WORLD COMET RESEARCH GROUP

International Journal of Applied Data Science & Modern Computing
Vol. 6, No. 1, 2023
Doi: 10.67228/30713498/IJADSMC-2023PI0Q6T
PP. 01-13

Original Article

Deep Learning Approaches for Speech Recognition Systems

Dr. Lydia Languish

School of Business, University of Sydney, Australia.

Received: 02-02-2023 Revised: 02-27-2023 Accepted: 03-02-2023 Published: 03-04-2023

ABSTRACT

Automatic Speech Recognition (ASR) has evolved from rule-based and statistical methods to deep learning approaches, achieving near human-level performance under certain conditions. Traditional HMM-GMM models face limitations in handling long-term dependencies, speaker variability, and noise. Modern architectures such as DNNs, CNNs, RNNs, LSTMs, and Transformers provide improved representation learning and feature extraction from speech signals. This paper presents a comprehensive review of deep learning-based ASR systems, covering their evolution, acoustic feature extraction, end-to-end modeling, and training techniques. A generalized ASR pipeline is proposed, including preprocessing, feature encoding, model design, optimization, and decoding. Performance is evaluated using metrics like Word Error Rate (WER) and Character Error Rate (CER). The study highlights key challenges such as low-resource languages, real-time processing, domain adaptation, and model interpretability. It concludes that deep learning is the standard for ASR, with future research focusing on self-supervised learning, multilingual models, and efficient edge deployment.

KEYWORDS

Automatic Speech Recognition, Deep Learning, Acoustic Modeling, End-to-End ASR, LSTM, Transformer, Speech Processing.

ISSN: 3071-3498
PP: 01-13.

1. INTRODUCTION

1.1. Background

Speech recognition refers to the computing method of converting orating speech to a relative textual signal, which facilitates natural and effective human machine communication. The initial systems of speech recognition were heavily reliant on template matching as well as rule based phoneme recognition strategies. These methods were based on pre-determined speech patterns and linguistic rules, which were very dependent on the variations of the speakers, noise and variations of pronunciation. Consequently, these systems were typically limited in size of vocabulary and closed systems and used in applications dependent on the speaker, which inhibited their useful usage. Statistical modeling was one of the first emerged milestones in the development of the Automatic Speech Recognition (ASR). Specifically, the Hidden Markov Models (HMMs) became the prevailing framework because of its capacity to provide a probabilistic way of modeling sequential and time-related aspects of speech signals. HMM-based systems, with speech represented as a series of hidden states, representing linguistic units, made recognition more flexible and robust than otherwise, allowing speech made up of rules through rule-based approaches before. This breakthrough formed the basis of large vocabulary continuous speech recognition and helped that more scalable ASR systems could be developed. In spite of this success, classical ASR systems are strongly defined with complex and modular pipelines, which involve feature extraction, acoustic modeling, pronunciation modeling and language modeling stages. This is because, each of the modules is customized and optimized individually and as such, the modules may not perform well together and therefore the global performance may not be optimal. In addition, they are very dependent on hand-created acoustic characteristics like Mel-Frequency Cepstral Coefficients (MFCCs) that though good in tracking the low-level spectral features of speech, they are unable to capture high-level semantic and contextual features of speech. Such constraints have inspired the movement towards the use of deep learning strategies which allow joint optimization and end-to-end learning, eventually leading to the high recognition rates and robustness of the system.

1.2. Importance of Deep Learning Approaches

Fig 1 - Importance of Deep Learning Approaches

ISSN: 3071-3498
PP: 01-13.

1.2.1. Automatic Feature Learning

Among the most major positive features of the deep learning methods in speech recognition, it is possible to note their capability to automatically learn feature representations at multiple hierarchies based on the effect of direct speech signals or minimally processed ones. Deep neural networks are able to produce low-level spectral patterns and high-level phonetic and linguistic abstractions, unlike traditional ASR systems, which are based on handcrafted features (like MFCCs). This automatic capability of learning models helps model complex acoustical variations near speaker differences, accents, speaking rates and environmental noise.

1.2.2. Improved Modeling of Temporal and Contextual Dependencies

The machine learning models that are most successful at capturing temporal correlations in sequential data include Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers. The speech signals have very high temporal correlations and the deep models have the ability of capturing both short-term and long-range context characters. This leads to a higher accuracy in recognition of continuous speech, better effect of coarticulation and prediction of word sequences as opposed to the traditional statistical models.

1.2.3. End-to-End Learning Capability

Deep learning has made it possible to construct end-to-end ASR systems that do not need separate pronunciation or language model and can associate acoustic inputs directly with textual outputs. End-to-end models achieve high levels of optimization in all the units in the same framework without generating additional complexity of the system or propagation of errors between modules. This general learning is achieved by resulting in better global optimization and increased recognition performance especially in large scale speech recognition tasks.

1.2.4. Scalability and Performance Gains

The other major input of deep learning methods is that they can scale to large data sets and current hardware computing systems. Deep models have the capability of using large quantities of labeled and unlabeled speech data with substantial improvements in the Word Error Rate. Moreover, research and deployment has been accelerated and trained complex deep architectures have become feasible with advances in parallel computing via GPUs and TPUs.

1.2.5. Robustness and Generalization

ASR systems that are designed on deep learning have high levels of robustness and generalization in a wide range of acoustic conditions. These models are more effective at adapting to noise, reverberation and domain shifts than standard methods, in addition to having data-driven learning and regularization methods. Thus, deep learning has turned out to be the foundation of the current state of the art speech recognition systems, age capable of providing the confidence of working in real-world scenarios, including voice assistants, transcription systems and conversational AI systems.

ISSN: 3071-3498
PP: 01-13.

1.3. Approaches for Speech Recognition Systems

Various methodological approaches have been developed towards speech recognition systems and have in turn been a reflection of improvement in computational models and learning paradigms. Early methods were mainly rule-based and template matching systems that used template- matching systems that were based on predetermined phonetic rules, and stored phonetic templates in order to identify the spoken words. Though these were easy to interpret and decode and the speaker variability and noise were limited, the pronunciation variation limited the application of these methods to small vocabularies and controlled conditions. As statistical learning developed, probability methods gained prominence in speech recognition. The standard ASR pipelines were built on Hidden Markov Model (HMM)-based systems usually with a Gaussian Mixture Models (GMM) added to the framework. GMMs were used to model the acoustic feature distributions in these systems and the HMMs were used to capture the temporal arrangement of speech signals. This method facilitated continuous speech recognition with a large vocabulary and it was much stronger than the previous methods. Nonetheless, the use of handcrafted features and system components optimization by itself restricted the overall performance. Discriminative modeling of specific speech tasks, including phoneme classification and segmentation, was made more effective with the introduction of machine learning-based algorithms, including Support Vector Machines (SVMs) and Conditional Random Fields (CRFs). Although these techniques were more accurate than purely statistical techniques, they did not scale or run efficiently with large-scale speech recognition systems. Later, the paradigm of deep learning-based techniques has taken place in ASR. Deep neural networks, convolutional networks, recurrent designs, and Transformer-based models are designed to automatically learn features and provide an effective model of more complex dependencies in time and context. End-to-end models like CTC and attention-based encoder-decoder models further simplify the model by taking the speech signals and mapping them directly to the text. These new technologies are far more effective than the traditional systems and are in the state of art today, with extensive deployment in practical speech-enabled applications becoming commonplace.

2. LITERATURE SURVEY

2.1. Traditional Speech Recognition Approaches

The Hidden Markov Models (HMMs) with the Gaussian Mixture Models (GMMs) were the key acoustic model used in Traditional Automatic Speech Recognition (ASR). HMMs were useful in this context in order to model the time structure of speech signals because GMMs modeled the probability distribution of handcrafted acoustic features including Mel-Frequency Cepstral Coefficients (MFCCs). The systems performed moderately and were interpretable, and it is due to this that they dominated a number of decades. Their dependence on manual feature engineering however constrained their capacity to capture complicated and nonlinear acoustic variation due to the variation of speakers, background noise, and pronunciation however. Other classifiers like Support Vector Machines (SVMs) and Conditional Random Fields (CRFs) were also studied as phoneme/segment-level recognition models, but were computationally too complex and too lacked scalability to be used on large vocabulary general speech recognition tasks.

ISSN: 3071-3498
PP: 01-13.

2.2. Deep Neural Network-Based Acoustic Models

The introduction of Deep Neural Networks (DNNs) in ASR signified a major change in generative to discriminative model of acoustic modeling. DNN-HMM hybrid systems substituted GMMs with DNNs to determine the state posterior probability, which resulted in significant decreases of Word Error Rate (WER) on benchmark datasets. Subsequent architectural developments enhanced performance: Convolutional Neural Networks (CNNs) added local receptive fields and weight sharing which improved allowed it to handle a frequency shift and speaker variations, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks efficiently learnt long-term temporal variations in speech signals. Although these hybrid systems were much more accurate, they still required a complicated pipeline with individual acoustic, pronunciation and language models, which made system design and optimization even more complicated.

2.3. End-to-End Speech Recognition Models

End-to-end ASR models attempt to simplify the speech recognition pipeline, simply mapping acoustic features to text sequences with no phonetic lexicon or separate language model. Connectionist Temporal Classification (CTC) can be used with large scale training since it doesn't require frame-wise annotations in order to be trained on the sequence. Encoders and decoders built on attention helped to raise recognition accuracy even higher, as they learnt to dynamically match the input speech frames with output characters or subwords. Transformer-based ASR models, by using self-attention-based models to capture long-range dependencies and training them in a highly parallelized way, have since shown state-of-the-art performance. Although end-to-end models reduce systems complexity, they often need large annotated, and large computing power to exhibit effective training.

2.4. Comparative Summary of Literature

The existing ASR approaches can be compared to show the trade-offs among the interpretability, accuracy, and scalability. HMM-GMM systems are full of transparency and formed models but have low scalability and the representative ability. DNN-HMM hybrids are also more accurate in recognition with complex multi-stage pipelines. Models based on LSTMs are efficient in speech sequence temporal modeling, but sequential processing makes them computationally expensive. Transformer-based models, by comparison, are efficient to parallelize and model context better, but are very data-intensive and cannot be used in low-resource language scenarios or in real-time deployment settings.

3. METHODOLOGY

3.1. Overall System Architecture

A typical generic deep learning-based Automatic Speech Recognition (ASR) system is a program that consists of organized processes that convert speech signals in the form of a raw audio format into text. All of the stages are critical in terms of making the recognition process robust, accurate, and scalable.

ISSN: 3071-3498
PP: 01-13.

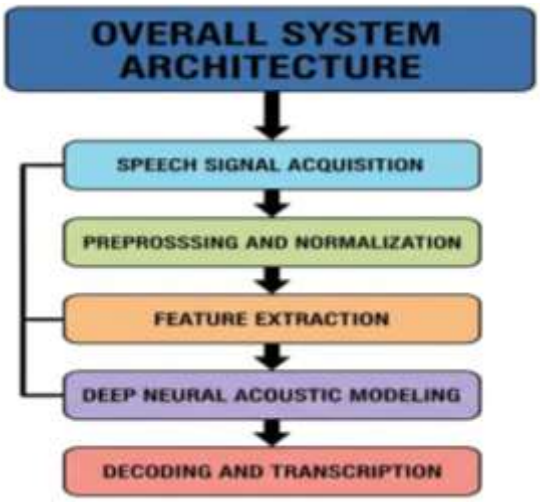

Fig 2 - Overall System Architecture

3.1.1. Speech Signal Acquisition

The first component of the ASR pipeline is the acquisition of speech signals by recording an analog speech signal with a microphone or audio sensor and converting it to the digital domain by analog-to-digital conversion. The performance of systems is directly affected by the quality of its acquisition since sensitivity of the microphones, sampling rate, background noise, and distortions on a channel affects the quality of the recorded speech. Common ASR systems run at the sampling rates of 16 KHz or more in order to retain speech-related frequency content. Strong acquisition mechanisms are necessary particularly in contexts of real-world where speaker variability and acoustic interference are engaged.

3.1.2. Preprocessing and Normalization

Preprocessing leads to the improvement of the quality of the received speech signal and minimization of undesired variations. This processing step normally involves noise elimination, silence detection, pre-emphasis filter, and amplitude equalisation. Other methods like voice activity detection (VAD) are used to detect speech portions and remove non-speech sections to enhance computational efficiency. Normalization aids in making input signals across speakers and recording situations standardized thus allowing the downstream neural models to better generalize.

3.1.3. Feature Extraction

The step of feature extraction converts the preprocessed speech signal into small and discriminative speech features befitting the learning algorithms. Popular ones include Mel-Frequency Cepstral Coefficients (MFCCs), filter bank energies, and log-Mel spectrograms, which are perceptually significant spectral attributes of speech. Temporal context may be added either by laying several frames one atop or by introducing delta and delta-delta coefficients. A good feature extraction technique saves data dimensionality without losing linguistic information that is important in proper recognition.

ISSN: 3071-3498
PP: 01-13.

3.1.4. Deep Neural Acoustic Modeling

The main part of the modern ASR systems is deep neural acoustic modeling, where the neural networks acquire complex mappings between acoustic features and linguistic units, like phonemes, subwords, or characters. All of these architectures, including DNNs, CNNs, RNNs, LSTMs, and Transformer-based models, are common in modeling the nonlinear relationship and temporal dependency within speech signals. It is these models that are trained with large scale labeled datasets with the aim of minimizing recognition errors and thus drastically outperforming conventional statistical acoustic models.

3.1.5. Decoding and Transcription

The last step in the ASR pipeline is decoding and transcription, during which the results of the acoustic model are translated into logical-text sequences. Other decoding algorithmics like beam search combine acoustic scores with language model probability to determine the best possible sequence of words. This step is used directly in end-to-end systems to generate character or subword sequences and is not explicitly models pronunciation. The resulting transcription can be used as the end product, which can then be applied to voice assistants, transcription services, and speech-driven interfaces.

3.2. Feature Extraction

The process of feature extraction is an essential part of an Automatic Speech Recognition (ASR) system since it transforms unstructured and inaccurate speech waveforms into informative and succinct representations that can be used in deep learning. The recently common acoustic features are Mel-Frequency Cepstral Coefficients (MFCCs), log-Mel filter bank energies as well as spectrogram-based representations. These features are set to receive the perceptually important attributes of human speech and deease redundancies and noise sensitivity. The most popular feature among them is MFCCs which is highly adopted because of its effectiveness and the efficiency of computing. Having a speech signal, x(t), which is a continuous time, the computation process of MFCC starts with dividing the signal into short overlapping frames, because speech may be supposed to be quasi-stationary in a short time. All the frames are initially windowed, usually with a Hamming window to reduce spectral leakage. This is then followed by Fast Fourier Transform (FFT) that transforms time-domain signal into frequency domain which displays its spectral contents. The magnitude spectrum of the FFT is a distribution of energy of the frequency components. The frequency spectrum is then subjected to a bank of Mel-scaled filters to match human auditory perception, paying more emphasis on lower frequencies and more compression on higher frequencies. The energy of the filter banks in the logarithm form is then calculated in order to represent the logarithmic sensitivity of the human hearing to the intensity of sound. The last step is to use the Discrete Cosine Transform (DCT) on the log-Mel energies, which is used to decorrelate the features and condense the most important ones into a few coefficients. Simply put, the MFCCs are derived through DCT of the logarithm of the magnitude of the FFT signal of the speech signal. Filter bank based Log-Mel features undergoes the same process but does not include the DCT step, instead,

ISSN: 3071-3498
PP: 01-13.

localized spectral information is retained which is desirable especially to the convolutional and Transformer based models. Spectrograms are a representation of speech in timefrequency form and are commonly utilized in end-to-end ASR systems. In general, the feature extraction is crucial in enhancing the recognition accuracy by giving strong and discriminative speech signal representations.

3.3. Deep Learning Model Design

The design of deep learning models is the backbone of up-to-date Automatic Speech Recognition (ASR) systems, in which the goal is to train a powerful mapping between applying a sequence of acoustic features and an equivalent sequence of textual transcriptions. Where X = {x 1, x 2, and so on (x T) are the sequences of input feature vectors derived out of a speech signal with x T being the coefficients and features used to characterize a particular sound in a short period of time and T being the number of sequences. The task of the deep learning model is to extract complex time and spectral patterns in such sequences of features and encode them into meaningful units of linguistic information. This model is conditionally trained on a probability model P (Y X; 0) representing the probability that the model produces output transcription Y given that the input features are X conditioned by a single parameter 0. Rudely, this term implies that this model approximates the likelihood of a specific word or sequence of characters, with the observed speech characteristics and the parameters of the learned model. These parameters are weights and biases of the neural network layers which are optimized in the course of training with large corpus of labeled speech. There are various deep learning networks that are used to approximate this probability function. The emphasis of the deep neural networks used in reinforcement is to learn discriminative transformations of the features, and the approach used by the convolutional neural network is the exploitation of the local spectral correlation and resistance to the frequency fluctuations. Temporal dynamic models like the RNN and the LSTM networks are more applicable to speech recognition as they capture the connections between the observations over time, hence the network is able to memorise the information across time. Transformer-based models, which use self-attention mechanisms to obtain long-range dependencies and process input sequences in parallel, have been developed more recently and achieve better scalability and performance. The model can be trained by minimising a loss function which quantifies the difference between predicted and ground-truth transcriptions, and modalities like stochastic gradient descent or adaptive learning algorithms can be used. The deep model has its parameters updated in an iterative way, by optimizing the conditional probability of the optimal transcriptions. Therefore, a structured deep learning model can be an important boost to recognition accuracy and resiliency in a sheer amount of acoustic settings.

3.4. Training Objective

The deep learning-based Automatic Speech Recognition (ASR) systems can be trained to achieve a specific quality of the transcriptions based on the training objective that measures the characteristics of model parameters. In models based on Connectionist Temporal Classification, denoted CTC models have the goal function formulation in order to deal with the problem of

ISSN: 3071-3498
PP: 01-13.

matching input acoustic frames and output sequence of labels when frame level annotations are not known. CTC unlike the traditional models that need exact time matching between speech frames and phonemes or characters allows training end-to-end with the final transcription only. The loss function of the CTC can be defined as L CTC = -logsum 0VA of the valid alignment paths with the input feature sequence X. Loss makes the model penal handed over in un-probabilistic form to an event where it places a low probabilistic score of the correct transcription, and with all possible frame-label sequences that can be folded onto the target output Y. In this case, every alignment path π is an indication of a potential mapping of the input frames to output symbols, with special blank labels that consider the variable length segments of speech. B -1(Y) appears to be the set of all the such paths that represent the transcription Y without redundant symbols and blanks labels. In training, the model calculates the likelihood of every alignment path when the speech features of the input are provided and adds these likelihoods to achieve the overall probability of the correct transcription. The CTC loss is the negative logarithm of this likelihood and it is minimized by gradient-based approaches. The model can learn both speech and text acoustic representation and alignment by simultaneously learning the text-speech correlation. A major strength of the CTC objective is that it is resistant to the rate of speech and pronunciation variation that it does not insist on frame-to-label correspondence. But CTC makes the conditioning independence of the output labels given the input, and this may restrict its capacity to capture long-range dependencies in linguistics. Although it has this shortcoming, CTC is a very popular training goal in ASR as it is simple, both efficient and empirically effective in large scale speech recognition tasks.

3.5. Evaluation Metrics

The data that can be used to evaluate the performance of Automatic Speech Recognition (ASR) systems are important evaluation metrics. Of the existing metrics employed in the study of speech recognition, Word Error rate (WER) is currently the most commonly utilized one because of its simplicity and high relationship with human perception of quality of transcription. WER is a metric that quantifies the difference between the system-generated transcription and the corresponding ground-truth reference transcription by counting at the word level errors. Defining the Word Error rate is an assessment of the total errors per the total words in the reference transcription. The calculation of WER in normal words is to sum up the number of substitutions (S), deletions (D), and insertions (I) and divide the result by the total number of words (N) in the reference text. Substitutions: Within a word in the reference is substituted with the wrong word in the hypothesis, deletions: With the help of the recognized output, a word in the reference was deleted, and insertions: In addition to the words, which appear in the reference, extra words were inserted, but they were not found in the reference. The higher is the WER, the higher the recognition performance, a zero value signifies an ideal transcription. WER is an all-purpose score of recognition error by representing various form of word level error in a single measure. It is also beneficial specifically in comparing diverse ASR models, training approaches, or feature representations with equal test circumstances. WER however does not put in consideration the semantic similarity between words or types of various errors. Indicatively, the replacement of a synonym is punished in

ISSN: 3071-3498
PP: 01-13.

the same manner as the one of an irrelevant word. Also, WER is prone to slight differences in word segmentation and use of punctuations. In spite of these shortcomings, WER is still considered the standard measure of evaluation in ASR studies and practice. It is commonly complemented with other measures including Character Error Rate (CER), real-time factor (RTF) to give a more holistic measure of the system performance, particularly when it is in a multilingual and real time recognition setting.

4. RESULTS AND DISCUSSION

4.1. Experimental Performance Comparison

The experimental analysis involves the comparison of performance of the various speech recognition models through the use of Word Error Rate (WER) as the main measure. The findings reveal the gradual advancements through the development of deep learning structures.

Table 1: **Experimental Performance Comparison**

Model	WER (%)
DNN-HMM	18.5
LSTM	9.7
Transformer	5.4

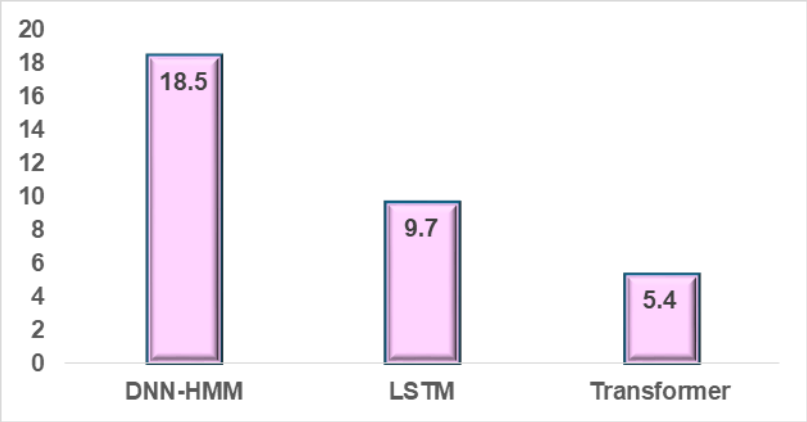

Fig 3 - Experimental Performance Comparison

4.1.1. DNN-HMM Model

The Word Error Rate of the DNN-HMM model stands at 18.5% and this reflects the weaknesses of the hybrid architecture that implements deep neural networks and the classical Hidden Markov Models. Although the performance of DNNs is much better in acoustic modeling than the versions based on GMMs, the general pipeline is quite complicated because it uses both pronunciation and language models. The frame-level assumptions of independence and limited contextual modeling also have an impact on the error rates which are higher in continuous and spontaneous speech conditions.

ISSN: 3071-3498
PP: 01-13.

4.1.2. LSTM-Based Model

A significant accuracy in the recognition is observed as the LSTM-based model ends with a Word Error rate of 9.7%. The number of errors can be reduced with the help of LSTM, which is able to learn long-term temporal dependencies in speech signals and thus better models coarticulation and contextual variation. The LSTM networks can deal with variable length input sequences and different speaking rates as they retain memory over time steps. Nevertheless, they are more complex to compute and more difficult to train since they are sequentially processed.

4.1.3. Transformer-Based Model

Transformer-based model has the lowest Word Error Rate of 5.4 and thus it performs better in terms of recognition compared to the other checked models. This has been enhanced mainly by the mechanism of self attention which enables the model to absorb global contextual relationship within the range of the entire speech sequence. Transformers are also capable of parallel computation because of their design in contrast to recurrent architectures, which leads to both higher training and improved scalability. Transformer models are generally data-intensive and computationally intensive though their performance is high.

4.2. Discussion

It is evident in the results of the experiment that Transformer-based models are more productive than the models of recurrent architecture like LSTMs and hybrid DNN-HMM frameworks. The main cause of this enhancement is that the Transformer can generate the model of the global contextual relationships using self-attention systems. Transformers attend to all input frames at once as opposed to recurrent networks that process speech signals in sequence and use the memory state to encode the effects of temporal dependencies in the speech. This allows the more important modeling of long range dependencies especially where speech recognition tasks are concerned with coarticulation, the ability to change the rate of speaking as well as the pronunciation patterns given the context. Transformer-based models are also characterized by another important benefit the training effectiveness. Transformers is Starred, since self-attention is parallelizable and thus can use the latest hardware accelerators like GPUs and TPUs to achieve both faster convergence and better scalability. Such efficiency allows it to be possible to train more profound and broad placing models that lead to improved recognition. By comparison, sequential architectures like LSTMs have sequential computation limits, which makes them slow to train and can be unable to scale to very long input sequences. Transformer-based ASR systems have significant problems despite these benefits. Their effectiveness is extremely sensitive to the presence of large-scale labelled data because self-attention models cannot learn powerful representations without large amounts of data. Transformers in low resources settings will not perform well or overfit as compared to basic recurrent models. Also, the computational and memory needs of Transformers are much larger especially in training and this can restrict their application in resource constrained contexts and in real time applications. In general, there is a trade-off of performance and requirements of the resources pointed out in the discussion. Although Transformer-based models can establish new

ISSN: 3071-3498
PP: 01-13.

standards in recognition accuracy and efficiency, LSTM and DNN-HMM models are still applicable to the cases with a limited amount of data or necessitating reduced computational power. These results underscore the need to base the ASR architecture choices on application needs, availability of data, and deployment factors and not only on the performance measures.

5. CONCLUSION

In this paper, a syntactic overview of deep learning-based solutions in Automatic Speech Recognition (ASR) was given, which massively highlighted the development of architectures, methodological schemes, and experimental performance results. Starting with conventional statistical models and proceeding to hybrid DNN-HMM models to the current end-to-end and Teformer-based models, the paper has highlighted the manner in which deep learning has truly transformed the ASR research environment. Deep neural networks have greatly enhanced recognition accuracy in a wide variety of speech circumstances by establishing automatic feature learning and efficient modeling of highly complex acoustic-linguistic interaction. The discussion showed that the end-to-end learning paradigms ensured that the complexity of systems has been lowered by eliminating human-created pronunciation dictionaries and language models that have been trained independently. LSTMs and Transformers have demonstrated impressive skills with regard to bubble-up capturing of change and extended context in support of substantial decrease in Word Error Rate. Specifically, Transformer-based models were found to be the best solutions because of their self-attention and the ability to train training in parallel, which contribute to the scalability and efficient optimisation on large datasets. These innovations have facilitated state of the art performance in both the academic standards as well as the real world speech enabling applications. Although these have been attained, there are a number of critical challenges that are unaddressed. Low resource languages have not been able to gain recognition performance compared to high-resource languages because of the scarcity of labeled data and linguistic diversity. Moreover, more complex deep learning models are associated with the issue of interpretability, transparency, and trustworthiness, particularly in areas where safety matters and regulation. The issue of efficient deployment is also a matter of concern as state-of-the-art ASR models can be computationally and memory-intensive, making them less applicable to real-time and edge-based systems. The future line of research is anticipated to be guided towards such issues by endeavoring into the areas of self-supervised and unsupervised learning methodology, which may make use of extensive amounts of untagged speech data to lessen annotation reliance. The multilingual and cross-lingual modeling approaches will probably be important in the development of a more inclusive ASR systems that can support a variety of languages and dialects. Also, future system development of lightweight and energy-efficient systems, such as model compression technology and knowledge distillation, will be required to allow real-time deployment on resource-limited devices. Comprehensively, further innovation of deep learning algorithms is likely to keep enhancing the accuracy, accessibility, and realistic implementation of the speech recognition technologies.

ISSN: 3071-3498
PP: 01-13.

REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.
[2] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Upper Saddle River, NJ, USA: Prentice Hall, 2001.
[3] S. Young et al., "The HTK Book," Cambridge Univ. Engineering Dept., Cambridge, U.K., 2006.
[4] J. Keshet and S. Bengio, "Automatic speech and speaker recognition: Large margin and kernel methods," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 6, pp. 533–550, 2011.
[5] F. Jelinek, Statistical Methods for Speech Recognition, Cambridge, MA, USA: MIT Press, 1997.
[6] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
[7] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 14–22, 2012.
[8] T. N. Sainath et al., "Convolutional neural networks for large-scale speech tasks," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 119–131, 2017.
[9] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE ICASSP, 2013, pp. 6645–6649.
[10] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," Proc. Interspeech, 2014, pp. 338–342.
[11] A. Graves et al., "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in Proc. ICML, 2006, pp. 369–376.
[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Proc. ICLR, 2015.
[13] C.-C. Chiu et al., "State-of-the-art speech recognition with sequence-to-sequence models," Proc. IEEE ICASSP, 2018, pp. 4774–4778.
[14] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
[15] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," in Proc. Interspeech, 2020, pp. 5036–5040.