Journal of Engineering Science and Technology Review 17 (6) (2024) 1-8

Research Article

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

Speech Recognition Method Based on GraphRAG

Wei Zhao[1,*] and Rongsheng Zhao[2]

[1]Hebei Minzu Normal University, Hebei 067000, China
[2]PetroChina Company Limited, Jilin 138000, China

Received 2 October 2024; Accepted 19 December 2024

Abstract

Speech recognition based on large neural network models requires vast computing resources and storage, and this requirement severely limits their direct application in resource-constrained edge devices. To considerably reduce the size and complexity of models while retaining the core performance of the original models, this study proposed a novel graph representation augmented with knowledge distillation (GraphRAG) speech recognition method based on adaptive multilevel distillation pruning (AMDP). First, graph neural networks were used to build and optimize knowledge graph embeddings to enhance the model's semantic understanding of speech signals. Second, an adaptive multilevel distillation mechanism was introduced to build a teacher–student architecture composed of multiple models of different sizes. The large bottom-layer model served as the main teacher, and the student models at each layer sequentially learned from the teacher model of the previous layer. Third, after each round of distillation, the student model was refined and pruned to further compress the model size while maintaining high recognition accuracy. Last, detailed experimental tests were conducted on multiple benchmark speech recognition data sets to verify the effectiveness of AMDP-GraphRAG technology. Results demonstrate that, (1) AMDP-GraphRAG consumes minimal computing time and memory while ensuring a low word error rate (WER). WER with and without a graph is 8% and 10%, respectively. The memory usage of the baseline is about 500 MB, whereas the memory usage of AMDP-GraphRAG is about 350 MB, which is the lowest among all the values produce by all the compared models. (2) The AMDP-GraphRAG model can maintain or improve recognition accuracy while reducing the number of parameters. The baseline has 120 million parameters, and AMDP has only 10 million parameters. Moreover, the baseline's WER is 3.5%, and that of AMDP-GraphRAG is 2.8%. (3) Atlas information can enhance the generalization ability of the models. Models with graph information can adapt quickly and show improved recognition accuracy. This study can considerably reduce the scale of models, decrease computational complexity and resource consumption, and serve as a reference for improving the application of speech recognition technology in edge devices.

Keywords: GraphRAG, Speech Recognition, Distillation, Pruning

1. Introduction

With the rapid development of artificial intelligence (AI) technology in recent years, deep learning, as one of the core driving forces of AI, has demonstrated breakthrough progress in many fields, particularly in speech recognition [1][2]. The accuracy and robustness of speech recognition systems have improved remarkably because of the powerful representation learning capabilities of deep learning algorithms and the support of massive data [3][4]. These systems are widely used in smartphones, smart homes, self-driving cars, and medical assistance apparatuses because they improve people's life experience and work efficiency.

However, such massive technological progress creates a challenge, that is, the high computing resource and storage requirements of large neural network models severely limit their direct application to resource-constrained edge devices. To overcome this problem, the research community and various industries have actively explored model compression and optimization technologies that retain the core performance of the original model while considerably reducing the size and complexity of the model, thereby achieving efficient, lightweight deployment [5]. Knowledge distillation and model pruning, which are mainstream technical methods, have attracted much attention and achieved remarkable results [6]. Knowledge distillation allows the teacher model (a large model with superior performance) to guide the learning of the small student model, which can inherit the key knowledge of the teacher model and achieve balance between performance and scale. By removing redundant parameters or structures in the model, model pruning directly reduces the volume and calculation amount of the model to further improve its operating efficiency.

This study proposed an innovative model compression technology called adaptive multilevel distillation–graph-based refined model aggregation and generalization (AMD-GraphRAG), which integrates an adaptive multilevel distillation strategy and a graph-based accurate model pruning method to further improve the performance and efficiency of lightweight speech recognition models. The core of AMD-GraphRAG technology is its multilevel distillation mechanism that can dynamically adjust the distillation strategy in accordance with the different levels of the model and the importance of features to ensure the effective transmission of key information. With graph-based model pruning technology, a detailed analysis of the internal structure of the model was conducted to identify and prune unnecessary or low-contribution components and further optimize the model structure. The application of AMD-GraphRAG technology is expected to substantially reduce

*E-mail address: she-shiv@163.com

ISSN: 1791-2377 © 2024 School of Science, DUTH. All rights reserved.
doi:10.25103/jestr.176.01

the size of the model, computational complexity, and resource consumption while maintaining or even enhancing the accuracy of speech recognition, thus facilitating the widespread application of speech recognition technology in edge devices. The study's results promote the popularization and development of smart Internet of Things (IoT) devices and provide strong technical support for realizing a highly intelligent and convenient lifestyle.

2. State of the art

This study focused on the application of graph embedding and deep learning in speech recognition and proposed a novel speech recognition technology on the basis of adaptive multilevel distillation pruning (AMDP). Hershey et al. proposed a deep clustering method, which has received widespread attention as a deep learning-based speech separation technology [5]. The method uses a neural network to learn the correlation between mixed speech and high-dimensional embedding space, clusters the embedding vectors in accordance with their corresponding speech sources, and employs the clustered embeddings to estimate independent speech sources. Specifically, a neural network is trained to graph mixed speech to a high-dimensional embedding space, and embedding vectors belonging to the same speech source are grouped together. Once the individual speech sources are estimated, they are separated using clustered embeddings. The permutation invariant training method proposed by Kolbaek et al. uses a neural network to generate a set of permutations for each speech source and selects the permutation with the smallest reconstruction error[1]. The task is challenging because of the diversity of speech patterns and accents and the potential interference from environmental noise. Given the advancement of machine learning algorithms, deep learning-based technology has been extensively applied in speech recognition and has achieved good results on various speech recognition benchmarks.

The connection temporal classification method proposed by Graves et al. is widely recognized as a major deep learning speech recognition approach [6].The method uses neural networks to learn direct mapping from the acoustic features of speech to text transcription, eliminating the need for explicit alignment. The Listen, Attend, and Spell technology proposed by Chan et al. is a common method that uses an attention mechanism to focus on different segments of the input speech signal during decoding [7]. With the advancement of deep learning technology, the performance of speech separation systems has substantially improved. Michelsanti et al. conducted an exhaustive analysis of contemporary methods and strategies for speech separation and used deep learning in their review [8]. In addition, Subramanian et al. proposed an end-to-end speech separation method that realizes direct mapping from mixed speech to a single speech source without intermediate processing stages[9]. Malik et al. described a joint study on speech separation and recognition in their review [10]. Despite the continuous development of research on the use of deep learning in speech separation and recognition, further work is still needed to improve the performance of systems [11]. Zhang Shaohua proposed a dual-channel network model on the basis of the squeeze–excitation attention mechanism and deep convolution to solve the problem of low recognition accuracy resulting from the inability of speech recognition to fully extract speech features[12].To address the Conformer encoder's insufficient ability to extract fine-grained local features of speech, Hu Conggang proposed a Conformer Chinese speech recognition method that integrates maximum pooling, it max-pools the output of the gated linear unit in the encoder convolution module to extract the fine-grained local features of a multiframe speech signal corresponding to a character [13]. Ou Jiale established a simple multimodal joint modeling framework that regards the joint modeling of speech translation and text translation as multilingual neural machine translation modeling, introduced modality-aware relative position encoding into the self-attention layer, and used a modality-aware single encoder to simultaneously realize speech and text translation [14]. Tian Sanli proposed a method called WLformer that integrates discrete wavelet transform (DWT) with end-to-end speech recognition to solve the problem of high computing resource usage of current end-to-end speech recognition models [15]. The model introduces the proposed signal compression module on the basis of DWT. This module compresses the representation by removing the high-frequency components with minimal information in the middle layer representation of the model, thereby reducing the computational resource consumption of the model.

To improve the accuracy of the mixed Chinese and English speech recognition system, Zhang Cong adopted a new E-Branchformer model that uses parallel branches to simultaneously extract global and local information and added depth convolution to the merging module to enhance the information fusion effect [16]. On the basis of the speech recognition optimization method called mel-frequency cepstral coefficients and the hidden Markov model (HMM), Guo Jiaqi introduced the expectation–maximization algorithm to optimize HMM and overcome the difficulty experienced by traditional HMM in recognizing complex speech environments [17]. In view of the limited modeling ability of convolutional neural networks (CNNs) for time series in automatic speech recognition process and the high computational complexity of deep CNN, Zhang Xuhang proposed a speech recognition model that integrates dilated CNN and a bidirectional long short-term memory network [18]. Kure and Dhonde proposed a solution that combines mel-frequency overlap transform with deep CNN for muscle rigidity speech recognition. The method demonstrates superior spectral–temporal representation of speech signals and outperforms traditional state-of-the-art techniques. This approach highlights the potential of using deep learning architectures to improve the accuracy and robustness of speech recognition systems. Meanwhile, Avila et al. explored the use of deep neural networks for speech emotion recognition on mobile devices and adopted modulation spectrum feature pooling [19]. Their experimental results showed that the proposed emotion recognition system outperforms the baseline algorithm in the 2016 Audio–Video Emotion Challenge in terms of the consistency correlation coefficient. This study highlights the adaptability of deep learning techniques to mobile platforms and their effectiveness in capturing subtle emotional information in speech.

Kristomo and Nugroho discussed speech signal classification through the fusion of time- and frequency-domain features [20]. They proposed three feature sets that use DWT, wavelet packet transform, and statistical methods to classify stop consonant word speech signals. Their work highlights the importance of multidomain feature extraction in improving speech signal classification accuracy and provides a powerful framework for integrating multiple signal processing techniques.

Comparison of these studies shows that deep learning technology is increasingly being used to solve various challenges in speech recognition. Kure and Dhonde and Avila et al. employed deep neural networks. Although their application domains differed (one for muscle rigidity speech recognition and the other for speech emotion recognition), their studies improve system performance through complex neural network architectures and demonstrate that deep learning is a versatile tool in the field of speech recognition. However, these studies had limitations, which can serve as future work directions. Although Kure and Dhonde and Avila et al. achieved considerable improvements, the computational complexity and resource requirements of deep learning models remain a challenge, especially when these models are applied on mobile devices in real time. Another issue is the reliance on specific data sets and the lack of generalization across different speech types and contexts. Kristomo and Nugroho addressed this problem to some extent by proposing multidomain feature extraction methods, but combining these methods with deep learning models to achieve broad applicability remains an unsolved problem.

To address these issues, this study optimized deep learning models to reduce the computational overhead, explored hybrid models that combine deep learning with knowledge graphs to enhance generalization capabilities, and developed a comprehensive evaluation framework that considers different speech types and environmental conditions [21][22].These advancements improve the performance of speech recognition systems and expand their application to real-world scenarios [23][24].

The remainder of this study is structured as follows. Section 3 introduces knowledge distillation-enhanced graph-based speech recognition (KDGS). Section 4 verifies the effectiveness of AMD-GraphRAG technology and presents detailed experimental tests on multiple benchmark speech recognition data sets. Section 5 summarizes this study. AMDP-GraphRAG has advantages in performance optimization, recognition accuracy, computing time, and memory usage. It can remarkably reduce computing time and memory usage while ensuring a low word error rate (WER).

3. Methodology

With the rapid development of natural language understanding and speech recognition, knowledge graph, as an effective semantic representation tool, has become an important bridge connecting text information and structured data. However, how to effectively integrate knowledge graph and speech recognition technology in practice remains challenging. This study explored a new framework that combines knowledge distillation strategies (i.e., KDGS) and experimentally verified its effect on improving speech recognition performance.

3.1 Knowledge graph embedding

For GraphRAG, graph neural networks (GNNs) were used to build and optimize knowledge graph embeddings for enhancing the model's semantic understanding of speech signals.

Graph convolution encoder: GNN was applied to encode the predefined knowledge graph and generate a series of entity vectors containing global context information. The purpose of this step is to capture the differences in how closely each lexical item is related to other words in the corpus so that the implicit semantic connections can be fully considered in the subsequent processing stages.

Attention-guided acoustic modeling: The sound signal obtained above was combined with the acoustic features of the current frame to form a new input. A multihead self-attention layer was used to calculate weight matrix A between positions, and the weighted average method was applied to obtain the final context-aware vector. The GNNs iteratively updated the embedding vector of an entity so that it contained the characteristics of the entity itself and reflected its correlation with other entities in the graph.

Suppose that $G=(V,E)$ is a knowledge graph, where V is a set of nodes (entities) and E is a set of edges (relationships). For each node $v \in V$, the update formula of embedding vector $h_v^{(l)}$ in the l round of iteration is as follows:

$$h_v^{(l+1)} = \sigma(W_h \cdot Aggr(h_v^{(l)} ; \\ \tilde{h}_u^{(l)} \mid M_r(u,v)^{(l)}]) + b_h), \tag{1}$$

where σ is the activation function; W_h and b_h are the weight matrix and bias term, respectively; $Aggr$ is an aggregation function (e.g., mean, max, or attention-based aggregation); $\tilde{h}_u^{(l)}$ is the embedding of neighbor nodes u adjusted by relationship matrix $M_r(u,v)^{(l)}$; and $M_r(u,v)^{(l)}$ is weight matrix corresponding to relationship type $r(u,v)$.

3.2 AMD

In AMD-GraphRAG, an AMD mechanism is introduced to flexibly adjust the distillation process in accordance with the needs of different model levels. The mechanism mainly includes the following key components:

Multilevel teacher–student architecture: A teacher–student structure composed of multiple models of different sizes is established. The large model at the bottom layer serves as the main teacher, and the student model at each layer learns from the teacher model at the previous layer.

Soft label of the teacher model (i.e., probability distribution of output): This label is taken as one of the learning goals of the student model in combination with the traditional hard label (one-hot encoding of the real category). The objective function at this stage can be expressed as

$$L_{KD} = \lambda L_{CE} + (1-\lambda) L_{KL} \tag{2}$$

where L_{CE} is the cross entropy loss, L_{KL} is the Kullback–Leibler divergence (measures the distance between two probability distributions), and λ is the proportional coefficient that balances the contributions.

Graph integration: In the student model, the GNN of the knowledge graph is introduced to use graph information in enhancing the language understanding ability of the model. This process involves graph convolution operations, which can be expressed as

$$H^{l+1} = f(\tilde{D} - \frac{1}{2}\tilde{A}\tilde{D} + \frac{1}{2}H^l W^l) \tag{3}$$

where H^l is the node feature matrix of layer l, $\tilde{A}$ is the normalized adjacency matrix, $\tilde{D}$ is the degree matrix, W^l is the weight matrix, and f is the activation function.

Hierarchical distillation strategy: The distillation process of each layer is fine-tuned on the basis of the characteristics of the teacher model of the previous layer and the capabilities of the current student model to ensure that each distillation can effectively transfer knowledge.

Adaptive distillation temperature adjustment: Distillation temperature (temperature scaling) is an important factor that determines the effectiveness of the distillation process. In AMD-GraphRAG, the distillation temperature is dynamically adjusted based on the training progress and performance of the student model to determine the optimal distillation configuration.

3.3 Multilevel model pruning

In addition to adaptive multilevel distillation, multilevel model pruning is integrated into AMD-GraphRAG. After each round of distillation, the current student model is fine-tuned to further compress the model size while maintaining high recognition accuracy. The objective function of model pruning is defined as follows:

$$F(S) = \lambda_1 L_{task}(S) + \lambda_2 L_{prune}(S) \tag{4}$$

subject to

$$0 < \theta_p < \theta_m < 1 \tag{5}$$

where $L_{task}(S)$ is the loss function of student model S in the task, similar to cross entropy loss. $L_{prune}(S)$ is the penalty term for model pruning, and it can control model complexity. λ_1 and λ_2 are weights used to balance task and pruning losses, respectively. θ_p and θ_m represent the minimum and maximum pruning thresholds, respectively, and are used to control the pruning degree.

4. Result Analysis and Discussion

4.1 Experimental design and data set selection

Exhaustive experimental tests were conducted on multiple benchmark speech recognition data sets, including LibriSpeech and TED-LIUM Release 3, to verify the effectiveness of AMD-GraphRAG technology. The experimental results demonstrated the considerable improvement of AMD-GraphRAG in terms of model size and computational cost and confirmed the advantages of the technology in recognition accuracy.

Fig. 1 compares the WER values of the teacher model, the student model, and the control model without graph information on different test sets. It reveals the performance gain brought by graph information and knowledge distillation. The chart focuses on the performance of different models in terms of WER and the comparison of the teacher model, the student model, and the control model without graph information.

The error rates of the teacher model, the student model with graph information, and the student model without knowledge graph information were 12%, 8.5%, and 10.2%, respectively. The chart reflects the value of graph information. The model using graph information exhibited a remarkable advantage over the model without graph information in reducing WER. This result shows that graph information can provide additional knowledge or context to the model, thereby helping improve recognition accuracy.

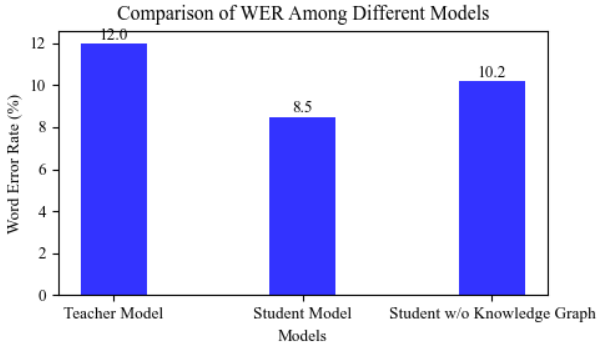

Fig. 1. Word error rate comparison

4.2 Analysis of the effect of knowledge distillation

By transferring the knowledge of a trained large model (teacher model) to a small model (student model), the student model, even though it is small, can achieve performance that is close to or even better than that of the original large model. This situation proves the effectiveness of knowledge distillation technology and shows that the method can considerably reduce model size without sacrificing performance.

As shown in the figure, graph information enhanced the model's ability to understand text, and the knowledge distillation methods achieved efficient miniaturized model construction. The combination improved the accuracy and efficiency of the model and made the final product suitable for deployment in resource-constrained environments. This finding has important implications for developing speech recognition solutions for practical applications.

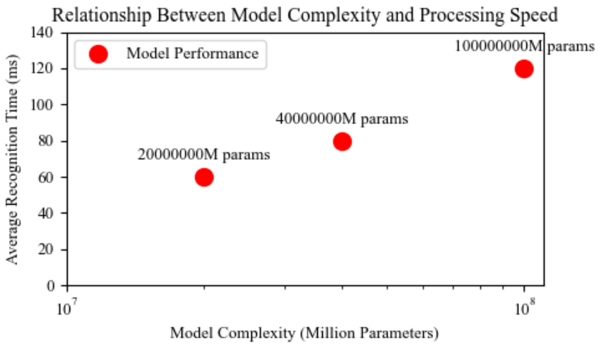

Fig. 2. Model size and computing speed

Fig. 2 shows model complexity (in terms of the number of parameters) versus average recognition time. It highlights the effectiveness of knowledge distillation in reducing model size while maintaining or even increasing the processing speed. The graphs focus on the relationship between model complexity (measured as the number of parameters) and average recognition time. The chart results were analyzed.

The average recognition time of the 10^8M, 4×10^7M, and 2×10^7M models was 120, 80, and 2×10^7 ms, respectively. The chart shows that when the number of model parameters decreased (i.e., the model became increasingly compact), the average recognition time decreased. This result shows that simplifying the model through appropriate technical means does not necessarily lead to performance degradation; it can also improve operating efficiency in many cases.

4.3 Analysis of the role of knowledge distillation

Knowledge distillation is a process of transferring knowledge from a large, complex model (teacher model) to a small, concise model (student model). This approach can effectively reduce model size without sacrificing accuracy. The miniaturized model, after knowledge distillation, remarkably reduced the demand for computing resources and processing time while maintaining or approaching the original high performance level.

For application scenes that require fast responses, such as voice assistants on mobile devices or real-time translation services, models that process fast and consume minimal resources are particularly important. Therefore, knowledge distillation technology provides strong support for the development of efficient, lightweight speech recognition solutions.

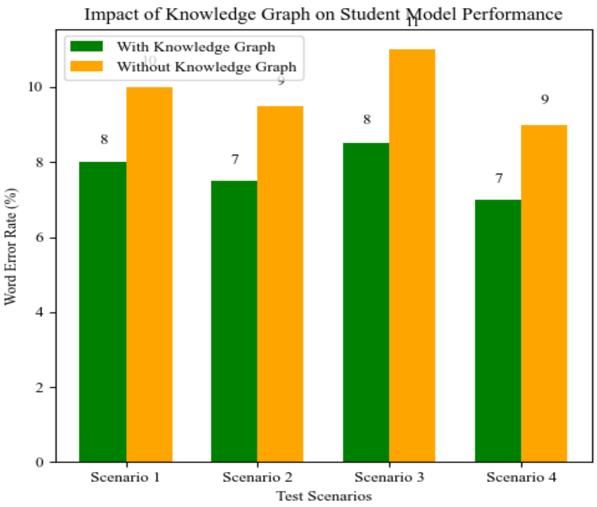

Fig. 3. Effect of graph information on model performance

The bar chart in Fig. 3 compares the performance of student models with and without graph information support in various test scenarios. It highlights the important role of graph information in improving speech recognition accuracy. The figure visually displays the performance of the student model in different test scenarios with and without the support of graph information in the form of a bar chart. Graph information refers to additional auxiliary information about the speaker, language environment, and background noise aside from the audio signal during the speech recognition process. It can substantially improve the model's understanding and speech signal recognition accuracy.

As shown in the figure, when the model was supported by graph information, its performance in all the test scenarios improved remarkably. The models showed considerable differences in the following aspects.

Recognition accuracy: The recognition accuracy of the models with and without graph information exhibited a notable difference. In Scenario 1, WER with graph information was 8%, and WER without graph information was 10%. In Scenario 2, WER with and without graph information was 7% and 9%, respectively. In Scenario 3, WER with and without graph information was 8% and 11%, respectively. In Scenario 4, WER with graph information was 7%, and WER without graph information was 9%. The models with graph information support showed high accuracy in all the test scenarios, indicating that graph information can provide additional contextual information to the model and help the model understand and process speech signals.

Scene adaptability: Graph information could also improve the scene adaptability of the models. In the different test scenarios, such as quiet environments, environments with background noise, and environments where multiple people are talking simultaneously, the models with graph information support could adapt to environmental changes and maintain high recognition accuracy.

Robustness: When faced with various challenges, such as the lack of atlas information and degradation of signal quality, the models supported by graph information showed high robustness. These models maintained stable performance even under nonideal conditions.

Generalization ability: Graph information enhanced the generalization ability of the models. When faced with new scenes or speakers, the models with graph information could adapt quickly and showed improved recognition accuracy, which is particularly important in practical applications.

To sum up, Fig. 3 highlights the important role of graph information in improving speech recognition accuracy by comparing the performance of models with and without graph information support in various test scenarios. This discovery is important for the design and optimization of speech recognition systems and reminds us that in practical applications, we should make full use of atlas information to improve the recognition performance of systems and user experience.

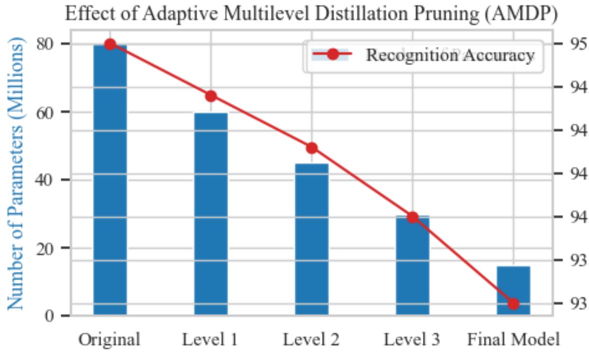

Fig. 4. Influence of graph information on model performance

The chart in Fig. 4 shows the change trends of the number of model parameters and recognition accuracy during the AMDP process. These trends are crucial to understanding how AMDP technology affects model size and performance. Analysis of the chart results revealed that as the AMDP process proceeded, the number of model parameters gradually decreased. The numbers of the original model, Level 1, Level 2, Level 3, and final model parameters were 80, 60, 40, 30, and 3 million, respectively. This finding shows that the pruning technique effectively simplified the model. This simplification helped reduce the storage requirements and computational cost of the model, making the model suitable for deployment on resource-constrained devices, such as mobile phones and embedded systems. The chart in Fig. 4 also shows that the model could still maintain high recognition accuracy even when the number of parameters was reduced. The AMDP method effectively removed the parts that had minimal effects on the final performance while retaining the important parameter structure. Analysis of the relationship between the number of model parameters and recognition accuracy indicated that the AMDP strategy maintained a sufficient performance level and achieved model light weighting.

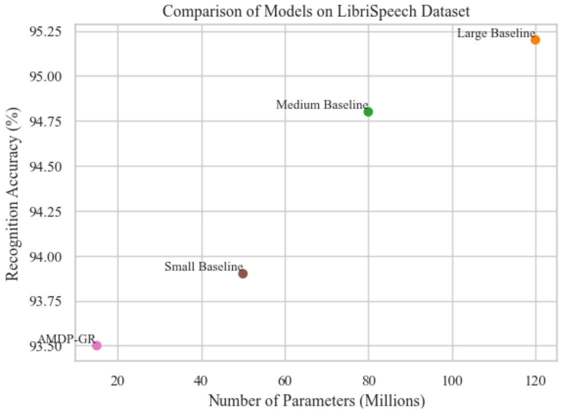

Fig. 5. Effect of graph information on model performance

The chart in Fig. 5 compares the performance of different models on the LibriSpeech data set. The chart also shows the number of parameters and recognition accuracy of each model on the LibriSpeech data set. The large baseline's volume was 120 million parameters, the medium baseline's volume was 80 million parameters, the small baseline's volume was 50 million parameters, and AMDP's volume was 10 million parameters. Typically, large models (with many parameters) are likely to have high accuracy because they can capture complex patterns. However, this trend is not absolute. The chart above shows that sometimes, through optimization techniques, such as knowledge distillation or pruning, small models can achieve performance that is close to or even surpasses that of large models, and their recognition accuracy is almost the same.

The baseline model was compared with the model version that uses new technologies (e.g., graph information enhancement and AMDP) to reveal the specific contributions of the aforementioned improvements to the enhancement of recognition accuracy. The AMDP-GraphRAG model could maintain or improve recognition accuracy while reducing the number of parameters.

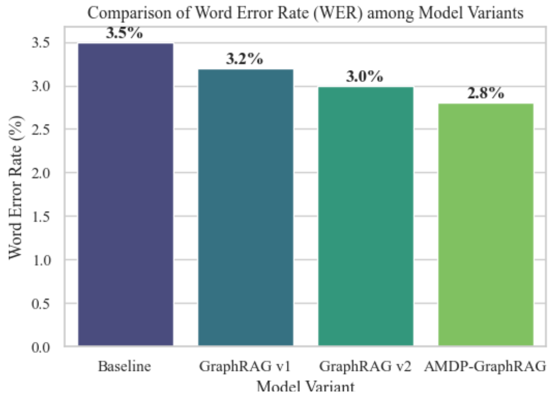

Fig.6. Comparison of the WER values of the baseline model and the GraphRAG variant

The histogram in Fig. 6 shows the WER gap between different model versions and intuitively reflects the role of AMDP technology in improving model performance. The WER of the models decreased with the upgrade of the GraphRAG version and the application of distillation pruning, which further proves the effectiveness of the technology. The WERs of the baseline model and different versions of GraphRAG are presented as a histogram. This intuitive representation helps understand the role of AMDP technology in improving model performance. The following text shows a specific analysis on the basis of the chart.

With the transition from the baseline model to the higher-level GraphRAG versions, a clear trend was observed: WER continuously decreased. The WER of the baseline was 3.5%, and the WER of GraphRAG v1 was 3.2%, which is slightly lower than that of the baseline. GraphRAG v2 had a WER of 3%, and AMDP-GraphRAG had the lowest WER of 2.8% among all the versions. This finding shows that each improvement step or the adoption of new strategies effectively improved the model's recognition accuracy. In particular, model performance considerably improved when the AMDP strategy was introduced. By presenting the WER gap between different model versions, the chart directly proves the effectiveness of the proposed technical solution. For example, if a specific version of GraphRAG exhibits a notable WER reduction compared with the previous version, then the new features or optimization measures added to this version can be considered to have played a positive role. The histogram in Fig. 6 shows not only the final result, but also the effect of each stage on the overall optimization process. It can help determine the specific steps that are critical to improving overall performance and provides valuable information for future work. Low WER means high speech recognition accuracy, which is crucial for developing reliable voice interaction systems. These results demonstrate that advanced technologies, such as GraphRAG and AMDP, can be used to develop efficient and accurate speech processing solutions without sacrificing user experience.

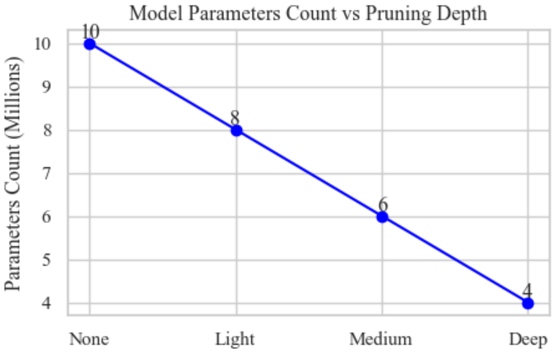

Fig. 7. Change trend of the model parameters with pruning depth

The line chart in Fig. 7 depicts the process where the number of model parameters decreases as the pruning depth increases. It highlights the important role of the AMDP strategy in model light weighting. This process is conducive to the deployment of models on low-power hardware platforms and identifies potential improvements in computing speed. Small models are easy to deploy on devices with limited computing resources, such as mobile devices and IoT terminals. A decrease in the number of parameters usually means reduced computational complexity, which helps improve model inference speed, reduce power consumption, and extend battery life. As shown in Fig. 7, when the pruning depth gradually increased from none to light, medium, and deep, the number of model parameters showed an obvious downward trend. For example, from none to light depth, the number of parameters decreased from 10 million to around 9 million. From light to medium depth, the number of parameters continuously decreased to around 8 million. From medium to deep, it decreased to around 6 million.

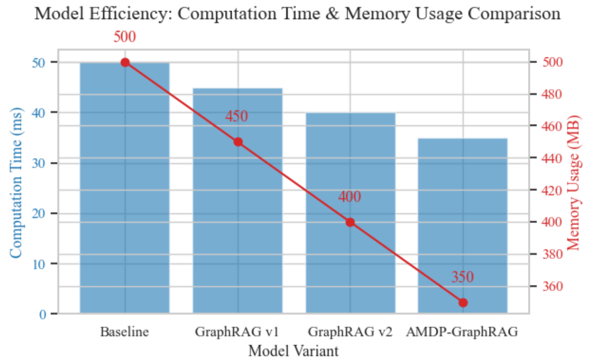

Fig. 8. Comparison of the computing speed and memory usage of different model versions

By combining model operation time and memory usage, the scatter plot in Fig. 8 shows the improvement in the efficiency and resource management of the different model versions. AMDP-GraphRAG achieved short computing time and low memory usage while ensuring low WER, revealing its advantages in the design of efficient speech recognition systems. The horizontal ordinate presents the different model variants, including GraphRAG v1, GraphRAG v2, the baseline, and AMDP-GraphRAG. The vertical ordinate shows the operation time and memory usage.

Computing time analysis: The computing time of the baseline was roughly 50 ms. The calculation time of GraphRAG v1 was approximately 48 ms, which was shorter than that of the baseline. The operation time of GraphRAG v2 was approximately 45 ms, which was also shorter than that of GraphRAG v1. AMDP-GraphRAG had the shortest computing time among all the models. Specifically, its computing time of 40 ms was about 5 ms shorter than that of GraphRAG v2, 8 ms shorter than that of GraphRAG v1, and 10 ms shorter than that of the baseline.

Memory usage analysis: The memory usage of the baseline was about 500 MB. The memory usage of GraphRAG v1 was 450 MB, which was slightly lower than that of the baseline. The memory usage of GraphRAG v2 was around 400 MB, which was also lower than that of GraphRAG v1. AMDP-GraphRAG had the lowest memory usage among all the models. Its memory usage of 350 MB was about 50 MB lower than that of GraphRAG v1, 100 MB lower than that of GraphRAG v2, and 150 MB lower than that of the baseline.

AMDP-GraphRAG showed obvious advantages in terms of calculation time and memory usage. While ensuring low WER, it considerably reduced the computing time and memory usage, making it highly competitive and useful for the design of efficient speech recognition systems. It is suitable for application in resource-constrained environments and improves the overall efficiency and performance of the system.

5. Conclusions

To explore speech recognition model compression and optimization methods and substantially reduce model size and complexity, this study combined model improvement comparison and experimental work to analyze a GraphRAG speech recognition method that is based on AMDP on the premise of retaining the core performance of the original model. The following conclusions were obtained.

(1) AMDG-GraphRAG achieved short computing time and low memory usage while ensuring low WER. WER with and without graph information was 8% and 10%, respectively. The memory usage of the baseline was about 500 MB, and the memory usage of AMDP-GraphRAG was 350 MB, which was the lowest among all the values.

(2) The AMDP-GraphRAG model could maintain or improve recognition accuracy while reducing the number of parameters. The parameter volumes of the baseline and AMDP were 120 and 10 million parameters, respectively. The WER of the baseline was 3.5%, and that of AMDP-GraphRAG was 2.8% (the lowest among all the WERs).

(3) Graph information could also enhance the generalization ability of the model. The models with graph information could adapt quickly and improve recognition accuracy.

The study demonstrated that AMDP-GraphRAG has obvious advantages in performance optimization, recognition accuracy, computing time, and memory usage. While ensuring low WER, it considerably reduces computing time and memory usage, making it highly competitive and useful for the design of efficient speech recognition systems. Moreover, it is suitable for application in resource-constrained environments and can improve the overall efficiency and performance of the system. The GraphRAG method focuses on the retrieval and generation of text information, but in practical applications, speech recognition often needs to be combined with other modal information, such as images and videos. Therefore, how to effectively integrate multimodal information and improve the multimodal understanding and generation capabilities of the model is an important direction for future research.

References

[1] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no.10, pp.1901-1913, Oct.2017.

[2] G. Thimmarajayadava and H. S. Jayanna, “Enhancements in automatic Kannada speech recognition system by background noise elimination and alternate acoustic modelling,”. *Int. J. Speech Technol.*, vol. 23, no. 1, pp.149-167, Jan. 2020.

[3] K. Liao, “Combining Evidence from Auditory, Instantaneous frequency and random forest for anti-noise speech recognition,” in *Proc. 7th Int. Conf. Comput. Sci. Inform. Technol.*, Dubai, DXB, UAE, 2021, pp. 75-85.

[4] Z. Q. Gao, “Chinese Speech Enhancement and Adaptive Recognition Technology for Complex Language Environments,” in *ACM Trans. on Asian and Low-Resour. Lang. Inform. Process.*, to be published. Accessed: Jul. 6, 2024. [Online]. Available: https://doi.org/10.1145/3608950

[5] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: discriminative embeddings for segmentation and separation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, SH,CHN, 2016, pp. 31-35.

[6] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proc. 23rd Int. Conf. Mach. Learn.*, Pittsburgh, PA, USA, 2006, pp. 369-376.

[7]W. Chan, N. Jaitly, Q. Le , and O. Vinyals, “Listen, attend and spell: a neural network for large vocabulary conversational speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Shanghai, SH,CHN, 2016, pp. 4960-4964.

[8] D. Michelsanti *et al.*, “An overview of deep-learning-based audio-visual speech enhancement and separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1368-1396, Mar.2021.

[9] A. S. Subramanian et al., “Far-field location guided target speech extraction using end-to-end speech recognition objectives,” in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Barca, ES, 2020, pp. 7299-7303.

[10] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, “Automatic speech recognition: a survey,” *Multimed. Tools Appl.*, vol. 80, no. 3, pp. 9411-9457, Mar. 2021.

[11] Y.V. Koteswararao and C.B.R. Rao, “Multichannel speech separation using hybrid GOMF and enthalpy-based deep neural networks,” *Multimed. Syst.*, vol. 27, no. 2, pp.271-286, Jan. 2021.

[12] S. H. Zhang, Y. Feng, R. J. Yu, P. R. Xing, and Y. H. Ren, “Speech emotion recognition based on SE attention mechanism and deep convolution,” *Mod. Electron. Tec.*, vol. 47, no. 22, pp.64-70, Nov. 2024.

[13] C.G. Hu, L.P. Yang, Y.Q. Sun, H.L. Chen, and K.K. Han, “Chinese speech recognition method of Conformer based on max pooling,” *Comput. Eng.*, to be published. Accessed: Nov. 8, 2024. doi: 10.19678/j.issn.1000-3428.0070055. [Online]. Available: https://link.cnki.net/doi/10.19678/j.issn.1000-3428.0070055

[14] J.L. Ou, H.Y. Zan, and H.F. Xu, "End-to-end Speech-to-text Translation Based on Multi-modal Joint Modeling," *J. CHIN. Comput. Syst.*, to be published. Accessed: Nov. 8, 2024. [Online]. Available:https://link.cnki.net/urlid/21.1106.TP.20241108.1523.007

[15] S.L. Tian, L.X. Ye, S.S. Wu, Q.W. Zhao, and P.Y. Zhang, “Low computational cost speech recognition based on discrete wavelet transform and subband decoupling,” *Acta Acust.*, to be published. Accessed: Nov. 6, 2024. [Online]. Available: https://link.cnki.net/urlid/11.2065.o4.20241105.1713.001

[16] S.S. Tan, C. Zhang, Y.C. Ma, and Z. Liu, “A new mixed mandarin-english code-swithcing speech recognition strategy integrating convolution and attention mechanism,” *Comput. Appl. Software*, to be published. Accessed: Oct. 29, 2024. [Online]. Available: https://link.cnki.net/urlid/31.1260.TP.20241029.1614.004

[17] J.Q. Guo and J.T. Zhang, “Research on speech recognition optimization method based on MFCC and HMM,” *Audio Eng.*, vol. 48, no.10, pp.83-85, Oct. 2024.

[18] X. H. Zhang and L. H. Yan, “End to end speech recognition with dilated convolutional network,” *Henan Sci.*, vol. 42, no.10, pp.1405-1414, Oct. 2024.

[19] A. R. Avila *et al*, “Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks,” in *Proc. IEEE Int. Symp. Signal Process. Inform. Technol.*, Bilbao, ESBIL, ES, 2017, pp. 360-365.

[20] D. Kristomo and F. H. Nugroho, “Classification of speech signal based on feature fusion in time and frequency domain,” in *Proc. 4th Int. Semin. Res. Inform. Technol. Int. Syst.*, Yogyakarta, JOG, IDN, 2021, pp. 560-564.

[21] J. D. Chen, X. H. Zhou, Q. Qin, “Research on speech recognition of sanitized robot based on improved speech enhancement algorithm,” in *Proc. 5th Int. Semin. Artif. Intell., Netw. Inform. Technol.*, Nanjing, NKG, CHN, 2024, pp. 1641-1644.

[22] N. U. Kure and S. B. Dhonde, “Dysarthric speech recognition using deep convolution neural network,” in *Proc. IEEE Int. Conf. Inform. Technol., Electron. Intell. Commun. Syst.*, Bangalore, BGR, IN, 2024, pp. 1-5.

[23] N. S. Jong, M. Kiatweerasakul and P. Phukpattaranont, “Channel reduction in speech recognition system based on surface electromyography,” in *Proc. 15th Int. Conf. Electr. Eng./ Electron., Comput., Telecommun. Inform. Technol.*, Chiang Rai, CRI, TH, 2018, pp. 184-187.

[24] S. Lotliker *et al*, “Podcast hosting using spectral gating and speech recognition methodology,” in *Proc. Int. Conf. Recent Trends Electron., Inform., Commun. Technol.*, Bangalore, BLR, IN, 2021, pp. 579-583.