DOI: 10.2478/seeur-2020-0019

AUTOMATIC SPEECH RECOGNITION: A COMPREHENSIVE SURVEY

PhD. Candidate Amarildo Rista; Prof. Dr. Arbana Kadriu
Faculty of Contemporary Sciences and Technologies,
South East European University, Tetovo, North Macedonia
ar29102@seeu.edu.mk ; a.kadriu@seeu.edu.mk

ABSTRACT

Speech recognition is an interdisciplinary subfield of natural language processing (NLP) that facilitates the recognition and translation of spoken language into text by machine. Speech recognition plays an important role in digital transformation. It is widely used in different areas such as education, industry, and healthcare and has recently been used in many Internet of Things and Machine Learning applications. The process of speech recognition is one of the most difficult processes in computer science. Despite numerous searches in this domain, an optimal method for speech recognition has not yet been found. This is due to the fact that there are many attributes that characterize natural languages and every language has its particular highlights. The aim of this research is to provide a comprehensive understanding of the various techniques within the domain of Speech Recognition through a systematic literature review of the existing work. We will introduce the most significant and relevant techniques that may provide some directions in the future research.

Key words: Automatic Speech Recognition (ASR), End-to-End Systems, Hybrid Systems, Low Resource Language.

INTRODUCTION

Speech recognition is an interdisciplinary subfield of natural language processing (NLP) that enables the recognition and translation of spoken language into text by machine (Yu and Deng, 2016). It is considered as an important bridge in fostering better human–human and human–machine communication. The architecture of Automatic Speech Recognition (ASR) system (Yu and Deng, 2016) is composed by four main components: Signal processing and feature extraction, acoustic model (AM), language model (LM), and hypothesis search. The signal processing and feature extraction component takes as input the audio signal and converts it from time-domain to frequency-domain suitable for the acoustic models. The acoustic model takes as input the features generated from the feature extraction component, and generates an AM score for the variable-length feature sequence. The language model encodes prior information about the words that are likely to be spoken and takes the form of frequency distributions for types of words that indirectly encode syntax, semantics and pragmatics. On the other hand, the hypothesis search includes a combination of the AM, LM, and lexicon to decode the signal. The decoding process entails a search through alternative transcriptions of the signal in order to locate the most likely transcription. The major issue of the ASR system is to adjust to the variability of the speech signal (Juang, 1991). The issue arises due to linguistic, speaker and channel variability, which includes various attributes such as: phonetics, adverse environment conditions (clean, noisy); speaker attributes such as: age, gender, accents, speed of utterance, dialects; training process and voice recording device. An efficient ASR system must be able to identify all such types of factors to produce the text corresponding to input signal. In addition, the corpus, which requires special focus, plays an important role in the success of ASR systems. In this paper, a systematic literature review of the existing literature has been presented with the aim of understanding various techniques within the domain of Speech Recognition, which may provide some direction in the future research.

METHODOLOGY

This paper provides a systematic literature review procedure as described by Kitchenham et al. (2009) and Brereton et al. (2007) on approaches and techniques of Speech Recognition. This study introduces the most significant and relevant techniques about speech recognition that may provide some direction of how the architecture of this system should be designed. Because the investigated subject is very extensive, it is impossible to describe all relevant topics. Hence, some related subjects will be briefly mentioned. Based on the search carried out through Google Scholar, IEEE Xplore, ACM, Springer and Elsevier database, we have selected the most relevant and cited articles that describe these technologies. Referring to the keywords, title and content of articles, we have divided them into three sections: Hybrid Speech Recognition Systems; End-to-End Speech Recognition Systems and Speech Recognition Systems for Low Resource Languages. Referring to the advantages and disadvantages of each technology, we give some recommendations for future research.

SPEECH RECOGNITION SYSTEM

A. HYBRID SPEECH RECOGNITION SYSTEMS

A hybrid speech recognition system consists of the combination of Hidden Markov Models (HMMs), Gaussian Mixture Model (GMMs) with different types Neural Networks (NNs) including Deep Learning in order to take the advantages of each technique to improve speech recognition. The combination of deep neural networks DNNs and HMMs is one of the paradigms for ASR that takes advantage of DNN’s strong representation learning power and HMM’s sequential modeling ability (Yu and Deng,2015). This paradigm is called DNN-HMM hybrid system. In this paradigm, the dynamics of the speech signal is modeled with HMMs and the observation probabilities are estimated through DNNs. Each output neuron of the DNN is trained to estimate the posterior probability of continuous density HMMs’ state given the acoustic observations. Fig.1 shows the architecture of this framework.

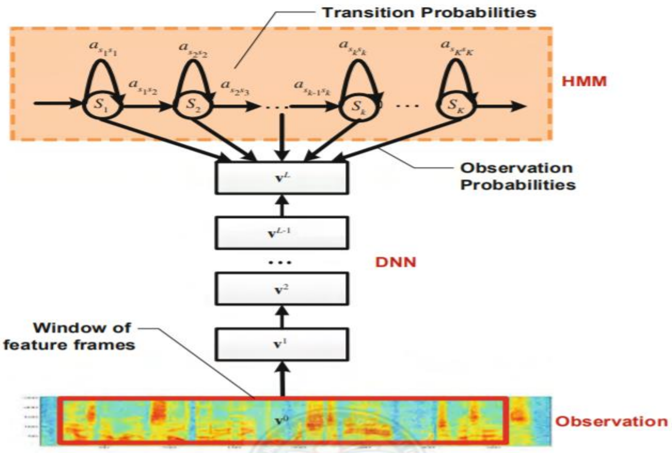

Figure 1. Architecture of the DNN-HMM hybrid system (Yu and Deng, 2015)

The HMM designs the sequential property of the speech signal, and the DNN designs the scaled observation likelihood of all the senones. This framework has two main advantages: training can be performed using the Viterbi algorithm, which finds the path that maximizes the possibility of observation between the values of HMM expressed in matrix form and decoding is generally too efficient. Shahin et al. (2014) performed a comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques, where it turns out that the hybrid DNN-HMM outperforms the conventional GMM-HMM for all experiments on both normal and disordered speech. The DNN-HMM represents accuracy above 85% when used with disordered speech. Another comparison between hybrid GMM-HMM model and DNN-HMM for the GlobalPhone (GP) multilingual text and speech database is presented by Tachbelie et al. (2020). Based on experiment results, the hybrid DNN-HMM frameworks outperform the GMM-HMM based frameworks regardless of the size of the training speech used. Overall, the achieved relative improvement ranges from 7.14% to 59.43%. Kadyan and Kaur (2020) presented a hybrid SGMM-HMM framework for acoustic modeling, which is compared with baseline GMM-HMM techniques. Fig.2 shows the architecture of the SGMM-HMM framework.

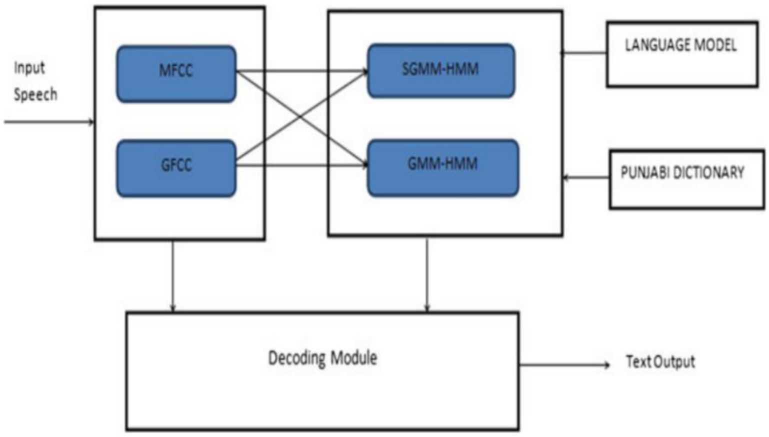

Figure 2. Workflow of SGMM-HMM-based Punjabi ASR system (Kadyan and Kaur, 2020)

First, GFCC and MFCC feature vectors (Wu and Cao, 2005) are extracted; then, these features are used to train on context-dependent (triphone) and context-independent (monophone) state with the help of provided acoustic observations. Initially, monophone state is trained to train the triphone state that helps in training the GMM-HMM model. This information state is further employed by the SGMM- HMM approach. The proposed system is evaluated on medium vocabulary Punjabi language. The clean data is initially used for training and testing of the proposed system with two hybrid classifiers SGMM-HMM and GMM-HMM. The system testing is performed with the help of two front-end approaches (MFCC and GFCC). Based on the obtained experiment results, the SGMM-HMM framework generates an improvement of 3–4% over the GMM-HMM approach. Another framework in this domain is presented by Dahl (2011) called Context-Dependent Pre-Trained Deep Neural Networks (CD- DNN-HMM). This framework is considered an improvement of CD-GMM-HMM hybrid system (Li and Deng, 2014). It is a robust and often helpful way to initialize deep neural networks generatively that can aid in optimization and reduce generalization errors. This framework is used for large-vocabulary speech recognition (LVSR) and offers better results compared to CD-GMM-HMM in the aspect of recognition accuracy, but training CD-DNN-HMMs is quite expensive compared to training CD-GMM- HMMs. This is mainly because the CD-DNN-HMM 'training algorithms are not easy to parallelize across computers and need to be carried out on a single GPU machine. CD-DNN-HMMs can be trained using the embedded Viterbi algorithm. One of the problems with CD-DNN-HMM and DNN-HMM systems is that they work with a limited vocabulary in the presence of noises. Tsunoo et al. (2019) proposed a method of end-to-end (E2E) adaptation, which adjusts the acoustic model (AM) and weighted finite-state transducer (WFST) (Allauzen et al., 2007). This framework converts a pre-trained WFST to a trainable neural network and adapts the system to the target environments/vocabulary by E2E joint retraining with the pre-trained AM. The training is done by combining the Viterbi decoding with a neural network module, based on a WFST as a joint E2E system that is to be optimized. Each AM posterior is mapped to WFST states associated with the posterior by a sparse affine transformation. Although this framework simplifies training and decoding pipelines, a unified model is hard to adapt when mismatch exists between training and test data. To solve this issue, Chen et al. (2019) proposed a class-based language model (CLM) (Kneser and Ney, 1993) that populates the classes with context-dependent information in real-time as well as a token passing decoder with efficient token recombination for E2E systems (Hall et al., 2015). Wang et al. (2020), proposed a low-latency end-to-end speech recognition framework, which consists of a Scout Network and a recognition network. The Scout Network can be implemented using any type of neural networks and aims to perform word boundary detection. The recognition network predicts the next sub word by utilizing the information from all the frames before the predicted boundary. In large vocabulary continuous speech recognition (LVCSR), it is difficult to construct language models that can capture the longer context information of words and ensure generalization and adaptation ability. To solve this issue, Sun and Chol (2020), proposed an approach that combines longer context information of recurrent neural networks (RNN) with adaptation ability of subspace Gaussian mixture model (SGMM). This approach is based on Tandem system (Hermansky et al., 2000). The Tandem system augments the input to a GMM-HMM system with features derived from the suitably transformed output of one or more neural networks. The Tandem features are typically trained to produce distributions over monophone targets to control the dimension of the augmented feature. SGMMs represent compactly featured space using few parameters and train sufficiently using small amounts of training data. To obtain the history feature vectors of a word with longer context information Recurrent Neural Networks (RNNs) are used. RNNs do not limit the size of the context. The context information can be reserved inside the network by RNN connections and the hidden states of RNNs depend on the entire input history. This approach exploits longer context information between words and allows task adaptation to a specific domain. Xu et al. (2019) proposed another DNN-HMM Hybrid System. This approach presents a semi-supervised training (SST) method based on features derived from a bottleneck hidden layer (Yu and Seltzer, 2011), which has a smaller size than that of other layers, instead of using the neural network outputs directly. The bottleneck layer creates a constriction in the network and forces the information pertinent to classification into a low-dimensional representation. This provides flexibility in choosing the training targets and the size of the augmented feature. Tanaka et al. (2019) presented a joint end-to-end and deep neural network hidden Markov model (DNN-HMM) hybrid automatic speech recognition (ASR) systems that share network components. In this approach, continuous vectors extracted by the part of DNN acoustic model are utilized as auxiliary features for the end-to-end ASR system, in order to improve the accuracy of ASR. Figure 3 shows the ASR procedure of joint end-to-end and DNN-HMM hybrid ASR systems. The ASR score of joint systems is calculated by linear interpolation with log probabilities calculated from the DNN-HMM hybrid system and end-to-end ASR system. The encoder-decoder models utilize internal outputs of DNN acoustic models, with bottleneck features being some of them (Yu and Seltzer, 2011).

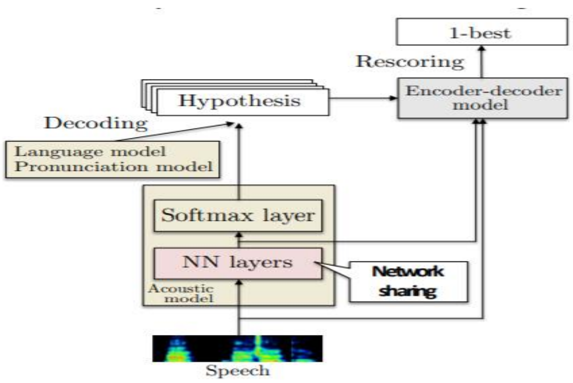

Figure 3. ASR procedure of joint end-to-end and DNN-HMM hybrid ASR systems (Tanaka et al., 2019)

The internal layer constructs a compressed continuous representation of the task-related information through the bottleneck. End-to-end system is trained with 40 mel-scale filter-bank features, delta and delta-delta features and the AdaDelta algorithm is used (Zeiler, 2012), to optimize the model parameters. As an acoustic model the convolutional neural network (CNN)-LSTM is used (Tsoi, 2006) in the DNN-HMM hybrid ASR system. The LSTM output is fed from the softmax layer. While the DNN-HMM hybrid ASR system includes a WFST based decoder (Hori et al., 2007). In the joint DNN-HMM hybrid ASR systems, the hyper- parameters of end-to-end and DNN-HMM hybrid systems are the same as shown in fig 3. The internal outputs of LSTM in the acoustic model are used as the input features of the encoder-decoder network. Based on experiment results, it is concluded that the shared network helps the end-to-end ASR system to predict characters more accurately.

B. END-TO-END SPEECH RECOGNITION SYSTEMS

A highly cited study in this domain has been conducted by Chiu et al. (2018). In this study, a speech-recognition with sequence-to-sequence framework is proposed. This framework is based on Listen, Attend, and Spell (LAS) model (Chan et al., 2016), which is a neural network that learns to transcribe speech utterances to characters. This model learns all the components of a speech recognizer jointly. Fig. 4 shows the components of this framework.

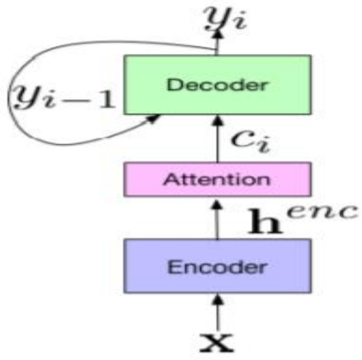

Figure 4. Components of the LAS end-to-end model (Chiu et al.,2018)

The encoder takes the input features and maps them to a higher-level feature representation. The attender determines which encoder features should be attended to in order to predict the next output symbol. Moreover, the speller produces a probability distribution over the current sub-word unit. To improve performance, this framework uses the sub-word units as a word-piece model, ranging from graphemes all the way up to entire words (Schuster and Nakajima, 2012). The word pieces are “position-dependent”, and are segmented deterministically and independently of context using a greedy algorithm (Wu et al., 2016). To reduce noise the multi-head attention (MHA) approach is used (Vaswani et al., 2017), which includes multiple heads, where each head can generate a different attention distribution. Despite the accuracy and performance that this framework has, overfitting remains one of the main challenges. Nguyen et al. (2020), proposed a solution that overcomes this problem. This solution consists of increasing the amount of available training data and the variety exhibited by training data with the help of data augmentation (Park et al., 2019). This approach is applied directly to the feature inputs of a neural network and consists of warping the features, masking blocks of frequency channels, and masking blocks of time steps. Authors have proposed two data augmentation methods. In the first model, before the LSTM layers in the encoder, a two-layer Convolutional Neural Network (CNN) with 32 channels is placed. While in the decoder, two layers of unidirectional LSTMs as language modeling for the sequence of sub-word units and the approach of Scaled Dot-Product (SDP) Attention are adopted (Vaswani et al., 2017), to generate context vectors from the hidden states of the two LSTM networks. Also, a simpler recurrent function in the decoder and an attention module are adopted (Weng et al., 2018). The adopted attention function learns an additional linear transformation for each input parameter and uses the multi-head mechanism together with Dropout and Layer-Norm for efficiently learning content-based attention (Vaswani et al., 2017). In the second model, LSTM is replaced with deep self-attention layers in both the encoder and decoder (Pham et al., 2019); however, in order to target the sequences, the BPE sub-word unit is used (Xu et al., 2019). The Deep Self-Attention framework is composed by an encoder, which uses the source sequence as input and then generates a high-level representation, and a decoder, which generates the target sequence. The decoder models the data as a language model. Both the encoder and decoder contain self-attentional sub-layers coupled with feed-forward neural networks and require neural components that should be able to learn the relationship between the time steps in the input and output sequence. To adapt the encoder to long speech utterances, the consecutive frames are grouped into one step (Sperber et al., 2018). After this, the input features are combined with sinusoidal positional encoding (Vaswani et al., 2017). This framework compared to previously proposed RNNs and CNNs networks has the advantage of parallelizing the layers over both mini-batch and time dimensions of the input and provides depth configurations to be trainable, by increasing performance. Liu et al. (2020), proposed another sequence-to-sequence recognition system, which is based on word embedding regularization (WER) approach (Unanue et al..2019) and fused decoding (Toshniwal et al..2018). It allows the decoder to consider the semantic consistency during decoding by absorbing the information carried by the transformed decoder feature. The pre-trained word embedding can serve as an additional target for sequence-to-sequence ASR regularization whereas, fused decoding mechanism utilizes the word embedding during decoding. This framework can significantly reduce ASR recognition error and has a negligible cost. Recently, multitask and transfer learning have found a widespread use in this domain. Multitask learning (Caruana,1997) is a machine learning technique that aims at improving the generalization performance of a learning task by jointly learning multiple-related tasks whereas, transfer learning (Pan and Yang,2010) aims at developing a reasonably performed system for a new task by retaining and leveraging the knowledge learned from one or more similar tasks. Inaguma et al. (2019) proposed a framework that uses an external language model (LM) under the transfer learning. First, a language-independent ASR system is built over an attention-based sequence-to-sequence framework (Bahdanau et al.,2015), which can learn soft alignments between input and output sequences of variable lengths. In order to effectively incorporate linguistic context of the target language LM fusion transfer is performed, where an external LM is integrated into the decoder network (Bahdanau et al.,2015). Authors have proposed three type of LM fusion: Shallow fusion (Zeyer et al.,2018) which uses the external LM only in the inference stage; Cold fusion (Sriram et al.,2018) uses the pre-trained LM during training of the S2S model to provide effective linguistic context and Deep fusion (Gulcehre et al.,2015) is used only for fine-tuning the gating part after parameters of both the pre-trained S2Smodel and RNNLM are frozen. This framework drastically reduces the performance gap from the hybrid systems. Denisov and Thang (2019), proposed a framework based on end-to-end automatic speech recognition (ASR) using Speaker Embedding and Transfer Learning. This framework is composed by two separate neural network models, the speaker embedding and the end-to-end ASR. The End-to-end ASR is based on a hybrid CTC/attention architecture (Watanabe et al.,2017), which takes as input the acoustic features of overlapped speech with speaker embedding vector of the target speaker and generates transcription of the target speaker's speech. This is used to update the parameters of end-to-end ASR model during the training or provided as the final output during the decoding. Whereas, speaker embedding is a fix dimensionality vector that represents speaker's features and is extracted from a reference recording of speaker's speech (Jia et al.,2018). Transfer (Seki et al.,2018) and Multi-condition training (Heigold et al.,2013) approaches are proposed to improve overlapped speech recognition. Based on experiment, it is concluded that the application of transfer learning plays a crucial role in the increasing number of speakers, as well as, significantly improves the performance of the system. Kubo and Bacchiani (2020), proposed an end-to-end speech recognition framework that uses multitask learning to improve generalization of the model by leveraging information from multiple labels. The multi-task learning model consists in simultaneous signal-to-grapheme and signal-to-phoneme conversions by sharing the encoder parameters. Differently from conventional multi-task learning systems, where the phonemes and grapheme sequences are independent, the proposed approach consists by a joint model based on iterative refinement approach (Lee et al.,2018), where dependency modeling is achieved by a multi-pass strategy. The iterative refinement approach going beyond multi-task models by using noisy hypotheses as a latent variable. Fig. 5 shows the block diagram of multi-sequence iterative refinement network.

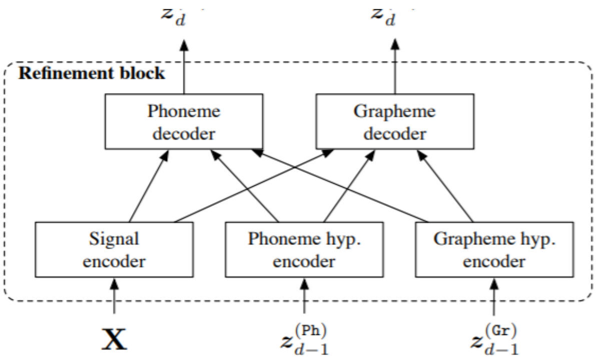

Figure 5. Block diagram of multi-sequence iterative refinement network (Kubo and Bacchiani,2020)

The encoder-decoder architecture has three encoders for handling three kinds of inputs: the acoustic feature sequence Xn and the latent variables Z for phoneme, the grapheme sequences and the auto-regressive recurrent neural net (RNN) decoders (Lee et al.,2018). The training of iterative refinement models is performed with supervision over the latent variable, because of the latent variables that are designed to be hypotheses of a phoneme and grapheme sequence, for which label information is supervised. With this supervised label information, every refinement block is trained to minimize the cross-entropy between the correct label distribution and the model prediction. The decoding process is considered as a multi-pass decoding technique based on Deliberation Networks (Xia et al.,2017). This approach is beneficial for improving the performance of attention-based end-to-end speech recognition.

C. ASR SYSTEMS FOR LOW RESOURCE LANGUAGE

Transfer learning is a very usable technique in low resource language recently, that generalizes models trained from a high to a low resource language (Wang and Zheng,2015). In speech recognition, an acoustic model trained for one high resource language can be used to recognize speech in a low resource language, with little or no re-training data.(Yu et al, 2019), proposed an end- to-end speech recognition framework based on transfer learning for the Low-Resource Tujia Language. This framework uses Chinese corpus as an extension of the Tujia language, to solve the problem of insufficient corpus in the Tujia language, constructing a cross-language corpus and an international phonetic alphabet (IPA) dictionary that is unified between the Chinese and Tujia languages. The scheme of proposed model is shown in fig.6. First, the Tujia language database, the extended Chinese corpus and the cross-language corpus are established through data preprocessing. Then, the extraction of the cross-language acoustic features, train shared hidden layer weights for the Tujia language and Chinese phonetic corpus are done by the convolutional neural network (CNN) (Ozeki and Okatani,2014), and bi-directional long short-term memory (BiLSTM) network connect in cascade with CTC networks (Zeghidour et8 al.,2018)., This framework adopts a baseline model based on Improved Deep Speech, 8in order to solve the problem of the low resources of the Tujia language and establish a bette8r speech recognition model (IDS) (Bai et al.,2015). During the training phase the input of 8the model is the spectrum of the Tujia language and Chinese phonetics. After the multi- layer convolutional layer is extracted, it enters the multi-layer BiLSTM, which completes the cross- language acoustic feature extraction and the shared hidden layer weight learning. The transfer learning is used to establish the model of the cross-language end-to-end Tujia language recognition system. This framework significantly reduces recognition error rate.

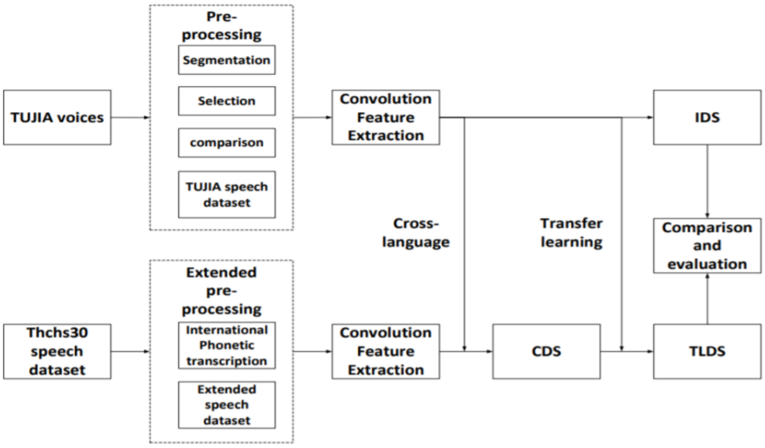

Figure 6. Model scheme (Yu et al., 2019)

Abad et al. (2020), proposed another framework based on transfer learning for low resource language. This framework uses a multi-lingual model in which several DNN layers are shared between languages. This architecture enables domain adaptation transforms learned for one well-resourced language to be applied to an entirely different low resource language. The training process is based on Multi-task learning referring as simultaneously learning multiple tasks from a single data set that contains annotations for different tasks. This approach exploits well-resource language data for improved acoustic modeling of the low-resource target domain improving the performance. Hsu et al. (2019), proposed the application of meta learning approach for low-resource automatic speech recognition (ASR). The aim of meta learning is to solve the problem of “fast adaptation on unseen data”, which is aligned with low-resource setting (Rusu et al., 2018). Model-agnostic meta-learning algorithm is used to optimize the process that meta learning training scheme follows, (MAML) (Finn et al.,2017), which learns initialization parameters from different languages. The initialization parameters found by MAML should adapt as many languages as possible. The architecture of this framework is based on CTC model (Graves et al., 2006). Meta ASR approach outperforms the state-of-the-art multitask pre-training approach on all target languages with different combinations of pre-training languages. One of the main problems in ASR systems is accuracy. Matsuura et al. (2020), proposed a framework that improves accuracy in low resource language. This framework converts the whole training speech data and makes it sound like the test speaker through CycleGANbased non-parallel voice conversion technology (Radford et al., 2015).

First, the source and the target acoustic features are prepared to train CycleGAN. Source features (S) are from original training data, whereas Target features (T) are from the target speaker who is in the test set and unseen in the training set. The CycleGAN is trained to minimize the loss to obtain a generator with which utterances in Source features are transformed to have characteristics of utterances in Target features. This approach is the most effective among these to mitigate the speaker sparsity problem on the low resource language, based on experiments results, comparing with conventional self-supervised adaptation and multilingual training, This approach brings significant improvement in transcriptions. Zhang et al. (2020), proposed an ASR framework based on transfer learning, which reduces significantly error rate. This framework transfers learning from a clean-trained dataset (WSJ) to a noisy trained dataset (CHiME4) for connectionist temporal classification models (CTC). The architecture is based on convolutional neutral networks (CNN), where the top layers of the CTC models are viewed as clean classifiers and the bottom layers considered as features extraction are random initialized or initialized using the weights of model clean, and trained on CHiME-4 with no learning rate re-scaling as is shown in fig.7.

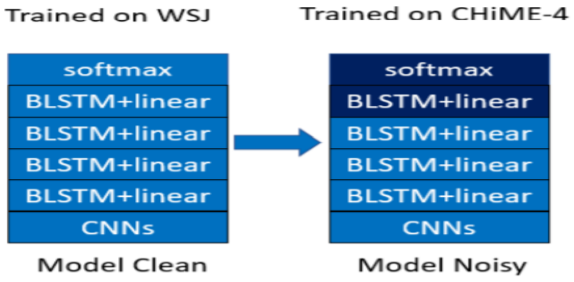

Figure 7. Process of the transfer learning from WSJ to CHiME-4 (Zhang et al.,2020)

This approach, based on experiment result, has significantly lower error rate compared to the conventional transfer learning method. Deep learning is recently, well known for its applicability in speech recognition. The advantage of deep learning for speech recognition stems from the flexibility and predicting power of deep neural network, and it makes a special contribution to low resource language. Chen and Yang (2020), proposed a framework for low resource language specifically Yi language that consists into two stages: training stage and recognition stage as is shown in fig.8.

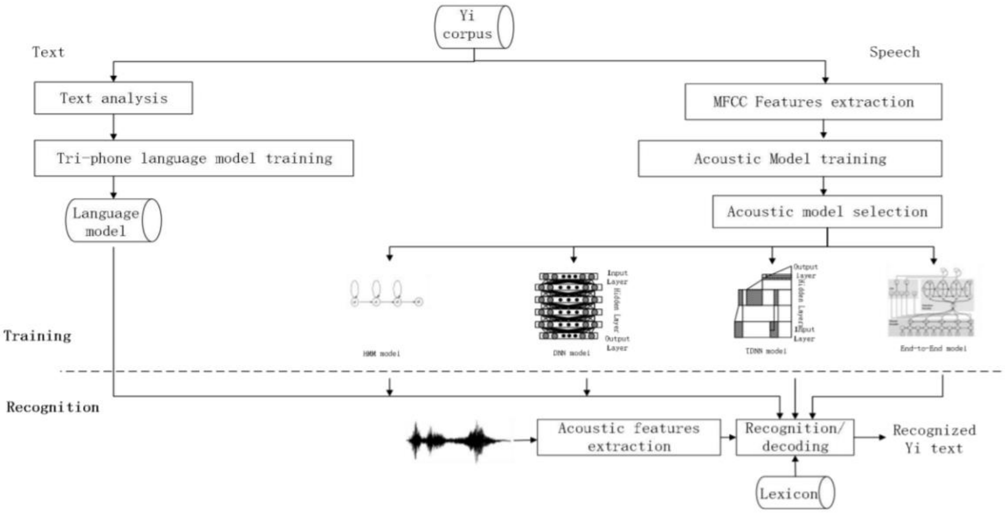

Figure 8. The framework of Yi speech recognition (Chen et al.,2020)

In the training stage, first the Yi language text is analyzed, to realize the language model modeling, and then the audio files are trained, based on the four acoustic models: Hidden Markov Gaussian Mixture models (HMM-GMM) (Rashmi et al.,2018), Deep Neural Network (DNN) (Karáfidt et al.,2018), Time-Delay Neural Network (TDNN) (Waibel et al.,1989) and end-to-end (Hori and Watanabe, 2016). While in the recognition stage, is obtained the Yi recognized text by combining the dictionary and acoustic feature vector for joint recognition and decoding. Based on experiment results MFCC-TDNN system in low resource Yi speech recognition has the best recognition word error rate of 16.65%, while MFCC-DNN has word error rate of 16.72%. Aye et al. (2020), presented another comparison of acoustic models such as: HMM-GMM, DNN, CNN and TDNN for low resource language. The TDNN technique represents better results, based on experiment results. Although the application of deep learning to ASR has resulted in dramatic reductions in word error rate (WER) for languages with abundant training data, ASR for low resource language has yet to benefit from deep learning to the same extent. To overcome this problem,Data Augmentation approach is used (Ragni et al.,2014), which aims to increase the quantity of data, available to train the system. Gokay and Yalcin (2019), proposed an ASR approach based on data augmentation and (TTL) for low resource language. As the data augmentation techniques are used, speed and volume perturbation (Ko et al., 2015) and their combination are applied to training data at the same time. While as the Speech synthesis are used two different approaches. The first approach, Google Translate Text to Speech (gTTS) is used as speech synthesizer and the second approach, an end-to-end Turkish TTS system is trained based on Deep Convolutional TTS

(DCTTS) architecture (Tachibana et al,2018). DCTTS system is composed by two networks as shown in the Fig.9.

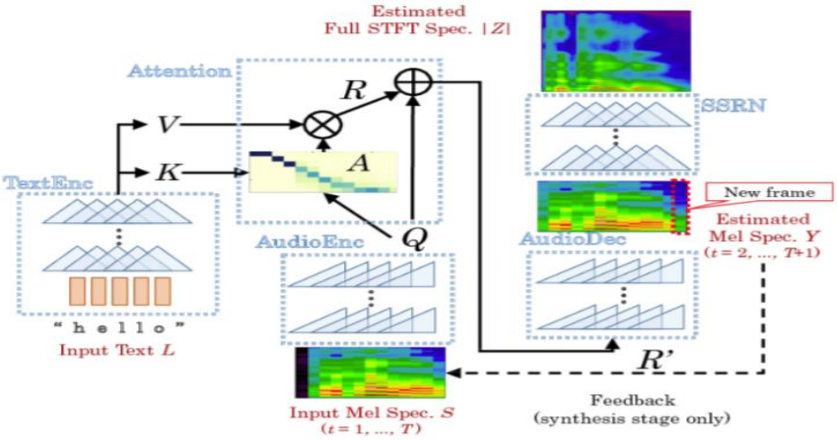

Figure 9. DCTTS architecture (Gokay and Yalcin, 2019)

The first is called Text2Mel which synthesizes mel spectogram according to text input and the second is called as Spectogram Super Resolution Network (SSRN) which converts mel spectogram to Short Time Fourier Transform (STFT) spectrogram. The ASR system is based on Speech2 architecture (Amodei et al., 2016) trained by using CTC (Graves et al., 2006). The CTC model can be coupled with a language model generated by KenLM toolkit (Heafield et al., 2013). Based on experiment results, it is concluded that augmentation or synthesis techniques improve speech recognition for low resource languages. Thai et al. (2020), proposed a full convolutional framework for acoustic modeling in ASR with a variety of established acoustic modeling approaches for low resource language. Fig.10 shows the overall architecture of this framework. The main block of this architecture is the Wide-Block that consists of several parallel streams, each consisting of bottleneck convolution layers before and after a normal convolution layer. The bottleneck layers reduce the complexity of the model by reducing the number of parameters required by the middle convolution operation. The acoustic model consists of two convolutional layers between the input feature vector and the first WideBlock.

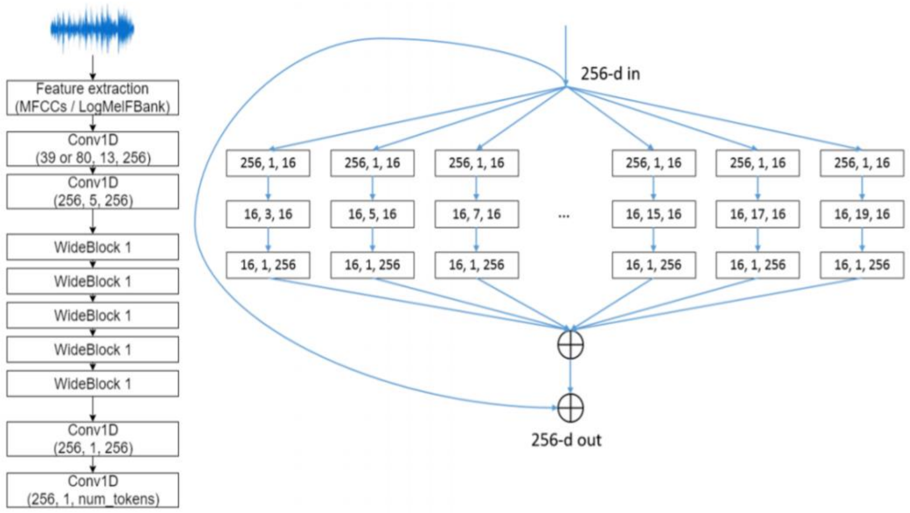

Figure 10. The overall architecture of convolutional framework (Thai et al., 2020)

These embedding layers convert input audio features into a vector of desired depth and temporal content. The final layer outputs a vector with size corresponding to the number of tokens to be predicted. The network is trained by the CTC loss function (Miao et al., 2016). This approach is compared with a Deep Neural approach as well as HMM-GMM approach. The Deep speech approach consists of a five-layer recurrent neural network with LSTM cells. The DeepSpeech is trained by the CTC, HMM/GMM framework is trained according to Kaldi toolkit (Povey, 2011). Training process of proposed framework goes through three stages: In the first stage, is trained an acoustic model on based LibriSpeech English corpus. In the second stage, weight initialization is from the model obtained in the first stage. The model is then trained on heavily augmented training data. In the third stage, the weights of the model from the second stage are used to initialize a model, which is trained only in unaugmented data. Based on experiment results, the proposed framework yields word error rates up to 40% lower than both standard GMM-HMM approaches and established deep neural methods, with a substantial reduction in training time.

DISCUSSION

This paper presents a comprehensive understanding about speech recognition as a part of NPL. It is mainly focused on the analysis of the recent technologies about this area. The analysis is focused on each technology, highlighting the advantages and disadvantages of them.

A. HYBRID SPEECH RECOGNITION SYSTEMS

The hybrid speech recognition systems consist of the combination of Hidden Markov Models (HMMs), Gaussian Mixture Model (GMMs) with different types Neural Networks (NNs) and recently are combined with end-to-end system, in order to take the advantages of each technique to improve speech recognition. Shahin et al. (2014) performed a comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques, while Tachbelie et al. (2020) compared these techniques again with different size of the training speech. In both cases, the DNN-HMM represents higher accuracy than GMM-HMM with a lower word error rate (WER). Dahl et al. (2011) proposed a hybrid framework called Context-Dependent Pre-Trained Deep Neural Networks (CD-DNN-HMM). This framework offers better results compared to CD-GMM-HMM in aspect of recognition accuracy, but training CD-DNN-HMMs is quite expensive compared to training CD-GMM-HMMs, while, Li et al. (2012) proposed a mixed bandwidth training data in the CD-DNN-HMM framework, which improves accuracy by reducing WER. Despite the good results that these techniques have in terms of accuracy, they work with a limited vocabulary on the presence of environmental noises. Combining these techniques with the deep learning approach, it helps overcome these problems and improve even more the accuracy. Sun and Chol (2020) proposed an approach that combines longer context information of recurrent neural network (RNN) with adaptation ability of subspace Gaussian mixture model (SGMM). This framework resulted with satisfactory WER and higher accuracy. Combining DNN-HMM approaches with end-to-end systems as proposed by Tanaka et al. (2019) it is considered as a new direction for future research in this domain, promising high accuracy and stability for a variety of datasets.

B. END-TO-END-SPEECH RECOGNITION SYSTEMS

Despite that the hybrid speech recognition systems have achieved the state-of-the-art in various benchmarks, the end- to- end systems have a wide use in ASR. This is due to their simplicity, flexibility and low cost. Chiu et al. (2018), proposed an end- to- end framework based on Listen, Attend, and Spell (LAS) model. This framework offers good accuracy with low word error (WER), but overfitting remains yet one of the main challenges. To overcome this problem, Nguyen et al. (2020), proposed a solution that consists in increasing the amount of available training data and the variety exhibited by training data with the help of data augmentation. Recently, multitask learning and transfer learning have found widespread use in end-to- end speech recognition systems by exceeding expectations. Inaguma et al. (2019), proposed a framework that uses an external language model (LM) under the transfer learning. This framework drastically reduces the performance gap from the hybrid systems. Lee et al. (2018), proposed an end-to-end speech recognition framework that uses multitask learning to improve generalization of the model by leveraging information from multiple labels. This framework significantly reduces word error rate (WER), by improving accuracy. Combining multitask and transfer-learning approaches with data augmentation is considered as a new direction for future research in this domain, promising high accuracy and stability for a variety of datasets. In addition, the combination of end-to-end systems with hybrid systems is another direction for future research in ASR domain.

C. LOW RESOURCE LANGUAGE

Despite of many studies that have been done about speech recognition for high resource language, in achieving satisfactory results, these techniques cannot be applied for low resource language, due to the limited resources that these languages have. Multitask and transfer learning is a good solution for low resource language. Yu et al. (2019), proposed an end-to-end method in which transfer learning is applied to establish the model of the cross-language end-to-end language recognition system. They concluded that the recognition error rate of the model that transfer learning uses, is reduced with 2.11%. While, the architecture proposed by Abad et al. (2020), enables domain adaptation transforms learned for one high resourced language to be applied to an entirely different low resource language. This approach outperforms other similar methods achieving up to a 29% relative word error rate (WER) improvement. Zhang et al. (2020), proposed an end-to end method that uses transfer learning from a clean dataset (WSJ) to a noisy dataset (CHiME4). This method gives up to 15.5% relative character error rate (CER) reduction compared to models trained only on CHiME-4. Applications of multitask and transfer learning in end-to end systems is considered as one important direction in the future research within this domain. Application of Time Delay Neural Networks in ASR systems is also a good solution for low resource languages. Chen et al. (2020), compared four acoustic models: hidden Markov model (HMM), deep neural network (DNN), time-delay neural network (TDNN) and end-to end trained with Yi low resource language and Aung et al. (2020) compared GMM-HMM, DNN, CNN, TDNN acoustic modeling trained with Myanmar low resource language. In both cases, TDNN is the best solution with lower word error rate (WER). This technology is another direction for future research. Recently, the usage of Data Augmentation approach improves ASR for low resource language, which aims to increase the quantity of data available to train the limited systems. Thai et al. (2019) presented three data augmentation techniques that are used to train various acoustic models. Combined with transfer learning, data augmentation reduces word error rate with 15%, compared to traditional frameworks. While Gokay and Yalcin (2019), used speed and volume perturbations data augmentation techniques to train data. Based on experiment results, it is concluded that augmentation or synthesis techniques improve speech recognition for low resource language. In this case, 14.8% relative WER improvement was obtained by using combination of augmented and synthetic data, while, Thai et al. (2020), presented a fully convolutional ASR for low resource languages. The model was trained on heavily augmented training data after a transfer learning from a high resource language. The combination of transfer learning from a high- resource language and data augmentation contributes to meaningful reductions in word error rate. By analyzing and comparing different technologies about speech recognition, based on the advantages and disadvantages of them, we conclude that a combination of TDNN approach with multitasking and transfer learning, as well as, data augmentation is the most efficient technology, and are considered as the future for low resource language.

CONCLUSION

In this paper, a systematic literature review of existing works has been presented, in order to provide a comprehensive understanding of speech recognition systems. The scope of this paper is targeted towards Hybrid speech recognition systems, End-to-End speech recognition systems and ASR Systems for Low Resource Language. Hybrid Speech Recognition Systems are widely used, due to the high accuracy they have. Referring to the analysis of many articles that we have done in this study about hybrid systems, we conclude that combining DNN-HMM approaches with end-to-end systems is considered as a new direction for future research in this domain, promising high accuracy and stability for a variety of datasets. Recently, speech recognition end-to-end systems are very promising. Compared to hybrid systems, these systems are simpler, more flexible, have lower costs and give satisfactory results. Referring to the papers we have analyzed in this study for end-to-end speech recognition systems, the combining multi task and transfer learning approaches with data augmentation is considered as a new direction for future research in this domain, promising high accuracy and stability for a variety of datasets. In addition, the combination of end-to-end systems with hybrid systems is another direction for future research in ASR domain. Whereas, referring to the analysis of the papers for low resource languages, we conclude that a combination of Time Delay Neural Networks (TDNN) approach with multitasking and transfer learning as well as data augmentation is the more efficient technology, and it is considered as the future for low resource language. However, taking into account the fact of the special features that each language has, these technologies remain to be studied in the future by analyzing their performance in different languages.

REFERENCES

Yu, D., & Deng, L. (2016). *AUTOMATIC SPEECH RECOGNITION*. Springer london limited.

Ruby, J. (2020). Automatic Speech Recognition and Machine Learning for Robotic Arm in Surgery. *American Journal of Clinical Surgery*, *2*(1), 10-18.

Juang, B. H. (1991). Speech recognition in adverse environments. *Computer speech & language*, *5*(3), 275-294.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, *51*(1), 7-15.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of systems and software*, *80*(4), 571-583.

Shahin, M., Ahmed, B., McKechnie, J., Ballard, K., & Gutierrez-Osuna, R. (2014). A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Tachbelie, M. Y., Abulimiti, A., Abate, S. T., & Schultz, T. (2020, May). Dnn-based speech recognition for globalphone languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8269-8273). IEEE.

Kadyan, V., & Kaur, M. (2020). SGMM-Based Modeling Classifier for Punjabi Automatic Speech Recognition System. In *Smart Computing Paradigms: New Progresses and Challenges* (pp. 149-155). Springer, Singapore.

Wu, Z., & Cao, Z. (2005). Improved MFCC-based feature for robust speaker identification. *Tsinghua Science & Technology*, *10*(2), 158-161.

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, *20*(1), 30-42.

Tsunoo, E., Kashiwagi, Y., Asakawa, S., & Kumakura, T. (2019). End-to-end adaptation with backpropagation through WFST for on-device speech recognition system. *arXiv preprint arXiv:1905.07149*.

Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohri, M. (2007, July). OpenFst: A general and efficient weighted finite-state transducer library. In *International Conference on Implementation and Application of Automata* (pp. 11-23). Springer, Berlin, Heidelberg.

Chen, Z., Jain, M., Wang, Y., Seltzer, M. L., & Fuegen, C. (2019, May). End-to-end contextual speech recognition using class language models and a token passing decoder. In *ICASSP*

2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6186-6190). IEEE.

Kneser, R., & Ney, H. (1993). Improved clustering techniques for class-based statistical language modelling. In *Third European Conference on Speech Communication and Technology*.

Hall, K., Cho, E., Allauzen, C., Beaufays, F., Coccaro, N., Nakajima, K., ... & Zhang, L. (2015). Composition-based on-the-fly rescoring for salient n-gram biasing.

Wang, C., Wu, Y., Liu, S., Li, J., Lu, L., Ye, G., & Zhou, M. (2020). Low latency end-to-end streaming speech recognition with a scout network. *arXiv preprint arXiv:2003.10369*.

Sun, R. H., & Chol, R. J. (2020). Subspace Gaussian mixture based language modeling for large vocabulary continuous speech recognition. *Speech Communication*, *117*, 21-27.

Hermansky, H., Ellis, D. P., & Sharma, S. (2000, June). Tandem connectionist feature extraction for conventional HMM systems. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)* (Vol. 3, pp. 1635-1638). IEEE.

Xu, H., Su, H., Chng, E. S., & Li, H. (2014). Semi-supervised training for bottle-neck feature based DNN-HMM hybrid systems. In *Fifteenth annual conference of the international speech communication association*.

Yu, D., & Seltzer, M. L. (2011). Improved bottleneck features using pretrained deep neural networks. In *Twelfth annual conference of the international speech communication association*.

Tanaka, T., Masumura, R., Moriya, T., Oba, T., & Aono, Y. (2019). A Joint End-to-End and DNN-HMM Hybrid Automatic Speech Recognition System with Transferring Sharable Knowledge. In *INTERSPEECH* (pp. 2210-2214).

Yu, D., & Seltzer, M. L. (2011). Improved bottleneck features using pretrained deep neural networks. In *Twelfth annual conference of the international speech communication association*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Hori, T., Hori, C., Minami, Y., & Nakamura, A. (2007). Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition. *IEEE Transactions on audio, speech, and language processing*, *15*(4), 1352-1365.

Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4774-4778). IEEE.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). 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) (pp. 4960-4964). IEEE.

Schuster, M., & Nakajima, K. (2012, March). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5149-5152). IEEE.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762.*

Nguyen, T. S., Stueker, S., Niehues, J., & Waibel, A. (2020, May). Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7689-7693). IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762.*

Weng, C., Cui, J., Wang, G., Wang, J., Yu, C., Su, D., & Yu, D. (2018, September). Improving Attention Based Sequence-to-Sequence Models for End-to-End English Conversational Speech Recognition. In *Interspeech* (pp. 761-765).

Pham, N. Q., Nguyen, T. S., Niehues, J., Müller, M., Stüker, S., & Waibel, A. (2019). Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377.*

Xu, H., Ding, S., & Watanabe, S. (2019, May). Improving end-to-end speech recognition with pronunciation-assisted sub-word modeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7110-7114). IEEE.

Sperber, M., Niehues, J., Neubig, G., Stüker, S., & Waibel, A. (2018). Self-attentional acoustic models. *arXiv preprint arXiv:1803.09519*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762.*

Liu, A. H., Sung, T. W., Chuang, S. P., Lee, H. Y., & Lee, L. S. (2020, May). Sequence-to-Sequence Automatic Speech Recognition with Word Embedding Regularization and Fused Decoding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7879-7883). IEEE.

Unanue, I. J., Borzeshi, E. Z., Esmaili, N., & Piccardi, M. (2019). ReWE: Regressing word embeddings for regularization of neural machine translation systems. *arXiv preprint arXiv:1904.02461*.

Toshniwal, S., Kannan, A., Chiu, C. C., Wu, Y., Sainath, T. N., & Livescu, K. (2018, December). A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 369-375). IEEE.

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41-75.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Inaguma, H., Cho, J., Baskar, M. K., Kawahara, T., & Watanabe, S. (2019, May). Transfer learning of language-independent end-to-end ASR with language model fusion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6096-6100). IEEE.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Zeyer, A., Irie, K., Schlüter, R., & Ney, H. (2018). Improved training of end-to-end attention models for speech recognition. *arXiv preprint arXiv:1805.03294*.

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017). Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H. C., ... & Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Denisov, P., & Vu, N. T. (2019). End-to-end multi-speaker speech recognition using speaker embeddings and transfer learning. *arXiv preprint arXiv:1908.04737*.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., ... & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558*.

Seki, H., Hori, T., Watanabe, S., Roux, J. L., & Hershey, J. R. (2018). A purely end-to-end system for multi-speaker speech recognition. *arXiv preprint arXiv:1805.05826*.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M. A., Devin, M., & Dean, J. (2013, May). Multilingual acoustic models using distributed deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8619-8623). IEEE.

Kubo, Y., & Bacchiani, M. (2020, May). Joint phoneme-grapheme model for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6119-6123). IEEE.

Lee, J., Mansimov, E., & Cho, K. (2018). Deterministic non-autoregressive neural sequence modeling by iterative refinement. *arXiv preprint arXiv:1802.06901*.

Xia, Y., Tian, F., Wu, L., Lin, J., Qin, T., Yu, N., & Liu, T. Y. (2017, December). Deliberation networks: Sequence generation beyond one-pass decoding. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 1782-1792).

Wang, D., & Zheng, T. F. (2015, December). Transfer learning for speech and language processing. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1225-1237). IEEE.

Yu, C., Chen, Y., Li, Y., Kang, M., Xu, S., & Liu, X. (2019). Cross-language end-to-end speech recognition research based on transfer learning for the low-resource Tujia language. *Symmetry*, *11*(2), 179.

Ozeki, M., & Okatani, T. (2014, November). Understanding convolutional neural networks in terms of category-level attributes. In *Asian Conference on Computer Vision* (pp. 362-375). Springer, Cham.

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., & Dupoux, E. (2018). End-to-end speech recognition from the raw waveform. *arXiv preprint arXiv:1806.07098*.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

Abad, A., Bell, P., Carmantini, A., & Renais, S. (2020, May). Cross lingual transfer learning for zero-resource domain adaptation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6909-6913). IEEE.

Hsu, J. Y., Chen, Y. J., & Lee, H. Y. (2020, May). Meta learning for end-to-end low-resource speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7844-7848). IEEE.

Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., & Hadsell, R. (2018). Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*.

Finn, C., Abbeel, P., & Levine, S. (2017, July). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning* (pp. 1126-1135). PMLR.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).

Matsuura, K., Mimura, M., Sakai, S., & Kawahara, T. (2020). Generative Adversarial Training Data Adaptation for Very Low-resource Automatic Speech Recognition. *arXiv preprint arXiv:2005.09256*.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Zhang, S., Do, C. T., Doddipatla, R., & Renals, S. (2020, May). Learning Noise Invariant Features Through Transfer Learning For Robust End-to-End Speech Recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7024-7028). IEEE.

Chen, Z., & Yang, H. (2020, June). Yi Language Speech Recognition using Deep Learning Methods. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (Vol. 1, pp. 1064-1068). IEEE.

Rashmi, S., Hanumanthappa, M., & Reddy, M. V. (2018). Hidden Markov Model for speech recognition system—a pilot study and a naive approach for speech-to-text model. In *Speech and Language Processing for Human-Machine Communications* (pp. 77-90). Springer, Singapore.

Karáfidt, M., Baskar, M. K., Veselý, K., Grézl, F., Burget, L., & Černocký, J. (2018, April). Analysis of multilingual blstm acoustic model on low and high resource languages. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5789-5793). IEEE.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing, 37*(3), 328-339.

Kim, S., Hori, T., & Watanabe, S. (2017, March). Joint CTC-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4835-4839). IEEE.

Aung, M. A. A., & Pa, W. P. (2020, February). Time Delay Neural Network for Myanmar Automatic Speech Recognition. In *2020 IEEE Conference on Computer Applications (ICCA)* (pp. 1-4). IEEE.

Ragni, A., Knill, K. M., Rath, S. P., & Gales, M. J. (2014, September). Data augmentation for low resource languages. In *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association* (pp. 810-814). International Speech Communication Association (ISCA).

Gokay, R., & Yalcin, H. (2019, March). Improving low Resource Turkish speech recognition with Data Augmentation and TTS. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 357-360). IEEE.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Tachibana, H., Uenoyama, K., & Aihara, S. (2018, April). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4784-4788). IEEE.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

Availablehttps://github.com/SeanNaren/deepspeech.pytorch accesed 25.05. 2020.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006, June). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning* (pp. 369-376).

Heafield, K., Pouzyrevsky, I., Clark, J. H., & Koehn, P. (2013, August). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 690-696).

Thai, B., Jimerson, R., Ptucha, R., & Prud'hommeaux, E. (2020, May). Fully Convolutional ASR for Less-Resourced Endangered Languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 126-130).

Miao, Y., Gowayyed, M., Na, X., Ko, T., Metze, F., & Waibel, A. (2016, March). An empirical exploration of CTC acoustic models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2623-2627). IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding* (No. CONF). IEEE Signal Processing Society.