Automatic Speech Recognition and Arabic Speech Corpora Survey

Fatma Zahra Besdouri[1], Ines Zribi[1],Lamia Hadrich Belguith[1]
[1]ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
fatma.basdouri@gmail.com, ineszribi@gmail.com, l.belguith@fsegs.rnu.tn

August 9, 2022

Abstract

Automatic Speech Recognition (ASR) systems rely on large amounts of annotated speech to learn accurate speech representation and recognition. Arabic includes various dialects with some annotated resources. Predominantly, the key goal of ASR is to allow natural collaboration between the end-user and the system through a series of inquiries and responses. The objective of this paper is to provide first, the basic step, the speech corpora created and second, the modern ASR systems which includes statistical approaches and deep learning approaches to have a general review on this field by highlighting challenges of the low resources' languages for instance Arabic language and its dialects for instance Tunisian dialect.

Keywords— Automatic Speech Recognition, Arabic Speech Corpora, Statistical models, End-to-End models

1 Introduction

Speech recognition is a multidisciplinary domain which comprises of statistics, signal processing, phonetics, linguistics, and machine learning. Speech recognition mainly consists of two stages: training and decoding. The process of the automatic speech recognition can be summarized as follows:

given a raw signal corresponding to an utterance, we need to identify the meaningful units in the sound stream. (Speech recognition is generally performed using one of three approaches that rely on the basic units of classification) This is the process of segmentation, which can be analyzed at multiple levels—phones, syllables, words, and collocations.

Given the variable nature of speech, which arises from different speaker characteristics, environmental conditions, and other factors, we need to find suitable abstract representations of the raw speech segments to aid generalization.

During the training, the leaning process can discover recurring patterns in the input signal and the relationships between those patterns. In speech, for example, the discovered patterns would not necessarily align with orthographically valid units like words in a dictionary. Such pairings, to obtain speech segments and clusters that are consistent with text by mapping each speech segment to its nearest neighbor in the text domain.

ASR systems are based on 2 types approaches: traditional approaches and deep approaches. Typical traditional ASR models are composed of three main components: an acoustic model, a pronunciation dictionary, and a language model. However, the end-to-end models based on an encoder–decoder architecture. And the output then, an external LM can be incorporated to improve the overall system performance.

A common and fundamental step in both of ASR types is acquiring a large amount of the training data. For the low resources' languages, formulating such large training corpora for acoustic modeling is more challenging. For instance, Arabic dialects.

The remaining paper is presented in the following manner: The subsequent section presents some speech corpora created especially for the under resourced language, for instance, Arabic language and its dialects. In Section 3, we details automatic speech recognition models that we divided them into the main 2 types of ASR models: traditional pipeline and the end-to-end models, also, we describe transfer learning. Finally, we present the conclusions in Section 5.

2 Speech corpora: Example of the Arabic language

Speech corpora fundamental and basic in Automatic Speech Recognition. In speech recognition systems, it is highly recommended and more accurate to use large collections of speech. In addition, several factors must be considered

when preparing a speech corpus, like varying acoustic environments, variations in pronunciation, dialects, accents, ages, etc.

Most of the created speech corpora are in English, including but not limited to, Librispeech. It is an ASR corpus based on public domain audio books. It is corpus freely available for download and suitable for training and evaluating speech recognition systems. The LibriSpeech corpus contains 1000 hours of speech sampled at 16 kHz.

However, for the low resource languages, it is inconveniently harder for individual researchers to achieve the task of creating a large body of speech compilations because it is an expensive and time-consuming task, particularly good quality datasets.

For instance, if we take the Arabic Language, accessibility of resources and benchmarks is challenging. On the top of that, for Arabic dialects is more challenging.

2.1 MSA Speech Resources

Beginning with Arabic versions of an international corpus, one of the earliest attempts to develop an Arabic speech corpus was made by [40] by producing OrienTel speech dataset for Arabic. The study depicts that the OrienTel dataset represents the first effort to collect speech data on an extensive scope. The participants of OrienTel collected standard and colloquial varieties of Arabic in Saudi Arabia, the UAE, Egypt, Palestine, Tunisia, and Morocco. OrienTel transcription and annotation conventions are largely based on conventions used by the Linguistic Data Consortium and ARPA). The distribution of male and female speakers is 50% each per database, the distribution of speaker age varies 16 to 60 and the distribution of recording environments varies between public places to quiet one (office or home). The corpus is publicly available through the European Language Resources Association (ELRA).

The second corpus is one of the open resources, Mozilla Common Voice. Common Voice's multi-language dataset is already the largest publicly available voice dataset of its kind. It is an audio dataset that consists of a unique MP3 and corresponding text file. The dataset consists of 7335 validated hours in 60 languages. Among them is the Arabic. The overall recorded hours are 145 hours, 87 of them are validated. The dataset also includes demographic metadata like age, sex, and accent. Indeed, the number of voices is 1,272 where 27% male and 18% female. The speakers are of all ages, mainly between 19 to 39 years old. It is publicly available voice dataset.

Thirdly, the Arabic speech data of the GlobalPhone project which produced a read speech corpus designed for the development and evaluation of large continuous speech recognition systems [36, 34]. The project provides a uniform, multilingual speech and text database. The Arabic corpus was produced using the Assabah newspaper covering national and international political news as well as economic news. It contains recordings of 78 speakers (35 males, 43 females) recorded in Tunisia, Palestine and Jordan. The age distribution is below 19, up to 50 and over and the recording environment is Office. However, It's paid resource.

Among the effort to create speaker-independent large vocabulary natural Arabic speech recognition system,[5] developed an MSA broadcast news speech recognition system. Out of 7.5 h of recorded speech the system trained on 7.0 h of matter and tested on the remaining half hour. The corpus contains 235 news items, 41 can be attributed to sports news while the remaining items examine primarily economic news. Among the items, female speakers created 88 items.

[22] presented a Holy Quran corpus of approximately 18.5 h in length. The main challenge is to develop a broadcast news corpus since the Holy Quran recordings already available. [22] indicated that it took approximately 732 working hours to build their Holy Quran corpus. This is an indication that preparing a speech corpus is a difficult task.

[1] developed an Arabic ASR system that relied on a phonetically rich and well-adjusted speech corpus. That work was based on 8043 utterances, which were gathered from eight (five male and three female) speakers and resulted in approximately 8 h of speech. The round-robin testing approach was applied.

Recently, several research studies tackled continuous Arabic speech as follows.[2] demonstrated a contribution to Arabic ASR; they used a continuous speech corpus that consists of 22.7 h to study the impact of phonological rules on Arabic speech recognition.

2.2 Dialectal Arabic Speech Resources

Specially formulating such large training corpora for acoustic modeling of local Arabic dialects is more challenging when compared to Modern Standard Arabic (MSA) [13].

[4] presented a Saudi-accent Arabic telephone speech database. It contains 96 h of speech, which were collected on a telephone network during 2002 and 2003 using 1033 native speakers (51% males, 49% females) [34].

[13] uses MSA acoustic models as multilingual models to decode Egyptian dialect. They chose the Nemlar broadcast news speech corpus for building the acoustic models. The corpus consists of 40 h of MSA news broadcasts. The total number of speakers is 259, with a lexicon of 62,000 words. The Nemlar corpus is monolingual and annotated corpus but unfree resource.

MSA Speech resources			
Corpus	Language	Hours	Speakers
Arabic Mozilla Common Voice	MSA	145	248
GlobalPhone[36]	MSA	12	74
[5] broadcast speech corpus	MSA	7.5	–
Holy Quran corpus [22]	MSA	18.5	–
[1]'s corpus	MSA	8	8
[2] 's corpus	MSA	22.7	–
Dialectal Arabic Speech resources			
Corpus	Language	Hours	Speakers
OrienTel [40]	MSA and Dialectal Arabic	–	004
[4]'s corpus	Saudi Arabic	96	1033
Nemlar broadcast speech corpus [13]	MSA and Egyptian Arabic	40	259
ALGASD [38]	Algerian Arabic	–	300
STAC[50]	Tunisian Arabic	5	plus 70
TARIC [28]	Tunisian Arabic	–	300
TunSpeech [29]	Tunisian Arabic	11	plus 10

Table 1: Arabic Speech Corpora Examples

[38] presented an MSA continuous speech corpus of 200 sentences, which are pronounced by 300 Algerian native speakers from eleven regions of Algeria. Recordings were made in quiet environments that were familiar to the speakers. The ALGASD corpus has numerous strengths: large number of speakers, high quality recordings, useful information about speakers (namely their region, initials, gender, age, and education level), and integration of regional phonetic variations.

To overcome the lack of Tunisian dialect spoken resources, [50] were created one of the most used corpora for the processing of Tunisian Dialect, it is STAC: The Spoken Tunisian Arabic Corpus. STAC corpus consists of almost 5 transcribed hours gathered from different Tunisian TV channels and radio stations. It contains spontaneous speech, less spontaneous speech and prepared speech and a big number of speakers (about 70 speakers) to make the dataset a representative sample of the Tunisian dialect.

In addition, we mention the TARIC corpus (Tunisian Arabic Railway Interaction Corpus) [28] which is a collection of audio recordings and transcriptions from dialogues in the Tunisian Railway Transport Network. The TARIC corpus was manually transcribed due to the absence of tools for automatic transcription for Tunisian Dialect. Then, a normalization step was applied to obtain coherent data using a standard orthography described in [49]. The number of statements is 18,657 and the number of words 71,684.

Finally, TunSpeech [29]which is paired text-speech dataset consisting of 11 hours. It builds from two different sources: firstly, Tunisian Parliament chamber of deputies (ARP) consists of 6 hours and 7 minutes of spontaneous conversations in a quite noisy environment. The second source is Books written in Tunisian dialect (TunBook). The dataset consists of about 4 Hours and 20 minutes of recorded books by 10 Tunisian native speakers (6 females, 4 males) in a non-noisy environment.

The table 1 presents details about the Arabic speech resources.

3 Automatic Speech Recognition models

3.1 Traditional Automatic Speech Recognition methods

The traditional speech recognition systems have relied on acoustic-phonetics knowledge. They are based on acoustic, language models and lexicon (pronunciation dictionary) as shown in the figure 1.

Beginning with the acoustic Model, it's a file that contains statistical representations of each of the various sounds that make up a word. A phoneme is a label given to each of these statistical representations. In the English language, there are approximately 40 distinct sounds that is suitable for speech recognition, resulting in 40 separate phonemes.

The AM calculates the probability of acoustic units (e.g phones, sub-word units etc.), which can be built using Gaussian Mixture Models (GMMs) [47]and Hidden Markov Models (HMMs) [23]. Typically, GMMs are used to calculate the probability distribution of phones in a single state, while HMMs are used to find the transition probability

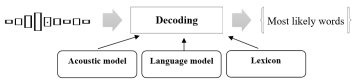

Figure 1: Traditional ASR Pipeline.

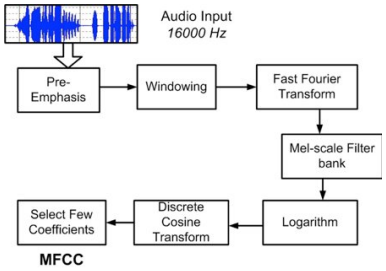

Figure 2: MFCC feature extraction. [39]

from one state to another. Each state corresponds to an acoustic event, such as a phone. The GMM-HMM model is trained by using the Expectation Maximization (EM) technique, and Viterbi decoding is used to find the optimal state sequence in HMMs.

With the advent of deep learning, the performance of ASR systems has improved. Artificial neural networks began to be used for acoustic modeling instead of GMM, leading to better results obtained in many studies [30] [37] [8]. Thus, the hybrid architecture, HMM-DNN, has become one of the most common models for continuous speech recognition. Secondly, to discriminate between words that sound similar, sounds are mapped with word sequences. We presume the audio sample is grammatically and semantically sound, even if it is not grammatically perfect or has skipped words. As a result, incorporating a language model (LM) into decoding can improve the accuracy of ASR. By combining linguistic knowledge from huge text corpora, LMs are used to increase the accuracy of acoustic models. LMs acquire implicit syntactic and semantic norms, which are subsequently utilized to re-score the acoustic model hypotheses. The LM is based on n-gram models.

The last component is the lexicon or the pronunciation dictionary which provides a mapping between a conventional symbolic transcript of speech, which can exhibit varying degrees of arbitrariness, and an acoustically / phonetically motivated one. In fact, it's used to map a sequence of phonemes into words and align the phonetic transcriptions that come from the AM with the raw text used in LMs. Those three components were trained separately, which made it difficult to manage and configure them, which can lead to a decrease in the efficiency of using them.

[27] build an ASR system for the Tunisian Dialect. It's based on an acoustic model of standard 3-states context-dependent triphone models where the GMM-HMM model has about 15K Gaussians for 2.5K tied states. The LM built using a total corpus of 100K words. The corpus is obtained from, firstly, from TARIC training data transcripts and secondly, blogs and Arabizi converted corpus. Next, the corpus normalized following the CODA recommendations [49]. They trained different 3-gram language models with modified Kneser-Ney discounting. These models were interpolated to create the final LM by minimizing the perplexity on a development corpus based on vocabulary size of 612 words. They included all development and test words in the lexicon to avoid out-of-vocabulary (OOV) words. This ASR reaches a word error rate of 22.6% on a held-out test set.

To create the [3]'s Arabic ASR system, the HMM Toolkit was used. The first step was to create an Arabic dictionary by composing the words to its phones. The speech feature vectors are then extracted using Mel Frequency Cepstral Coefficients (MFCC)(see Figure 2) of the speech samples. The system's training is then developed using triphones to estimate the parameters for a HMM. The experimental results showed that the overall system performance was 90.62%, 98.01% and 97.99% for sentence correction, word correction and word accuracy respectively [32] developed both hybrid and end-to-end ASR approaches exploring some techniques to improve the performance of Speech-to-Text tasks under IberSpeechRTVE 2020 Challenge. By using multi-condition data augmentation, we demonstrated that Hybrid DNN-HMMs may be adapted to the TV show domain. The WER was dramatically reduced when reverberated data was added to the training data (10% relative). In better conditions, a WER of 7.96% was recorded.

Figure 1 displays a block diagram of Traditional ASR.

3.2 End-to-End models

Modern ASR systems are fully end-to-end (E2E). Currently, end-to-end deep learning approaches have surpassed previously predominant solutions based on Hidden Markov Models. The E2E structure presents the system as a single neural network, unlike the traditional one, which has several independent elements [45][31].

The E2E system provides direct reflection of acoustic signals in the sequence of labels without intermediate states, without the need to perform subsequent processing at the output, which makes it easy to implement. E2E ASR models simplify the hybrid DNN/HMM ASR models by replacing the acoustic, pronunciation and language models with one single deep neural network, and thus transcribe speech to text directly.

Modern ASR systems are fully E2E. E2E deep learning presents the exciting opportunity to improve speech recognition systems continually with increases in data and computation. For instance, [7] describes an encoder–decoder architecture, where the input audio is processed using a cascade of convolutional layers to produce a compact vector. The decoder then takes the encoded vector as input and generates a sequence of characters. The output labels of the E2E system can be characters or sub-word units such as byte-pair encoding (BPE). An external LM can be incorporated to improve the overall system performance.

To optimize E2E ASR, several different objective functions can be used. For instance, Connectionist Temporal Classification (CTC) [14]. In recent work, [41] train a CTC-based model with word output targets, which was shown to outperform a state-of-the-art context-dependent-phoneme baseline on a YouTube video captioning task.

When Deep Speech [17] was proposed, the model end-to-end is trained using the CTC loss function [14], which allows the network to directly predict the sequences of characters from input audio. As result, CTC has shown promising results in Deep Speech [17]. For instance, we cite E2E deep learning approach used to recognize English and Mandarin Chinese speech [6]. Also, [29] proposed a Tunisian Dialectal End-to-end Speech Recognition based on Deep Speech, where after prediction, a CTC loss is computed to measure the prediction error. The best performances achieved a WER of 24.4%.

Moreover,[16] presented an E2E training of acoustic models using the lattice-free maximum mutual information (LF-MMI) objective function in the context of hidden Markov models. Comparing to the CTC in character-based and lexicon-free settings, their method showed 5% to 25% relative reduction in word error rates on different large vocabulary tasks while using significantly smaller models.

Moving to one of the most used objective functions, we mention the sequence-to-sequence (seq2seq) models which have been gaining in popularity in the automatic speech recognition (ASR) community [10]. The proposed encoder architecture presented in [18], is a fully convolutional seq2seq with a simple and efficient decoder. Key to the success of the proposed convolutional encoder is a Time-Depth Separable block structure, another neural network architecture, which allows the model to retain a large receptive field. The TDS model benefits more from the integration of an external strong convolutional LM. As a result, this neural convolution architecture improves by more than 22% relative WER over a strong RNN baseline on LibriSpeech.

To overcome seq2seq training in low resource conditions, recent work has shown that this problem can be addressed using unpaired data.

On the one hand, several works use only unpaired speech for unsupervised training. For instance, [42] proposed an E2E differentiable loss integrating ASR models and Text-To-Speech (TTS) as well by the straight-through estimator. Besides, the work in [21] used the same method and proposed an E2E differentiable loss integrating ASR and a Text-to-Encoder (TTE) model. Both works, showed that connecting ASR with TTS/TTE can handle unpaired speech data as well as reduce ASR recognition errors. On the other hand, other works concern methods that leverage unpaired text data, which focus on enhancing the decoder component of seq2seq ASR [35] or moving the encoder representation closer to text representation [19].

Besides seq2seq models, we find also objective functions like ASG [12], Differentiable decoding [11] and Transduction [33]

Moreover, different architectures of neural networks such the Transformer [43] models have been explored. Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution.

In previous studies, it was revealed that the combined use of Transformer models and an E2E model, like CTC, contributed to the improvement of the quality of the English and Chinese speech recognition system.

[48] have developed a framework for multilingual speech recognition based on low-resource languages.

This work used sequence-to-sequence based and attention-based models and a single transformer for recognition. The multilingual modelling unit used is Subwords, without the need of pronunciation lexicon. The ASR transformer represents the main architecture of this model. They employed a language specific softmax layer instead of a softmax layer to solve the problem of a few training data on low resource languages. With experiments on six languages from the CALLHOME corpora (Arabic, Mandarin, Japanese, English, German, and Spanish). The result of the Arabic dataset was 13.5% average WER.

To realize a faster and more accurate ASR system, [24] combined Transformer and the advances in RNN-based

Figure 3: Modern ASR Pipeline

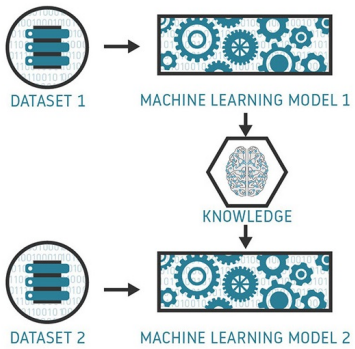

Figure 4: This frog was uploaded via the file-tree menu.

ASR. They integrated connectionist temporal classification (CTC) with Transformer for joint training and decoding. This approach makes training faster than with RNNs and assists Language Model integration.

Also, for the Mandarin Chinese language, [30] proposed the Transformer-based online E2E ASR model, which consists of, first, the state reuse chunk- Self-Attention Encoder (SAE) which splits a speech into non-overlapping isolated chunks of Nc central length, and second, the Monotonic Truncated Attention (MTA) based Self-Attention Decoder. Then, they integrated the proposed Transformer-based online E2E ASR model into the CTC/attention ASR architecture.

[15] proposed the convolution-augmented transformer for speech recognition, named Conformer. In this way, they tried to combine convolution neural networks and transformers to model both local and global dependencies of an audio sequence in a parameter-efficient way.

The end-to-end system's output labels can be characters or sub-word units. To enhance the functionality of the system, an external LM can be integrated. Figure 3 illustrates the end-to-end ASR pipeline.

Annotated speech data of sufficient quantity and quality to train end-to-end speech recognizers is scarce for most languages, low resourced specifically.

Nevertheless, there is demand for high-quality ASR systems for these languages. Dealing with this issue requires specialized methods.

On some problems where you may not have very much data, transfer learning can enable you to develop skillful models that you simply could not develop in the absence of transfer learning.

3.3 Transfer Learning

Machine learning techniques have been widely exploited in modern research on speech and language processing. Transfer learning is among the most interesting. This machine learning technique is used to enhance a model's performance in a data-scarce domain by cross-training on data from other domains or tasks.

Indeed, transfer learning involves all methods that use any auxiliary resources (data, model, labels, etc.) to improve model learning for the target task where a model developed for a task is reused as the starting point for a model in another task as shown in the figure 4.

Transfer learning comes in a variety of forms. The predominant one being applied to ASR is *heterogeneous transfer learning* [46] which implies training a base model on several languages (and tasks) simultaneously. While this achieves some competitive results [9, 25] it still requires large amounts of data to yield robust improvements [20].

A much more promising type of transfer learning is called *model adaptation* [46] is much more promising in terms of how much data is required for efficient retraining. This method involves firstly training a model on one (or more) languages, then retraining it entirely or in part on a different language that wasn't seen in the first training. In a manner similar to pre-training, the first language's parameters serve as the starting point.

[44] applied this method by initially training a multilayer perceptron (MLP) on a variety of languages with relatively ample data, such English, before achieving competitive performance on less data-rich languages like Czech and Vietnamese.

Based on model adaptation, ASR system proposed by [26] explored transfer learning as an approach for training ASR models under constrained GPU memory, throughput, and training data. They carried out a number of systematic experiments to adapte an English-trained Wav2Letter convolutional neural network into the German language. The cost of training ASR models in other languages is reduced since they demonstrate how this method enables quicker training on consumer-grade resources while using less training data to attain the same accuracy.

With its promising methods and efficient results, Transfer learning is popular in deep learning given the enormous resources required to train deep learning models or the large and challenging datasets on which deep learning models are trained.

4 Conclusion

We reviewed various research efforts in the direction of automatic speech recognition. In this section, we summarize the main takeaways from this review and outline some of the challenges. First, we reviewed the speech corpora, especially for under resourced languages for instance, Arabic language and its dialects. This step is fundamental requirement and challenge in ASR field. Then, we moved the ASR types: the traditional ASR and the End-to-end ASR. We detailed their specific components and presented some studies followed each method. Then, we highlighted the Transfer learning method, witch has become highly popular and promising given the enormous resources required to train Deep Learning models, especially in the areas of ASR as well as Natural Language Processing (NLP). In further work, we aim to create a speech corpus for the Tunisian Dialect which is an under resourced language by .

References

[1] M. A. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa. Phonetically rich and balanced text and speech corpora for arabic language. *Language resources and evaluation*, 46(4):601–634, 2012.

[2] F. S. Al-Anzi and D. AbuZeina. The impact of phonological rules on arabic speech recognition. *International Journal of Speech Technology*, 20(3):715–723, 2017.

[3] B. A. Al-Qatab and R. N. Ainon. Arabic speech recognition using hidden markov model toolkit (htk). In *2010 international symposium on information technology*, volume 2, pages 557–562. IEEE, 2010.

[4] M. Alghamdi, F. Alhargan, M. Alkanhal, A. Alkhairy, M. Eldesouki, and A. Alenazi. Saudi accented arabic voice bank. *Journal of King Saud University-Computer and Information Sciences*, 20:45–64, 2008.

[5] M. Alghamdi, M. Elshafei, and H. Al-Muhtaseb. Arabic broadcast news transcription system. *International Journal of Speech Technology*, 10:183–195, 12 2007.

[6] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[7] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, E. Elsen, J. H. Engel, L. Fan, C. Fougner, T. Han, A. Y. Hannun, B. Jun, P. LeGresley, L. Lin, S. Narang, A. Y. Ng, S. Ozair, R. Prenger, J. Raiman, S. Satheesh, D. Seetapun, S. Sengupta, Y. Wang, Z. Wang, C. Wang, B. Xiao, D. Yogatama, J. Zhan, and Z. Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595, 2015.

[8] H. A. Bourlard and N. Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012.

[9] D. Chen and B. K.-W. Mak. Multitask learning of deep neural networks for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1172–1183, 2015.

[10] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[11] R. Collobert, A. Hannun, and G. Synnaeve. A fully differentiable beam search decoder. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1341–1350. PMLR, 09–15 Jun 2019.

[12] R. Collobert, C. Puhrsch, and G. Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *CoRR*, abs/1609.03193, 2016.

[13] M. Elmahdy, R. Gruhn, W. Minker, and S. Abdennadher. Modern standard arabic based multilingual approach for dialectal arabic speech recognition. In *2009 Eighth International Symposium on Natural Language Processing*, pages 169–174. IEEE, 2009.

[14] 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 '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery.

[15] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang. Conformer: Convolution-augmented transformer for speech recognition, 2020.

[16] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur. End-to-end speech recognition using lattice-free mmi. In *Interspeech*, pages 12–16, 2018.

[17] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.

[18] A. Y. Hannun, A. Lee, Q. Xu, and R. Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. *CoRR*, abs/1904.02619, 2019.

[19] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. F. Astudillo, and K. Takeda. Back-translation-style data augmentation for end-to-end ASR. *CoRR*, abs/1807.10893, 2018.

[20] 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*, pages 8619–8623. IEEE, 2013.

[21] T. Hori, R. F. Astudillo, T. Hayashi, Y. Zhang, S. Watanabe, and J. L. Roux. Cycle-consistency training for end-to-end speech recognition. *CoRR*, abs/1811.01690, 2018.

[22] H. Hyassat and R. Abu Zitar. Arabic speech recognition using sphinx engine. *International Journal of Speech Technology*, 9(3):133–150, 2006.

[23] B. H. Juang and L. R. Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[24] S. Karita, N. Yalta, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration. pages 1408–1412, 09 2019.

[25] K. Knill, M. Gales, A. Ragni, and S. P. Rath. Language independent and unsupervised acoustic models for speech recognition and keyword spotting. 2014.

[26] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober. Transfer learning for speech recognition on a budget. *arXiv preprint arXiv:1706.00290*, 2017.

[27] A. Masmoudi, F. Bougares, M. Ellouze, Y. Estève, and L. Belguith. Automatic speech recognition system for tunisian dialect. 52, 2018.

[28] A. Masmoudi, M. E. Khemekhem, Y. Esteve, L. H. Belguith, and N. Habash. A corpus and phonetic dictionary for tunisian arabic speech recognition. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 306–310, 2014.

[29] A. Messaoudi, H. Haddad, C. Fourati, M. B. Hmida, A. B. Elhaj Mabrouk, and M. Graiet. Tunisian dialectal end-to-end speech recognition based on deepspeech. *Procedia Computer Science*, 189:183–190, 2021. AI in Computational Linguistics.

[30] H. Miao, G. Cheng, C. Gao, P. Zhang, and Y. Yan. Transformer-based online ctc/attention end-to-end speech recognition architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6084–6088. IEEE, 2020.

[31] M. Orken and D. Oralbekova. Modern trends in the development of speech recognition systems. *PHYSICO-MATHEMATICAL SERIES*, 4:42–51, 08 2020.

[32] J. M. Perero-Codosero, F. M. Espinoza-Cuadros, and L. A. H. Gómez. Sigma-upm asr systems for the iberspeech-rtve 2020 speech-to-text transcription challenge. In *IberSPEECH*, 2021.

[33] R. Prabhavalkar, K. Rao, T. Sainath, B. Li, L. Johnson, and N. Jaitly. A comparison of sequence-to-sequence models for speech recognition. 2017.

[34] L. R. Rabiner, R. W. Schafer, et al. Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, 1(1–2):1–194, 2007.

[35] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe. Multi-modal data augmentation for end-to-end asr. pages 2394–2398, 09 2018.

[36] T. Schultz. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*, 2002.

[37] F. Seide, G. Li, and D. Yu. Conversational speech transcription using context-dependent deep neural networks. In *Twelfth annual conference of the international speech communication association*, 2011.

[38] S. A. Selouani and M. Boudraa. Algerian arabic speech database (algasd): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, 35(2):157–166, 2010.

[39] A. Shaukat, H. Ali, U. Akram, et al. Automatic urdu speech recognition using hidden markov model. In *2016 International Conference on Image, Vision and Computing (ICIVC)*, pages 135–139. IEEE, 2016.

[40] R. Siemund, B. Heuft, K. Choukri, O. Emam, E. Maragoudakis, H. Tropf, O. Gedge, S. Shammass, A. Moreno, A. N. Rodriguez, et al. Orientel-arabic speech resources for the it market. In *LREC 2002 (Arabic workshop)*, 2002.

[41] H. Soltau, H. Liao, and H. Sak. Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition. *CoRR*, abs/1610.09975, 2016.

[42] A. Tjandra, S. Sakti, and S. Nakamura. Machine speech chain with one-shot speaker adaptation. *CoRR*, abs/1803.10525, 2018.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[44] N. T. Vu and T. Schultz. Multilingual multilayer perceptron for rapid language adaptation between and across language families. In *Interspeech*, pages 515–519, 2013.

[45] D. Wang, X. Wang, and S. Lv. An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8), 2019.

[46] D. Wang and T. F. Zheng. Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1225–1237, 2015.

[47] Y. Zhang, M. Alder, and R. Togneri. Using gaussian mixture modeling in speech recognition. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–613. IEEE, 1994.

[48] S. Zhou, S. Xu, and B. Xu. Multilingual end-to-end speech recognition with a single transformer on low-resource languages. *ArXiv*, abs/1806.05059, 2018.

[49] I. Zribi, R. Boujelbane, A. Masmoudi, M. E. Khemekhem, L. H. Belguith, and N. Habash. A conventional orthography for tunisian arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2355–2361, 2014.

[50] I. Zribi, M. Ellouze, L. H. Belguith, and P. Blache. Spoken tunisian arabic corpus "stac": transcription and annotation. *Research in computing science*, 90:123–135, 2015.