INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Arabic Code-Switching Speech Recognition using Monolingual Data

Ahmed Ali[1], *Shammur Chowdhury*[1], *Amir Hussein*[1,2], *Yasser Hifny*[3]

[1]Qatar Computing Research Institute, HBKU, Doha, Qatar
[2]Kanari AI, California, USA
[3]University of Helwan, Egypt

{amali, shchowdhury}@hbku.edu.qa, amir@kanari.ai, yhifny@fci.helwan.edu.eg

Abstract

Code-switching in automatic speech recognition (ASR) is an important challenge due to globalization. Recent research in multilingual ASR shows potential improvement over monolingual systems. We study key issues related to multilingual modeling for ASR through a series of large-scale ASR experiments. Our innovative framework deploys a multi-graph approach in the weighted finite state transducers (WFST) framework. We compare our WFST decoding strategies with a transformer sequence to sequence system trained on the same data. Given a code-switching scenario between Arabic and English languages, our results show that the WFST decoding approaches were more suitable for the intersentential code-switching datasets. In addition, the transformer system performed better for intrasentential code-switching task. With this study, we release an artificially generated development and test sets, along with ecological code-switching test set, to benchmark the ASR performance.

Index Terms: speech recognition, code-switching

1. Introduction

Multilingual speech recognition has made rapid progress in recent years, primarily due to the advances in deep learning and the availability of adequate training resources [1, 2, 3, 4].

In a study by [5], they discussed and compared the connectionist temporal classification (CTC) and the end-to-end lattice-free maximum mutual information (LF-MMI) for multilingual ASR. They illustrated that end-to-end LF-MMI is indeed competitive on a low-resourced multilingual task, comfortably outperforming a CTC baseline, owing specifically to its end-to-end nature, while the usage of context-independent phone labels made it attractive for multilingual ASR.

In a study by [6], they presented a single sequence-to-sequence ASR model trained on 9 different Indian languages, with very little overlap in their scripts. They took the union of language-specific grapheme sets and trained a grapheme-based sequence-to-sequence model jointly on data from all languages. They achieved 21% in performance over separately trained systems. A language-agnostic multilingual ASR system was studied by [7], which is the case for multicultural societies where several languages are frequently used together and often rendered with different writing systems. The system transforms all languages to one writing system through a many-to-one transliteration transducer. Thus, similar sounding acoustics were mapped to a single, canonical target sequence of graphemes, effectively separating the modeling and rendering problems. They achieved 10% relative reduction over a language-dependent system.

Most of the previous studies focused on multilingual scenarios without paying attention to code-switching, where there is an alternation between two or more languages within the same utterance. Code-switching in spontaneous speech is highly unpredictable and difficult to model [8]. English-Mandarin has been studied most extensively [9, 10, 11] in addition to other language pairs such as Frisian-Dutch [12], Hindi-English [13, 14, 15] and French-Arabic [16].

There have been initiatives in building and analysing code-switching in Arabic-English speech [17]. [18] studied code-switching in Sudanese Arabic and English through social media applications, 75% of their corpus was ready by 87 bilingual Arabic native speakers resulting in 2,289 audio files. They achieved 33% word error rate (WER) using ASR and language identification pipeline. [19] built the ArzEn corpus, speech corpus for Egyptian Arabic and English code-switching. ArzEn comprised of 12 hours transcribed along with meta-data from linguistic, sociological, and psychological perspectives. Their corpus was recorded by 38 participants. They reported Code-mixing Index (CMI) over the whole corpus of 0.12, and over the code-switching sentences is only 0.17. A recent study, in [20], proposed a multilingual strategy to model code-switching in Arabic ASR. With the E2E model they reported state-of-the-art results indicating the efficacy of such method for handling both cross-lingual and Arabic dialectal code-switching with little CS data present in QASR dataset [21].

However, most of the previous studies in Arabic English code-switching are lacking the following: (*i*) publicly available corpus to reproduce their results; (*ii*) real code-switching (ecological) data where participants are talking spontaneously in daily life, not in data collection projects; and mostly (*iii*) detailed and complete analysis of ASR behaviours and insights on AM and LM system in code-switching ASR. To the best of our knowledge, there are no public studies with reproducible results for Arabic-English code-switching in spontaneous speech using the available monolingual dataset.

Therefore, our contributions in this paper are: (*i*) Using global multilingual WFST approach (our baseline), (*ii*) Developing a novel approach combining the Kleene closure with multi-graph WFST to support multilingual and code-switching for hybrid ASR, and (*iii*) Comparing the two hybrid approaches with end-to-end transformer ASR in code-switching scenarios.

2. ASR Architecture

The hybrid HMM-DNN ASR architecture based on the weighted finite-state transducers (WFSTs) outlined in [22]. The training, development, and testing are the same as the Arabic MGB-2 [23] and the English TED-LIUM3 [24]. For more details, refer to Table 1.

Multilingual Hybrid ASR: For the hybrid ASR, we trained a Time Delay Neural Network (TDNN) [25] using sequence discriminative training with the LF-MMI objective [26] with

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-2231

Table 1: *Data used for acoustic modeling.*

Type	Hours	Programs	#segments
AR Train	1,200h	2,214	370K
AR Dev	10h	17	5,800
AR Test	10h	17	5,600
EN Train	450h	2,351	268K
EN Dev	1.6h	8	507
EN Test	2.6h	11	1,150

the alignments from a context-dependent Gaussian mixture model-hidden Markov model (GMM-HMM). The input to the TDNN is composed of 40-dimensional high-resolution MFCC extracted from 25 msec frames and 10 msec shift along with 100-dimensional i-vectors computed from 1500 msec. Five consecutive MFCC vectors and the chunk i-vector are concatenated, forming a 300-dimensional features vector each frame.

A universal grapheme set has been investigated to build end-to-end multilingual ASR systems [27, 6]. Modeling graphemes implicitly models spelling, which reduced the amount of entries in the lexicon. However, graphemes can differ immensely from language to language, and languages may have nothing in common in terms of graphemes (e.g. Arabic and English in our case). Thus, we propose a multilingual architecture that merges all graphemes from multiple languages, keeping the language identity at the grapheme level. A multilingual *4*-gram language model is learned over the transcription for all the languages.

End to End ASR Architecture: Our end-to-end ASR system is based on the transformer architecture [28], which consists of two sub-networks: an encoder model and a decoder model. The encoder transforms the input filter bank speech features $\mathbf{X}$ to a latent representation $\mathbf{H}$. Given the $\mathbf{H}$ and previous predicted tokens outputs $Y_{1:i-1}$, the decoder generates the next token Y_i. Th **Encoder** is a multi-blocked architecture, where each block consists of a multihead self-attention (MA) module and positionwise feed-forward (FF) module. The **Decoder** is similar to the encoder. However, it has an additional masked self-attention layer. The masked self-attention enforces the decoder to attend only to earlier positions in the output sequence.

During the training, the transformer ASR predicts the target sequence $\mathbf{Y}$ of tokens from acoustic features $\mathbf{X}$. For text tokenization, the word-piece byte-pair-encoding (BPE) [29] is used with BPE size of 5000. The total loss function $\mathcal{L}_{asr}$ is a multi-task learning objective that combines the decoder cross entropy (CE) loss $\mathcal{L}_{\text{ce}}$ and the CTC loss [30] $\mathcal{L}_{\text{ctc}}$.

3. Multilingual WFSTs

A weighted finite-state transducer (WFST) is a generalization of the finite automata where each transition has an input label, an output label, and an optional weight [22]. In this framework, it is possible to combine and compose multiple sources of knowledge to construct the output graph in a unified way. The output graph or network can be optimized using the determination and minimization algorithms. The speech recognition decoders based on WFSTs are employed in modern systems [31]. In WFST based decoding, the decoder is used to search the optimal solution constrained by a big WFST. This big WFST graph is composed of a language model (LM), an acoustic model (AM), and lexicon FSTs. Our implementation is based on the Kaldi speech recognition engine [32], where the final decoding graph $HCLG$ is composed of

$$HCLG = H \circ C \circ L \circ G \quad (1)$$

where H is the HMM definitions FST (its input symbols are acoustic modeling units (transition-ids) and the output symbols are the context-dependent phones). C is the context dependency FST (its input symbols represent context-dependent phones and its output symbols are monophones). L is the pronunciation lexicon FST (its input symbols are monophones and its output symbols are words). G is the language model or the grammar finite state acceptor (FSA). The HMM transition probabilities, HMM emission probabilities, and LM probabilities are encoded in the graph via the weights. The optimal path is found by searching this graph using the Viterbi algorithm [33]. It represents the most likely word sequence given the acoustic features extracted from the input audio.

In this work, we present two approaches to build true multilingual speech recognition systems based on one graph. This graph will encode all the knowledge sources for all languages. We assume that the acoustic model and pronunciation models are shared between all languages in this setup. Figure 1 summarizes the two approaches developed in this work.

3.1. Global G approach

In this approach, a language model is trained on a multilingual corpus [34]. The corpus consists of the concatenation of the text for each language. This approach does not require text that contains code-switching where the text is a mix of different languages. The language model is converted into G FSA and composed with other sources to build the $HCLG$ cascaded graph. The graph is used for decoding, and the output of this process is a bilingual/multilingual text, depending on the spoken audio.

3.2. Multi-graph approach

Assume we built a $HCLG$ graph for each language, then it is possible to search the graphs in parallel using a union operation [35]. However, searching the languages in parallel during decoding does not allow the transition between languages.

To overcome this problem, we propose to add transition arcs between the languages using the Kleene closure. Hence, the decoder can switch between languages during the decoding.

For example, assume we have simple $HCLG$ graphs for English and Arabic languages as shown in Figure 2. Then each of these graphs is composed with sigma matcher FST (shown in Figure 4) to get rid of the paths which only have epsilons on (i.e. no symbol). This is necessary to avoid errors during decoding in the Kaldi framework. The two languages can be searched in parallel using a union operation. Finally, a closure operation would allow the transition between languages during the search process. The two operations are shown in Figure 3.

Allowing the transitions between languages during the search means there is no need for explicit code-switching data to train the language model (main advantage of this approach). This is a desired property of the developed multilingual system.

4. Code-Switching Phenomena

The two most common types of code-switching are: (*i*) intersentential (switching between-utterances): the alternation in a single discourse between two languages, where the switching occurs after a sentence in the first language has been completed and the next sentence starts with a new language; and (*ii*) intrasentential (within utterances): the alternation in a single discourse between two languages, where the switching occurs within the same sentence.

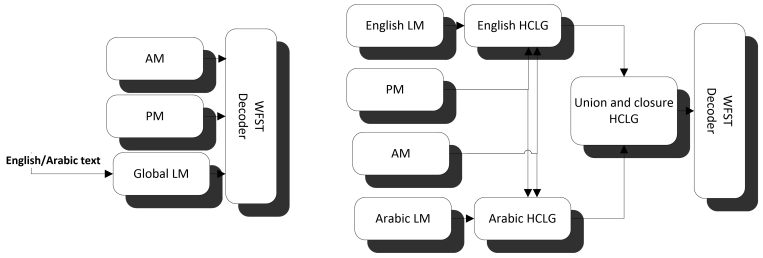

Figure 1: *Two approaches to build graphs used for multilingual speech recognition decoding. Global G approach on the left and multi-graph approach on the right. The acoustic model (AM) and the pronunciation model (PM) are shared between the two approaches.*

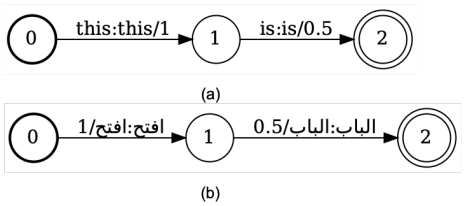

Figure 2: *Simple graphs for English and Arabic languages. (a) A simple $HCLG$ graph for English language. (b) A simple $HCLG$ graph for Arabic language.*

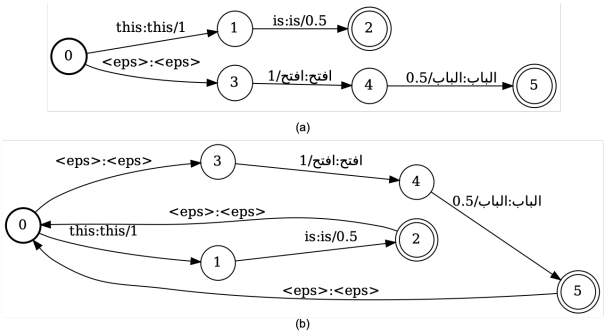

Figure 3: *Multilingual graph building using union and closure operations. (a) Generating the union of the two graphs to allow the search in the parallel graphs. (b) Connecting final states to the initial states via a closure operation to allow switching between languages during the search.*

These phenomena in spontaneous speech are highly unpredictable and difficult to model for both NLP and speech modules [36], thus are of great interest to the research community. One of the main reasons for such a gap in Arabic is the scarcity of both publicly available linguistic and speech resources, or lack of recipe to build models and test code-switching scenarios between Arabic and English.

4.1. Intersentential Code-Switching Corpus

We concatenate the audio files of Arabic and English, we did not add any extra silence between sentences. It is worth mentioning that during the concatenation of the audio files, it was not possible to maintain same speakers during the transition between languages. Since the number of utterances in the English TED-LIUM3 dev/test is lower than the corresponding number in the MGB-2 dev/test as shown in Table 1, we randomly picked sentences from the MGB-2 to match the TED-LIUM3. Despite the fact that this can be seen as artificial setup, there are many practical scenarios where non-native speakers are interviewed to comment on news, such as multinational events in Broadcast news domain [23].The *EAE corpus* refers to concatenating English audio from TED-LIUM3 with Arabic Audio from MGB-2 and English Audio from TED-LIUM3. This is to simulate $EN \longrightarrow AR \longrightarrow EN$ intersentential code-switching scenario. Same for *AEA corpus*. More details of the EAE and AEA are present in Table 2.

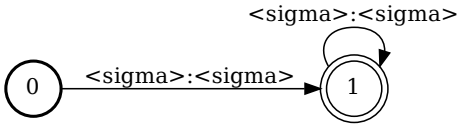

Figure 4: *Sigma special FST type.*

Table 2: *Dev and test data used for code-switching evaluation.*

Corpus ID	Total Hours	#segments
Dev $\mathcal{EAE}$	4.5h	506
Dev $\mathcal{AEA}$	4.1h	506
Test $\mathcal{EAE}$	7.8h	1,154
Test $\mathcal{AEA}$	8.0h	1,154
ESCWA	2.8h	845

4.2. Intrasentential Code-Switching: ESCWA Corpus

We studied an eight-hour corpus collected over two days of meetings of the United Nations Economic and Social Commission for West Asia (ESCWA) in 2019. From the data, we observed that more than 2.8 hours (in Table 2) of the collected speech demonstrate intrasentential code-switching, where the alternation between Arabic and English[1] is happening within the same sentence. We use *Code-Mixing Index* (CMI) to report the level of code mixing in the ESCWA corpus[2].

We reported the corpus level CMI by simply averaging the utterance level switching.[3] The details of the level of code-switching in ESCWA corpus are presented in Table 3, and an example of segment in 45-100% range is presented in Figure 5.

5. Experiments and Results

Table 4 shows the WER of the monolingual, the multilingual, and the transformer systems. The decoding in the multilingual system is based on the Global G approach, where the language

[1]In few cases, as with Algerian, Tunisian, and Moroccan native-speakers, the switch is between Arabic and French.

[2]Data can be downloaded from `https://arabicspeech.org/resources/`

[3]Ignoring the switches between the utterances.

Table 3: *Details of code-switching level of ESCWA data using CMI range. word/Utt. represents the average word count per utterance, CA is the mean number of code alternation points in utterances, #. is the number of utterances and Avg. dur is average duration in seconds in that particular CMI range.*

Utterance Level CMI				
CMI Range	CA	word/Utt.	Avg. dur	#.
0-15%	2.6	26.8	12.4	485
15-30%	4.7	24.2	11.4	289
30-45%	7.4	23.2	10.2	68
45-100%	7.7	14.7	8.6	3
Corpus-CMI:	28			

Table 4: *WER in % for hybrid ASR (monolingual, multilingual) and end-to-end transformer systems.*

Type	Monolingual	Multilingual	Transformer
AR Dev	17.2	17.1	15.49
AR Test	16.3	17.5	17.61
EN Dev	9.2	9.9	9.68
EN Test	9.1	9.8	8.29

model graph is built from a multilingual text. It is noticeable that the transformer system WER results are slightly better than the hybrid system.

We tested four methods to evaluate our multilingual decoding strategies: (*i*) the Global G approach, where the language model is trained on a multilingual corpus; (*ii*) the multi-graph approach, where we build $HCLG$ graph for each language, combined using union operation;(*iii*) the end-to-end transformer architecture;and (*iv*) finally, assuming that we have the oracle language code-switching, we build mono-language models for Arabic and English and decoded the split dev/test sets. This setup provides an upper bound on the code-switching results reported using the global G and multi-graph approaches.

Results are shown in Table 5. For intersentential code-switching, the WFST based decoding strategies are doing much better than the transformer system. This can be explained as the attention mechanism has never encountered a long Arabic sentence followed a long English sentence in the training data set. In addition, the multi-graph approach led to the best results. Hence, the WFST decoding strategies may be more suitable for the intersentential code-switching tasks. On the other hand, the transformer system results for intrasentential code-switching task evaluated using the ESCWA dataset are better than the WFST decoding strategies. The transformer system can respond quickly to the rapid change in the language during talking. This is not the case for WFST decoding systems, where it is difficult to switch the language model quickly during the decoding. In general, high WER for the ESCWA dataset is expected since there is acoustic conditions mismatch between the train and test datasets. Also, the data is dominated by dialectal Arabic with intrasentential code switching. Such intrasentential phenomena can affect the performance of ASR more than intersentential code-switching, as clearly observed from the ASR performance.

Intrasentential Example: Figure 5 presents an example from the ESCWA corpus with dialectal Arabic along with English. The 16-words sentence has a duration of 8.5 seconds with 9 switching points. In fact, it can be argued that there are 10 switching points, since the word data is written in Arabic script, while it is an English word. The hybrid speech-to-text system in the global graph approach and the multi-graph approach are

Table 5: *Decoding performance for intersentential and intrasentential code-switching tasks. WER in %.*

Decoding scenario	Dev WER	Test WER
$\mathcal{EAE}$: Global G WFST	20.2	23.3
$\mathcal{EAE}$: Multi-graph WFST	19.9	23.0
$\mathcal{EAE}$: Transformer	52.9	51.1
$\mathcal{EAE}$: Oracle language diarization	12.9	13.5
$\mathcal{AEA}$: Global G WFST	24.2	29.0
$\mathcal{AEA}$: Multi-graph WFST	23.8	28.7
$\mathcal{AEA}$: Transformer	54.2	54.5
$\mathcal{AEA}$: Oracle language diarization	16.0	17.2
ESCWA Global G WFST	63.2	
ESCWA Multi-graph WFS	65.2	
ESCWA transformer	**50.1**	

not able to get more than two switching points, while the end-to-end transformer is able to detect nine switching points. Table 6 shows the global-graph WFST hybrid system is achieving the lowest WER 62.5%, while the end-to-end achieves 81.2%. However, by looking at the character error rate (CER) and code-switching points, it is clear that the end-to-end is outperforming both of the hybrid systems considerably. This can be addressed using transliterated WER.

Intrasentential code-switching (**ESCWA Corpus**)	
Original Reference (Arabic and English)	هو بغطي ال governance والبنية التحتية والداتا وال skills وال education وال government وال public services
Translated Reference (only in English)	He covers the governance and the infrastructure and the data and the skills and the education and the government and the public services
Reference (Buckwalter Format)	hw bgTy Al governance wAlbnyp AltHtyp wAldAtA wAl skills wAl education wAl government wAl public services
ASR Systems	
Multi-graph WFST Hybrid ASR	*** *** *** governance Albnyp AltHtyp with data will skills will education government him into a public synthesis
Global-graph WFST Hybrid ASR	*** *** *** governance Albnyp AltHtyp data will skills will education government to a public services
Transformer End-to-End ASR	hw ygTy Algovernance wAlbnyp AltHtyp wAl data wAlskills wAleducation wAlgovernment to will bublic selfishs

Figure 5: *Example of intrasentential code-switching from ESCWA Corpus. Text is in Arabic, English and Buckwalter.*

Table 6: *WER and CER in % for the intrasentential code-switching example in Figure 5*

System	WER	CER
Multi-graph WFST	81.3	40.0
Global-graph WFST	62.5	26.7
Transformer	81.1	13.3

6. Conclusions

In this paper, we proposed an innovative new strategy for multilingual ASR speech recognition. In particular, we implemented three strategies to recognize code-switching speech. The first strategy is to decode the speech using a global language model built from a multilingual text. This approach serves as our baseline system. Our innovative framework deploys a multi-graph approach in the weighted finite state transducers (WFST) framework. Using a closure operation, our decoder can switch between languages during the decoding process, and the output of this process is a bilingual/multilingual text that depends on the spoken audio. The third strategy is to decode the speech using a powerful transformer system. Given a code-switching scenario between Arabic and English languages, the WFST decoding approaches were more suitable for the intersentential code-switching datasets. Moreover, the transformer system was doing well for the intrasentential code-switching task. The transformer system seems to very slow in training and decoding.

7. References

[1] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *ICASSP*, 2009.

[2] J. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[3] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP*, 2013.

[4] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, A. Rastrow, R. C. Rose, and S. Thomas, "Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models," in *ICASSP*, 2010.

[5] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of multilingual ASR using end-to-end LF-MMI," in *ICASSP*, 2019.

[6] 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 *ICASSP*, 2018.

[7] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," *arXiv preprint arXiv:2004.09571*, 2020.

[8] S. Sitaram, K. R. Chandu, S. K. Rallabandi, and A. W. Black, "A survey of code-switched speech and language processing," *arXiv preprint arXiv:1904.00784*, 2019.

[9] 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.

[10] ——, "Language modeling for mixed language speech recognition using weighted phrase extraction." in *Interspeech*, 2013.

[11] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for Mandarin-English code-switch conversational speech," in *ICASSP*, 2012.

[12] E. Yilmaz, M. Andringa, S. Kingma, J. Dijkstra, F. Kuip, H. Velde, F. Kampstra, J. Algra, H. Heuvel, and D. A. van Leeuwen, "A longitudinal bilingual frisian-dutch radio broadcast database designed for code-switching research," 2016.

[13] A. Pandey, B. M. L. Srivastava, R. Kumar, B. T. Nellore, K. S. Teja, and S. V. Gangashetty, "Phonetically balanced code-mixed speech corpus for Hindi-English automatic speech recognition," in *LREC 2018*, 2018.

[14] K. Taneja, S. Guha, P. Jyothi, and B. Abraham, "Exploiting monolingual speech corpora for code-mixed speech recognition," *Interspeech*, 2019.

[15] 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.

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

[17] 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.

[18] M. O. E. Elfahal *et al.*, "Automatic recognition and identification for mixed sudanese arabic–english languages speech," Ph.D. dissertation, Sudan University of Science & Technology, 2019.

[19] I. Hamed, N. T. Vu, and S. Abdennadher, "Arzen: A speech corpus for code-switched egyptian arabic-english," in *Proceedings of The 12th Language Resources and Evaluation Conference*, 2020, pp. 4237–4246.

[20] S. A. Chowdhury, A. Hussein, A. Abdelali, and A. Ali, "Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic Asr," *Proc. Interspeech 2021*, 2021.

[21] 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.

[22] M. Mohri, "Weighted automata algorithms," in *Handbook of weighted automata*, 2009.

[23] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition," in *SLT*, 2016.

[24] F. Hernandez, V. Nguyen, S. Ghannay, N. Tomashenko, and Y. Estève, "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation," in *International Conference on Speech and Computer*. Springer, 2018, pp. 198–208.

[25] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[26] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016.

[27] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," in *ICASSP*, 2018.

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

[29] 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.

[30] 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.

[31] M. Mohri, F. Pereira, and M. Riley, "Speech recognition with weighted finite-state transducers," in *Springer Handbook of Speech Processing*, 2008.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *ASRU*, 2011.

[33] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[34] C. Liu, Q. Zhang, X. Zhang, K. Singh, Y. Saraf, and G. Zweig, "Multilingual ASR with massive data augmentation," *arXiv preprint arXiv:1909.06522*, 2019.

[35] E. Yılmaz, S. Cohen, X. Yue, D. van Leeuwen, and H. Li, "Multi-graph decoding for code-switching ASR," *arXiv preprint arXiv:1906.07523*, 2019.

[36] 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.