International Journal of Communication and Information Technology

E-ISSN: 2707-6628
P-ISSN: 2707-661X
www.computersciencejournals.com/ijcit
IJCIT 2024; 5(2): 122-127
Received: 12-10-2024
Accepted: 18-11-2024

Charmy Khapra
Guru Gobind Singh Indraprastha University, Delhi, India

Corresponding Author:
Charmy Khapra
Guru Gobind Singh Indraprastha University, Delhi, India

A survey on end-to-end speech recognition systems

Charmy Khapra

DOI: https://doi.org/10.33545/2707661X.2024.v5.i2b.100

Abstract
Interest towards end to end SCR systems over the past few years have been growing because the system designs are more integrated and the methods have lower error rates than previous techniques. This survey also seeks to deliver a comprehensive understanding about these systems, analyse the improvement, status and potential developments in this field. Integrated systems avoid the lengthy process of converting sound to script by directly mapping the acoustic signal to the text string without the need for elaborate feature extraction. Some of the central sections that are discussed focuses towards the "Recurrent Neural Network (RNN)", "Convolutional Neural Network (CNN)", and transformer that have played crucial role in understanding optimization and advancement of accuracy for the recognition of speech. The survey also specifically discusses how these benchmarks datasets and evaluation methods are used to compare performance of these systems and offers information about these benchmarks. Some of the limitations that have been highlighted include; accent variation, noise and computational complexity with regard to the approaches used to address such concerns. Furthermore, the study delves into the implementation of end-to-end speech recognition systems in many fields, specifically focusing on the practical use of these systems in healthcare, telecommunications, and customer service industries. In order to explicate the path of "end-to-end systems of speech recognition" and their ability to transform the way man interacts with machines, this survey brings a review of the present literature and advancements.

Keywords: Recurrent neural networks, convolutional neural networks, human-computer interaction, end-to-end systems

1. Introduction
Integrative speech recognition systems are hybrid models that have recently been employed in the ASR domain as the improved end-to-end strategy as in comparison with the traditional approaches. End-to-end systems are free from cascaded processing stages and handbook extracted features unlike the traditional ASR architectures which possess numerous stages. This has been fostered by Neuronal Networks like the "Recurrent Neuronal networks (RNNs)", "Convolutional Neuronal networks (CNNs)" among others but not forgetting the recent transformers.
These models have shown very high improvements in terms of accuracy of the recognition as well as the robustness with different datasets and languages. This survey's objective is to present what is currently known about end-to-end systems for speech recognition, including its main elements, potential issues, and possible development trajectories. Thus, along with investigating the recent advancements and their usage in practice, this research aims to explore how end-to-end systems may reshape the existing perspective on the Human-Computer Interaction.

1.1 Background
The advances in the field of speech recognition have seen major developments where the modern more efficient and effectual approach of using machine learning techniques to develop speech recognition systems were developed to replace rule based systems. Most recognizers are implemented as a three-stage process, as example the language model, acoustic model, and feature extraction, which are usually fine-tuned by utilizing domain data and prior knowledge about the field. These conventional systems mainly face difficulties where there is a need for extensions and flexibility, especially in terms of different accents, languages, and harsh conditions (Audhkhasi *et al*., 2017a) [2].

Deep learning has indeed revolutionized ASR and has directed to project of end-to-end ASR systems. Such systems do not contain a number of processing stages as traditional ASR systems do; however, they share a single model which translates an input audio into text output. It also makes the architecture less complex and improves performance by utilizing large sale pre-trained neural networks with the large datasets. RNN and CNN paved the way to this transformation and transformer models have added a new boost to the progression (Prabhavalkar *et al*., 2017) [9].

Nevertheless, end-to-end systems have their drawbacks, e. g. they consume more computational resources and can be affected by differences in the people's speech and sounds in the environment. These question, this paper aims to discuss them, specifically, with details in the methods and technology advancements that enable the removal of these barriers, and thus captures the current and future roadmap of the end-to-end SR (Zweig *et al*., 2017) [12].

2. Literature Reviews

According to the Hori T, Prabhavalkar R, Schlüter R & Watanabe S, (2023) [8]. This research explores that Compared to the modelling which does not include deep learning, it has been seen that deep learning has effectively reduced the word mistake rate by over 50% relatively in the last ten years of research in ASR. Subsequent to this change a few "all-neural ASR architectures have been introduced".

"These so-called end-to-end (E2E) models provide fully neural, highly integrated ASR models which" do not largely depend on the ASR-specific knowledge, but rather they rely on generic machine learning.

They also learn from data more frequently and effectively, although they also tend to over-learn from data. As deep learning became all-pervading and more effective, and the general model architectures have been improved, E2E models are at present the most widely used strategy in the ASR approach.

"This survey wants to present a classification of E2E ASR models and associated developments, and" discuss most of their attributes and how they compare to conventional ASR frameworks rooted in HMMs. This paper conveys all the specific aspects of E2E ASR such as modelling, training, decoding and interfacing of outside language models; additionally it underlines and outlines performance and deployment directions and lastly makes a prognosis of future development.

As pointed out by Lv S, Wang D and Wang X, (2019) [11]. This research focuses on the importance of On-Break Voice Recognition "in the field of machine learning" specifically regarding Big Vocabulary Continuous Speech Recognition. Many people actually have known that "the Hidden Markov Model (HMM) Gaussian Mixed Model (GMM) is the speech recognition framework for a very long time".

Recently, the HMM deep neural network (DNN) model and deep learning end to end model have outperformed the HMM-GMM model. Despite using deep learning methodologies, the results of both models are similar. Nevertheless, the end-to-end strategy offers advantages such as concurrent training and the elimination of the need for data alignment. In contrast, the HMM-DNN strategy has several drawbacks, such as the need for forced data segmentation alignment, independent hypothesis development, and separate training of each module inherited from HMM.

Thus, an important area of the study of speech recognition is the "end-to-end model". This work focuses on the analysis of the end-to-end model as a process of transformation. This study first introduces the overview of the E2E model and HMM-based model, their advantages and shortcomings, and states that the speech recognition research is on the move to the latter type.

The paper then makes in-depth theoretical and experimental comparisons between three distinct end-to-end models: Three are "CTC based, RNN transducer, and attention based techniques". It also emphasizes on the theories, advancement, and the research trends of each of the models. Finally, the advantages and disadvantages of each is presented, followed by the possible enhancements for the end-to-end model in the future. An application of symmetry in computer science is used in automatic Mode of speech recognition; something that is a pattern recognition job.

Orken M, Dina O, Keylan A, Tolganay T, Mohamed O, (2022) [7] mentioned that this paper aims at studying the Transformer model, which is widely used in the voice recognition sector and also can be paralleled and has its self-attended internal component. Its benefits includes fast learning speed and the fact that it does not operate sequentially as the recurrent neural networks.

When establishing the Kazakh voice recognition automated system, the "Transformer models and an end-to-end model based on CTC were considered." Beside it is commonly known that Kazakh is rather limited in the amount of information to be used for voice recognition and is among the agglutinative languages. Some works reported "that the Transformer model improves system performance for low-resource languages". Our studies showed that the use of Transformer and the connectionist temporal classification significantly improved the results obtained by the Kazakh SR system. Comparing the results in different cases, the low CER values were obtained by using an integrated language model: 3. 7% in the clean environment.

2.1 Research Gap

"However, end to end speech recognition system has notable gaps in the research front". The first key issue is the ability of the system to process various accents/dialects and specifically noise which tend to directly lead to reductions in the overall performance. Also, these models require high computational capacity to execute which hampers their implementation on resource-limited devices. The fourth deficiency is in the availability of training data, specifically big, diverse datasets that are necessary for training of such models when address underrepresented languages. Closing these gaps is therefore important for improved scalability, accuracy, and fairness of end-to-end SR systems for application in real-world diverse scenarios.

2.2 Research Question

- What kind of differences are possible when the systems are compared concerning their approach to various accents and dialects?
- What are the key computation problems of "end-to-end ASR systems"?
- What strategies are there that could be applied to the training datasets "for end-to-end speech recognition systems to incorporate more minority languages".

2.3 Research Objectives

- "In order to assess the performance of E2E speech recognition, in stressing different accents and dialects".
- To propose a set of specific research questions so that the typical computational difficulties aforesaid systems encounter can be recognized and critiqued.
- In order to provide proposals on how the training datasets can be enhanced for inclusion of the less popular languages.

2.4 Research Limitation

This research is limited by the ethnic of diverse and large dataset which may cause a restricted generalization of the result to all the languages and accents. Also, there is a concern that since "end-to-end speech recognition technology is" a swiftly "developing field", conclusions made to this work may be rendered obsolete later by new advancements. Another factor that might prevent further experimentation with large scale model is the limitation of computational resources.

3. Research Methodology

The research methodology of this survey regarding "end-to-end speech" recognitions systems entails a review of the current literature concerning the systems, models, and approaches. The structure of the study will be based on the analysis of the peer-reviewed journal articles, conference materials, and other practical technical papers showing the current tendencies and difficulties in the field. To compare the efficacy of traditional and end-to-end systems, comparative analysis will be done across the cases; "while to gain a better understanding of the" effectiveness of the methods proposed in practice, relevant cases will be used for analysis.

3.1 Research Method and Design

The research technique of this survey on "the end-to-end speech recognition system" uses secondary and qualitative research methods to conduct an extensive study of the subject. Secondary research is a process of collecting commercial literature by reviewing academic and industry documents to obtain a general understanding of the available methodology and model, as well as their usage. This approach enables consolidation of information gathered from different sources, thus building strong analytical framework (Soltau *et al*., 2017) [10].

This involves the use of qualitative methodologies in which open ended interviews are conducted among the professionals in the field. Such structured interviews bring out rather subtle factors such as technologies that enable a particular practice, peculiarities of practice use, and its future development. This naturalistic qualitative research strategy seeks more descriptive and inclusive information that supports and expands on results obtained from the existing literature.

Thus, the work aims to implement the methodological approaches of secondary analysis and qualitative research to investigate latent patterns, personal views, and the absence of existing "end-to-end systems for speech recognition", and, thus, complement the understanding of its effectiveness and further development (Li *et al*., 2015) [6].

3.2 Research Approach and Analysis

"The survey on end-to-end speech recognition systems discovered that the field is anchored on deep learning models like recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformers". These systems cut down the amount of effort that is required for feature learning by mapping the audio signals directly on the text. It has been established that the proposed methods outcompete traditional methods in terms of precision and program speed particularly for people with different accents and dialects. But, the evaluation reveals inherent issues like, high computational complexity, robustness to noise, small training corpus especially for low resource languages.

Continuing studies are focused on the fine-tuning of these aforesaid systems by using data augmentation methods, improving the algorithms and advanced technologies of hardware. The survey reveals the possibilities "of the end-to-end SR systems in terms of reshaping the human–computer interface in different fields" and discusses the further development of the problem with regard to the improved performance in terms of noise immunity and Gestalt principles, as well as practical applicability to real-life scenarios (Kim *et al*., 2017) [4].

4. Data Analysis and Interpretation

4.1 Handling diverse accents and dialects in end-to-end speech recognition systems

When comparing the way that end-to-end systems acknowledge any variations in the accents and dialects in contrast to the traditional systems, the investigation of the data analysis exhibits valuable findings. It is quite conventional to face problems in pronunciation variations and other language differences and it is very cumbersome to fine tune and accommodate to different dialects.

"On the other hand, end to end systems are able to use deep learning models such as RNNs and transformers which can handle patterns of speech naturally". This allows them to handle different accents much better without lowering the features for this specific modality and not readjusting for the next language.

The interpretation of this conclusion suggests that the ability of end-to-end systems in managing Th. accents and other dialects highly depends on the section's variety and adequacy of the training sample. Models trained with datasets that cover a broad variety of accents are observed to generalize and conglomerate better in field conditions.

However, there are still some difficulties which are noticed most of all with the less popular dialects or languages – the question is that the availability of the training materials may be insufficient, which in its turn leads to the lowering of the recognition grade and mistakes.

Furthermore, the assessment highlights current work to improve the system's capability by incorporating data modification methodologies mimicking accent differences, along with the addition of more diverse speakers. These methods presume to expand the initial training sets and enhance the system's versatility when working with speakers from various linguistic backgrounds. Furthermore, efforts to improve transfer learning, domain adaptability, and other optimization approaches are being researched to enhance the system's performance and its ability to handle new accents when there is a severely limited number of corresponding samples available.

In conclusion, end-to-end systems provide better ways of coping with all the different forms of accent and dialect than the traditional methods, "the current research area focuses

on the general diversification of the training data and the best ways of adapting an STT system". The implication of the analysis is that future developments of the variations depend on the diversification and elaboration of the database to include all the variants of the linguistic expressions of the language under consideration to improve on the accuracy and applicability of the systems on the international arena.

Thus, according to the specified view, the analysis of the produced data is more important, given its "influence on the progress of the end-to-end speech recognition technologies to become more inclusive", with improved focus on performance reliability (Hori *et al*., 2017a) [4].

Computational Challenges in "End-to-End Speech Recognition Systems".

When evaluating the difficulties appearing throughout the end-to-end speech recognition system, several crucial components can be identified that define the system's efficiency and possibilities. One of the well-known issues considered here is the dramatic computational requirements that stem up from the complex structures "of deep learning technologies including recurrent neural networks (RNNs)", "convolutional neural networks (CNNs)", and transformers.

These models involve large-scale parameter estimates and computational operations such as matrix operations in the process of both training and inference. Therefore, running such systems on limited devices, or real time applications may proof to be costly or even impossible.

More analysis on the data indicates the effects of model complexity and its size on further worsening the computational problems. Models which have more layers and more parameters in general provide better performance, but possess the requirement for more computing resources. This trade-off requires solution for optimizing the models, usually through techniques such as model pruning, quantization and more efficient use of hardware.

Furthermore, it, again, stresses that data preprocessing and feature extraction are helpful in reducing computational loads. Where the end-to-end systems are enhanced due to a combination of these tasks within an end-to-end system, there are pre-processing sub-tasks that are time-consuming including noise reduction and feature normalization, especially when dealing with noisy inputs or input of variable quality.

Interpreting the data also put focus on the continuous exploration into ways of solving these difficulties through further enhancement on the algorithms and hardware systems. For example, improvements in the parallel computing structures, utilization of specific chips "such as graphic processing units (GPUs)" and "tensor processing units (TPUs)", and distributed computing platforms increase the performance and speed of the training and deployment steps of a model.

Finally, end-to-end ASR models have several advantages in terms of accuracy and model simplicity over the popular architectures, but they also imply serious computational difficulties that prevent the creation of large-scale and scalable models. According to the analysis, it is recommended that the future studies emphasize further refinement of model structures, improvement of algorithmic speed, and utilization of computing technologies in the improvement of performance at the same time increasing effectiveness of computations. Thus, solving these problems helps researchers work on the creation of the end-to-end systems that will provide simple, rational and effective vocal recognition for as many kinds of tasks and in various conditions as possible.

4.2 Improving training datasets for underrepresented languages in end-to-end speech recognition

It is important to understand that the difficulty of enhancing training datasets for underrepresented languages "in end-to-end speech recognition optimization" is consequently apparent when examining "the current state of the art. The first problem found is the lack" of large scale and variegated corpora which truly represent the multilingual scenarios of the given number of languages and dialects.

Most of the current datasets are mostly formulated in large languages, ignoring those languages that are minor or low-resource languages, and dialects. Hence models trained on such limited corpora sometimes perform very poorly and with lower accuracy in the under-represented languages.

As with the previous analysis interpreting the data further brings out the need to embrace data augmentation techniques in order to counter these limitations. Data augmentation in forms of synthetic data, transfer learning from similar languages and crowd sourcing are instrumental in enhancing training datasets for low resource languages. These approaches try to replicate language changes accounting for the generalization of the models and increasing the recognition rates when used in other languages (Orken *et al*., 2022) [7].

Furthermore, the analysis points at the presence of bottom-up approaches and partnerships in the increase and management of datasets for languages with limited resources. Intersectoral collaboration between academia, industry, and the linguistic populations provide ways of obtaining speech data and progressing in the creation of more inclusive speech recognition systems.

Also, the data further under evaluation strategies' interpretation highlight that there is constant research towards crafting better assessment standards and measures specific to minority languages. "These metrics try to evaluate the effectiveness of end-to-end systems in a more realistic way for different languages so the quantity and quality of the information that has been delivered will inform about the improvements and failures of the model".

Altogether, it is obvious that it is possible to improve training data sets and datasets of underrepresented languages in the end-to-end ASR systems further; however, the difficulties mentioned above still exist. The discussion of the findings indicates that the future research directions should involve an increase in the efforts aimed at the collection of the data, the improvement of the methods used in data augmentation, and the development of the equal ground for all languages and "dialects for the advancements in the field of the speech recognition technology".

Their solutions target to address these challenges in a bid to enhance put into operation of such systems as well as enhance inclusiveness with a view of catering for the worlds discuss diverse linguistic communities.

5. Results

By evaluating the survey on "end-to-end speech recognition systems and" the related findings adopted from big data analysis, several improvements, issues, and future trends are identified in the corresponding research area. Key findings raise awareness of how deep learning models like RNNs,

CNNs, and transformers paved the way to revolutionary ASR. They help in avoiding complicated processing stages that would otherwise have to be incorporated in building end-to-end systems that translate spoken language to text (Prabhavalkar *et al*., 2023) [8].

From the survey, authors were able to demonstrate how "end-to-end systems are more effective as compared to the conventional" ASR techniques especially in terms of efficiency and the capability to work with different intonations, accents and dialects.

These systems are also characterized by great scalability, and relying on the multiple-terabyte data sets and the neural networks, they demonstrate a high level of tolerance to variations in the patterns of speech and language, ideal for multilingual and multicultural environments. This adaptability is essential for improving users' perception and broadening the usage of the products and services offered in different areas and among differently aged people.

However, the survey also mentions major issues that affect the use and the enhancement of span-wide speech recognition systems. A major concern is that, deep learning models become tough computational models such that they are challenging to manage. All these models need plenty of computational resources to be trained and to be deployed, which reduces the scalability and the real-time requirements especially when the platform is used on low-memory devices.

The second recognized implication can be associated with the quality and the differentiation "of the training datasets". "As for the end-to-end systems which can effectively learn from the masses of data of the most commonly spoken languages", the problem of scarcity of annotated resources for under-represented languages and dialects persists. This may contribute to unevenness in the recognition accuracy and performance between two linguistic groups and show that there is more awareness needed in data-gathering and augmentation processes.

The interpretation of the data also marks continued research towards addressing these issues. Other approaches, including data augmentation, transfer learning and model compression are being applied in order to strengthen the performance and also the optimization of these "end-to-end systems". "These approaches are meant to expand the range of offered training datasets, enhance the models' ability to perform across different linguistic contexts, and increase their computational effectiveness", which in turn enables their adoption for various applications (Wang *et al*., 2019) [11].

Therefore, it can be concluded that despite "development of the end-to-end systems as the next step in the ASR advancement", the presented survey revealed that the field has multiple dimensions that are constantly evolving and experiencing both progress and pitfalls. These systems are also capable of using deep learning capabilities, which help in the system's ability to identify spoken words in as many languages and dialects as it is possible.

But, optimization of computational requirements and diversity of datasets are the key areas which require further improvement to bring the models into universal usability that can be opened for people in different countries as equally as possible. Following this, the future research directions regarding end-to-end ASR systems include system performance enhancement, dataset diversification, and technological development for popularization of ASR systems among society.

6. Conclusion

Therefore, the end-to-end speech recognition system is the major breakthrough of automatic speech recognition technology with simplified structures and higher recognition rates. It has been possible for these systems to be accurately efficient by the integration of other neural network models like RNNs, CNNs, and transformers. However, there are several constraints on the identified major areas in ASR, which include speech variability particularly in the different speech patterns, noise interferences, and also the level of computations the system has to perform.

Thus, the further work on these topics is needed, focusing on the improvement of models, their enhancement and the collection of larger, more diverse datasets. Thus, with further advancements, these systems can have a huge impact on the interaction between humans and computers, therefore improving the functionality of speech recognition for different applications and languages. This survey presents the status quo, shows the voids, and provides weights on potential research configurations for the ambitious goal evident in the current advancement of end-to-end SR systems.

6.1 Future Scope

The future developments for E2E ASR systems are the improvements in the model's ability to handle various accents and noisy background as well as optimize the resources needed for the implementation of the model for devices with constrained capacity, and expanding the training database for the languages that are limited or underrepresented in order to augment performance in a number of applications.

7. References

1. Audhkhasi K, Kingsbury B, Ramabhadran B, Saon G, Picheny M. Building competitive direct acoustics-to-word models for English conversational speech recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018 Apr 15-20, Calgary, AB, Canada. IEEE, 2018, p. 4759-63. DOI: 10.1109/ICASSP.2018.8462432.
2. Audhkhasi K, Ramabhadran B, Saon G, Picheny M, Nahamoo D. Direct acoustics-to-word models for English conversational speech recognition. arXiv preprint arXiv:1703.07754, 2017, p. 959-63. Available from: https://arxiv.org/abs/1703.07754.
3. Hori T, Watanabe S, Zhang Y, Chan W. Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM. arXiv preprint arXiv:1706.02737, 2017. Available from: https://arxiv.org/abs/1706.02737.
4. Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017 Mar 5-9, New Orleans, LA, USA. IEEE, 2017, p. 4835-9. DOI: 10.1109/ICASSP.2017.7953091.
5. Li J, Ye G, Zhao R, Droppo J, Gong Y. Acoustic-to-word model without OOV. In: Proceedings of the 2017 IEEE Automatic speech recognition and understanding

workshop (ASRU), 2017 Dec 16-20, Okinawa, Japan. IEEE, 2017, p. 111-7.
DOI: 10.1109/ASRU.2017.8268978.
6. Li J, Zhang H, Cai X, Xu B. Towards end-to-end speech recognition for Chinese Mandarin using long short-term memory recurrent neural networks. In: Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, 2015 Sep 6-10, Dresden, Germany. ISCA, 2015, p. 3615-9.
7. Orken M, Dina O, Keylan A, Tolganay T, Mohamed O. A study of transformer-based end-to-end speech recognition system for Kazakh language. Sci Rep. 2022;12(1):8337.
8. Prabhavalkar R, Hori T, Sainath TN, Schlüter R, Watanabe S. End-to-end speech recognition: A survey. IEEE/ACM Trans Audio Speech Lang Process, 2023.
9. Prabhavalkar R, Rao K, Sainath TN, Li B, Johnson L, Jaitly N. A comparison of sequence-to-sequence models for speech recognition. In: Proceedings of Interspeech, 2017 Aug 20-24, Stockholm, Sweden. ISCA, 2017, p. 939-43.
10. Soltau H, Liao H, Sak H. Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition. arXiv preprint arXiv:1610.09975, 2017, p. 3707-11. Available from: https://arxiv.org/abs/1610.09975.
11. Wang D, Wang X, Lv S. An overview of end-to-end automatic speech recognition. Symmetry. 2019;11(8):1018.
12. Zweig G, Yu C, Droppo J, Stolcke A. Advances in all-neural speech recognition. In: Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017 Mar 5-9, New Orleans, LA, USA. IEEE, 2017, p. 4805-9. DOI: 10.1109/ICASSP.2017.7952903.