A Survey on End-to-End Speech Recognition Techniques for Improving Model Performance

1st 57122404 WangShi Ying
Southeast University
Nanjing, China
City, Country
3415669069@qq.com

2nd 09022107 LiangYao Xin
Southeast University
Nanjing, China
City, Country
1208134539@qq.com

3rd 57122315 CuiJi Yuan
Southeast University
Nanjing, China
City, Country
1147904689@qq.com

4th 09022108 WangYe
Southeast University
Nanjing, China
City, Country
2686208951@qq.com

***Abstract*—The model of end-to-end speech recognition solves the problem of traditional speech recognition schemes in which each part of the task is independent and cannot be jointly optimized, and the new neural network structure can better utilize and adapt to the new hardware (e.g., GPUs) parallel computing capabilities, with faster computing speeds and better extraction of semantic features. We read more than a dozen related journal papers to conduct research on algorithm training and optimization of end-to-end speech recognition models, with a focus on analyzing four high-level papers . In the process we gained a deeper understanding of end-to-end speech recognition and had many thoughts about its application and significance in the future.**

***Index Terms*—End-to-End, Speech Recognition, Data Augmentation**

I. Introduction

The current traditional speech systems rely on carefully designed processing pipelines, and when used in noisy environments, these traditional systems often perform poorly. And most ASR (Automatic Speech Recognition) systems involve acoustic, pronunciation, and language model components, which require separate training and high cost of money and time. At the same time, development requires organizing pronunciation dictionaries, and defining phoneme sets for specific languages requires professional knowledge, which is time-consuming Since the adoption of hybrid modeling based on deep neural networks (DNN) ten years ago, the accuracy of automatic speech recognition (ASR) has significantly improved. This breakthrough mainly involves using DNN to replace traditional Gaussian mixture models for acoustic likelihood assessment, while retaining all modules such as acoustic models, language models, and dictionary models, thus forming a hybrid ASR system. Recently, the speech community has made a new breakthrough by transitioning from hybrid modeling to end-to-end (E2E) modeling, using a single network to directly convert input speech sequences into output marker sequences. This breakthrough is more revolutionary as it overturns the modular modeling that has been used in traditional ASR systems for decades.

The end-to-end model has several main advantages over traditional hybrid models.

Firstly, the end-to-end model uses a single objective function that is consistent with the ASR objective to optimize the entire network, while traditional hybrid models optimize each module separately and cannot guarantee global optimization. Moreover, end-to-end models have been proven to outperform traditional hybrid models in both academia and industry.

Secondly, due to the end-to-end model directly outputting characters or even words, the speech recognition process is greatly simplified. In contrast, the design of traditional hybrid models is complex and requires a large amount of ASR expert experience knowledge.

Finally, due to the use of a single network in the end-to-end model, which is more compact than traditional hybrid models, the end-to-end model can be deployed on high-precision, low latency devices.

Our research direction is end-to-end speech recognition based on different methods, including deep convolutional neural networks, deep learning, and streaming end-to-end speech recognition for mobile devices, as well as related training and optimization models. The following is a rough overview of some of our research papers on end-to-end speech recognition.

II. Association

The first paper focuses on an end-to-end deep learning approach that can be used to recognize speech in English or Chinese, which are very different languages. By replacing a pipeline of hand-designed components with neural networks, end-to-end learning allows us to deal with diverse speech, including noisy environments, accents, and different languages. The key to this approach is the application of high-performance computing techniques so that experiments that previously took weeks can now be completed in days. This allows us to iterate faster and discover better architectures and algorithms. The results show that in some cases our system competes with manual transcription on standard datasets. Finally, we use a technique called Batch Dispatch to deploy the system using GPUs in a data center to serve users with low latency in an online environment. The introductory section describes the current state of the Automatic Speech Recognition (ASR) field, noting that hand-designed domain knowledge has been applied to existing ASR processes. By using end-to-end deep learning to train ASR models, the training process can be simplified and engineering steps such

TABLE I
THE CONNECTIONS BETWEEN THE FOUR PAPERS WE STUDIED

PUBLICATION	FEATURE
Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin	Proposing an end-to-end deep learning-based approach to handle diverse speech, including noisy environments, accents and different languages, by replacing the traditional component pipeline with a neural network. This approach uses large-scale training sets and high-performance computing techniques during training to accelerate the model.
Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition	Focusing on how to improve the robustness of end-to-end speech recognition systems against noise. The authors propose a joint training method based on a gated loop fusion algorithm to solve the speech distortion problem.This method enhances the robustness of the model to noisy speech, thus improving the recognition accuracy.
You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation	Looking at how data enhancement techniques can be used to improve the performance of automatic speech recognition (ASR), especially when dealing with low-resource tasks. The authors employ text-to-speech (TTS) techniques to build a TTS system and use synthetic speech extension data to train a recognition model.
Component Fusion: Learning Replaceable Language Model Component for End-to-End Speech Recognition System	Describing an approach called "Component Fusion" that integrates externally trained neural network language models into an attention-based ASR system. This approach solves the problem of limited learning of language models in ASR systems based on attention mechanisms, and thus improves the recognition accuracy.
Connection	These four papers are all research work around end-to-end speech recognition technology, and together they explore how deep learning methods can be used to improve the performance of automatic speech recognition systems.

as bootstrapping/alignment/clustering/HMM mechanisms required to build state-of-the-art ASR models can be eliminated. The authors benchmarked the system on several publicly available test sets with the ultimate goal of achieving human-level performance. To this end, the authors also measured the performance of crowdsourced workers on each benchmark test for comparison. The results show that our best Chinese speech system is more accurate than a typical Chinese speaker in transcribing expressions similar to short speech queries.

The second paper focuses on three mainstream approaches to improve the noise robustness of end-to-end speech recognition (ASR). The first approach is to add speech enhancement components to the ASR front-end, including spectral subtraction, Wiener filtering, and deep neural network-based speech enhancement methods. However, these methods cannot be optimized to the final objective, tend to lead to suboptimal solutions, and produce excessively smooth speech that is prone to speech distortion. The second approach is to use multiconditional training (MCT), which utilizes different types of data (clean and noisy speech) to train speech recognition models. However, MCT has increased complexity and computational cost and performs poorly under mismatched conditions. The third method is the joint training method, which jointly trains speech enhancement and speech recognition simultaneously. In order to solve the speech distortion problem, the thesis employs a gated recurrent fusion (GRF) algorithm to dynamically fuse noise and enhancement features. Meanwhile, the joint training algorithm is used to optimize the enhancement and speech recognition. Experimental results show that the proposed method reduces the relative character error rate (CER) by 10.02% compared to the traditional joint enhancement and transformer method that only uses enhancement features. In particular, the proposed method performs better in low signal-to-noise ratio situations. The contributions of the paper are twofold: first, the gated loop fusion algorithm is used to dynamically fuse noise and enhancement features to solve the speech distortion problem; second, the speech transformer and mono voice enhancement are applied to the joint training framework for the first time. Experimental results show that the proposed method achieves better performance on the AISHELL-1 Mandarin dataset, with a relative CER reduction of 10.02%. Especially, the proposed method performs better in the low signal-to-noise ratio case.

The third paper focuses on analyzing the disadvantages of end-to-end automatic speech recognition at this stage and introduces the use of data enhancement techniques to improve the performance of automatic speech recognition (ASR), especially when dealing with low-resource tasks. The authors present an approach that uses text-to-speech (TTS) techniques to build TTS systems and extends the data with synthetic speech to train recognition models. Comparing the impact of semi-supervised learning techniques, the authors explore the effect of using vocoders on the final ASR performance and compare the method with other work. The results show that the authors' method has a competitive WER of 4.3% on LibriSpeech test-clean and 13.5% on test-other. The whole

method can be considered as an effective application of end-to-end automatic speech recognition based on data augmentation.

The fourth paper focuses on a method called "Component Fusion" for integrating externally trained neural network language models (NN LMs) into an attention mechanism-based automatic speech recognition system. Traditional neural network-based ASR systems include multiple components, such as acoustic models, lexicons, and language models. In contrast, end-to-end ASR systems based on attention mechanisms integrate all the necessary ASR components into a single neural framework and have achieved state-of-the-art results on several speech tasks. However, the Language Model (LM) component of the Attention Mechanism-based ASR system has to be learned implicitly from transcribed speech data, which limits the possibility of utilizing large text corpora for improved language modeling. To address this issue, the authors propose the "Component Fusion" approach, which allows externally trained NN LMs to be combined with an ASR system based on attention mechanisms. In the training phase, an additional LM component is added to the ASR system and replaced by an externally trained NN LM in the decoding phase. The experimental results show that the Component Fusion approach outperforms two previous LM fusion approaches.

III. INTRIDUCTIONS TO PAPERS

Traditionally, the machine learning process usually consists of multiple steps, including data preprocessing, feature extraction, model training and evaluation. Each step usually requires manual design and optimization. But end-to-end machine learning eliminates these steps. In end-to-end machine learning, a neural network model can perform data preprocessing, feature extraction, and output prediction simultaneously. With end-to-end training, the model can learn higher-level representations and features from the raw data.

Automatic speech recognition, especially large vocabulary continuous speech recognition, is an important topic in the field of machine learning. For a long time, Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM) has been the dominant speech recognition framework. But after development, HMM-deep neural network (DNN) models and end-to-end models using deep learning have achieved performance beyond HMM-GMM. Both models use deep learning techniques and have comparable performance. However, the HMM-DNN model itself is limited by various disadvantages such as forced data segmentation alignment, independent assumptions, and separate training of multiple modules inherited from the HMM, whereas the end-to-end model has the advantages of simplified modeling, joint training, direct output, and no need for forced data alignment. Therefore, end-to-end modeling is an important research direction in speech recognition.

A. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin [1]

The research in this paper shows in detail the benefits of end-to-end deep learning in the field of language recognition. Through their work, the authors demonstrate that an end-to-end deep learning approach can be applied to two distinctly different speech sounds - English and Mandarin. Moreover, their system allows for low-cost deployment in online environments and low latency when serving large-scale users by using batch scheduling techniques with GPUs in the data center.

They improved the model in terms of architecture and algorithms. The architecture uses Recurrent Neural Networks (RNN), which is a type of neural network capable of processing sequential data. The entire architecture consists of several components which include convolutional input layer, recurrent layer (which can be unidirectional or bidirectional), fully connected layer and softmax layer. In this architecture, the input audio data is passed through the convolutional layer to extract features and then passed to the recurrent layer for sequence modeling. The looping layer can pass the information in chronological order (one-way looping layer) or it can consider the contextual information before and after at the same time (two-way looping layer). Next, the output features are passed through a fully connected layer for further processing and finally through a softmax layer for character sequence prediction.

To train this architecture, the authors used the CTC (Connectionist Temporal Classification) loss function, which is a commonly used loss function for sequence modeling tasks. By optimizing the CTC loss function, the model can predict the corresponding character sequence directly from the input audio.

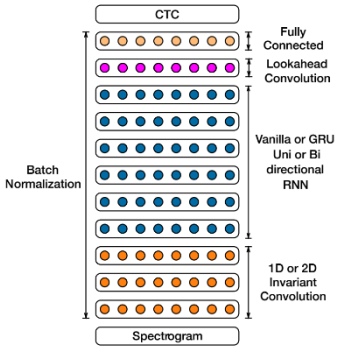

Fig. 1. Architecture of the deep RNN

Overall, because the authors' their end-to-end model predicts characters directly, new language-specific development is no longer necessary when performing language porting. By making minor changes, a Mandarin recognition system can be created quickly.

Since the ability to evaluate data and model assumptions depends on the speed of training, the authors they created a highly optimized training system based on a high performance computing (HPC) infrastructure. First, the authors focused

on carefully optimizing the most important routines used for training. This was demonstrated by creating a custom AI1-Reduce code for OpenMPI to sum gradients across GPUs on multiple nodes, developing a fast implementation of CTC for GP, and using a custom memory allocator These techniques combined to achieve a final result of being able to maintain an overall 45% of the theoretical peak performance on each node. And, they developed a fast GPU implementation that reduced overall training time by 10-20%.

Data training is very important in end-to-end speech learning. During the construction of the dataset, for a given audio-text pair (x, y), the most likely alignment is computed as

$$\ell^* = \underset{\ell \in \mathrm{Align}(x,y)}{\arg\max} \prod_{t}^{T} p_{\mathrm{ctc}}\left(\ell_t \mid x; \theta\right)$$

This is essentially a Viterbi alignment found using an RNN model trained using CTC, which produces an accurate accompanying alignment when using a two-integer RNN. In addition, they provided a large amount of training data, using 11,940 hours of labeled speech containing 8 million segments to train the English model, and 9,400 hours of labeled speech containing 11 million segments to train the Mandarin model.

The authors of this paper have greatly enhanced the capabilities of speech recognition systems through end-to-end deep learning. Their results confirm and exemplify the value of end-to-end deep learning methods for speech recognition in multiple environments. End-to-end learning will continue to be promoted.

B. Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition [2]

In the previous post End-to-End Language Learning, it was clear that end-to-end language learning has a great advantage when dealing with one or even multiple languages. But there is a prerequisite for this advantage, which is that it must be a clean speech signal. If it's a noisy environment, performance can drop dramatically. And in real environments, recorded speech is always subject to a lot of interference.

In order to solve this problem, there are three mainstream approaches, the first one is to add speech enhancement components to the ASR front-end, the second one is to use multi-conditional training (MCT) to improve the robustness of ASR noise, and the third one is the joint training method.

However, all three methods have their own shortcomings. The authors of this article propose a gated recursive fusion (GRF) with a joint training framework for achieving robust end-to-end ASR.

$$\begin{aligned} |\widetilde{X}| &= \mathrm{Enhancement}(|Y|) \\ \widetilde{O}_{\text{enhanced}} &= \mathrm{Fbank}(|\widetilde{\mathrm{X}}|) \\ \mathcal{L}_{\mathrm{ASR}} &= -\ln P\left(S^* \mid \widetilde{O}_{\text{enhanced}}\right) \end{aligned}$$

Above is the schematic diagram and formula of the traditional joint training method, Enhancement(-) denotes the speech enhancement function. function, which aims to find out the

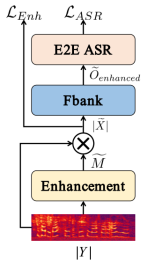

Fig. 2. The schematic diagram of conventional joint training method for robust end-to-end ASR

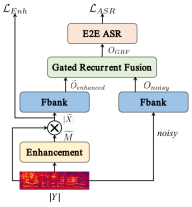

Fig. 3. The schematic diagram of authors proposed joint training method for robust end-to-end ASR.

clean target speech from the noisy input —Y—, $|\tilde{X}|$ refers to the enhanced speech, Fbank(-) denotes the function to extract the Fbank feature, which converts $|\tilde{X}|$ to Oenhanced, and LASR is the loss function of end-to-end ASR.

So, the traditional joint training method is divided into two parts: a speech enhancement part and a speech recognition part. First, noise and clean parallel data are used to train the speech enhancement model. Second, only the enhanced speech is used as an input feature for the speech recognition model. Finally, the enhancement (LEnh) and the total loss of speech recognition (LASR) are applied to optimize the whole model. Thus, the augmented and ASR models can be trained jointly. However, this approach only uses the enhancement features as inputs to the speech recognition model, which still suffers from the speech distortion problem to some extent.

To address this problem, this paper proposes a gated recursive fusion (GRF) method with a joint training framework for robust end-to-end automatic speech recognition.The GRF algorithm is used to dynamically combine the noise features and enhancement features. As a result, GRF not only removes the noise signal from the enhancement features, but also learns the original fine structure from the noise features, thus mitigating speech distortion.

This approach consists of three parts: speech enhancement, GRF and speech recognition.

First, a mask-based speech enhancement network is used to enhance the quality of the input speech. This network learns to generate enhanced speech signals based on the spectral characteristics and noise information of the input speech to reduce noise interference and improve speech intelligibility.

Secondly, GRF (Generative Relational Feature) is introduced to solve the problem of speech distortion. GRF is a generative model that reconstructs speech corrupted or distorted by noise by learning the probability distribution of the speech

signal. By using GRF, we can better restore the quality of the original speech and reduce the impact of noise, distortion and other problems on speech recognition performance.

The third part is to improve the performance of speech recognition (ASR), the authors have used a state-of-the-art speech transformer algorithm as a speech recognition component. This speech transformer algorithm converts the input speech into a more semantic and phonetically expressive underlying representation to better support ASR tasks. By introducing this powerful speech transformer as a recognition component, we can improve the performance and accuracy of the overall speech recognition system.

Finally the authors use a joint training framework to achieve optimization between these three components. This means that the speech enhancement, GRF and speech recognition components are trained in the same model and the performance of the whole system is maximized by sharing features and joint optimization. This joint training strategy can effectively improve the robustness and accuracy of the system. The authors conducted experiments on a Mandarin speech corpus named AISHELL-1. The experimental results show that the authors their proposed method achieves a reduction of 10.04% in the character error rate (CER) relative to the traditional joint enhancement and transformation method using only enhanced features. Especially at low signal-to-noise ratio (0 dB), the new method demonstrates better performance.

The authors of this paper have made significant progress in the field of speech enhancement, GRF and speech recognition, providing new ideas and methods to improve the performance of speech processing and recognition.

C. You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation [3]

This is a paper about using text-to-speech data augmentation to improve end-to-end speech recognition, and in the title, it mentions some keywords, such as "Text-To-Speech Data Augmentation", which refers to a technique that converts text to speech and applies it to train an end-to-end speech recognition model, and "End-To-End Speech Recognition", which is a method that directly converts speech input to text output without an intermediate step. End-To-End Speech Recognition", which refers to a technique that converts text to speech and is used to train end-to-end speech recognition models, and "End-To-End Speech Recognition", which is a method of speech recognition that directly converts speech input to text output without intermediate steps, simplifying the system architecture. It uses deep learning techniques to fuse acoustic features and language models into a unified neural network model. This approach is better able to handle diverse speech inputs and language variants. After studying this paper using a three-step reading method, I gained some knowledge.

First of all, Automatic Speech Recognition (ASR) technology is a recent achievement in scientific research, and the end-to-end method studied in this thesis is the only method for it besides the hybrid method. The performance of both methods is similar, but the end-to-end method shows a relative disadvantage when there are fewer resources, so we can realize the performance improvement of the end-to-end method by data augmentation techniques. This paper focuses on acquiring additional data in this way, in the traditional sense it includes three approaches, semi-supervised learning, transfer learning and active learning, in this paper the authors propose a new idea called speech synthesis technology (TTS). This technique is mainly applied to have applications in human-computer interaction with ASR data enhancement. The aim of this thesis is to compare the efficiency of the above newly proposed method, i.e., TTS technique, with the traditional method, semi-supervised learning, for application in ASR, which consists of the following preparatory work:

The ASR direction contains three segments: acoustic modeling, language modeling (LM), data preprocessing and enhancement, as well as some specialized methods such as Transformer, LSTM prediction, etc., and the TTS direction uses a dual-network setup (divided into a dual network of synthesizers and vocoders) approach containing preprocessing (G2P vocal tract processing), synthesizers (Tacotron modeling), and voice changers The three segments are used by the authors to predict and improve their own experiments based on analyzing the strengths and weaknesses of previous experiments, and ultimately drawing conclusions to refine the study.

This study investigated speech recognition using text-to-speech for data augmentation. The researchers used a 100-hour subset of LibriSpeech to simulate a low- to medium-resource environment. They used GMVAE-Tacotron as a speech synthesizer and a modified LPCNet as a vocoder to generate 360 hours of synthetic speech containing random rhymes that were modeled by a variational autoencoder. By adding the synthesized speech, they succeeded in improving the robust end-to-end ASR benchmark results by 39% in relative word error rate (WER) on the test clean dataset and 21% on the test other datasets. The researchers' approach excelled in absolute WER and relative WER improvement and achieved competitive results on the LibriSpeech train-clean-100 dataset, with a WER of 4.3% on the test-clean dataset and 13.5% on the test-other dataset. In addition, their experiments showed that using TTS enhancement was more successful on the test-clean dataset and less effective on the test-other dataset compared to semi-supervised learning. In the future, the researchers plan to address the issue of accent domain transfer to improve performance on the test-other dataset and close the gap with the test-clean dataset. In the future, they also plan to evaluate their approach in a non-simulated, low-resource environment.

D. COMPONENT FUSION: LEARNING REPLACEABLE LANGUAGE MODEL COMPONENT FOR END-TO-END SPEECH RECOGNITION SYSTEM

Recently, attention-based end-to-end automatic speech recognition (ASR) systems have shown promising results. However, these systems still have limitations in effectively utilizing large textual corpora, as discussed in the third paper. To

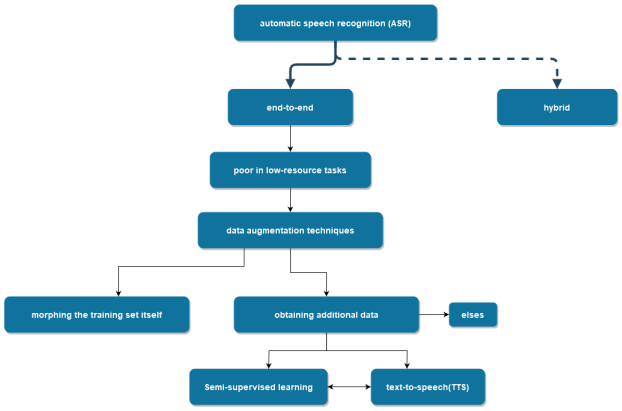

Fig. 4. Structural organization of the third article

address this issue, the paper proposes a novel approach called ”component fusion,” which involves incorporating externally trained neural network (NN) language models (LMs) into the ASR system.

The main body of the paper is divided into six parts. The introduction provides an overview of the traditional neural network-based ASR system and compares it with the attention-based ASR system discussed in this paper. It highlights the shortcomings of the attention-based system in effectively utilizing text corpora for LM training. The introduction also analyzes the methods previously used by other researchers and introduces the paper's own approach, the component fusion technique. Part II of the paper introduces the Attention-based ASR base model, outlining its key features and components. Part III provides a comprehensive review of existing methods for incorporating external LMs, trained from large text corpora, into attention-based end-to-end systems. The methods discussed include baseline, Shallow Fusion, Deep Fusion, and Cold Fusion.

Part IV describes the proposed ”Component Fusion” approach in detail. In this approach, the external LM is trained on the transcribed part of speech, enabling quick convergence and improved performance on the training data. The advantage of this approach is that it enhances LM recognition performance and allows for the replacement of the external LM with a completely different LM, facilitating fast domain adaptation. Part V presents the experimental results. Two datasets were collected, consisting of monosyllabic Mandarin and English datasets as source data, and a Mandarin-English code dataset as target data. The performance of incorporating LMs into the attention model was explored in both out-of-domain and in-domain scenarios. Positive conclusions were obtained, highlighting the effectiveness of component fusion. Part VI summarizes the work. The authors of this research introduce a method called Component Fusion, which aims to integrate an external language model (LM) into an attention-based model for automatic speech recognition (ASR). This approach enables the utilization of large-scale text training data and facilitates quick adaptation to specific domains within an end-to-end ASR system.

The effectiveness of Component Fusion is evaluated in two scenarios: out-of-domain and in-domain. In both cases, Component Fusion consistently outperforms other fusion methods, such as Deep Fusion and Cold Fusion, with Shallow Fusion. The evaluation results consistently show that Component Fusion achieves the best performance compared to other fusion techniques.

In summary, this research proposes the Component Fusion approach as a solution to the limitations in utilizing large textual corpora in attention-based end-to-end ASR systems. By integrating an externally trained LM into the ASR system, the proposed method enables improved LM recognition performance and rapid adaptation to specific domains. The experimental results demonstrate the superiority of Component Fusion over other fusion techniques in various scenarios. This research contributes to advancing the field of automatic speech recognition and provides valuable insights into the integration of external LMs into attention-based ASR systems. The Component Fusion approach presented in this paper paves the way for advancements in attention-based end-to-end ASR systems, enabling better utilization of large textual corpora, improved LM recognition performance, and efficient domain adaptation. Its potential impact extends to future ASR research and development, providing a valuable contribution to the field and setting the stage for further advancements in automatic speech recognition technology.

IV. FUTURE AND THOUGHTS

A. *Significance*

These works are of great significance in the field of Automatic Speech Recognition (ASR), mainly in the following aspects. First, by adopting new methods and techniques, such as end-to-end learning methods, deep neural networks, and data augmentation, the recognition accuracy of speech recognition systems can be effectively enhanced to overcome the shortcomings in the traditional component pipeline, thus improving the adaptability of speech recognition systems under diverse speech environments, noise, and accents. Second, these works focus on improving the robustness of speech recognition systems so that they can still recognize speech correctly in the face of noisy environments and other disturbances. This is important for speech recognition systems in real-world application scenarios, such as cell phone assistants, smart homes, and other domains. In addition, due to the limited nature of training data, how to expand the training data is a key issue to improve the performance of speech recognition systems. One of the papers proposed an approach to synthesize speech using text-to-speech (TTS) technology and use it as extended data for training models, which utilizes a large amount of synthesized speech data for training and can improve the performance of speech recognition systems. Finally, one of the papers proposes an approach to integrate externally trained neural network language models into an ASR system based on attention mechanisms. This integration can provide more

TABLE II
COMPARISON OF ENGLISH WER FOR REGULAR AND NOISY DEVELOPMENT SETS ON INCREASING TRAINING DATASET SIZE.

Fraction of Data	Hours	Regular Dev	Noisy Dev
1%	120	29.23	50.97
10%	1200	13.80	22.99
20%	2400	11.65	20.41
50%	6000	9.51	15.90
100%	12000	8.46	13.59

robust and enriched language models, which can improve the recognition performance.

Overall, the significance of these works is to improve the accuracy, robustness, and adaptability of automatic speech recognition systems under complex conditions, which further extends the scope and reliability of speech recognition technology in practical applications.

B. Weakness

- High data requirements:
  as mentioned in DEEP SPEECH2, end-to-end speech learning typically requires large amounts of labeled data for training. For general large-scale speech datasets, such as LibriSpeech, thousands to millions of hours of speech data are needed for training. Again, for example, the model for end-to-end speech recognition in the first paper has the following results with different training data: As you can see from the table, for every 10-fold increase in training set size, the WER decreases by a relative 40%. Thinking in the other direction, if the training size is too small, then the results will be greatly reduced. And it may be difficult to find a sufficient amount of training data for a faculty-specific domain or dialect.
- Lack of interpretability:
  Since the models used in end-to-end speech learning are typically complex neural networks, their internal representations and decision-making processes are often black-boxed and lack interpretability. This makes it difficult to understand how the models perform speech processing and make predictions. This can lead to problems such as difficulty in debugging and improving the model , and potentially difficulty in gaining user trust and acceptance. Not to mention the fear of this unexplainable "monster" that is now being created by the rapid development of AI.
- High model complexity:
  End-to-end speech learning often uses complex models such as deep neural networks, e.g. Recurrent Neural Networks (RNN). These deep neural network models typically have a large number of parameters and require more computational resources and training time to train. Training deep end-to-end models may require the use of large-scale datasets as well as running on high-performance computing devices, such as GPUs, to speed up the training process. This can become a challenge for some resource-constrained application scenarios or devices.

In the course of our research, we found many papers that improved the relevant defects, such as "You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation". For example, in "You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation", they built a TTS system on top of ASR training data, and then extended the data with synthesized speech in order to train recognition models. This system improves on the disadvantage of end-to-end speech recognition that requires large amounts of training data. When the amount of training data is low, this approach allows the end-to-end model to reach the quality of a hybrid system. I believe that with the efforts of countless outstanding researchers, the shortcomings of end-to-end speech recognition will continue to be improved and become more powerful and adaptable to more changing situations.

C. Prospect

First of all, after the proposal of various models, speech recognition technology will improve the accuracy and robustness, due to the cost and limitations of the acquisition of labeled data, its existing shortcomings of not being able to process information on the basis of a large amount of data are weakened by the application of various enhancement models, and the advantages that it originally had are further expanded, and future research will focus on how to make full use of methods such as migration learning and unsupervised learning to make full use of limited data resources, whether it is data augmentation, noise algorithms or data synthesis techniques, can improve the accuracy and robustness of the technology in a variety of complex situations such as noisy environments, helping to accelerate the deployment of speech recognition systems and the expansion of the range of applications. The application of this technology will become more common and add convenience to people's lives.

Secondly, speech modeling has been developing rapidly at present, and for the ever-changing Internet era, multimodal fusion has been adopted and applied in various fields, and with the rapid development of virtual assistants, smart homes and smart driving, speech recognition may be fused with other perception modalities (e.g., images, gestures, etc.), enabling the system to better understand the user's intentions and needs, and to provide more personalized and intelligent services.

Finally, due to the application of big data and artificial intelligence, people are more concerned about privacy and security issues, and when the technology is upgraded to a certain threshold, future research will be more committed to designing more reliable and secure speech recognition systems that protect users' privacy and personal information, and that ensure the ability to explore the user's preferences, habits, and contextual information while ensuring both personalization and adaptability to provide customized and personalized speech recognition services, while at the same time adequately

protecting personal privacy and preventing unauthorized access to data during transmission and storage.

ACKNOWLEDGMENT

Thanks to Mr. Zhou's guidance, we have a deeper understanding of how to read and research a paper. In this process, we understood how to conduct preliminary research on a field and obtained a lot of related knowledge, which provided us with more opportunities to see and think!

REFERENCES

[1] Dario Amodei,Rishita Anubhai,Eric Battenberg,Carl Case,Jared Casper,Bryan Catanzaro,Jingdong Chen,Mike Chrzanowski,Adam Coates,Greg Diamos,Erich Elsen,Jesse Engel,Linxi Fan,Christopher Fougner,Tony Han,Awni Hannun,Billy Jun,Patrick LeGresley,Libby Lin,Sharan Narang,Andrew Ng,Sherjil Ozair,Ryan Prenger,Jonathan Raiman,Sanjeev Satheesh,David Seetapun,Shubho Sengupta,Yi Wang,Zhiqian Wang,Chong Wang,Bo Xiao,Dani Yogatama,Jun Zhan,Zhenyao Zhu, " Deep Speech 2: End-to-End Speech Recognition in English and Mandarin " ICML 2016: 173-182

[2] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu and Z. Wen, "Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 198-209, 2021, doi: 10.1109/TASLP.2020.3039600.

[3] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov and S. Rybin, "You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation," 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Chengdu, China, 2020, pp. 439-444, doi: 10.1109/CISP-BMEI51763.2020.9263564.

[4] C. Shan et al., "Component Fusion: Learning Replaceable Language Model Component for End-to-end Speech Recognition System," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5361-5635, doi: 10.1109/ICASSP.2019.8682490.

[5] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng,"Deep Speech: Scaling up end-to-endspeech recognition"arXiv:1412.5567, 2014.

[6] P. Tzirakis, J. Zhang and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5089-5093,doi:10.1109/ICASSP.2018.8462677.

[7] Y. Zhang, W. Chan and N. Jaitly, "Very deep convolutional networks for end-to-end speech recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017, pp. 4845-4849, doi: 10.1109/ICASSP.2017.7953077.

[8] Y. He et al., "Streaming End-to-end Speech Recognition for Mobile Devices," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6381-6385, doi: 10.1109/ICASSP.2019.8682336.

[9] C. Shan et al., "Component Fusion: Learning Replaceable Language Model Component for End-to-end Speech Recognition System," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5361-5635, doi: 10.1109/ICASSP.2019.8682490.

[10] Ying Zhang,Mohammad Pezeshki,Philemon Brakel,Saizheng Zhang,Cesar Laurent Yoshua Bengio,Aaron Courville, " Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks " arXiv:1701.02720

[11] Wang, D.; Wang, X.; Lv, S. An Overview of End-to-End Automatic Speech Recognition. Symmetry 2019, 11, 1018. https://doi.org/10.3390/sym11081018

[12] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen and S. Zhang, "Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring From BERT," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1897-1911, 2021, doi: 10.1109/TASLP.2021.3082299.