Towards One Model to Rule All: Multilingual Strategy for Dialectal Code-Switching Arabic ASR

Shammur Absar Chowdhury[1], *Amir Hussein*[1,2], *Ahmed Abdelali*[1], *Ahmed Ali*[1]

[1]Qatar Computing Research Institute, Qatar
[2]KanarI AI , California, USA

{shchowdhury, aabdelali, amali}@hbku.edu.qa, amir@kanari.ai

Abstract

With the advent of globalization, there is an increasing demand for multilingual automatic speech recognition (ASR), handling language and dialectal variation of spoken content. Recent studies show its efficacy over monolingual systems. In this study, we design a large multilingual end-to-end ASR using self-attention based conformer architecture. We trained the system using Arabic (Ar), English (En) and French (Fr) languages. We evaluate the system performance handling: (i) monolingual (Ar, En and Fr); (ii) multi-dialectal (Modern Standard Arabic, along with dialectal variation such as Egyptian and Moroccan); (iii) code-switching – cross-lingual (Ar-En/Fr) and dialectal (MSA-Egyptian dialect) test cases, and compare with current state-of-the-art systems. Furthermore, we investigate the influence of different embedding/character representations including character *vs* word-piece; shared *vs* distinct input symbol per language. Our findings demonstrate the strength of such a model by outperforming state-of-the-art monolingual dialectal Arabic and code-switching Arabic ASR.

Index Terms: multilingual, multi-dialectal, code-switching, conformer, E2E, speech recognition

1. Introduction

Multilingual ASR [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] have shown remarkable improvement over monolingual system and has recently gained immense popularity. A single model capable of learning from many languages has been a motivation for multi- and cross-lingual speech communities for many decades. With the recent success of end-to-end models over the hybrid systems in monolingual settings [8], along with the availability of large multilingual speech datasets, more focus has shifted towards leveraging multiple languages for a more robust, language-agnostic all-in-one system.

Studies like [5, 4] used seq2seq and RNN-Transducer [11] models for Indian languages, and used language id as an additional input to improve the performance. A recent study [2], designed a multilingual system covering 51 languages and suggested that such a system can benefit the performance of low-resourced languages.

However, training such a model to cover all the languages ($\approx 7K$), is computationally challenging, and is hard to maintain. Furthermore, having a single model supporting all languages may not be a practical need for regional ASR – e.g., Indic and Arabic languages – where the language with-in is very rich in dialects and also has been heavily influenced by a handful of languages due to globalization and/or colonization.

In such multilingual (and multidialectal) communities, code-switching – a speaker switches from one language (dialects) to another within an utterance (intrasentential) – is also a very common phenomenon. Recently, in addition to multilingual ASR, attention has been paid to design code-switching (CS) ASR. Studies have been conducted for Mandarin-English [12], Hindi-English [13], and French-Arabic [14] and little to no prior work in dialectal code-switching.

One of the major challenges in designing CS ASR is the lack of CS training data. This drawback restricts the exploitation of End-to-End (E2E) systems. However, recent studies like [13] models Hindi-English CS using E2E attention model, and [15] uses context-dependent target to word transduction, factorized language model and code-switching identification. The authors in [16] proposed two symmetric language-specific encoders to capture the individual language attributes in a transformer-based architecture.

Unlike the aforementioned studies, we investigate how to design a multilingual ASR and leverage it to enhance the performance of the ASR in dialectal and code-switching contents, without any external language identification input or code shift recognition. For this study, *Arabic* is an appropriate language due to its uniqueness as a shared language with 22 countries and having more than 20 mutually incomprehensible dialectal variety with modern standard Arabic (MSA) being the only standardized dialect [17], thus posing a unique set of challenges [18]. Furthermore, the dialects in North African regions (e.g. Morocco, Algeria) are heavily influenced and adapted to the French language. French and English also act as lingua-franca with non-Arabs or non-native dialectal speakers and thus commonly use in spoken contents like broadcast news or other spoken contents. Therefore, in this study, we :

(a) Designed an *end-to-end ASR* system supporting dialectal Arabic (**Ar**), English (**En**) and French (**Fr**) languages. We trained the model using a self-attention based conformer architecture and benchmark the system in: *(i)* Monolingual contents: Ar, En, and Fr only, *(ii)* Dialectal Arabic (DA) contents: Egyptian, Moroccan and MSA, *(iii)* Code-switching contents: cross-lingual (Ar⇔En/Fr) and in dialectal CS (MSA⇔Egyptian dialect).

(b) Compared *different character representation space* (e.g., shared *vs* distinct input symbol for Latin languages, character *vs* word-piece tokenization) to investigate its influence on model performance.

(c) Analyzed the effect of inconsistent ASR output – resulting in the same word being transcribed using different writing systems, on the ASR performance measure. For this, in addition to the word error rate, WER, we benchmark the CS ASR results with transliterated WER (TW).

This is the first study to benchmark the performance of a multilingual ASR for dialectal and code-switching Arabic test sets. The proposed model outperforms current Arabic state-of-the-art E2E

http://dx.doi.org/10.21437/Interspeech.2021-1809

ASR. Without any adaptation, the model gives a comparable performance in heavily dialectal datasets. Our finding also suggests, with little to no code-switching in training data, the multilingual strategy is capable of modeling both cross-lingual and dialectal code-switching without any language/dialect identification input. With this study, we release two new code-switching datasets. This study is the potential benchmark for future dialectal and code-switching Arabic ASR.

2. End-to-End Acoustic Model

To train a dialectal code-switching Arabic ASR with multilingual support, we adopted an end-to-end convolution-augmented transformer (conformer)[1] architecture, consisting of a number of conformer encoders and transformer decoders, as proposed in [19]. In the architecture, given input feature, the decoder predicts the output $\widehat{Y}_t$, at t time step conditioning on the final latent representation of the encoder and the previous output target sequence, i.e. $\widehat{Y}_{1\ldots t-1}$, in an auto-regressive manner.

Input: From raw signals, we extract 83-dimensional feature frames consisting of 80-dimensional log Mel-spectrogram and pitch features [20] and apply cepstral mean and variance normalization (CMVN). The acoustic features, X is then transformed into sub-sampled sequence $\widehat{X} \in \mathbb{R}^{O_l \times D_i}$ (O_l: length of the output sequence and D_i is input feature dimension to the encoder) using the convolution sampling layer.

Encoder-Decoder: We then pass the features to the conformer encoders – each comprise of a stack of four modules including a positionwise feed-forward (FFN) module, multihead self-attention (MHSA) module, a convolution operation (CONV) module, and another FFN module in the end. As for the decoder, we used transformers – each with an extra masked self-attention layer in addition to a MHSA and a feed-forward layer.

Training Loss: To improve the system robustness, we train the system using a multi-task learning objective. We combined the decoder cross entropy (CE) loss $\mathcal{L}_{ce} = -log(P_d(Y|X))$ and the CTC loss [21] $\mathcal{L}_{ctc} = -log(P_{ctc}(Y|X))$, with weighting factor, α, given the posterior probability of output sequence Y of X acoustic input: $\mathcal{L} = \alpha\mathcal{L}_{\text{ctc}} + (1-\alpha)\mathcal{L}_{\text{ce}}$

3. Corpus

3.1. Monolingual Datasets

English Data: For training, validating and testing the ASR, we use TEDLIUM3[2], LibriSpeech[3] (clean and other) and Multi-Genre Broadcast (MGB-1) [22] dataset. Details of train/dev/test split along with genre is presented in Table 1.

French Data: To incorporate the French language, due to its influence in North African Arabic countries, we add a subset from CommonVoice-French[23] dataset to the train/dev set and also test it on their official test data. See details in Table 1.

MSA and Dialectal Arabic Data: For the study, we use a subset of 642 hours of speech data, from a multi-genre broadcast dataset, QASR [24] dataset of 2k hours, collected from Aljazeera Arabic news channel's archive, spanning over 11 years from 2004 until 2015. The data includes lightly supervised transcriptions, covering contents mostly in MSA ($\approx$ 70%) and the rest in DA from regions like Egypt, Gulf, Levantine, and North Africa.

Table 1: *Data description for train, dev and test set. C: means CMI value. The duration: in speech hours. [*] hours of data for multiple reference.*

Datasets		Train (hrs)	Dev (hrs)	Test (hrs)
En	TEDL / LibriC/ LibriO/ MGB1	50/50/50/50	1/0.5/0.5/–	2.62/5.4/5.1/–
Fr	CV	100	2	20.64
Ar	QASR	642	-	-
	MGB2	-	4	9.57
	MGB3	4 [16.14]	2	5.78
	MGB5	10.2 [38.47]	1.3	1.4
CS	ESCWA.CS	-	-	2.8 (C: 28)
	QASR.CS	-	-	5.9 (C: 30.5)
	DACS	-	-	1.5 (C: 36.5)
TOTAL		996.61	11.3	51.14

A small percentage of the dataset also contains intrasentential code-switching. We quantify the amount of code-switching present in the subset using corpus level *Code-Mixing Index* (CMI), motivated by [25, 26]. See in Table 1 for details.

Furthermore, we add two real ecological dialectal datasets, collected from YouTube, distributed over different genres[4]:

- Egyptian MGB3 [27]: A $\approx$ 16 hours dataset from 80 videos.
- Moroccan MGB5[28]: A $\approx$ 13 hours of speech from 93 YouTube videos.

3.2. Intrasentential Code-Switching

Between-Language Code Switching Data: To evaluate the ASR performance in intrasentential code-switching instances, we test the ASR using two code-switching test sets:

- **ESCWA.CS**:[5] $\approx$ 2.8 hours of speech code switching corpus collected over two days of meetings of the United Nations Economic and Social Commission for West Asia (ESCWA) in 2019. The data includes intrasentential code alternation between Arabic and English. In cases of Algerian, Tunisian, and Moroccan native-speakers, the switch is between Arabic and French. Our initial analysis shows that such phenomena is present in $\approx$35% of the dialectal Arabic speech. Further investigation indicates that on average 22% of the segments is English/French content and other 78% is dialectal Arabic. The corpus level CMI of ESCWA.CS is 28.
- **QASR.CS**:[6] $\approx$ 5.9 hours of code switching extracted from the Arabic broadcast news data (QASR) to test the system for code switching. An average of 30.5 CMI-value is observed in the corpus level. The dataset also include some instances where the switch is between Arabic and French, however this type of instances are very rare occurrence.

Dialectal Code Switching Data: Dialectal code-switching (DCS) is a common phenomena in Arab region. However, it is more challenging and under studied research. Therefore to evaluate DCS from MSA to Egyptian dialect and vice versa, we tested the ASR with dialectal Arabic code-switching dataset (**DACS**)[7] [25] of $\approx$ 1.5 hours of speech. The CMI value for the overall corpus is 36.5. In addtion to DCS, the data also includes few instances of cross-lingual CS. These code alteration tokens are transliterated in Arabic in the distributed corpus. For more details about the data mentioned in the section, refer tot Table 1.

[1] We also explored transformer encoders and obtained a similar pattern with all test sets. However, as the conformer outperforms the transformer encoder, for brevity, we are reporting all the results with conformer ASR.

[2] `https://openslr.magicdatatech.com/51/`

[3] `https://openslr.magicdatatech.com/12/`

[4] For Example: Cooking, Sports, Drama, TEDx talks among others.

[5] `https://arabicspeech.org/escwa`

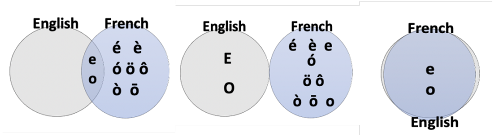
[6] `https://arabicspeech.org/qasr`

[7] `https://github.com/qcri/Arabic_speech_code_switching`

4. Experimental Setup

4.1. Data Preparation

For the acoustic modeling, we first augment the raw speech data with the speed perturbation, with speed factors of 0.9, 1.0, and 1.1 [29]. We then use the augmented audio to extract the input features (log Mel-spectrogram with pitch) and again augmented with specaugment approach [30]. For transcription, we first cleaned the data (*i*) removing all punctuation except the % and @ due to its verbatim usage, (*ii*) removing diacritics (for Arabic), (*iii*) transliterating all Arabic digits to Arabic numerals (e.g. ١ to 1) and (*iv*) converting all the Latin characters to lower case.



(a) Default (b) District (c) Forced shared

Figure 1: *Illustration of the different character space.*

4.2. Model Parameters

We trained the end-to-end ASR using a Noam [31] optimizer for 50 epochs with a learning rate of 5 with $20,000$ warmup steps and dropout-rate of 0.1. The trade-off weights, α, for $\mathcal{L}$, we use a value of α =0.3. The number of encoder, decoder layers and attention-heads differed based on the choice of large/small architecture. As for the text tokenization, we used word-piece byte-pair-encoding (BPE) [32] for the multilingual ASR.

Large ASR Architecture: For the large ASR (with $\approx 1,000$ hours of speech), we used 12 encoder layers and 6 decoder layers each with 2,048 encoder/decoder units from FFN and 8 attention heads with 512 transformation dimensions. For the architecture, we used 31 CNN kernals. For multilingual dialectal ASR, we opt for an BPE of $10K$.

Small ASR Architecture: For exploring the influence of different character space representation, we designed small-scale ASR using 8/4 encoder/decoder layers each with 2048 FFN units. As for the attention modules, we used 4 attention head with a dimension of 256. For the task, we opt for 15 CNN module kernals.

4.3. Monolingual baseline models

To compare the performance of multilingual ASR with monolingual performance using the same English (200 hours) and French (100 hours) data, we trained individual monolingual ASR using the small conformer architecture described in Section 4.2. For the experiments we used a bpe size of 500, motivated by the ESPnet recipe for TEDLIUM data. For the Arabic monolingual and dialectal models, we used reported study by [33], using E2E transformer models trained with MGB2-train data.

4.4. Shared/Distinct Character Space

To explore the impact of using a shared or distinct character set, in the ASR performance, when dealing with multilingual and code-switching data, we experimented with different input character sets. We investigate three settings: (a) natural shared character (see Figure1(a) Default); and (b) distinct character (see Figure1(b) Distinct). Furthermore, we also compared with a strict variation of *(a)*, where (c) all the nearby characters are mapped to one character representation (see Figure1(c) Forced shared).

We randomly picked a subset of 100 hours of training speech, maintaining a similar distribution as the full dataset (i.e. 70% Arabic, 20% English, and the rest 10% French, motivated by analysing ESCWA meeting recordings). We then trained ASR systems with each of the settings (in Figure1) and tested on the dev and test sets. We trained the model using the small conformer architecture and parameters (mentioned in Section 4.2) and a BPE size of 1000 (i.e. $\frac{1}{10}$ of the large ASR). In addition to BPE, we also use character-based encoding to compare the performance of default setting (Figure1a: Default), using a vocabulary size of 120 characters.

4.5. Evaluation Measures

We benchmarked the ASR using the conventional word error rate (WER). Furthermore, we also reported the transliteration WER (TW) for the code-switching test sets. We hypothesize that transliterating the En and Fr recognized tokens into Arabic, will help to disambiguate some code-switching errors introduced by the multilingual writing systems supported by the ASR. For this, we created a simple Global Mapping File (GLM) to transliterate, using system proposed in [34] and the recognized outputs. For such baseline evaluation, we only considered English words and ignored any mixed-code tokens.

Table 2: *Reported WER and TW on monolingual, dialectal Arabic datasets along with code-switching datasets. E2E:MGB2 represent transformer ASR trained with 1200hrs of MGB2. E2E:MGB2+TEDL: E2E:MGB2 + TEDLIUM3-train.* **: with added LSTM LM. Adapt: pretrained E2E:MGB2, adapted on the in-domain dialectal data.*

Tests	Multilingual		SOTA
	WER	**TW**	**WER (*Model*)**
MGB2	**12.1**	–	12.5 (*E2E:MGB2**)
MGB3	**35.2**	–	36 (*Adapt*)
MGB5	61.3	–	57.2 (*Adapt*)
QASR.CS	**26.4**	*25.8*	–
ESCWA.CS	**37.7**	*37.3*	50.1 (*E2E:MGB2+TEDL**)
DACS	**25.0**	**25.0**	30.7 (*E2E:MGB2*)

5. Results and Discussion

5.1. ASR Performance

The performance of the designed E2E multilingual ASR on dialectal Arabic and code-switching test sets are reported in Table 2. The results are benchmarked using several state-of-the-art models discussed in [33].

From WER, we observed that the proposed multilingual ASR outperforms the monolingual E2E transformer ASR (E2E:MGB2) – trained with MGB2 (2× more) data – on the MGB2-test set by 0.4% absolute WER. We also noticed an increase in performance (by 0.8%) in MGB3-test compared to E2E ASR [33] adapted specifically for Egyptian dialect. The multilingual model gives a comparable performance to the adapted ASR (on MGB5 train), when tested on MGB5-test set.

One of the main drawbacks of a monolingual model is that it can not handle cross-lingual code-switching tasks. Therefore, we benchmarked the ASR on intrasentential code-switching tests using a bilingual transformer model [35], trained using MGB2 and TEDLIUM3 train set. From the ESCWA.CS data, we see a significant decrease in WER when using the proposed ASR. This gain in performance reflects the capability of the multilingual model to handle both dialectal and code-switching data.

As for the dialectal code-switching (DACS-test set), we

Table 3: *Reported WER for En and Fr testsets using monolingual ASR ; MLMD ASR; and the current SOTA (ESPnet) models using large conformer (LC) architecture. represents the WER obtained using transformer LM in addition to AM.*

WER	Monolingual		Multilingual	SOTA
	En (200hrs)	Fr (100hrs)	(1khrs) Ar,En,Fr	LC
TedL	18.9	–	19.8	7.6
LibriC	8	–	8.6	2.1*
LibriO	16.1	–	16.5	4.9*
CV:Fr	–	16.2	18.2	14.8

noticed our multilingual ASR significantly outperforms a large monolingual E2E:MGB2 ASR by absolute 1.8% WER. This again indicates the strength of the latent dialectal representation learned in the proposed ASR.

In addition to the Arabic test sets, we evaluate the performance of the proposed ASR using monolingual En and Fr test sets, reported in Table 3. We compared the proposed multilingual ASR with its monolingual counterparts (see Section 4.3) and also reported state-of-the-art results on these test sets. We observed the multilingual ASR gives a comparable performance w.r.t. monolingual model, however, it is significantly accurate in spotting the language and no confusion was seen between different scripts for the En and Fr test sets.

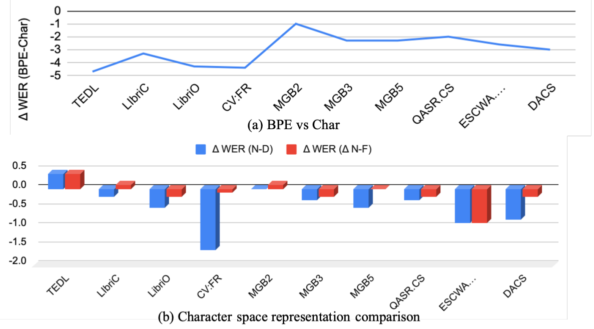
5.2. Error Analysis and Transliteration WER

We studied different types of error in dialectal and code-switching test sets. We noticed that in the dialectal dataset (mainly Moroccan: MGB5), most substitutions are due to inconsistencies between the dialectal and MSA orthographic and linguistic rules. We noticed substitutions in relative pronoun اللي ("Ally" - "which") that does not exit in MSA and is replaced by its closest match لي ("ly" - "for me"). We also noticed large numbers of substitutions due to the presence of vowels like Alif/Ya, such as the case of نحنا ("nHnA" - "we") and نحن ("nHn" - "we") orرح ("rH" - "will") and راح ("rAH" - "will"). Such errors are easily fixed with a GLM designed for Maghrebi dialects[8]. It is commonly noted in dialectal Arabic that some substitutions are frequent in one or more dialects[36] such as the case of "ث" or "ذ" that are substituted by "ت" or "س" and "ز" respectively. Another type of error, we noticed is partial/full transliteration of a word, e.g. "Artificial" to "ارت ificial" and "Drones" to "الدرونز". Such errors are noticed often and motivates the use of transliterated WER (TW). From the reported TW results in Table 2, we noticed transliterating all outputs to one language smoothes out the rendering error and offers a better perspective of the output. However, this still does not handle partial transliterated ASR output (e.g., "ارت ificial"). Thus indicating a need for a better CS evaluation metric.

5.3. Effect of Different Character Representation Space

We examined the effect of different strategies representing the Latin character space, and different tokenization techniques (BPE *vs* Char) (see Figure 2a-b). We noticed using distinct character set (D) (Figure 1b), WER increases significantly ($\Delta WER(N-D)$ is negative, meaning WER(D)>WER(N)) in the monolingual Fr. A similar observation is seen with the code-switching data, specially ESCWA.CS. We hypothesize such a change in performance is due to the presence of Fr CS and North African regional dialectal instances in ESCWA.CS data and the same is seen in MGB5 data which is also heavily influenced by Fr. Moreover, our result suggests that using the natural distinction (default settings in character space), the model is able to capture language-agnostic, yet discriminating representations.

In addition to character representation, we also noticed BPE is better in representing multilingual embeddings than character-based tokenization. Thus, validating our choice for using BPE as the tokenizer of the proposed ASR.



(a) BPE vs Char

(b) Character space representation comparison

Figure 2: *Change in WER* $\Delta = WER(N) - WER(*)$ *Figure 2 (a): BPE* vs *Char; Figure 2(b): naturally shared character (N)* vs *distinct (D) and forced merged (F) character set.*

5.4. Key Observations

Using the multilingual training strategy, we observed improved performance when tested on Modern Standard and Dialectal Arabic test sets. These findings indicate the strength of shared latent (dialectal) language representation in such an ASR. The WER in monolingual English and French along with intrasentential code-switching settings shows the proposed model is able to distinguish the input language with more success and has the potential to be the one system to replace them (monolingual, CS ASR) all. The advantage of using a shared language representation for modeling is also reflected from our study, using different character representation scripts. Our evaluation, using TW, also points out the drawbacks of using WER, as a measure for CS, in offering a better understanding of the model. Thus, signaling a need for better evaluation metrics for code-switching.

6. Conclusions

In this paper, we presented the first comprehensive study comparing multilingual ASR strategy to develop an E2E Arabic dialectal and code-switching ASR. The study benchmarked the performance of the proposed multilingual ASR for dialectal and code-switching Arabic test sets. The multilingual E2E conformer model outperforms the current Arabic state-of-the-art E2E transformer ASR. Without any fine-tuning/adaptation, the model gives a comparable performance in dialectal datasets and is fully capable of handling dialectal and cross-lingual code-switching instances without the need for any/large training data.

Moreover, we explored the strength of such multilingual ASR in learning better latent representation – gained from the multiple languages and its shared character space. With this study, we also release two new cross-lingual code-switching datasets. This dataset has the potential to benchmark future dialectal code-switching Arabic ASR.

[8] We present WER after post-processing with such GLM, with a gain of $\approx 5\%$ in WER.

7. References

[1] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8239–8243.

[2] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[3] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.

[4] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," *Proc. Interspeech 2019*, pp. 2130–2134, 2019.

[5] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[6] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 521–527.

[7] H. Bourlard, J. Dines, M. Magimai-Doss, P. N. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, no. 5, pp. 885–915, 2011.

[8] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8619–8623.

[9] H. Lin, L. Deng, D. Yu, Y.-f. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary asr," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4333–4336.

[10] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey *et al.*, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4334–4337.

[11] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[12] Y. Li and P. Fung, "Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints," in *ICASSP*, 2013.

[13] G. Sreeram and R. Sinha, "Exploration of end-to-end framework for code-switching speech recognition task: Challenges and enhancements," *IEEE Access*, vol. 8, pp. 68 146–68 157, 2020.

[14] D. Amazouz, M. Adda-Decker, and L. Lamel, "Addressing code-switching in French/Algerian Arabic speech," in *Interspeech*, 2017.

[15] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[16] X. Zhou, E. Yılmaz, Y. Long, Y. Li, and H. Li, "Multi-encoder-decoder transformer for code-switching speech recognition," *arXiv preprint arXiv:2006.10414*, 2020.

[17] E. S. Badawi, M. Carter, and A. Gully, *Modern written Arabic: A comprehensive grammar*. Routledge, 2013.

[18] A. Ali, S. Chowdhury, M. Afify, W. El-Hajj, H. Hajj, M. Abbas, A. Hussein, N. Ghneim, M. Abushariah, and A. Alqudah, "Connecting arabs: bridging the gap in dialectal speech recognition," *Communications of the ACM*, vol. 64, no. 4, pp. 124–129, 2021.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[20] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 2494–2498.

[21] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006.

[22] P. Bell, M. J. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The mgb challenge: Evaluating multi-genre broadcast media recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 687–693.

[23] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[24] H. Mubarak, A. Hussein, S. A. Chowdhury, and A. Ali, "Qasr: Qcri aljazeera speech resource a large scale annotated arabic speech corpus," in *ACL*, 2021.

[25] S. A. Chowdhury, Y. Samih, M. Eldesouki, and A. Ali, "Effects of dialectal code-switching on speech modules: A study using egyptian Arabic broadcast speech," *Proc. Interspeech*, 2020.

[26] B. Gambäck and A. Das, "Comparing the Level of Code-Switching in Corpora," in *LREC*, 2016.

[27] A. Ali, S. Vogel, and S. Renals, "Speech recognition challenge in the wild: Arabic mgb-3," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 316–322.

[28] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1026–1033.

[29] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[30] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[32] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," *arXiv preprint arXiv:1808.06226*, 2018.

[33] A. Hussein, S. Watanabe, and A. Ali, "Arabic speech recognition by end-to-end, modular systems and human," *arXiv preprint arXiv:2101.08454*, 2021.

[34] F. Dalvi, Y. Zhang, S. Khurana, N. Durrani, H. Sajjad, A. Abdelali, H. Mubarak, A. Ali, and S. Vogel, "QCRI live speech translation system," in *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 61–64. [Online]. Available: https://www.aclweb.org/anthology/E17-3016

[35] A. Ali, S. A. Chowdhury, A. Hussein, and Y. Hifny, "Arabic code-switching speech recognition using monolingual data," *Proc. Interspeech 2021*, 2021.

[36] Y. Samih, M. Attia, M. Eldesouki, A. Abdelali, H. Mubarak, L. Kallmeyer, and K. Darwish, "A neural architecture for dialectal Arabic segmentation," in *Proceedings of the Third Arabic Natural Language Processing Workshop*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 46–54. [Online]. Available: https://www.aclweb.org/anthology/W17-1306