Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition

Yonatan Belinkov[1,2], *Ahmed Ali*[3], *James Glass*[1]

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA
[2]Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA, USA
[3]Qatar Computing Research Institute, HBKU, Doha, Qatar
{belinkov,glass}@mit.edu, amali@qf.org.qa

Abstract

End-to-end neural network systems for automatic speech recognition (ASR) are trained from acoustic features to text transcriptions. In contrast to modular ASR systems, which contain separately-trained components for acoustic modeling, pronunciation lexicon, and language modeling, the end-to-end paradigm is both conceptually simpler and has the potential benefit of training the entire system on the end task. However, such neural network models are more opaque: it is not clear how to interpret the role of different parts of the network and what information it learns during training. In this paper, we analyze the learned internal representations in an end-to-end ASR model. We evaluate the representation quality in terms of several classification tasks, comparing phonemes and graphemes, as well as different articulatory features. We study two languages (English and Arabic) and three datasets, finding remarkable consistency in how different properties are represented in different layers of the deep neural network.

Index Terms: speech recognition, end-to-end, phonemes, graphemes, analysis, interpretability

1. Introduction

Traditional automatic speech recognition (ASR) systems employ a modular design, with different modules for acoustic modeling, pronunciation lexicon, and language modeling, which are trained separately. In contrast, end-to-end (E2E) models are trained to convert acoustic features to text transcriptions directly, potentially optimizing all parts for the end task. Unfortunately, they are also less interpretible: identifying what different parts do and what properties they capture is less straightforward.

It is a common problem in many neural network models besides E2E ASR. Therefore, a line of work is concerned with deciphering the information captured by learned representations in neural models that are trained on some downstream task [1]. Previous work analyzed different neural representations and various properties, such as evaluating how phonetic information is captured in neural acoustic models [2, 3]. However, E2E ASR models are still relatively under-explored.

In previous work [4], we anlyzed DeepSpeech2 [5] E2E models, from the perspective of the phonetic information that is learned in different layers. However, that work only considered TIMIT as a source of phonetic information. In this paper, we extend this analysis to multiple languages (English and Arabic) and three different datasets, as well as explore additional properties (e.g., phonemes vs. graphemes). We find that over many different configurations—languages, datasets, linguistic properties—the E2E models exhibit strikingly similar behavior across layers. We also investigate the drop in representation quality at the top layers, attributing part of it to the focus on graphemes and long-distance information.

Limitations: Potential limitations are the restrictions to a specific E2E architecture and to frame classification. Future work can explore other architectures and larger segments.

2. Related Work

2.1. Analysis of Representations

Several studies analyzed what phonetic information is encoded in acoustic models using clustering and classification methods to [2, 3]. Others correlated the behavior of gates in recurrent neural networks with phoneme boundaries [6, 7] or visualized skip connections in speech enhancement models [8]. Various phonetic and speaker features were investigated in speaker embeddings [9, 10], and properties like style and accent were analyzed in a convolutioanl ASR performance prediction model [11]. Another line of work is concerned with developing and analyzing joint audio-visual models [12, 13, 14, 15].

Recent work [16] clustered neurons in convolutional E2E ASR and found that lower layers encode phonemes better than graphemes. Most related, our previous investigation of E2E ASR [4] used the same E2E model and analyzed phoneme representations only on English in TIMIT. In contrast, here we explore two different languages (English and Arabic) and three datasets. We also consider new aspects such as different phonetic features and representing past and future information.

2.2. E2E and Arabic ASR

Recently, E2E ASR has attracted attention in both academia and industry. The E2E system is based on a single deep neural network that can be trained from scratch to directly transcribe speech into labels (words, phonemes, etc.) [5, 17, 18]. It integrates disjoint modules, developed from traditional hybrid methods, into one network. While attention-based models [19, 20] address the ASR problem as sequence mapping using an encoder-decoder architecture, the connectionist temporal classification (CTC) [21, 17] objective function performs frame-level classification with specialized time-aggregation.

Previous work made various attempts to reduce word error rate (WER) in Arabic ASR on the MGB-2 dataset [22]. While initial work used phoneme-based systems [23], recent work, and the wining submissions, are grapheme-based [24, 25]. These efforts reduced WERs from 32% to 14%. Since most systems focused on traditional ASR with separate acoustic, pronunciation, and language models, Arabic E2E is still unexplored. More broadly, it is important to understand the difference between grapheme and phoneme modeling in E2E setups.

Table 1: *Statistics of annotated datasets.*

(a) *Number and total duration of utterances.*

	Utterances			Hours:Minutes		
	Train	Dev	Test	Train	Dev	Test
Librispeech	1920	241	240	3:40	0:26	0:30
TedLIUM	345	79	76	1:05	0:15	0:14
MGB-2	1990	288	295	3:12	0:31	0:33

(b) *Label set sizes.*

	Phonemes	Graphemes	Place	Manner
English	40	28	9	7
Arabic	34	37	12	9

3. Methods

3.1. Analysis by Classification

To analyze the representation quality in the E2E ASR model, we adopt the paradigm of classification or probing tasks [1, 26, 27]. First, we train the E2E model in the normal fashion, on pairs of utterances and transcriptions. Then, we run the trained model on a dataset with frame-level annotations, such as phoneme labels, and record activations from different layers of the E2E model. These activations are fed to a classifier that is trained to predict the labels. A separate classifier is trained for every annotation type (say, phonemes or graphemes) and layer.

To maintain consistency with our previous analysis of DeepSpeech2 [4], the classifier is a simple feed-forward neural network with one hidden layer of size 500. The input and output sizes are determined by the feature representation from the E2E model and by the label set size, respectively. It is non-contextual, taking only the current frame representation, although context may be captured in the representation itself via the ASR model. The classifier is trained for 30 epochs and the model with the best validation loss is used for evaluation.

The code for running our experiments is publicly available.[1]

3.2. Classification Tasks

We consider the following classification tasks:

- **Phonemes**: for every frame, predict an aligned phoneme.
- **Graphemes**: for every frame, predict an aligned grapheme.
- **Phonetic features**: for every frame, predict the place or manner of articulation of its aligned phoneme.

3.3. Obtaining Labels

Our objective here is to estimate the correct timing of a sequence of phonemes for a speech signal given verbatim transcription. We use triphone HMM models with speaker information similar to [28]. It is worth noting that timings from the Viterbi-alignment results are not as precise as the manually-aligned TIMIT data. Therefore, we consider running phoneme classification using a bigger window to overcome potential shift in the timing as shown in Section 5.1. We also note that although using forced-aligned phonemes is a possible limitation, the experimental results show consistent patterns with our previous analysis using manual segmentation from TIMIT [4].

4. Experimental Setup

4.1. E2E ASR Systems

We use standard, large-scale datasets for training the E2E ASR models. For English, we use a pre-trained reimplementation [29] of DeepSpeech2 [5]. This model has 2 convolutional layers and 5 recurrent long short-term memory (LSTM) layers, trained with CTC. It was trained on Librispeech [30].

For Arabic, we train our own model using the implementation in [29], with the same architecture. We use the MGB-2 dataset [22], which comprises 1,200 hours of broadcast videos from the Aljazeera Arabic TV channel. We exclude sentences longer than 30 seconds and sentences failed to align with the seed models trained in [24]. This give us a total training data of 700 hours, which consists of more than 312K sentences.

[1] `http://github.com/boknilev/asr-repr-analysis`

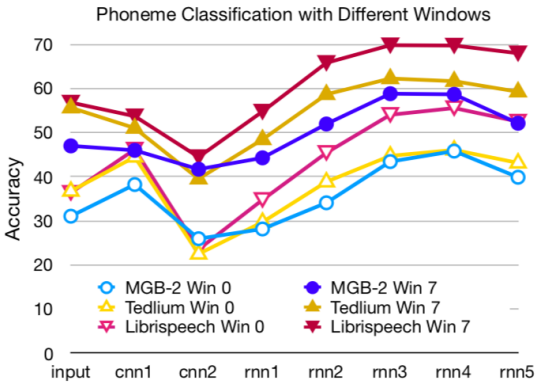


Figure 1: *Phoneme classification accuracy in different datasets.*

4.2. Modular ASR Systems

To obtain phoneme and grapheme labels, we employed traditional ASR systems with separate components for phoneme or grapheme modeling. This allows us to run forced-alignment and get annotated data to be used by the classifier.

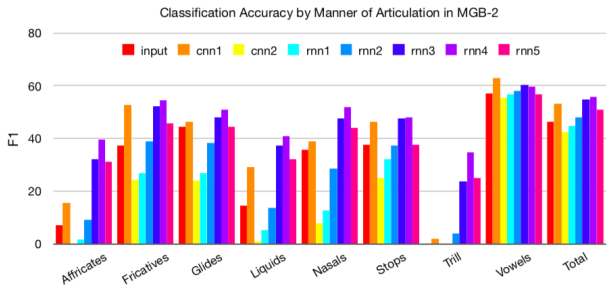
For English, we train time-delay neural network acoustic-models using the implementation in [31] on the standard TedLIUM corpus [32], comprising more than 210 hours and 92K sentences. For the phoneme system, we use the official pronunciation dictionary that has more than 152K words and 40 phonemes/monophones. Meanwhile, the English grapheme-based lexicon is formed from the 26 alphabet letters /a-z/, and was constructed similar to the system described in [33].

For Arabic, we train the same architecture as the English system on the MGB-2 corpus using the implementation in [24]. The phoneme system used the phonetic lexicon shared in the MGB-2 challenge [22], while the grapheme lexicon used the same word list with 1:1 word-to-character mapping to keep the same vocabulary size. Both lexicons have more than 950K words. For a detailed comparison of phoneme vs. grapheme Arabic ASR, see Section 7.2.3 in [34].

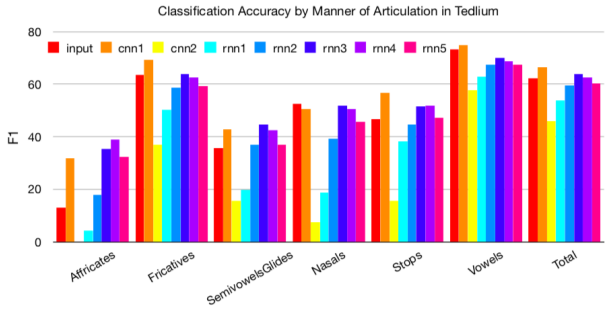
4.3. Classification Data

Given the modular ASR systems, we annotated a subset of each dataset with forced-aliged phonemes and graphemes. For articulatory features, we mapped phonemes to place and manner of articulation.[2] We used TedLIUM and Librispeech for English, and MGB-2 for Arabic. We made sure to train and evaluate the classifier on a portion of the data not used for training the ASR models. Table 1 provides statistics on the annotated datasets.

[2] We used TIMIT and Wikipedia (`English_phonology`) for English mappings and another mapping for Arabic: `http://sites.middlebury.edu/arabiclingusitics2014/files/2014/02/class6_phonetics_1.pdf`. When classifying by place, we set the vowels as a separate group.

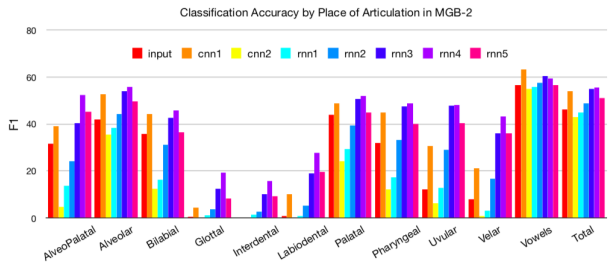


(a) *MGB-2 (Arabic).*

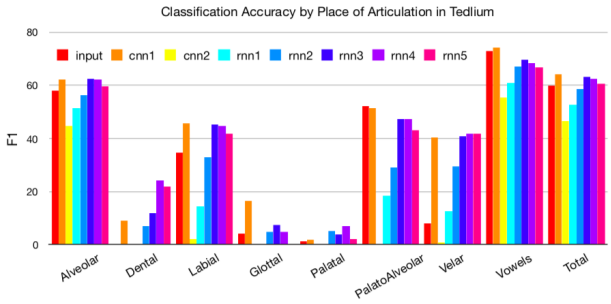


(b) *TedLIUM (English).*

Figure 2: *Classification by manner of articulation.*



(a) *MGB-2 (Arabic).*



(b) *TedLIUM (English).*

Figure 3: *Classification by place of articulation.*

5. Results

5.1. Results in Different Datasets

Figure 1 shows the result of phoneme classification in three datasets: The Arabic MGB-2 and English TedLIUM and Librispeech. The overall layer-wise trend is similar in all cases: the first convolutional layer improves representation quality above the input spectrograms, while the second convolutional layer leads to a large drop. In the LSTM layers, there is steady increase until the last layer, where performance drops. This pattern is consistent with our previous analysis of phoneme representations in DeepSpeech2 based on TIMIT classification [4].

We also compare classification using only the current frame vs. using a window of +/-7 frames around it. As Figure 1 shows, while using additional context always helps performance on the classification tasks, the layer-wise patterns do not change, consistent with [4]. Interestingly, throughout the recurrent layers the difference between using a window or not becomes smaller. For instance, on Librispeech, the difference is around 20% at layers rnn1–2, decreasing to 14–15% at layers rnn3–5; a similar pattern is found in the other datasets. This indicates that the top recurrent layers capture more context, thereby reducing the benefit from a large context at the input.

5.2. Phonetic Features

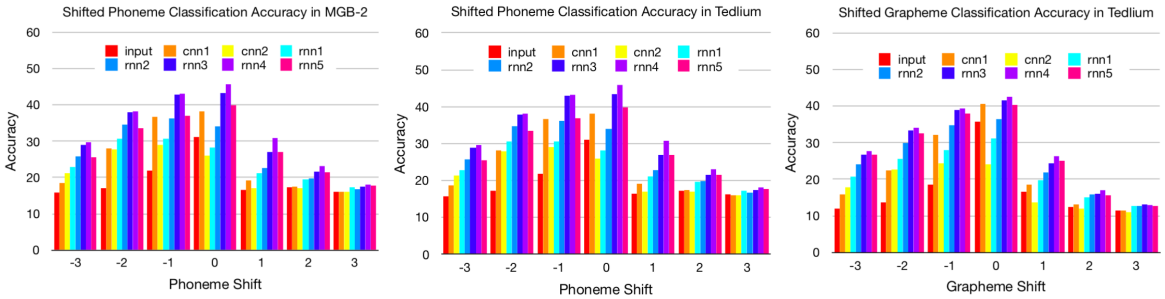
In this section, we analyze the representation quality from the perspective of articulatory features. We map each phoneme to either its place or its manner of articulation.

Figure 2 shows the manner classification results, summarized by F1 score (harmonic mean of precision and recall) for each manner of articulation. In most cases, the common layer-wise pattern recurs. Some manners are easier to classify than others: especially vowels, which are very different from consonants, and also fricatives, nasals, and stops. Affricates are more difficult, perhaps due to their composite nature. The Arabic liquid (/l/) and trill (/r/) are also hard. Comparing English and Arabic, the results are fairly consistent, as shown by the similar shape of the two sub-plots, although the labels do not entirely overlap. To test this quantitatively, Table 2 (right) shows the Pearson correlation across layers between several manners of articulation in English and Arabic. In most clases, there are high positive correlations, up to 0.97 in the case of English semivowels/glides and Arabic glides. Vowels are less correlated, which is not surprising given the limited vowel inventory in Arabic (only 3 vowels) compared to English.

Table 2: *Cross-language correlation in layer-wise classification accuracy by manner and place of articulation.*

Place			Manner		
TedLIUM	MGB-2	r	TedLIUM	MGB-2	r
Glottal	Glottal	0.16	Vowels	Vowels	0.74
Palatal	Palatal	0.68	Fricatives	Fricatives	0.85
Vowels	Vowels	0.71	Semi/Glides	Liquids	0.88
Labial	Labiodent.	0.73	Stops	Stops	0.91
Palato-Alveolar	Alveo-Palatal	0.88	Nasals	Nasals	0.93
Dental	Interdent.	0.89	Affricates	Affricates	0.93
Velar	Velar	0.91	Semi/Glides	Glides	0.97
Alveolar	Alveolar	0.93	Total	Total	0.82
Labial	Bilabial	0.98			
Total	Total	0.87			

Turning to place of articulation, Figure 3 exhibit similar layer-wise patterns in classifying each place. Some places are easier to classify: alveolar, alveo-palatal, labial, and velar consonants. Glottal and dental/interdental consonants are more difficult. Again, these results are quite consistent for the two languages, although the place labels also do not entirely overlap. Looking at the correlations (Table 2, left), several consonant groups behave very similarly in the two languages:



(a) *MGB-2, phonemes (Arabic).* (b) *TedLIUM, phonemes (English).* (c) *TedLIUM, graphemes (English).*

Figure 4: *The effect of predicting past and future phonemes and graphemes.*

labial/bilabial ($r = 0.98$), alveolar (0.93), velar (0.91), and dental/interdental (0.89). This is striking as these groups do not always overlap: English labials include /b/, /p/, /v/, /f/, and /m/, while Arabic bilabials include /b/, /m/, and /w/, yet their correlation is very high. Cases of low(er) correlations are the glottal (0.16), palatal (0.68), and labial/labiodental consonants (0.73). In the glottal case, this may be explained by Arabic having glottals /ʔ/ and /h/, while the English phoneme set only has /h/.

5.3. Phonemes vs. Graphemes

How can we explain the drop in representation quality towards the top layers of the model? One possibility is that a model that was trained on acoustics-to-characters "forgets" some of the phonetic information at the top layer(s), close to the output. To test this, we performed several grapheme classification tasks.

Figure 5 shows the results. Evidently, the layer-wise patterns are very similar to the phoneme case, although grapheme classification tends to be slightly easier. Interestingly, the gaps between grapheme and phoneme classification are somewhat larger at the top recurrent layers than in intermediate layers. This suggests that the top layers are indeed more geared towards graphemic than phonetic information. However, the drop at the top layer is still apparent and cannot be explained solely by phoneme/grapheme differences. Table 3 compares the top layer drop in in phoneme and grapheme classification. In all three datasets, the (relative) drop is smaller when predicting graphemes than phonemes. This again indicates that the top layers are more concerned with graphemes than with phonemes.

Table 3: *Relative drop from the penultimate to ultimate layer in phoneme vs. grapheme classification.*

	TedLIUM	Librispeech	MGB-2
Phonemes	6.36%	5.49%	12.94%
Graphemes	5.05%	4.35%	10.60%

5.4. Predicting the Future or Past

Another possible explanation for the drop at the top layer has to do with capturing long-distance information. Previous work [4] conjectured that "higher layers in the model are more sensitive to long distance information that is needed for the speech recognition task, whereas the local information that is needed for classifying phones is better captured in lower layers." We investigate this conjecture by predicting past or future phonemes. We simply shift the labels in the datasets by $k \in \{-3, -2, -1, 0, 1, 2, 3\}$ phonemes, and retrain the classifier. The results are shown in Figures 4a and 4b.

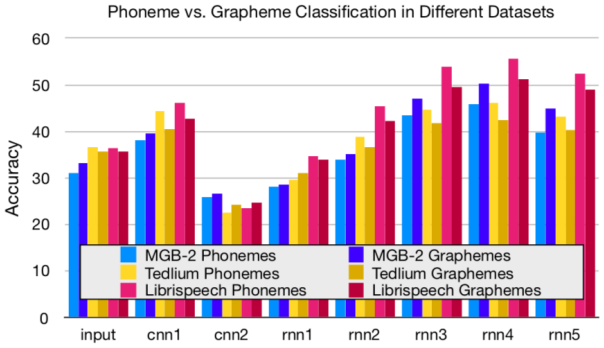


Figure 5: *Phoneme and grapheme classification accuracy.*

We find that predicting the future is much more difficult than predicting the past, as performance quickly drops when predicting even only one phoneme into the future, but only moderately degrades when predicting up to three phonemes into the past. This can be explained by the use of unidirectional LSTM layers in the models we experiment with. This holds in both languages (compare Figures 4a and 4b) and in both phoneme and grapheme classification (compare Figures 4b and 4c).

The drop in accuracy at the top layer is still apparent. In the case of Arabic phonemes (Figure 4a), this drop is more moderate when predicting future phonemes: in relative terms, we see a drop of 2–6% when predicting 2 or 3 phonemes into the future, but 14% drop when predicting into the past. In English phonemes (Figure 4b), there is only a very mild drop (1.4%) when predicting 3 phonemes into the future. Thus, the top layer may be losing less long-distance information about the future than the past. This is not always consistent, however, as there is a substantial drop (12%) when predicting 2 phonemes into the future in English. In the case of graphemes, the top layer drop is fairly consistent in all shifts: predicting 3 graphemes into the future or past results in a similar drop of around 3%.

6. Conclusion

In this work, we analyzed an E2E speech recognition model in terms of phonetic and graphemic representations. We observed consistent behavior in layer-wise quality across languages, datasets, output labels, and articulatory features. This suggests that such models may benefit from sharing information, for example using multilingual systems as in a recent E2E codeswitching ASR model [35]. In the future, we plan to extend the analysis to other E2E models, such as attentional sequence-to-sequence [19, 20] or purely convolutional models [18].

7. References

[1] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Transactions of the Association for Computational Linguistics (TACL)*, 2019.

[2] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *Interspeech*, 2015.

[3] ——, "On the role of nonlinear transformations in deep neural network acoustic models," in *Interspeech*, 2016, pp. 803–807.

[4] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems (NIPS)*, December 2017.

[5] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep Speech 2: End-to-end speech recognition in english and mandarin," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 173–182.

[6] Y.-H. Wang, C.-T. Chung, and H.-y. Lee, "Gate activation signal analysis for gated recurrent neural networks and its correlation with phoneme boundaries," in *Interspeech*, 2017.

[7] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016.

[8] J. F. Santos and T. H. Falk, "Investigating the effect of residual and highway connections in speech enhancement models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*, 2018.

[9] S. Shon, H. Tang, and J. Glass, "Frame-level speaker embeddings for text-independent speaker recognition and analysis of end-to-end model," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1007–1013.

[10] S. Wang, Y. Qian, and K. Yu, "What does the speaker embedding encode?" in *Interspeech*, 2017, pp. 1497–1501.

[11] Z. Elloumi, L. Besacier, O. Galibert, and B. Lecouteux, "Analyzing learned representations of a deep ASR performance prediction model," in *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2018, pp. 9–15.

[12] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 613–622.

[13] A. Alishahi, M. Barking, and G. Chrupała, "Encoding of phonology in a recurrent neural model of grounded speech," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017, pp. 368–378.

[14] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.

[15] J. Drexler and J. Glass, "Analysis of audio-visual features for unsupervised speech recognition," in *International Workshop on Grounding Language Understanding*, 2017.

[16] A. Krug, R. Knaebel, and S. Stober, "Neuron activation profiles for interpreting convolutional speech recognition models," in *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*, 2018.

[17] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[18] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, "Wav2Letter++: A fast open-source speech recognition system," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6460–6464.

[19] X. Wang, R. Li, S. H. Mallid, T. Hori, S. Watanabe, and H. Hermansky, "Stream attention-based multi-array end-to-end speech recognition," *arXiv preprint arXiv:1811.04903*, 2018.

[20] 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.

[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 (ICML)*, 2006, pp. 369–376.

[22] "The MGB-2 challenge: Arabic multi-dialect broadcast media recognition."

[23] A. Ali, Y. Zhang, P. Cardinal, N. Dahak, S. Vogel, and J. Glass, "A complete kaldi recipe for building arabic speech recognition systems," in *2014 IEEE spoken language technology workshop (SLT)*. IEEE, 2014, pp. 525–529.

[24] "QCRI advanced transcription system (qats) for the Arabic multi-dialect broadcast media recognition: MGB-2 challenge."

[25] P. Smit, S. R. Gangireddy, S. Enarvi, S. Virpioja, and M. Kurimo, "Aalto system for the 2017 arabic multi-genre broadcast challenge," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 338–345.

[26] Y. Adi, E. Kermany, Y. Belinkov, O. Lavi, and Y. Goldberg, "Fine-grained analysis of sentence embeddings using auxiliary prediction tasks," in *International Conference on Learning Representations (ICLR)*, 2017.

[27] D. Hupkes, S. Veldhoen, and W. Zuidema, "Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure," *Journal of Artificial Intelligence Research*, vol. 61, pp. 907–926, 2018.

[28] J.-P. Hosom, "Automatic phoneme alignment based on acoustic-phonetic modeling," in *Seventh International Conference on Spoken Language Processing*, 2002.

[29] S. Naren, "deepspeech.torch," https://github.com/SeanNaren/deepspeech.torch, 2016.

[30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[31] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[32] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus." in *LREC*, 2012.

[33] Y. Wang, X. Chen, M. J. Gales, A. Ragni, and J. H. Wong, "Phonetic and graphemic systems for multi-genre broadcast transcription," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[34] A. Ali, "Multi-dialect Arabic broadcast speech recognition," Ph.D. dissertation, The University of Edinburgh, 2018.

[35] N. Luo, D. Jiang, S. Zhao, C. Gong, W. Zou, and X. Li, "Towards end-to-end code-switching speech recognition," *arXiv preprint arXiv:1810.13091*, 2018.