Beyond Orthography: Automatic Recovery of Short Vowels and Dialectal Sounds in Arabic

Anonymous ACL submission

Abstract

This paper presents a novel Dialectal Sound and Vowelization Recovery framework, designed to recognize borrowed and dialectal sounds within phonologically diverse and dialect-rich languages, that extends beyond its standard orthographic sound sets. The proposed framework utilized quantized sequence of input with(out) continuous pretrained self-supervised representation. We show the efficacy of the pipeline using limited data for Arabic, a dialect-rich language containing more than 22 major dialects. Phonetically correct transcribed speech resources for dialectal Arabic is scare. Therefore, we introduce ArabVoice15, a first of its kind, curated test set featuring 5 hours of dialectal speech across 15 Arab countries, with phonetically accurate transcriptions, including borrowed and dialect-specific sounds. We described in detail the annotation guideline along with the analysis of the dialectal confusion pairs. Our extensive evaluation includes both subjective – human perception tests and objective measures. Our empirical results, reported with three test sets, show that with only one and half hours of training data, our model improve character error rate by $\approx 7\%$ in ArabVoice15 compared to the baseline.

1 Introduction

Self-supervised learning (SSL) paradigm has transformed speech research and technology, achieving remarkable performance (Baevski et al., 2020; Chen et al., 2022) while reducing the dependency on extensively annotated datasets (Radford et al., 2023). The SSL models excel at discerning the underlying acoustic properties in both frames and utterance level (Pasad et al., 2021, 2023; Chowdhury et al., 2023) irrespective of language. Phonetic information is sailent and preserved even when these continuous representations are mapped to a finite set of codes via vector quantization (Hsu et al., 2021a; Sicherman and Adi, 2023; Wells et al., 2022; Kheir et al., 2024). This allows the learning paradigm to leverage unlabeled data to discover units that capture meaningful phonetic contrasts.

Leveraging insights from acoustic unit discovery (Park and Glass, 2008; Versteegh et al., 2015; Dunbar et al., 2017; Eloff et al., 2019; Van Niekerk et al., 2020), unsupervised speech recognition (Baevski et al., 2021a; Da-Rong Liu and shan Lee, 2018; Chen et al., 2019; Da-rong Liu and yi Lee, 2022; Baevski et al., 2021b), and phoneme segmentation (Kreuk et al., 2020; Bhati et al., 2022; Dunbar et al., 2017; Versteegh et al., 2015) have utilized quantized discrete units for various purposes. These include *(i)* pretraining the SSL model (Baevski et al., 2020; Hsu et al., 2021a), *(ii)* employing acoustic unit discovery as a training objective (van Niekerk et al., 2020), and *(iii)* utilizing discrete labels for training phoneme recognition and automatic speech recognition (Chang et al., 2023; Da-rong Liu and yi Lee, 2022; Da-Rong Liu and shan Lee, 2018).

Inspired by previous research, we employ SSL representations and vector quantization to recognize acoustic units in phonologically diverse spoken dialects, extending beyond their standard orthographic sound sets. We introduce a simple yet potent network leveraging SSL and a discrete codebook to recognize these non-orthographic dialectal and borrowed sounds with minimal labeled data.

Arabic is an appropriate language choice for the task. The language has a rich tapestry of dialects, each with its unique characteristics in phonology, morphology, syntax, and lexicon (Ali et al., 2021). These dialects[1] differ not only among themselves but also when compared to Modern Standard Arabic (MSA). While MSA prevails in official and educational domains, Dialectal Arabic (DA) serves as the means for daily communication. The diver-

[1] There are 22 Arab countries, and typically, there is more than one dialect spoken in each Arab country (ex: rural versus urban areas)

sity in pronunciation and phoneme sets for DA goes beyond standardized MSA sound sets. Moreover, to add to the challenges, DA follows no standard orthography. Therefore, despite the abundance of DA speech data in online platforms, accurately (phonetically correct) transcribed resources are scarce, categorizing DA among the low-resource languages.

To bridge this gap, we introduce the Arabic "*Dialectal Sound and Vowelization Recovery*" (DSVR) framework. The proposed framework exploits the frame-level SSL embeddings and quantizes them to create a handful of discrete labels using k-means model. These discrete labels are then fed (can be in combination with SSL embeddings) as input to a transformer-based dialectal unit and vowel recognition (DVR) model.

We show its efficacy for *(a)* dialectal and borrowed sound recovery; and *(b)* vowelization restoration capabilities with only 1 hour 30 minutes of training data. We introduced Arabic dialectal test set – "**ArabVoice15**", a collection of 5 hours of dialectal speech and verbatim transcription with recovered dialectal and borrowed sounds from 15 Arab countries. For vowelization restoration, we tested on 1 hour of speech data, sampled from CommonVoice-Ar (Ardila et al., 2019), transcribed by restoring short vowels. Our paper describes the phonetic rules adopted, special sounds considered along with detailed annotation guidelines for designing these test sets. Furthermore, we evaluate the quality of the intermediate discrete labels using human perceptual evaluation, in addition to other purity and clustering-based measures.

We observed that these discrete labels can capture speaker-invariant, distinct acoustic, and linguistic information while preserving the temporal information. Consequently, encapsulating the discriminate acoustic unit properties, which can be used to recover dialectal missing sounds. Our empirical results suggest that DSVR can exploit unlabeled data to design the codebook and then with a small amount of annotated data, a unit recognizer can be trained.

Our contribution involves: *(i)* Proposed Arabic Dialectal Sound and Vowelization Recovery (DSVR) framework to recognize dialectal units and restore short vowels; *(ii)* Developed annotation guidelines for the verbatim dialectal transcription; *(iii)* Introduced and benchmark ArabVoice15[2] test set – a collection of dialectal speech and phonetically correct verbatim transcription of 5 hours of data. *(iv)* Released a small subset of CommomVoice - Arabic (Ardila et al., 2019) data with restored short vowels, dialectal and borrowed sounds.

This study addresses the crucial challenge of identifying and understanding these phonetic intricacies, acknowledging their essential role in improving the performance of speech processing applications like dialectal Text-to-Speech (TTS) and Computer-Assisted Pronunciation Training applications. To the best of our knowledge, this study is the first to attempt to automatically restore vowels, borrowed and dialectal sounds for rich spoken dialectal Arabic language with very limited amount of data. Moreover, the study also introduce the very first dialectal testset with phonetically correct transcription representation.

[2]Will be made publicly available upon acceptance.

2 Arabic Sounds

The exploration of phonotactic variations across Arabic dialects, including MSA and other regional dialects offers a rich field of study within the domain of Arabic linguistics. These variations are not merely lexical, but phonetic and in many cases deeply embedded in the phonological rules that dictate the permissible combinations and sequences of sounds within each dialect (Biadsy et al., 2009).

2.1 Related Studies

Limited research investigated dialectal sounds in Arabic transcribed speech. (Vergyri and Kirchhoff, 2004) deployed an EM algorithm to automatically optimize the optimal diacritic using acoustic and morphological information combination. (Al Hanai and Glass, 2014) employed automated text-based diacritic restoration models to add diacritics to speech transcriptions and to train speech recognition systems with diacritics. However, the effectiveness of text-based diacritic restoration models for speech applications is questionable for several reasons, as demonstrated in (Aldarmaki and Ghannam, 2023), they often fail to accurately capture the diacritics uttered by speakers due to the nature of speech; hesitation, unconventional grammar, and dialectal variations. This leads to a deviation from rule-based diacritics. Recently, (Shatnawi et al., 2023) developed a joint text-speech model to incorporate the corresponding speech signal into the text based diacritization model.

Grapheme to Phoneme (G2P) has been studied thoroughly by many researchers across multiple languages. Recent approaches in G2P include data-driven and multilingual (Yu et al., 2020; Garg et al., 2024) mapping from grapheme sequence to phoneme sequence. However, previous work in Arabic G2P is comprised of two steps: (*i*) Grapheme to vowelized-grapheme (G2V) to restore the missing short vowels and (*ii*) Vowelized-grapheme to phoneme sequence (V2P). The first step is often statistical and deploys techniques like sequence-to-sequence; (Abdelali et al., 2016; Obeid et al., 2020) are used widely for restoring the missing vowels in Arabic. The second step is relatively one-to-one and can be potentially hand-crafted rules for MSA as well as various dialects, refer to (Biadsy et al., 2009; Ali et al., 2014) for more details. MSA Arabic speech recognition phoneme lexicon can be found here[3]

The distinction between MSA and regional dialects is nuanced; viewing them as separate is over-simplified. Arabs perceive them as interconnected, leading to diglossia, where MSA is for formal contexts and dialects for informal ones, yet with significant overlap and blending (Ali et al., 2016b). (Chowdhury et al., 2020) studied dialectal code-switching in the Egyptian corpus in the Arabic Dialect Idenitifcation (ADI) Challenge in the MGB-3 challenge (Ali et al., 2017), which has been manually labeled per utterance. In this study, the researchers annotated the corpus per token, considering both the linguistic and the acoustic cues. They showed that what has been labeled as Egyptian sentences, when studied per tokens; the corpus showed roughly $2.6K$ Egyptian words verus $9.3K$ MSA. Here is a brief overview of Arabic phonology and its dialectal sounds.

2.2 MSA and Dialectal Phonlological Variations

Arabic dialects exhibit phonological differences when compared to MSA, these differences might be noted across various aspects of pronunciation and phonology, such as consonants, vowels, and diphthongs. It's suggested that Arabic generally encompasses around 28 consonants, alongside three short vowels, three long vowels, though these numbers could vary slightly depending on the dialect in question. The consonant pronunciation of ث [θ], ذ [ð], ظ [ðˤ], ج [dʑ], ض [dˤ], and ق [q] cover most of the variations across Arabic dialects. Here are some examples of phones that vary between MSA and various Arabic dialects.

- Interdental Consonants: In particular ث [θ]/, ذ [ð] found in MSA are pronounced differently. For example, in Egyptian Arabic, they are often pronounced as س [s].
- The voiceless stop constant ق [q] is a good example across Arabic dialects, In many cases, it will be pronounced as glottal stop ء [ʔ] in Egyptian dialect and voiced velar ج [dʑ] in Gulf and Yemeni dialects.
- Long and short vowels might exhibit a reduction in duration or even drop in duration in various dialects. In some dialects, the difference between long and short vowels may be subtle to notice.
- The difference in stress between Arabic dialects can lead to different meanings.

The phonological differences and examples mentioned above do not cover all variations but highlight several distinctions between Arabic dialects and MSA. A depiction of certain MSA sound variations is presented in Appendix A.1.

3 Methodology

Figure 1 gives an overview of our proposed *Dialectal Sounds and Vowelization Restoration Framework*. The goal of the pipeline is to recover (verbatim) dialectal sound and short vowel units, using frame-level representation. Given an input speech signal $X = [x_1, x_2, \cdots, x_T]$ of T frames, the frame-level representation (Z) is first extracted from a *multilingual SSL pretrained* model.

We subsampled frame-level vectors ($\widetilde{Z} \subset Z$) to train a simple *Vector Quantization* (VQ) model using k-means for getting a Codebook $\mathbb{C}_k$, with k categorical variables. Each cluster, in the codebook, is then associated with a code Q_i^k and a centroid vector G_i^k. Using the $\mathbb{C}_k$ codebook, we infer the discrete sequences codes $\hat{Z}$ corresponding to the input Z. $\hat{Z}$ is the input of our *Dialectal Units and Vowel Recognition* (DVR) module.

3.1 Pretrained Speech Encoder

The XLS-R[4] model is a multilingual pre-trained SSL model following the same architecture as wav2vec2.0 (Baevski et al., 2020). It includes a CNN-based encoder network to encode the raw

[3] https://catalog.ldc.upenn.edu/LDC2017L01

[4] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

audio sample and a transformer-based context network to build context representations over the entire latent speech representation. The encoder network consists of 7 blocks of temporal convolution layers with 512 channels, and the convolutions in each block have strides and kernel sizes that compress about 25ms of 16kHz audio every 20ms. The context network consists of 24 blocks with model dimension 1024, inner dimension 4096, and 16 attention heads.

The XLS-R model has been pre-trained on around $436,000$ hours of speech across 128 languages. This diverse dataset includes parliamentary speech ($372,000$ hours in 23 European languages), read speech from Multilingual Librispeech ($44,000$ hours in 8 European languages), Common Voice ($7,000$ hours in 60 languages), YouTube speech from the VoxLingua107 corpus ($6,600$ hours in 107 languages), and conversational telephone speech from the BABEL corpus ($\approx 1,000$ hours in 17 African and Asian languages).

We opt for the smallest XLR-S ($317M$ parameters) to minimize computational requirement. Our preliminary analysis revealed limitation in the XLR-S in differentiating between acoustic sounds, such as د [d]/ ض [dˤ] and ت [t]/ ط [tˤ] present in MSA and DA. Consequently, we primed the model towards Arabic sounds by finetuning with 13 hours clean avaliable MSA data (Ardila et al., 2019) for ASR task. We restricted the training to 5 epoch to prevent the risk of catastrophic forgetting of the pretrained representation (Goodfellow et al., 2013).

3.2 Vector Quantization

Vector Quantization (Makhoul et al., 1985; Baevski et al., 2020) is a widely used technique for approximating vectors or frame-level embeddings through a fixed codebook size. In our Vector Quantization (VQ) modules (see Figure 1), we pass forward a sequence of continuous feature vectors $Z = \{z_1, z_2, \ldots, z_T\}$ and then assign each z_t to its nearest neighbor in the trained codebook, $\mathbb{C}_k$. In other words, each z_t is replaced with the code $Q_i^k \in \mathbb{C}_k$ assigned to the centroid G_i^k . The resultant discrete labels are quantized sequence $\hat{Z} = \{\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_T\}$. These labels are expected to facilitate better proninciation learning and incorporate distinctive phonetic information in the subsequent layers.

Training the Codebook For quantization, we utilized the k-means clustering model. We selected a random subset of frame-level representation for training the cluster model. Moreover, to select wide varieties of sound unit, we forced-aligned the available/automatic transcription of the datasets (see Section 5.1) with a GMM-HMM based ASR models. Using the timestamps, we then select SSL frame representations that aligned with wide varieties of sound labels.[5] We trained the codebook for different $k = \{128, 256, 512\}$

3.3 Dialectal Units and Vowel Recognition (DVR) Model

We explored two variants of DVR – discrete and joint Model (as seen in Figure 2). The discrete DVR takes only the discrete $\hat{Z}$ labels from the VQ as input, where as the joint module concatenate both the $\hat{Z}$ and Z inside the subsequent layer. The resultant embeddings (for both model) are then passed to the transformer layers and the head feed-forward layer. The DVR model is optimized with character recognition objective to identify arabic units.

3.4 Baselines

As baselines, we initially opt for two architecture. For the first, we have extracted the frozen frame-level representation from the XLS-R model and then passed it to a output head. The second, we used the frozen frame-level representation to pass to the feedforward layer followed by the transformers and output head. The second architecture use similar encoder as the DVR model (see Figure 2 Baseline). For brevity, we reported with the results of the second architecture (SSL frame-level representation with transformer-based encoder) as the baseline of the paper.

4 ArabVoice15 Dataset

Spoken DA remains a low-resource language primarily due to the scarcity of transcription that can faithfully capture the diverse regional and borrowed sounds in the standard written format. Such lack of data posses significant challenge for speech and linguistic research and evaluation. In this study, we address this challenge by designing and developing ArabVoice15 test set. Furthermore, we have also enhanced a subset of the existing Arabic Commonvoice (Ardila et al., 2019), Ar:CV$_R$ dataset with restored vowels, borrowed and dialectal sounds. In the following sections, we will discuss the datasets,

[5] $10k$ sample frames for each sound label.

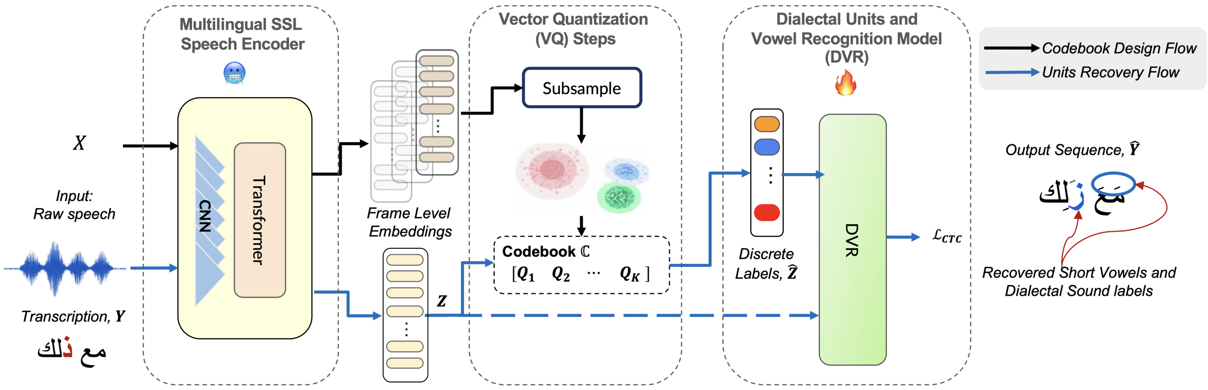

Figure 1: Proposed Arabic Dialectal Sound and Vowelization Recovery (DSVR) Framework

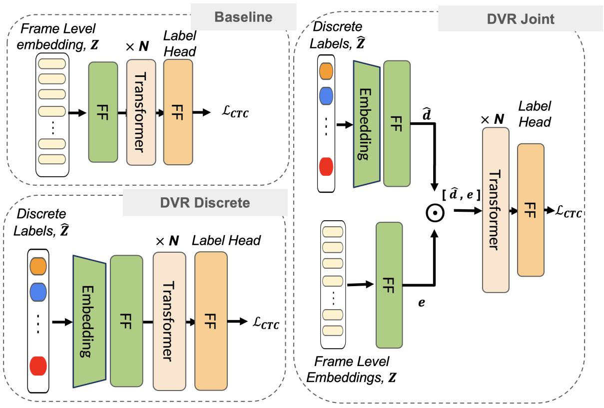

Figure 2: Baseline and DVR – Discrete and Joint Model

preprocessing steps along with in detail annotation guidelines.

ArabVoice15 is a collection of 5 hours of speech utterances randomly selected from testset of ADI17 (Ali et al., 2019) dataset, widely used for dialect identification task. For the ArabVoice15, we selected a total of 2500 utterance, $\approx 146(\pm 3.6)$ utterance from each of the 15 Arab countries including: Algeria (ALG), Egypt (EGY), Iraq (IRA), Jordan (JOR), Saudi Arabia (KSA), Kuwait (KUW), Lebanon (LEB), Libya (LIB), Morocco (MOR), Palestine (PAL), Qatar (QAT), Sudan (SUD), Syria (SYR), United Arab Emirates (UAE), and Yemen (YEM). The average utterance duration: 7-8 seconds. As for $Ar:CV_R$, we randomly extracted 21.38 hours from the Ar:CV trainset, which we then manually annotated at both verbatim and vowelized level (test $\approx$ 1hr).

Data Verbatim Pre-Processing We present a set of rules employed for data normalization, aiming to reduce annotators' tasks through a rule-based phonemic letter-to-sound approach in Arabic, as detailed in (Al-Ghamdi et al., 2004). For vowelization, we initially applied diacritization (aka vowelization or vowel restoration) module present in the Farasa tool (Abdelali et al., 2016). We then applied the following rule-based phonemic letter-to-sound function to our dataset. This step also removed any Arabic letters that are not traditionally pronounced in spoken conversation.

- For ا [a:] : (i) If it appears within a word (not at the beginning) and is followed by two consonants, we delete it. For example, كتب الكتاب [ktb a:lktb] becomes كتب لكتاب [ktb lktb]. (ii) If it occurs at the beginning in the form of the definite article ال, we replace it with [ʔa]. For example, العلم [a:lmʕlm/] becomes ءلعلم [ʔalmʕlm].
- For ل [l] : We removed the Shamsi (Sun) [l], that refers to [l] in ال followed by a Sun consonant (نلكططصصزرضذدثت). For example: الرحمان [a:lrħman] becomes ارحمان [a:rħman]
- For آ, we replaced it wherever it occurred in the text with ءا [ʔa:].
- For Hamza shapes (ئ إ ؤ أ ء), we normalized them to ء [ʔ].
- For ى ا, we normalized them to ا [a:/].
- For Tanwin diacritics (اٌ اٍ اً [/un/, /in/, /an/]) at the end of a phrase, we replaced it with a short vowel, and elsewhere, we turned it into اَن, اُن, اِن [/un/, /in/, /an/] to match the typical verbatim sounds.

Annotation Guideline We gave extensive training to an expert transcriber, a native speaker from

Dataset	Source of Data	Train (#hrs)	Test (#hrs)	Annotated with
Ar:CV$_R$$^+$	Subset from Arabic Common Voice (Ardila et al., 2019) Train split	1 hr (*total 19 hrs)	1 hr	Restored short vowels, dialectal and borrowed sounds
AR:TTS-data	Subset collected from available test-to-speech speech corpus (2 speakers, one from Egypt and Levantine region) (Abdelali et al., 2022, 2024; Dalvi et al., 2024)	30 mins	–	–
EgyAlj	in-house, source Aljazeera Arabic channel, containing MSA and Egy content	–	1.8 hrs	Semi-supervised transcription, manually restored short vowels, dialectal and borrowed sounds.
ArabVoice15$^+$	A small subset for ADI17 (Ali et al., 2019) test set	–	5 hrs	Transcribed with dialectal and borrowed sound in consideration

Table 1: Train and Test dataset used for Dialectal Units and Vowel Recognition (DVR) model. * present total hours of data available and used to show the effect of training data size. $^+$ test data will be made available to the public.

Egypt, to provide the written form for each word and its verbatim transcription. For example, if the word is قَلَم [qalam] (pen), and the speaker said كَلَم [kalam], then the transcriber writes [qalam/kalam]. This is the summary of the annotation guidelines:

- For sounds that are not in MSA and have been borrowed from foreign languages, the following special letters[6] are used:
  - چ [g] as in the word جوجل "google" which is written as جوجل [ju:jl] / چوچل [gu:gl].
  - ڤ [v] as in the word ڤيديو "video" which is written as فيديو [fi:dyu:] / ڤيديو [vi:dyu:].
  - پ [p] as in the word إسپراي "spray" which is written as سبراي [sbra:y] / سپراي [spra:y].
- For dialectal sounds that are missed in MSA, the following special letters are used:
  - گ (Gulf /Qaf/) as in the word عگال which is written as عقال / عگال.
  - The Egyptian/Syrian/Lebanese ق [q] is pronounced mostly as ء [ʔ] as in قال [qa:l] / ءال [ʔa:l].
  - ڟ (Egyptian/Lebanese /Z/) as in the word بيظهر is written as بيظهر / بيڟهر.

There are few words with special spellings that do not precisely reflect their pronunciation. In these cases, the transcriber writes both, as in the word هذا [hadha] / هاذا (/ha:dha/). Numbers and some special symbols (ex: the percentage sign %) are written in letters and are being judged according to speakers' pronunciation.

[6]The special letters used in the annotation process do not belong to the Arabic alphabet; instead, we borrowed them from Farsi sharing similar Arabic shapes, these letters were employed to represent distinct dialectal sounds.

Quality Control: Detection of possible annotation errors was done automatically and doubtful cases were returned to the transcriber for review. In addition, a manual inspection of random sentences (10%) from each file was performed. Any file below 90% accuracy was returned for full correction.

5 Experimental Design

5.1 Training Datasets and Resources

Datasets: Unsupervised Codebook Generation To train the codebook, we randomly selected utterances from publicly available resources. For Arabic sounds, we opt for utterances from official CommonVoice train set along with Arabic TTS data. Moreover, to add borrowed/special sounds missing in MSA phonetic set (e.g., /g, v, p/), we included publicly available English datasets like LibriSpeech (Panayotov et al., 2015), and TIMIT (Garofolo et al., 1993). For the subsampling process, we opt for hybrid ASR systems[7] for Arabic and Montreal Forced-Aligner[8] for the English.

Datasets: Spervised DVR Model To train the DVR model, we opt for a small training dataset to showcase our the efficacy of our proposed framework in low-resource setting. The details of dataset used for DVR is presented in Table 1. For the training, we utilize dataset transcribed with restored vowels, borrowed and dialectal sounds. We used 1 hour 30 minutes of training data in this study.

5.2 Model Training

The Models, presented in Figure 2, are optimized using Adam optimizer for 50 epochs with an early stopping criterion. The initial learning rate is

[7]Trained on Arabic CommonVoice

[8]https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner.git

1×10^{-4}, and a batch size of 16 is employed. The loss criterion is CTC loss, utilized for predicting verbatim sequences. The input dimension for the SSL frame-level representation is $d = 1024$, the dimension of the discrete labels $d = k$. For all the architectures in Figure 2, the dimension of feedforward (FF) layer is $d = 512$. For the DVR joint, the output from the FFs $(\hat{d}, e)$ are concatenated to form $[\hat{d}, e]$ of dimension $d = 1024$. These outputs are then passed to 2 transformer encoders each with 8 attention heads. Following, the encoded information is then projected to output head of dimension $V = 39$ equivalent to the characters supported by the models. The total number of trainable parameters are Baseline:7.634M; DVR discrete:7.110M; and joint: 33.346M.

5.3 Evaluation Measures

We used **Davis-Bouldin index** (DBindex) to select the k value for our codebook. The DBindex is widely used in clustering performance evaluation (Davies and Bouldin, 1979), and is characterized by the ratio of within-cluster scatter to between-cluster separation. A lower DBindex value is better, signifying compact clustering. Following, we adapted the approach of (Hsu et al., 2021b) to evaluate the codebook quality using **Phone Purity**, **Cluster Purity**, and **Phone-Normalized Mutual Information** (PNMI). These measures use frame-level alignment of characters with discrete codes assigned to each frame. Phone purity measures the average frame-level phone accuracy, when we mapped the codes to its most likely phone (character) label. Cluster purity, indicates the conditional probability of a discrete code given the character label. PNMI measures the percentage of uncertainty about a character label eliminated after observing the code assigned. A higher PNMI indicates better quality of the codebook. Moreover, we assessed the codebook quality by **human perception** tests as mentioned in the following section. As for evaluating the dialectal sounds and short vowel recognition model, we reported Character Error Rate (CER) with and without restoring short vowels.

Human Perception Test Setup We performed cluster quality analysis for $k = \{128, 256, 512\}$ following the steps of (Mao et al., 2018; Li et al., 2018). For our study, we defined each clusters (demoted by a code) as either Clean or Mix. Clusters are considered as Clean when 80% of its instances are matched to one particular character, where as for Mix clusters, the instances are mapped to different characters.[9] We hypothesise that the Mix clusters represent examples which can resembles closely to either two of canonical sound unit $/l1/$ and $/l2/$, or a mix of both $/l1\_l2/$. We randomly selected 52 examples from each perceived Mix Clusters. We asked the four annotators (2 native and 2 non-native Arabic speakers) to categorize it into these four classes: more similar to $/l1/$, more similar to $/l2/$, a mix of both, or neither.

6 Results and Discussion

Number of discrete codes in Codebook We reported the DBindex for the codebook sizes $k = \{128, 256, 512\}$ in Table 2. We observed lower DBindex with $k = 256$ indicating better codebook quality. We further evaluated the codebook quality and reported purity measures with the Ar:CV$_R$ testset only for brevity and CER with all the testsets. Our CER results shows the efficacy of the selected $k = 256$ for most of the test sets. We observed that increasing codebook size improves the purity and the PNMI. We noticed, the gain in cluster stability between $k = 256$ *vs* $k = 516$ is not very large with respect to the performance and computational cost. Hence we selected the codebook $\mathbb{C}$ of size $k = 256$ for all the experiments.

k	*128*	*256*	*512*
$\mathbb{C}$ size k selection criterion			
DBindex (↓)	2.59	**2.57**	2.7
Purity Measures: Ar:CV$_R$ testset			
Phone Purity (↑)	0.600	0.641	0.672
Discrete Code Purity (↓)	0.436	0.289	0.236
PNMI (↑)	0.343	0.418	0.495
CER (↓): Borrowed and Dialectal Unit Recognition			
Ar:CV$_R$	0.149	0.108	0.107
EgyAlj	0.246	0.206	0.218
ArabVoice15	0.465	0.447	0.462
Average	0.287	**0.254**	0.262

Table 2: Quality evaluation of discrete codes based on DBindex, purity measures and CER for 3 test sets.

Perceptual test of Codebook We averaged annotator judgments across four categories for all Mix clusters, revealing no clear majority and highlighting the listeners' difficulty in categorically labeling audio within these clusters. In aligned with Mao et al. (2018); Li et al. (2018), we also conclude that these mixed labels genuinely exist and cannot be

[9] Only characters above 20% frequency are considered.

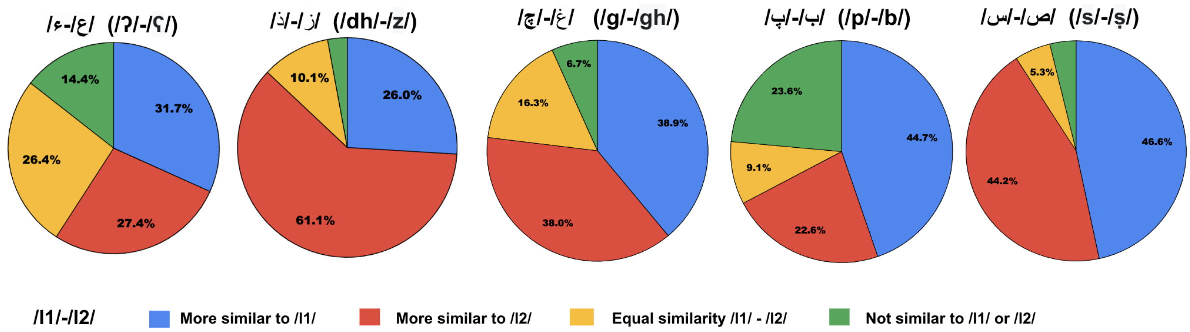

Figure 3: The statistical results of perceptual tests of different sounds using cluster with $k = 256$

CER	Z	D_D	D_J	Z	D_D	D_J
Training:	1hr 30min			5hr 30min		
Ar:CV$_R$	0.113	0.108	**0.094**	**0.095**	0.110	0.099
EgyAlj	0.252	**0.206**	0.231	0.257	**0.245**	0.248
AraVoice15	0.536	**0.447**	0.464	0.485	**0.477**	0.491
Training:	3hrs 30min			~20 hrs		
Ar:CV$_R$	0.103	0.108	**0.096**	**0.099**	0.108	0.101
EgyAlj	0.270	**0.241**	0.253	0.264	0.244	**0.227**
AraVoice15	0.497	**0.470**	0.483	0.492	0.478	**0.457**

Table 3: Reported CER performance for borrowed and dialectal unit recognition task with Baseline (Z), DVR Discrete (D_D) and DVR Joint (D_J) models, for all three test sets and different training data sizes.

CER	Farasa	Z	D_D	D_J
Ar:CV$_R$	0.279	0.123	0.278	**0.118**
EgyAlj	**0.250**	0.279	0.395	0.274

Table 4: Reported CER for Farasa, Baseline (Z), DVR Discrete (D_D) and DVR Joint (D_J) models for two test sets. Training set of 1 hour 30 minutes.

precisely characterized by any conventional given label. We present some of our findings of the perceptual test in Figure 3 for 5 different Mix clusters with average judgment per category.

Dialectal Unit Recognition Performance We reported the performance of the proposed DVR discrete and joint model in Table 3 for borrowed and dialectal unit recognition task. Our results shows the efficacy of the DVR models over the baseline specially for dialectal test sets (ArabVoice and EgyAlj). We observed for borrowed and dialectal unit recognition, the discrete model outperforms the joint model significantly. Breakdown of the performance for 15 countries are presented in Appendix A.2.

Impact of Training Data size Table 3 also shows the impact of the training data size. We observed for dialectal unit recognition, our DVR discrete model outperforms the other two models significantly with limited data sets of $\{1hr30min, 3hr30min, 5hr30min\}$. We see an improvement in performance from 1hr30min to 3hr30min settings. However, beyond a certain data threshold, the improvements plateaued.

Performance for short vowel restoration For short vowel restoration (in Table 4), we observed that the added frame-level embeddings (in DVR joint) improve the recognition performance. We also observed that the baseline model performs comparably with DVR joint. This indicates that the restoration of short vowels benefits from high dimensional fine-grained information compare to using few discrete codes. We also compared the CER with Farasa – state-of-the-art text-based dicretization tool (Abdelali et al., 2016). We observed the acoustic models outperform Farasa by a large margin, especially for common voice subset. However, Farasa excelled in formal content – news content presented in EgyAlj testset.

7 Conclusion

In this study, we propose a novel dialectal sound and short vowel recovery framework that utilizes a handful of discrete codes to represent the variability in dialectal Arabic. We also observed with only 256 discrete labels, the borrowed and dialectal sound recognition model outperforms both baseline and joint (discrete code with frame-level SSL representation) models by $\approx 7\%$ CER improvement. For restoring vowels, we noticed SSL embeddings play a bigger role. Our findings indicate the efficacy of the discrete model with small training datasets. To foster further research in dialectal Arabic, we introduced, benchmarked, and released ArabVoice15 – a dialectal verbatim transcription dataset containing utterances from 15 Arab countries. In the future, we will apply the framework to more dialects and other dialectal languages.

Limitations

The diversity of representation and the size of ArabVoice15 could limit the conclusion to generalize in all Arabic dialects due to variability in dialectal sounds. Although the annotator was an expert transcriber and received extensive training, their dialect may have led to some bias in judgment.

Ethics Statement

For the research work presented in this paper on the Dialectal Sound and Vowelization Recovery (DSVR) framework, we have adhered to the highest ethical standards. All the speech/audio data used in this study were already publicly available. The human perception tests for our evaluation process were designed with a commitment to fairness, inclusivity, and transparency. The participants were selected keeping in mind balancing gender and nativity. Listeners were fully briefed on the nature of the research and their rights as participants, including the right to withdraw at any time without consequence. However as we mentioned in the limitation section, we cannot guarantee any human bias toward any dialectal sound or preference.

References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Ahmed Abdelali, Nadir Durrani, Cenk Demiroglu, Fahim Dalvi, Hamdy Mubarak, and Kareem Darwish. 2022. Natiq: An end-to-end text-to-speech system for arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, pages 394–398, Abu Dhabi, UAE. Association for Computational Linguistics.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Malta. Association for Computational Linguistics.

Mansour M Al-Ghamdi, Husni Al-Muhtasib, and Moustafa Elshafei. 2004. Phonetic rules in arabic script. *Journal of King Saud University-Computer and Information Sciences*, 16:85–115.

Tuka Al Hanai and James R Glass. 2014. Lexical modeling for arabic asr: a systematic approach. In *INTERSPEECH*, pages 2605–2609.

Hanan Aldarmaki and Ahmad Ghannam. 2023. Diacritic recognition performance in arabic asr. *arXiv preprint arXiv:2302.14022*.

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

Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah. 2021. Connecting arabs: Bridging the gap in dialectal speech recognition. *Communications of the ACM*, 64(4):124–129.

Ahmed Ali, Najim Dehak, Patrick Cardinal, Sameer Khurana, Sree Harsha Yella, James Glass, Peter Bell, and Steve Renals. 2016b. Automatic dialect detection in arabic broadcast speech. In *Interspeech 2016*, pages 2934–2938.

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. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1026–1033. IEEE.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic mgb-3. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 316–322. IEEE.

Ahmed Ali, Yifan Zhang, Patrick Cardinal, Najim Dahak, Stephan Vogel, and James Glass. 2014. A complete kaldi recipe for building arabic speech recognition systems. In *2014 IEEE spoken language technology workshop (SLT)*, pages 525–529. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021a. Unsupervised speech recognition. *Advances in Neural Information Processing Systems*, 34:27826–27839.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2021b. Unsupervised speech recognition. In *NeurIPS*.

Alexei Baevski, Yuhao 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*, 33:12449–12460.

Saurabhchand Bhati, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. 2022. Unsupervised speech segmentation and variable rate representation learning using segmental contrastive predictive coding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Fadi Biadsy, Julia Bell Hirschberg, and Nizar Y Habash. 2009. Spoken arabic dialect identification using phonotactic modeling.

Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, and Shinji Watanabe. 2023. Exploration of efficient end-to-end asr using discretized input from self-supervised learning. *arXiv preprint arXiv:2305.18108*.

Kuan-Yu Chen et al. 2019. Completely unsupervised phoneme recognition by a generative adversarial network harmonized with iteratively refined hidden markov models. In *Interspeech*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020. Effects of dialectal code-switching on speech modules: A study using egyptian arabic broadcast speech.

Shammur Absar Chowdhury, Nadir Durrani, and Ahmed Ali. 2023. What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis. *Computer Speech & Language*, 83:101539.

Hung yi Lee Da-Rong Liu, Kuan-Yu Chen and Lin shan Lee. 2018. Completely unsupervised phoneme recognition by adversarially learning mapping relationships from audio embeddings. In *Interspeech*.

Yi-chen Chen Sung-feng Huang Shun-po Chuang Da-yi Wu Da-rong Liu, Po-chun Hsu and Hung yi Lee. 2022. Learning phone recognition from unpaired audio and phone sequences based on generative adversarial network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating llms benchmarking.

David L Davies and Donald W Bouldin. 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.

Ewan Dunbar et al. 2017. The zero resource speech challenge 2017. In *ASRU*.

Ryan Eloff, André Nortje, Benjamin van Niekerk, Avashna Govender, Leanne Nortje, Arnu Pretorius, Elan Van Biljon, Ewald van der Westhuizen, Lisa van Staden, and Herman Kamper. 2019. Unsupervised acoustic unit discovery for speech synthesis using discrete latent-variable neural networks. *arXiv preprint arXiv:1904.07556*.

Abhinav Garg, Jiyeon Kim, Sushil Khyalia, Chanwoo Kim, and Dhananjaya Gowda. 2024. Data-driven grapheme-to-phoneme representations for a lexicon-free text-to-speech. *arXiv preprint arXiv:2401.10465*.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Nawar Halabi and Mike Wald. 2016. Phonetic inventory for an arabic speech corpus.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021a. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021b. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Yassine El Kheir, Ahmed Ali, and Shammur Absar Chowdhury. 2024. Speech representation analysis based on inter- and intra-model similarities. In *Explainable Machine Learning for Speech and Audio Workshop, ICASSP*.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. In *Interspeech*.

Xu Li, Shaoguang Mao, Xixin Wu, Kun Li, Xunying Liu, and Helen Meng. 2018. Unsupervised discovery of non-native phonetic patterns in l2 english speech for mispronunciation detection and diagnosis. In *INTERSPEECH*, pages 2554–2558.

John Makhoul, Salim Roucos, and Herbert Gish. 1985. Vector quantization in speech coding. *Proceedings of the IEEE*, 73(11):1551–1588.

Shaoguang Mao, Xu Li, Kun Li, Zhiyong Wu, Xunying Liu, and Helen Meng. 2018. Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6244–6248. IEEE.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Alex S. Park and James R. Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Sara Shatnawi, Sawsan Alqahtani, and Hanan Aldarmaki. 2023. Automatic restoration of diacritics for speech data sets. *arXiv preprint arXiv:2311.10771*.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP*.

Benjamin Van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. *arXiv preprint arXiv:2005.09409*.

Benjamin van Niekerk, Leanne Nortje, and Herman Kamper. 2020. Vector-quantized neural networks for acoustic unit discovery in the zerospeech 2020 challenge. In *Interspeech 2020*, pages 4836–4840.

Dimitra Vergyri and Katrin Kirchhoff. 2004. Automatic diacritization of arabic for acoustic modeling in speech recognition. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*, pages 66–73.

Maarten Versteegh et al. 2015. The zero resource speech challenge 2015. In *Interspeech*.

Dan Wells, Hao Tang, and Korin Richmond. 2022. Phonetic analysis of self-supervised representations of english speech. In *Interspeech*.

Mingzhi Yu, Hieu Duy Nguyen, Alex Sokolov, Jack Lepird, Kanthashree Mysore Sathyendra, Samridhi Choudhary, Athanasios Mouchtaris, and Siegfried Kunzmann. 2020. Multilingual grapheme-to-phoneme conversion with byte representation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8234–8238. IEEE.

A Appendix

A.1 Sound Analysis

In Figure 5, we have depicted potential confusion between specific sounds in MSA and Arabic dialects. Utilizing a Hidden Markov Model-Time Delay Neural Network (HMM-TDNN) model[10], trained with MGB-2 (Ali et al., 2016a) for Arabic, we aligned randomly selected samples from the original datasets of CommonVoice Arabic and EgyAlj. For the English dataset TIMIT, we used the provided ground truth alignment.

After aligning speech signals with their original unvowelized character-based transcriptions, we matched frame-level features extracted from XLS-R (see Section 3.1) with their corresponding characters. In Figure 5.A, we randomly selected 1000 samples associated with ز [z] and 1000 samples associated with ذ [ð] from CommonVoice Arabic. Despite CommonVoice Arabic being considered as clean MSA speech data with good pronunciation, we observed that some samples of ذ [ð] were clustered with ز [z], primarily explained by the speakers getting influenced by their dialectal variations, as discussed in Section 2.

Figure 5.B displays the selection of three characters: ت [t], ة [t/h], ه [h]. Notably, ة is at times pronounced as [t] and at other times as [h]. Although rule-based methods (Halabi and Wald, 2016) can predict when it will correspond to which sound, applying these rules in everyday spoken language,

[10] https://kaldi-asr.org/models/m13

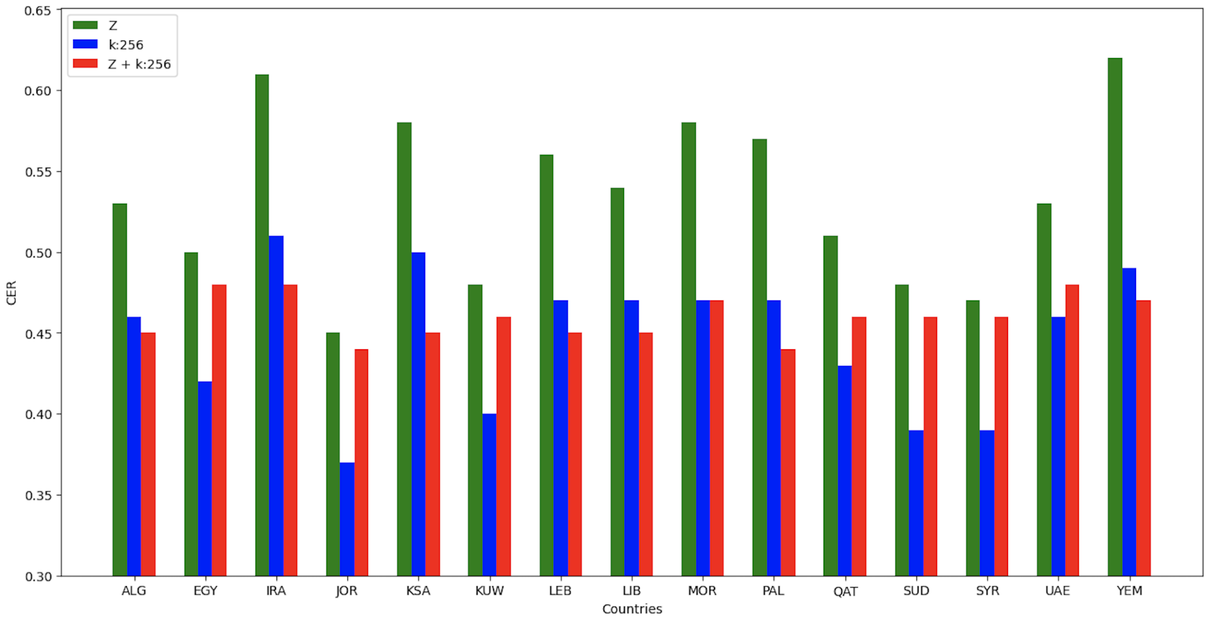

Figure 4: Reported CER for test utterances from 15 Arab countries for three models Baseline (Z), DVR discrete (k:256) and DVR joint (Z+k:256)

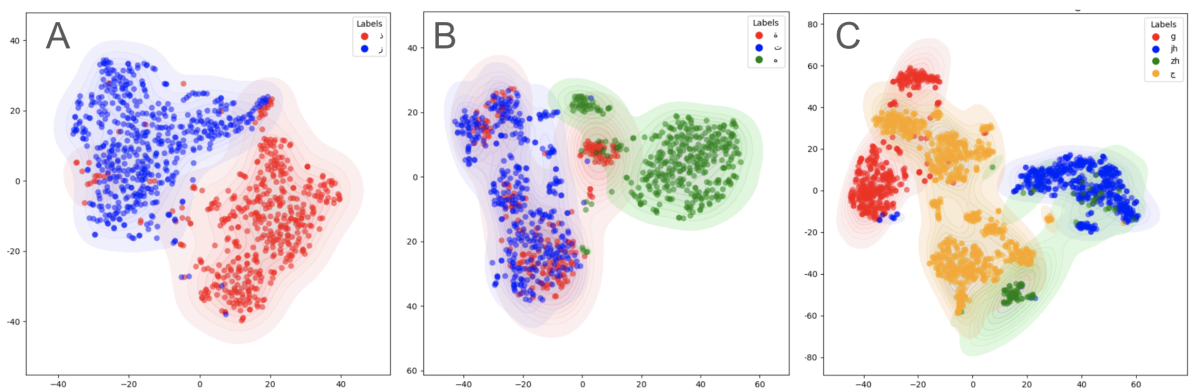

Figure 5: 2D t-SNE Projection of Frame-Level Presentations Extracted Randomly from Finetuned Arabic XLS-R. A. Pairs (ذ ز) [ð z]. B. Sounds (ه ة ت) [h t]. C. Pairs (ج [dʑ], zh [ʑ], g).

where people don't follow rule based pronunciation, proves challenging. The figure reveals two main clusters for [t] and [h], with vectors associated with ة scattered between these clusters, highlighting the aforementioned point.

Figure 5.C illustrates the selection of four labels: Arabic ج [[dʑ], and English phonemes (zh, g, jh) [ʑ, g, dʑ]. We selected 1000 Arabic samples of ج from CommonVoice Arabic and EgyAlj, along with 500 samples for each of the English phonemes. It became apparent that the Arabic sound ج is distributed across different English pronunciations (zh, g, and jh), indicating dialectal variations in the pronunciation of ج.

A.2 Country-wise DVR performance

In this section, we present the aforementioned results discussed in Section 6. Figure 4 displays CER results for the Baseline (Z), SVR Discrete (k:256), and DVR joint (Z+k:256) models trained on 1H30min of data, tested on AraVoice15. We analyze the CER results for each dialect individually. Our observations reveal that SVR Discrete (k:256) and DVR joint (Z+k:256) consistently outperform the Baseline (Z) across all dialects, exhibiting a substantial performance gap in MOR, YEM, PAL, and IRA dialects. Moreover, SVR Discrete (k:256) and DVR joint (Z+k:256) exhibit similar performance across the majority of the 15 dialects (10/15), with notable disparities observed in JOR, SUD, SYR, where a discernible performance gap is evident.