of sound file durations, offers a robust foundation for our study. The average duration of each sound file is 6 seconds, with a maximum of 22 seconds and a minimum of 0.6 seconds.

In the initial stages, inconsistencies were identified in the sampling sizes across the sound files, which led to conflicts during the training phase. These discrepancies were subsequently rectified during pre-processing, thereby ensuring uniformity across the dataset and facilitating a more streamlined training process.

3.2. **Model**

The Wav2Vec 2.0 model serves as a comprehensive architecture for speech recognition tasks. It incorporates a multi-layer convolutional feature encoder, which ingests raw audio input and produces latent speech representations. These representations are then fed to a Transformer that develops context-aware encodings capturing information from the entire sequence. A unique aspect of Wav2Vec 2.0 lies in its use of a quantization module, which discretizes the feature encoder's output into a finite set of speech representations using product quantization. This model has the capability to capture dependencies over the entire sequence of latent representations end-to-end, providing robust speech recognition capabilities. The design of the

Wav2Vec 2.0 model facilitates efficient processing of raw audio data, successfully capturing important features of the speech that prove beneficial for a range of downstream tasks.

3.3. **Fine Tuning**

Wav2Vec 2.0 provides a multitude of pre-trained checkpoints for fine-tuning, including a multilingual variant, XLSR, trained over 56,000 hours across 53 languages. For our task of Arabic syllable recognition, we fine-tuned this model by minimizing the Connectionist Temporal Classification (CTC) loss, enabling the model to deliver effective results.

In the fine-tuning process, we employed a specially tailored vocabulary from our text data consisting of 3086 tokens, representing the unique syllables in the data set, including padding, unknown, and word delimiter tokens. This vocabulary was designed to encapsulate the diverse elements of the speech data effectively. We utilized the wav2vec tokenizer and feature extractor to process the data. The tokenizer converted our data into a token format that the wav2vec model could interpret, while the feature extractor transformed the raw audio data into suitable features for the model's input. Notably, the first component of XLSR, a stack of Convolution Neural Networks (CNN) layers for extracting acoustically meaningful but contextually independent features, required no further fine-tuning. This combination of a custom vocabulary and wav2vec-specific tools, along with the effective utilization of pre-training, enabled efficient fine-tuning of the model for Arabic syllable recognition.

Data augmentation techniques were incorporated during the fine-tuning process to enhance the model's robustness and generalization capabilities. These techniques included Gaussian Noise injection with an amplitude range of 0.001 to 0.012, TimeStretch with a rate between 0.8 and 1.25, and PitchShift with semitone alteration ranging from -4 to 4.

The fine-tuning was conducted under the following hyper parameters:

Table 2: used hyper parameters

Hyperparameter	Value
Space augment	True
Attention Dropout	0.001
Hidden Dropout	0.008
Feature Proj Dropout	0.0
Mask Time Prob	0.1

Layerdrop	0.0001
Per Device Train Batch Size	2
Gradient Accumulation Steps	8
Num Train Epochs	20
Learning Rate	1e-5
Eval Steps	500

The combination of effective utilization of pre-training, strategic data augmentation, and carefully selected hyper parameters, enabled efficient fine-tuning of the model for Arabic syllable recognition.

3.4. **Language models and decoding**

Our system incorporates a 5-gram language model (LM) trained on our training transcriptions as well as a large corpus of specifically curated Arabic syllables. This strategy enables effective capture of the statistical properties of Arabic syllable sequences, which is crucial for our task. The hyper parameters for the language model were optimized based on the performance on our validation set. For the test performance measurement, we employed a beam search decoder with a beam size of 1500, ensuring a comprehensive search over potential outputs.

4. Results

Our model was evaluated under various conditions, as shown in Table 3, to investigate the impact of different components and conditions on its performance.

Table 3: Experiments results

Experiment	WER
Model without Data Augmentation	0.175
Model with Data Augmentation	0.125
Model without 5-gram Language Model (LM)	0.125
Model with 5-gram Language Model (LM)	0.066

The use of data augmentation techniques significantly improved performance, with the Word Error Rate (WER) decreasing from 0.175 to 0.125. Furthermore, the incorporation of the 5-gram Language Model (LM) resulted in a further reduction of the WER to 0.066. The evaluation of the model's performance on text with and without Tashkeel provides insight into the model's ability to handle Arabic text in different forms.
Our final model, which incorporates data augmentation techniques and the 5-gram Language Model (LM), was evaluated under various conditions, as shown in Table 4.

Table 4: Results with different metrics

Metric	With Tashkeel	Without Tashkeel
WER	0.06624	0.05959
CER *	0.03044	0.03994
MER	0.06577	0.05916
WIL	0.11201	0.09928
WIP	0.88798	0.90071

*Notice for CER: With Tashkeel, CER count the letter with Tashkeel as one character

In this table, WER stands for Word Error Rate, CER for Character Error Rate, MER for Match Error Rate, WIL for Word Information Lost, and WIP for Word Information Preserved. The results suggest that our model performs effectively in Arabic syllable recognition regardless of the presence or absence of Tashkeel.
Below are samples transcription based on test set:

Table 5: transcription sample

Reference	**Prediction**
مِمْ مَاْ عَ مَ لَ عَ لَى تَوْ لِيْ دِ عَ وَاْ ءِ دَ سِ يَاْ حِيْ يَ تِنْ تُ قَدْ دَ رُ بِ عَرْ بَ عَ تِ مِلْ يَاْ رَاْ تِنْ وَ ثَ لَاْ ثِ مِ ءَ تِ مِلْ يُ وُنْ دُوْ لَاْرْ	مِمْ مَاْ عَ مَ لَ عَ لَى تَوْ لِيْ دِ عَ وَاْ ءِ دَ سِ يَاْ حِيْ يَ تِنْ تُ قَدْ دَ رُ بِ عَرْ بَ عَ تِ مِلْ يَاْ رَاْ تِنْ وَ ثَ لَاْ ثِ مِ ءَ تِ مِلْ يُ وُنْ دُوْ لَاْرْ
وَ اِثْ نَيْ نِ وَ سِتْ تِيْ نَ فِلْ مِ ءَهْ	وَ ءِثْ نَيْ نِ وَ سِ تِيْ نَ فِلْ مِ ءَهْ

5. Discussion

This work presents an effective approach to syllable-based Arabic speech recognition using a model that incorporates data augmentation techniques and a 5-gram language model. However, two key points emerge for future enhancements:

- Dataset Errors: The Arabic News Corpus, prepared via automated software, contains an estimated 3% error rate. Refining this process could improve data accuracy and potentially enhance model performance.
- Language Model Complexity: While our 5-gram language model contributed significantly to performance, it is conceivable that a more complex model, such as a transformer-based one, could yield improved results with an increased data volume.

In order to showcase the competitiveness of our approach, we conducted a further comparison with existing models. Specifically, we assessed the performance of our two-step model, employing the wav2vec model and a 5-gram language model for syllable recognition, followed by mT5 model for syllable-to-word conversion. This was compared with Google's API, which offers a direct speech-to-text conversion service. For a fair comparison, we evaluated both systems on the first 100 audio samples from the test split of the Common Voice Arabic dataset ("common_voice · Datasets at Hugging Face"). This comparison is done without Tashkeel, so it can give better indication relative to Google API. The mT5 model is a highly versatile pre-trained text-to-text transformer with extensive multilingual capabilities. This model is essentially a multilingual adaptation of T5, having undergone pre-training on a dataset encompassing 101 languages. Consequently, it is a commendable choice for text-to-text tasks. The mT5- small model achieved accuracy for Arabic Language reaching 65.2% (Xue et al., 2020).

Table 6: Comparison with Google API.

metric	Our pipeline	Google API
WER	0.412	0.459
CER	0.135	0.127
MER	0.409	0.446
wil	0.630	0.680
wip	0.369	0.319

Despite training on a significantly smaller dataset of approximately 15 hours of speech, our model outperforms Google's API in key metrics such as WER, MER, and WIP. However, Google's API shows marginally better CER, areas our model could potentially improve upon. The disparity in training data highlights our model's potential and efficiency, suggesting further enhancements with larger, more diverse training datasets.

To foster transparency and reproducibility in research, our models are publicly available on the Hugging Face model hub. The model trained with a 5-gram language model can be accessed with Syllables to text model in the last two references.

6. Conclusion

We introduce a ground-breaking approach to Arabic speech recognition by incorporating three key elements: syllable-based segmentation, a 5-gram language model, and the Wav2Vec-2 architecture. Through meticulous experimentation and evaluation, we have demonstrated the remarkable efficacy of this novel method.

The use of syllable-based segmentation addresses the inherent complexity of the Arabic language's rich morphological structure and enables the capture of fine-grained phonetic details. This unique approach contributes to a substantial improvement in recognition accuracy, paving the way for more reliable and practical speech recognition systems in Arabic.

Moreover, the integration of a 5-gram language model effectively handles the linguistic intricacies of Arabic, including word-level variations and contextual dependencies. This linguistic modelling enhances the overall system performance and ensures more precise recognition results even in challenging linguistic contexts.

The adoption of the Wav2Vec-2 architecture as the acoustic model further enhances the system's robustness. By leveraging unsupervised pre-training on extensive unlabelled speech data and fine-tuning on a carefully curated labelled dataset for Arabic speech recognition, the system acquires resilient acoustic representations, making it capable of coping with variations in pronunciation and ambient noise.

Our comprehensive experiments have yielded impressive Word Error Rates (WER) of 0.06624 with Tashkeel and 0.05959 without Tashkeel, underscoring the exceptional performance of our proposed approach.

In conclusion, this paper's contributions are twofold: the introduction of a novel method that harnesses the power of Arabic syllables, a 5-gram language model, and the Wav2Vec-2 architecture for speech recognition, and the validation of its efficacy through extensive evaluation. We believe that this research has a transformative impact on the field of Arabic speech recognition and will inspire further advancements in the development of robust and accurate speech recognition systems for Arabic and other complex languages.

7. References

AbuZeina, D., Al-Khatib, W., Elshafei, M., & Al-Muhtaseb, H. (2011). Cross-word Arabic pronunciation variation modeling for speech recognition. *International Journal of Speech Technology*, *14*(3), 227–236. https://doi.org/10.1007/s10772-011-9098-0

Alghamdi, M., Almuhtasib, H., & Elshafei, M. (2004). Arabic phonological rules. *King Saud University Journal: Computer Sciences and Information*, *16*, 1-25.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

Sayed, F. (2015.). *A few surprising facts about the Arabic language*. British Council. https://www.britishcouncil.org/voices-magazine/surprising-facts-about-arabic-language

Huang, X., Acero, A., Hon, H. W., & Reddy, R. (2001). *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR.

Elshafei, M. (1991). Toward an Arabic text-to-speech system. *The Arabian Journal for Science and Engineering*, *16*(4B), 565-583.

Jasmin,S., Samuel, A. A., & Rajan, R. (2022). A study on conventional and syllable-based approaches for automatic speech recognition in Malayalam. *Sādhanā*, *47*(4). https://doi.org/10.1007/s12046-022-02058-z

Fantaye, T. G., Yu, J., & Hailu, T. T. (2019, December). Syllable-based Speech Recognition for a Very Low-Resource Language, Chaha. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence* (pp. 415-420).

Aşlyan, R. (2011). Syllable Based Speech Recognition. *Speech Technologies*, 263.

Majewski, P. (2008). Syllable based language model for large vocabulary continuous speech recognition of polish. In *Text, Speech and Dialogue: 11th International Conference, TSD 2008, Brno, Czech Republic, September 8-12, 2008. Proceedings 11* (pp. 397-401). Springer Berlin Heidelberg.

Zhou, S., Dong, L., Xu, S., & Xu, B. (2018). Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese. *arXiv preprint arXiv:1804.10752*.

Kirchhoff, K. (1996, October). Syllable-level desynchronisation of phonetic features for speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 4, pp. 2274-2276). IEEE.

Wu, S. L., Kingsbury, E. D., Morgan, N., & Greenberg, S. (1998, May). Incorporating information from syllable-length time scales into automatic speech recognition. In *Proceedings of the 1998 IEEE*

International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181) (Vol. 2, pp. 721-724). IEEE.

Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., Alotaibi, B., & Fayed, Z. T. (2022). Deep investigation of the recent advances in dialectal Arabic speech recognition. *IEEE Access*, *10*, 57063-57079.

Ali, A., Vogel, S., & Renals, S. (2017, December). Speech recognition challenge in the wild: Arabic MGB-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 316-322). IEEE.

Mahrous, M., Nousir, A., Wael, A., & Elshafei, M. (2021). Speech recognition using Attention Based Bidirectional RNNs for Egyptian Arabic dialect. *CIE556, Term Paper.*

Alsayadi, H. A., Abdelhamid, A. A., Hegazy, I., & Fayed, Z. T. (2021). Arabic speech recognition using end-to-end deep learning. *IET Signal Processing*, *15*(8), 521-534

Hussein, A., Watanabe, S., & Ali, A. (2022). Arabic speech recognition by end-to-end, modular systems and human. *Computer Speech & Language*, *71*, 101272.

Azmi, M. M., & Tolba, H. (2008, July). Syllable-based automatic Arabic speech recognition in a noisy environment. In *2008 International Conference on Audio, Language and Image Processing* (pp. 1436-1441). IEEE.

Common_voice · datasets at hugging face. common_voice · Datasets at Hugging Face. (2021). https://huggingface.co/datasets/common_voice/viewer/ar/test(accessed Jul. 30, 2023).

Hugging face – the AI community building the future. Hugging Face – The AI community building the future. (2023.).

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., ... & Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

https://huggingface.co/IbrahimSalah/Arabic_speech_Syllables_recognition_Using_Wav2vec2/settings (accessed Jul. 30, 2023).

Hugging face – the AI community building the future. Hugging Face – The AI community building the future. (2023).

https://huggingface.co/IbrahimSalah/Arabic_Syllables_to_text_Converter_Using_MT5/settings (accessed Jul. 30, 2023).

8. نبذة تعريفية عن الباحث أو الباحثين

بيانات الباحث إبراهيم عبدالعال،طالب حديث التخرج في في (قسم الاتصالات وتكنولوجيا المعلومات) (كلية الهندسة) في جامعة العلوم والتكنولوجيا بمدينة زويل بـ(مصر). تدور اهتماماته البحثية حول تعرف على الكلام, تعلم متعمق , استعمال الذكاء الصناعي مع اللغة العربية. محمد عبد الواحد،طالب حديث التخرج في في (قسم الاتصالات وتكنولوجيا المعلومات) (كلية الهندسة) في جامعة العلوم والتكنولوجيا بمدينة زويل بـ(مصر). تدور اهتماماته البحثية حول تعرف على الكلام, تعلم متعمق , استعمال الذكاء الصناعي مع اللغة العربية. مصطفى الشافعى ، أستاذ/ مدير برنامج هندسة الاتصالات وتكنولوجيا المعلومات بمدينة زويل: جامعة العلوم والتكنولوجيا بجمهورية مصر العربية . حصل على درجة الدكتوراة من جامعة مكجيل بكندا عام 1982 وتدور اهتماماته البحثية حول استعمال الذكاء الصناعي في اللغة العربية والروبوتات والتطبيقات الصناعية.	**AUTHOR BIODATA** Ibrahim Abdalaal is a recent graduate student in the Department of Communications and Information Engineering, College of Engineering, University of Science and Technology in Zewail City. His research interests include Speech Recognition, Deep Learning and Using Artificial Intelligence with Arabic Language. Mohamed Abdelwahed is a recent graduate student in the Department of Communication and Information Engineering, College of Engineering, University of Science and Technology in Zewail City. His research interests include Speech Recognition, Deep Learning and Using Artificial Intelligence with Arabic Language. Moustafa Elshafei is a Full Professor and Director of the Department of Communications and Information Engineering, College of Engineering, University of Science and Technology in Zewail City. He obtained his Ph.D. from McGill University, Canada, in 1982. His research interests include Arabic Speech Recognition and Artificial Intelligence applications in Robotics and Industrial Automation.

معرف أوركيد

(ORCID: 0000-0002-8205-666X) :

Email: Moelshafei@zewailcity.edu.eg.