Arabic Speech Recognition of zero-resourced Languages: A Case of Shehri (Jibbali) Language

Norah Alrashoudi[1], Omar Said Alshahri[2], Hend Al-Khalifa[1]

[1] Department of Information Technology, College of Computer and Information Science, King Saud University, Riyadh, Saudi Arabia.

[2] Islamic Sciences Institute, Diwan of the Royal Court, Salalah, Sultanate of Oman.

omar9297@gmail.com, author3@hhh.com

Abstract

Many under-resourced languages face data scarcity issues due to a lack of standardized writing systems, making ASR training more challenging and costly. However, there's a growing need to adapt ASR for indigenous languages to support language documentation, preservation, and the development of learning materials for these communities. Shehri or Jibbali, a spoken language in Oman, lacks extensive annotated speech data. This paper aims to investigate transfer learning techniques to develop an ASR model for this under-resourced language. We collected a Shehri (Jibbali) speech corpus and utilized transfer learning by fine-tuning pre-trained ASR models on this dataset, including Wav2Vec2.0, HuBERT and Whisper. Evaluation using word error rate (WER) and character error rate (CER) showed that the Whisper model, fine-tuned on the Shehri (Jibbali) dataset, significantly outperformed other models, with the best results from Whisper-medium achieving 3.5% WER. This demonstrates the effectiveness of transfer learning for resource-constrained tasks, showing high zero-shot performance of pre-trained models.

Keywords: Automatic Speech Recognition (ASR), Speech Processing, Transfer Learning, Zero-Resource Languages, Indigenous Languages

1. Introduction

Languages with rich linguistic resources often have extensive corpora and annotated speech data which facilitate the development of accurate and robust Automatic Speech Recognition (ASR) systems. In contrast, languages with limited or zero resources face data scarcity issues, making it a challenge to train ASR models effectively. These languages often lack a standardized writing system, or their written form may be limited to a small number of experts (SIL International, 2022), (R. Coto-Solano et al, 2022). This complicates the transcription process, making it more challenging and costly compared to widely spoken languages. Despite these challenges, there is an increasing need to adapt ASR to work effectively on indigenous languages. One of the main reasons is to support indigenous communities in documenting their languages and preserving their linguistic heritage. Moreover, such adaptations enable these communities to develop learning materials for their languages and facilitate their continuous use (R. Coto-Solano et al, 2022).

One such under-resourced language is Shehri, also known as Jibbali, spoken in Oman. Shehri lacks extensive annotated speech corpora, making conventional supervised training approaches difficult to apply for building an ASR system. However, with a dropping number of fluent speakers, particularly among younger generations, there is a need to develop technological tools that can help document the language.

This study aims to investigate the application of transfer learning techniques to develop an initial ASR capability for Shehri (Jibbali) language without requiring a large, annotated dataset. Therefore, the main contributions of this work are as follows:

1) Collection of a Shehri (Jibbali) speech dataset,

2) Fine-tuning pre-trained ASR models like Wav2Vec2.0, HuBERT, and Whisper on the Shehri dataset, and

3) Evaluation of the adapted models on Shehri (Jibbali) using word error rate and character error rate metrics.

The rest of the paper is structured as follows: Section 2 provides an overview of the Shehri (Jibbali) language. Section 3 discusses related work in under-resourced languages ASR. Section 4 describes the methodology adopted including dataset collection and model fine-tuning. Section 5 presents the results and analysis. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Shehri (Jibbali) Language

The Dhofar Governorate (محافظة ظفار) is situated in the southernmost region of the Sultanate of Oman, bounded to the east by the Al Wusta Governorate, and to the north and northwest by the Rub' al Khali desert, while sharing its southwestern border with Yemen (Oman Encyclopedia, 2013, p. 2321). Additionally, it shares a frontier with Saudi Arabia to the northwest. It encompasses ten administrative divisions, including Salalah, Taqah, Mirbat, Sadah, Shalim, the Halaniyat Islands, Thumrait, Muqshin, Al Mazunah, Dhalkut, and Rakhyout (Ministry of Information, 2020, p. 67).

The population stands at 416,458 individuals (as per the 2020 census). Covering an area of approximately 99,300 km² (Oman Encyclopedia, 2013, p. 2321), the Dhofar Governorate presents a rich linguistic tapestry despite its relatively modest size. It hosts a diverse array of contemporary South Arabian languages, including Shehri (Jibbali), Mahri, Bathari, and Hobyot, alongside Arabic dialects with close affinities to North

Arabian, encompassing both urban and Bedouin variants (Al-Kathiri et al., in press). The Shehri (jibbali) language is referred to by different names among its speakers, either exclusively among Shehri speakers or exclusively among Jibbali speakers. However, in research studies, it also sometimes appears with a combination of both, in order to address ambiguities (Al-Hafeezh, 1987; Johnstone, 1981; Rubin, 2014). This language shares 25 letters with Standard Arabic, which are: (/ʔ/ أ, /b/ب, /t/ت, /θ/ ث, /dʒ/ ج, /ħ/ ح, /x/ خ, /d/ د, /ð/ ذ, /r/ ر, /z/ ز, /s/ س, /ʃ/ ش, / tˤ/ ط, /ðˤ/ ظ, /ʕ/ ع, /ɣ/ غ, /f/ ف, /k/ ك, /l/ ل, /m/ م, /n/ ن, /h/ ه, /w/ و, and /j/ ي). It has retained old methods of pronouncing some letters (Al-Mashani, S, 2017). Researchers have proposed various alphabets for these letters, and after consulting with experts Watson and Al-Kathiri, the appropriate alphabet was settled upon Al-Kathiri et al., (2024), which is as follows: (/ɬ/ڛ , /ɮ/ ڸ, /tʃʼ/ ض, /g/ ج, /ʒʷ/ ج, /ʃ/ش , /ʃʷ/ش , /sʼ/ ص, /ɬʼ/ ض, and /kʼ/ ق). The common letters between Shihri (Jibbali) and Classical Arabic are similar in pronunciation, with some differences in letter characteristics. Some of them correspond to Arabic in both articulation and characteristics, while the other common letters correspond to Arabic in articulation but differ in some characteristics. For further details on the characteristics of these alphabets, refer to Al-Mashani, S,(2017); Watson & Al-Kathiri, (2022).

3. Literature Review

Previous studies have explored the challenges and opportunities in implementing ASR systems for languages with limited resources. R. Coto-Solano et al, (2022) and Gupata et al., (2020) analyzed the challenges involved in the transcription of spoken audio recordings in indigenous languages. In addition, Stan et al., (2022) conducted a comprehensive analysis of the challenges involved in developing ASR systems. Their work highlighted issues such as a lack of annotated data, phonetic variations, and the importance of cultural context in acoustic modeling. Some recent advancements in ASR have leveraged self-supervised learning (SSL) techniques to address resource constraints. A study conducted by Chen et al., (2023) demonstrated the efficiency of the SSL in adapting pre-trained models to indigenous languages, mitigating the need for extensive language-specific training data. In the context of multi-lingual ASR, Arisaputra et al., (2024) evaluated the performance of the XLS-R model on various low-resource languages. They incorporated a 5-gram KenLM into the optimized model and it has resulted in a significant decrease in the Word Error Rate (WER). In addition, Zellou et al., (2024) investigated a cross-language ASR transfer approach for the low-resource Tashlhiyt language, which shares similar phonological inventories with Arabic. Their experiment utilized a commercially available Arabic ASR system without any modifications for the target language, resulting in approximately 45% accurate word transcription. Furthermore, Woldemariam et al., (2020) investigated the efficiency of transfer learning to improve the performance of ASR for the under resourced Semitic language (Amharic). They utilized Deep Neural Network (DNN) acoustic models pre-trained on English and Mandarin as source languages, adapting them to Amharic. Experimental results demonstrate significant improvements through transfer learning compared to the baseline Amharic model. The best enhancements were observed with models transferred from English, achieving WER reductions of 5.75% and 8.06%. In contrast, the Mandarin model achieved a WER reduction of 14.65%, while the baseline only improved by 38.72%.

4. Methodology

The methodology of this study aims to improve an ASR model for a zero-resource language, Shehri (Jibbali) language. We collected speech data from Shehri (Jibbali) speakers and constructed a dataset for the training of ASR model. The study leveraged the efficiency of transfer learning to adapt a pre-trained ASR model to our specific task.

Transfer learning involves leveraging knowledge from pre-trained models on large-scale datasets and adapting them to perform specific tasks or domains with smaller, task-specific datasets (N. Das et al., 2021). This approach allows ASR systems to benefit from the generalization and feature extraction capabilities learned from the pre-training phase, improving performance and reducing the need for extensive labeled data in the target domain (Neyshabur et al., 2020). Transfer learning in ASR typically involves fine-tuning pre-trained models on task-specific data.

In the following subsections, we provide details about our Shehri (Jibbali) speech dataset, give an overview of the fine-tuned models, present our Shehri (Jibbali) ASR model, and explain the model evaluation criteria.

4.1 Dataset

The dataset for Shehri (Jibbali) language speech was collected from 30 speakers, including 23 males and 7 females. Speakers represent the eastern and central parts of Oman, and the western part adjacent to the central part (As'aib region), due to the similarity of dialects in these regions. Informed consent was obtained from all subjects involved in the study.

Each speaker uttered 15 sentences and repeated each sentence 5 times, with a total of 75 utterances for each speaker. In the selection of sentences, we focused on Shehri (Jibbali) phonemes that are not represented in the Arabic language, to ensure the model can effectively distinguish these unique sounds. Table 1 represents the selected sentences with their corresponding Arabic and English translations. The total duration of the dataset is 1 hour and 54 minutes, with an average duration of 3 seconds for each file. The dataset is publicly available through a GitHub repository[1].

English Translation	Arabic Translation	Shehri (Jibbali) Sentences
Light the fire and fetch the firewood	أشعل النار، وأحضر الأثافي	اعلق پسوط، بغد هير اوقودر
When the pot is full, turn off the water.	إذا امتلأ القدر أغلق الماء	هير اصفريت ميلوت قفل اميه
Watch out for the children so they don't fall into the water	انتبه للأطفال كي لا يسقطوا في الماء	اقول لقيلون او يهبي عق اميه
Let them just play on the beach	دعهم فقط يلعبوا على الشاطئ	قلع هوم بس ينحوج ظير حض
If you exit 'Madinat al-Haqq' you will see my car on the road	إذا خرجت من "مدينة الحق" سترى سيارتي على الطريق	هير تيروفك بخيضول أتيپنأ سيارهي ظير اورم
This man often opposes me	هذا الرجل كثيرا ما يعارضني	اغيج ذان يكين ار ديجحود تو
Did the funeral arrive or is it still there?	هل الجنازة وصلت أم ؟لا زالت هناك	اچينوزت بروت زحوت من دعوت لحك
I will go and see if the ewe has given birth or is still in labor	سأذهب وأنظر هل ربّت الغنمة أم لا زالت	الغد لپنأ اووز بيروت غيچوت من دعوت
Do you think they are still asleep until now?	أتظنهم لا زالوا ؟نائمين إلى الآن	تعمورهم دعود ؟دشيف اد ناصنو
If you want it strong, put a lot of tea	إذا أردته جيدا ضع الكثير من الشاي	هير عك تش إصلح ازد شاهي حور
Your whole head is gray	رأسك كله شيب	ارشك بير كلش پسوب
He went in the morning to fetch provisions and has not returned yet	ذهب صباحا ليحضر الزاد ولم يعد بعد	بير اغد كحصف هير خصور بعود اوزحم لو
There appeared among them a wise man	ظهر فيهم رجل حكيم	ضهر عمقوهم غيج بيصير
If your foot falls asleep, don't walk on it	إذا تنملت رجلك فلا تمشي عليها	هير حيضوت فعمك اوتركت ليس لو
They spent their day searching, but they found nothing	ظلوا يومهم يبحثون ولكن لم يجدوا شيئا	قـهب يوهم ديغولق ار هيس او كسئ پسي لو

Table 1 : Selected sentences for Shehri (Jibbali) speech dataset

For the training of Shehri (Jibbali) ASR model, we split the data into 80% for the training and 20% for the testing. Before splitting the dataset, we shuffled it to ensure a random distribution of the data. Table 2 represents details of the data division.

Subset	**Utterances**	**Duration**
Training	1800	1 hr. and 31 min
Testing	450	22 minutes
Total	2250	1 hr. and 54 min

Table 2: Dataset Summary of the Training and Testing Subset Statistics

4.2 Fine-Tuned Models

In this study, we selected three large-scale pre-trained models, including Wav2Vec2.0, HuBERT, and Whisper. We provide an overview description for each model in the following subsections.

4.2.1 Wav2Vec2.0

Wav2Vecc2.0 is a speech representation framework based on self-supervised learning, enabling the extraction of rich features from raw audio data without the need for annotations or labeled data. The framework was pre-trained on a large quantity of unlabeled data and leveraged Transformer architecture to achieve remarkable performance in speech-related tasks (Baevski et al., 2020). Wav2Vec2 employs a multi-stage architecture consisting of several key components, as shown in Figure 2. The speech features are extracted from raw audio using a CNN, followed by a Transformer layer for contextualized representation aggregation. Self-supervised training involves discretizing the output of the feature encoder into a finite set of speech representations using product quantization. Wav2Vec2.0 offers multiple models with varying parameters and training datasets. The base model, called 'Wav2vec2-base-960h', was trained with over 94 million parameters on 960 hours of the Librispeech corpus, which is designed for native English speakers. Additionally, a large-scale multi-lingual pre-trained model known as 'Wav2Vec2-XLS-R-300M' was pre-trained with up to 300 million parameters on 436,000 hours of unannotated speech data collected from diverse corpora spanning 128 languages.

[1] https://github.com/iwan-rg/Shehri-Jibbali-Speech-Dataset

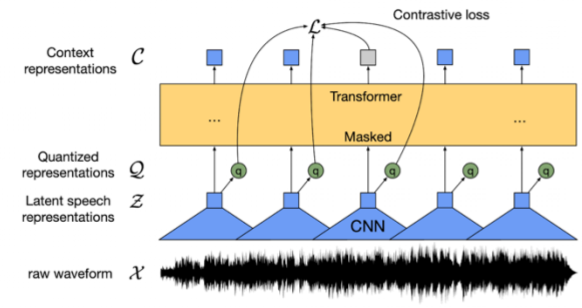

Figure 2: The architecture of Wav2Vec2.0 (Baevski et al., 2020)

4.2.2 HuBERT

Hidden unit BERT (HuBERT) is a self-supervised speech representation framework learned by masked prediction of hidden units (Hsu et al., 2021). Figure 3 shows the architecture of HuBERT framework. HuBERT integrates an offline clustering step for BERT-like pre-training with noisy labels. It utilizes a BERT model on masked continuous speech features to predict predetermined cluster assignments, focusing the predictive loss on masked regions to learn robust high-level representations. This setup enables simultaneous learning of acoustic and language models from continuous inputs, addressing acoustic modeling challenges and capturing long-range temporal relations in learned representations. HuBERT model was pre-trained on either standard Librispeech 960h or the Libri-Light 60k hours on three model sizes, including Base (90M parameters), Large (300M parameters), and X-Large (1B parameters).

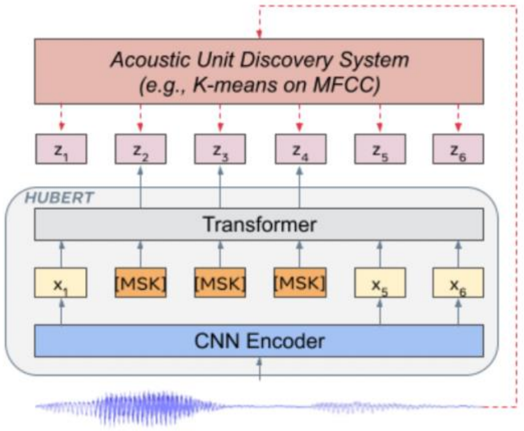

Figure 3: The architecture of HuBERT (Hsu et al., 2021)

4.2.3 Whisper

Whisper is a large-scale speech representation framework based on a weakly-supervised approach, that was pre-trained on 680,000 hours of labeled audio data, including English speech and multilingual data covering 96 languages to perform two different tasks: speech recognition and speech translation (Radford et al., 2022). The English-only models were trained on speech recognition tasks, whereas the multi-lingual models were trained on speech recognition and speech translation tasks to predict the transcription of different languages. Figure 4 illustrates the architecture of the Whisper, featuring an encoder-decoder Transformer chosen for reliability and scalability. Audio data is resampled to 16,000 Hz and transformed into an 80-channel log-magnitude Mel spectrogram representation. The encoder consists of a stem with two convolutional layers followed by sinusoidal position embeddings and pre-activation residual blocks. The decoder utilizes learned position embeddings and tied input-output token representations. Both the encoder and decoder have the same width and number of transformer blocks for consistency in processing the input and generating the output. Whisper was pre-trained on several models with different numbers of parameters, ranging from 39M to 1.5B parameters.

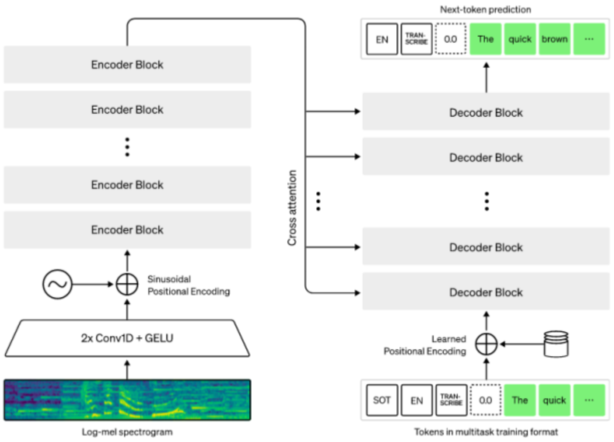

Figure 4: The architecture of Whisper (Radford et al., 2022)

4.3 Shehri (Jibbali) ASR Models

To implement an ASR model for the Shehri (Jibbali) language, we utilized a transfer learning approach by fine-tuning several pre-trained models, including Whisper (Radford et al., 2022), HuBERT (Hsu et al., 2021), and Wav2Vec2.0 (Babu et al., 2021) on our constructed speech dataset for Shehri (Jibbali) language. For Wav2Vec2.0, we selected the XLS-R model as a large-scale model for cross-lingual speech data that was trained on 436K hours of unannotated speech data including 128 different languages. For HuBERT, we selected the large model that was trained on both Libri-Light 60k and LibriSpeech 960 hours of speech data. For Whisper, we trained the base, tiny, small, medium, and large-v3 models with varying numbers of parameters, ranging from 39M to 1.5B of parameters.

Training Details. Models were trained on NVIDIA Tesla T4 GPU with 54GB of memory and CUDA version 12.2. We utilized Huggingface trainer[2] to train each model, and PyTorch (version 2.1.0+cu121) to perform GPU-accelerated training. The pre-processing step was applied to both audio and textual data. The transcription texts contain some punctuation marks, such as '?' and ',', then we normalized the text by removing these marks before training. Additionally, the audio data was re-sampled to 16kHz and converted the raw waveform of the speech signals into a floating array. During the fine-tuning process, we selected similar configurations and hyperparameter settings for both the XLS-R and HuBERT models, because they were implemented on closely related architectures. According to (Babu et al., 2021), we trained these models with a learning rate of 3e-4, 500 warmup steps, 20 epochs, 16 for the batch size, and no weight decay. Table 3 shows the hyperparameters of all fine-tuned models.

For training Whisper models, we encountered issues related to the GPU and computational resources due to its huge number of parameters. To address these issues, we applied some parameter-efficient fine-tuning (PEFT) techniques for model optimization and improving the training process. PEFT is a technique employed in Natural Language Processing (NLP) and ASR to enhance the effectiveness of pre-trained language models on specific downstream tasks. It aims to decrease the hyperparameter numbers for the large-scale language models, which minimizes the computational resources and time compared to the training of the entire model (Z. Fu et al., 2023). We trained Whisper models using two PEFT methods, named Int8 matrix multiplication for Transformers at scale (LLM.int8) (Dettmers et al., 2022) and low-rank adaption of large language models (LORA) (E. J. Hu et al., 2023). LLM.int8 was utilized to lower the precision of floating-point data types, thereby reducing the memory required to store model weights. The LORA approach involves freezing the weights of the pre-trained model and incorporating trainable rank decomposition matrices into each layer of the Transformer architecture, which reduces the number of trainable parameters. After performing these methods, the number of parameters in Whisper models has reduced to utilize only 1% to 1.5% of all trainable parameters. For example, the number of parameters of the medium model has been reduced from 9.4 M to 773K parameters, which improves the performance of the training process using less memory and other computational resources.

Table 4 presents the number of model parameters, trainable parameters, and training time for each model. The '*All parameters*' represents the number of parameters for each pre-trained model, while the number of 'trainable parameters' refers to the number of parameters that are trainable during the training process. As shown in Whisper models, the number of trainable parameters was reduced after optimization and parameter reduction, while Wav2Vec2 and HuBERT remained unchanged. Additionally, the time consumed to train Wav2Vec2 and HuBERT models are closely similar, because they were trained on the same settings and have the same number of parameters. The training time of Whisper models is higher than Wav2Vec2 and HuBERT, despite their smaller sizes. However, the training time increased exponentially with the growth of model parameters.

Hyper-parameters	**XLS-R-Wav2Vec2**	**HuBERT**	**Whisper**
learning-rate	3e-4	3e-4	1e-3
warmup_steps	500	500	50
num_train_ epochs	20	20	10
batch_size	16	16	6
gradient_ accumulation_ steps	2	2	1

Table 3: Hyperparameter settings for fine-tuning Wav2Vec2.0, HuBERT, and Whisper

Model	**All Parameters**	**Trainable Parameters**	**Training Time**
XLS-R-Wav2Vec2	315 M	315 M	49 min
HuBERT-large	315 M	315 M	43 min
Whisper-tiny	39 M	589 K	57 min
Whisper-base	75 M	1.1 M	1 hr.
Whisper-small	244 M	3.5 M	1 hr. and 56 min
Whisper-medium	769 M	9.4 M	3 hrs. and 24 min
Whisper-large-v2	1.5 B	15 M	4 hrs. and 28 min

Table 4: Number of model parameters, training parameters, and the training time consumed for each mode

4.4 Model Evaluation

The evaluation measures of each model are word error rate (WER) and character error rate (CER) which are commonly used to evaluate the performance of ASR systems. Both are used to measure the rate of errors in transcribing the recognized speech compared to the reference (ground truth) transcription. WER measures the rate of errors in recognized speech at the word level, while CER measures errors at the character level. The

[2] https://huggingface.co/docs/transformers/main/trainer

WER and CER are calculated as the following equations (S. Young et al., 1995):

$$WER = \frac{S + I + D}{N} \#(1)$$

$$CER = \frac{S + I + D}{N} \#(2)$$

N refers to the number of labels, whereas the S, I, and D, are referring to the number of substitutions, insertions, and deletions of the recognized words or characters. A lower rate of WER or CER indicates better accuracy in the ASR system's transcription output.

5. Results and Analysis

Table 5 presents the achieved results among all models. XLS-R-Wav2Vec2 and HuBERT-large achieved similar performance since they followed the same architecture and model size, both gained a WER of 19%, while XLS-R-Wav2Vec2 achieved a lower CER at 6.5%. In contrast, Whisper models demonstrated superior performance with WER ranging from 5.5% to 3.5% and CER ranging from 4% to 2.6%. Among the Whisper models, Whisper-medium has the lowest WER and CER, while Whisper-tiny has the highest. There is a noticeable improvement in performance as the model size increases within the Whisper models, with Whisper-medium and Whisper-large-v2 achieving the lowest WER and CER among all models.

Overall, Whisper models consistently outperform XLS-R-Wav2Vec2 and HuBERT-large in terms of both WER and CER, with Whisper-medium demonstrating the best performance among all models. These results highlight Whisper's efficiency in recognizing the Jibbali language, even with a limited amount of training data.

Figure 5 shows the confusion matrix of the recognized and misrecognized characters obtained from the Whisper-medium model. The confusion matrix demonstrates the effectiveness of the mode in character recognition with a higher number of correctly recognized characters compared to the misrecognized characters.

Model	**WER (%)**	**CER (%)**
XLS-R-Wav2Vec2	19.2%	7.5%
HuBERT-large	19.4%	11.8%
Whisper-tiny	5.5%	4.0%
Whisper-base	4.3%	3.2%
Whisper-small	3.8%	3.0%
Whisper-medium	**3.5%**	**2.62%**
Whisper-large-v2	**3.5%**	**2.65%**

Table 5: ASR Model evaluation results based on error rate (WER %) and character error rate (CER %)

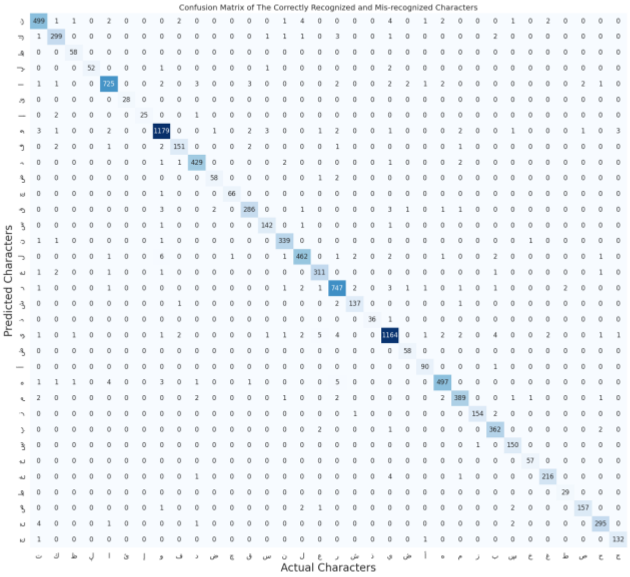

Figure 5 Confusion matrix of the recognized and misrecognized characters obtained from Whisper-medium mode

Table 6 represents examples of the transcribed text predicted from XLS-R-Wav2Vec2, HuBERT-large, and Whisper-medium models with the ground truth. These examples show how these models can identify Shehri (Jibbali) sounds that are not presented in the Arabic language.

The XLS-R-Wav2Vec and Whisper models were trained on a large amount of cross-lingual data, including Arabic. However, Arabic and Shehri (Jibbali) languages contain several similar sounds, as discussed in Section 2, which enabled these models to achieve high performance results. In contrast, the Shehri (Jibbali) language includes some unique sounds not presented in the Arabic language. Despite these unique sounds and the limited size of our dataset, the results obtained were high and accurate. This demonstrates the efficiency of the transfer learning approach for such resource-constrained tasks and the high performance of the pre-trained models applied in this study.

Ground Truth	XLS-R-Wav2Vec2	HuBERT-large	Whisper-medium
ضهر عمقوهم غيج بيصير	ضهر عمقوهم غيج بيص بيصير	ضهر عمقوهم غيج بيص بيصير	ضهر عمقوهم غيج بيصير
هير تيروفك بخيضول أتيبنأ سيارهي ظير اورم	هير تيروفك بخيضولأتيبنأ سيارهي ظير اورم	هير تيروفك بخيضول أتيبنأ سيارهي ظير اورم	هير تيروفك بخيضول أتيبنأ سيارهي ظير اورم
بير اغد كحصف هير خصور بعود اوزحم لو	بير اغد كحصفهير خصور بعود اوزحم لو	بير اغد كحصف هير خصور بعود اوزحم لو	بير اغد كحصف هير خصور بعود اوزحم لو

Table 6 : Examples of predicted transcriptions obtained from XLS-R-Wav2Vec2, HuBERT-large, and Whisper-medium models with the ground truth.

To analyze speech recognition errors resulting from various models from a linguistic perspective, we selected several examples of misrecognized transcription from different models to discover the reasons behind these failures. Table 7 represents different reference examples with their predicted transcription. In the first example, the model misrecognized and deleted the long vowel (/w/ و) in the word "خيضول" because the speaker pronounced it rapidly which reduces the pronunciation rate of the vowel sound to the extent of hiding it, leading to the appearance of the word "خيضول" without the vowel sound, as follows: "خيضل". In the second example, the sounds (/ħ/ ح) and (/f/ ف) are both voiceless consonants, which are produced without the vibration of the vocal cords. This characteristic increases in the pronunciation of the sound (/ħ/ ح), so that its pronunciation is close to the sound (/h/ ه). In this case, the model predicted the sound (/ħ/ ح) as (/f/ ف), making the word "قحل" instead of "قفل," where the speaker in the example pronounced more like "قهل". In the last three examples, the speakers were very fast in their pronunciations which made the models mis-recognized some sounds. This leads us to one of the challenges in constructing a speech dataset, which is ensuring that speakers pronounce sentences at a balanced pace, as the speaking rate affects training results, especially if the language is new to the trained model.

Ground Truth	Predicted Transcription
هير تيروفك بخيضول أتيبنأ سيارهي ظير اورم	هير تيروفك بخيضل أتيبنأ سيارهي ظير اورم
هير اصفريت ميلوت قفل اميه	هير اصفريت ميلوت قحل اميه
اغيج ذان يكين ار ديجحود تو	ار عكذا تيجحود ار
اغيج ذان يكين ار ديجحود تو	عكوز بير كلش او كحصير
هير حيضوت فعمك اوتركت ليس لو	ضهر عمقوهم غيج بيصير

Table 7: *Examples of misrecognized transcription resulted from different trained models*

6. Conclusion

This study presented a promising approach to developing an ASR system for the under-resourced Shehri (Jibbali) language using transfer learning techniques. By fine-tuning various speech pre-trained models like Wav2Vec2.0, HuBERT, and Whisper on the collected Shehri (Jibbali) speech dataset, the research demonstrated the capability of transfer learning methods to address the limitations in data availability that are typically faced for under-resourced languages. The evaluation results showed that the Whisper models significantly outperformed the other models that were evaluated, achieving word error rates as low as 3.5%. This highlights the efficiency of Whisper models in adapting to low-resource tasks even with limited training data.

While the results obtained were encouraging, there is still room for improvement. The performance of the models could be enhanced further by collecting a larger and more diverse Shehri (Jibbali) speech dataset containing a greater variety of speakers, accents, acoustic environments, and content. This would allow the models to learn from more varied data and generalize better. Additionally, future work could explore utilizing multilingual models that have been trained on languages that are closely related to Shehri (Jibbali) both linguistically and geographically. Such models may learn representations that transfer even better.

Overall, this research achieved promising outcomes and demonstrated that transfer learning is an effective solution for overcoming the computational challenges presented by under-resourced languages due to a lack of annotated data resources. With continued efforts to develop larger datasets and optimize model architectures, even more advanced ASR capabilities can be developed to support the documentation, preservation and technological empowerment of under-represented languages like Shehri (Jibbali). The approach presented in this study paves the way for applying similar techniques to other low-resource languages.

7. References

— (2003). Lisān Ẓufār al-ḥimyarī al-muʿāṣir. Dirāsah muʿjamiyyah muqāranah, Jāmiʿat al-Sulṭān Qābūs, Markaz al-dirāsāt al-ʿumāniyyah.

“Ethnologue: Languages of Africa and Europe, Twenty-Fifth Edition,” SIL International, 2022. Accessed: Feb. 18, 2024. [Online]. Available: https://www.sil.org/resources/publications/entry/93719

A. Babu *et al.*, “XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale,”, 2021 arXiv.org. Accessed: Aug. 08, 2023. [Online]. Available: https://arxiv.org/abs/2111.09296v3

A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 12449–12460. Accessed: Apr. 06, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision.” arXiv, Dec. 06, 2022. Accessed: Nov. 05, 2023. [Online]. Available: http://arxiv.org/abs/2212.04356

Āl Ḥāfiẓ, ʿAlī Muḥsin. (1987). Min lahajāt “mahrah” wa-ʾādābihā. Majallat an-nahḍah al-ʿumāniyyah. Muscat.

Al-Kathiri, A, Al-Mashani, A & Alshahri, O. (2024). Al-Luban Wal-Turath Al-Thaqafi. Ministry of Culture, Sports, and Youth, Literary Forum. Muscat, Sultanate of Oman.

Al-Kathiri, Amer; Al-Maashani, Abdulaziz; Al-Kathiri, Salem. (In press). Al-Wadu' al-lughawi fi Dhofar: Dirasat lughawiyat ijtimaiyah. Arab Journal of Humanities Sciences.

Al-Maashani, Saeed. (2017). Al-Ishtiqagh fi al-Lughah al-Jabaliiyah: Mawazanah bil Lughah al-Arabiyyah. Research and Cognitive Communication Center. Riyadh.

B. Neyshabur, H. Sedghi, and C. Zhang, “What is being transferred in transfer learning?,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 512–523. Accessed: Feb. 18, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html

C.-C. Chen, W. Chen, R. Zevallos, and J. E. Ortega, “Evaluating Self-Supervised Speech Representations for Indigenous American Languages.” arXiv, Oct. 08, 2023. Accessed: Jan. 22, 2024. [Online]. Available: http://arxiv.org/abs/2310.03639

E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models.” arXiv, Oct. 16, 2021. Accessed: Nov. 06, 2023. [Online]. Available: http://arxiv.org/abs/2106.09685

G. V. Stan, A. Baart, F. Dittoh, H. Akkermans, and A. Bon, “A Lightweight Downscaled Approach to Automatic Speech Recognition for Small Indigenous Languages,” in *14th ACM Web Science Conference 2022*, Barcelona Spain: ACM, Jun. 2022, pp. 451–458. doi: 10.1145/3501247.3539017.

G. Zellou and M. Lahrouchi, “Linguistic disparities in cross-language automatic speech recognition transfer from Arabic to Tashlhiyt,” *Sci. Rep.*, vol. 14, no. 1, Art. no. 1, Jan. 2024, doi: 10.1038/s41598-023-50516-3.

Johnstone. (1981). Jibbali Lexicon. London: Oxford University Press.

Ministry of Heritage and Culture. (2013). The Omani Encyclopedia (Vol. 6). Muscat: Ministry of Heritage and Culture.

N. Das, S. Bodapati, M. Sunkara, S. Srinivasan, and D. H. Chau, “Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning.” arXiv, Mar. 09, 2021. Accessed: Feb. 18, 2024. [Online]. Available: http://arxiv.org/abs/2103.05834

National Center for Statistics and Information. (2020). Census 2020. Muscat: National Center for Statistics and Information.

P. Arisaputra, A. T. Handoyo, and A. Zahra, “XLS-R Deep Learning Model for Multilingual ASR on Low- Resource Languages: Indonesian, Javanese, and Sundanese.” arXiv, Jan. 12, 2024. Accessed: Jan. 31, 2024. [Online]. Available: http://arxiv.org/abs/2401.06832

R. Coto-Solano *et al.*, “Development of Automatic Speech Recognition for the Documentation of Cook Islands Māori”, 2022.

Rubin, A. D. (2014). The Jibbali (Shahri) language of Oman: grammar and texts. In The Jibbali (Shaḥri) Language of Oman. Brill.

S. Young *et al.*, *The HTK Book*, vol. 3.4. 1995. Accessed: Nov. 05, 2023. [Online]. Available: https://www.inf.u-szeged.hu/~tothl/speech/htkbook.pdf

T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, “LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale.” arXiv, Nov. 10, 2022. doi: 10.48550/arXiv.2208.07339.

V. Gupta and G. Boulianne, “Speech Transcription Challenges for Resource Constrained Indigenous Language Cree,” in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds., Marseille, France: European Language Resources association, May 2020, pp. 362–367. Accessed: Jan. 22, 2024. [Online]. Available: https://aclanthology.org/2020.sltu-1.51

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,” *IEEEACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

Watson, J. C., & Al-Kathiri, A. A. A. (2022). A phonetically “unnatural” class in Central and Eastern Shehret (Jibbali). Kervan: International Journal of Afro-Asiatic Studies, 26(1), 129-159.

Y. Woldemariam, “Transfer Learning for Less-Resourced Semitic Languages Speech Recognition: the Case of Amharic,” in Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds., Marseille, France: European Language Resources association, May 2020, pp. 61–69. Accessed: Mar. 30, 2024. [Online]. Available: https://aclanthology.org/2020.sltu-1.9

Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, and N. Collier, “On the Effectiveness of Parameter-Efficient Fine-Tuning,” *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 11, Art. no. 11, Jun. 2023, doi: 10.1609/aaai.v37i11.26505.