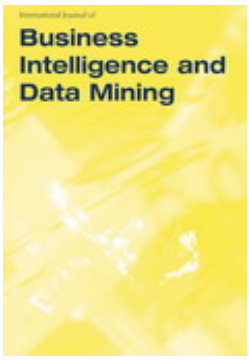


International Journal of Business Intelligence and Data Mining

ISSN online: 1743-8195 - ISSN print: 1743-8187
https://www.inderscience.com/ijbidm

Lightweight CNN-transformer hybrid network for English speech recognition

Yan Li, Weiguo Huang, Cui Gu

DOI: 10.1504/IJBIDM.2026.10077674

Article History:

Received:	17 September 2025
Last revised:	08 December 2025
Accepted:	13 January 2026
Published online:	30 April 2026

Lightweight CNN-transformer hybrid network for English speech recognition

Yan Li

School of Foreign Languages,
Hunan University of Arts and Science,
Changde 415000, Hunan, China
Email: maggie@huas.edu.cn

Weiguo Huang*

School of Information Engineering,
Hunan University of Science and Engineering,
Yongzhou 425199, Hunan, China
Email: huangweiguo@huse.edu.cn
*Corresponding author

Cui Gu

Department of Academic Affairs,
Changde College,
Changde – 415000, Hunan, China
Email: gu6084cui@126.com

Abstract: Speech recognition is the core technology for achieving human-computer interaction, among which English speech recognition has extremely high practical value in global communication scenarios. Although CNN-based speech recognition models are good at extracting local features, they cannot effectively capture global semantics. In contrast, transformer-based models outperform CNN in extracting global semantics, but their model parameters and computational complexity are high, making it difficult to deploy and run on resource constrained devices. Inspired by this, we proposes a lightweight CNN-transformer hybrid network (LwCTHNet) for English speech recognition. LwCTHNet effectively integrates local feature extraction, frequency domain detail supplementation, and global semantic capture capabilities by alternately stacking 3×3 convolution layers, wavelet enhanced convolution modules, and lightweight transformer modules. In addition, it also achieves multi-scale feature learning through skip connections and enhances feature discriminability by using a mixed loss function that combines cross entropy loss and contrastive loss. The experimental results on three English speech recognition datasets show that the proposed method not only has the minimum parameter size, but also achieves an approximately optimal word error rate. This indicates that the proposed LwCTHNet method has achieved a good balance in recognition performance, computational complexity, and parameter size.

Keywords: lightweight model; English speech recognition; transformer; multi-scale feature learning.

Reference to this paper should be made as follows: Li, Y., Huang, W. and Gu, C. (2026) 'Lightweight CNN-transformer hybrid network for English speech recognition', *Int. J. Business Intelligence and Data Mining*, Vol. 28, No. 7, pp.1–22.

Biographical notes: Yan Li currently serves as a Lecturer at the School of Foreign Languages, Hunan University of Arts and Science, China. She has published several academic articles in reputed journals. Her research interests mainly encompass AI-supported applied linguistics and cross-cultural communication.

Weiguo Huang is currently a Lecturer in the School of Information Engineering, Hunan University of Science and Engineering, China. He has published several academic articles in reputed journals. His research interests mainly include data mining, deep learning, and artificial intelligence.

Cui Gu is currently a Lecturer in Changde College, China. She has published several academic articles in reputed journals. Her research interests include computer assisted language learning and second language acquisition.

This paper was originally accepted for a special issue on 'Knowledge Discovery from Big Data to Spur Social Development' guest edited by Dr. Ziyang Guo, Dr. Chee-Onn Chow and Dr. Vijendra Singh.

1 Introduction

As the oldest and most traditional way of communication in the world, speech has been passed down and developed for hundreds of thousands of years. Long before the invention of writing, ancient humans already began to communicate using information such as different tones and frequencies of sounds. Moreover, it is not just humans, a large number of animals in nature also communicate with their unique speeches. With the continuous development of technologies such as the internet and artificial intelligence (Chen et al., 2025a), the process of globalisation has gradually accelerated, leading to increasing interactions between humans, as well as between humans and machines. This has given rise to a crucial issue: how to enable machines to accurately understand human speech. To address this problem, speech recognition technology has emerged. It has become a key enabler of natural human-machine interaction and plays an indispensable role in various fields such as speech command control.

As the most widely used language in the world, the research on English speech recognition technology holds extremely high practical value and significance. For instance, in the generation of real-time subtitles for international conferences, accurate English speech recognition can break down the barriers to real-time language communication, allowing participants from different native language backgrounds to instantly understand the content of speeches. In cross-border business and trade scenarios, an efficient English speech recognition system can automatically convert

customers' speech demands into text and conduct preliminary semantic analysis, significantly improving the response speed and processing efficiency of customer service. In the field of business English learning, English pronunciation assessment tools based on high-precision speech recognition can target scenario-specific expressions such as business negotiations, product introductions, and telephone communications, provide real-time feedback on learners' pronunciation issues, help them improve their oral English communication skills in a business context, and meet the needs of international business communication. Therefore, how to recognise English speech accurately and effectively has become a core issue that supports the efficient operation of business scenarios and promotes the development of cross-cultural business communication.

Existing speech recognition technologies can be roughly divided into two categories: traditional speech recognition methods and deep speech recognition methods. Traditional speech recognition methods are generally constructed based on template matching and statistical models, including the dynamic time warping (DTW) (Zhao and Itti, 2018), (Vintsyuk, 1968) algorithm, linear predictive coding (LPC) method (Anjum et al., 2020) and Gaussian mixture model-hidden Markov model (GMM-HMM) (Povey et al., 2011), etc. When processing speech signals, such methods require manual and elaborate design of feature extraction approaches in order to obtain valuable information for speech recognition. However, manually designing features is not only time-consuming and labor-intensive, but also often fails to comprehensively and accurately capture the complex features and patterns in speech signals. This results in significant limitations on recognition performance, leading to unsatisfactory recognition accuracy when faced with complex real-world application scenarios.

With the rapid development of deep learning technology (Wang et al., 2024; Chen et al., 2025b), it has also triggered a revolution in the field of speech recognition. This is because deep learning models possess powerful automatic feature learning capabilities, enabling them to automatically extract effective feature representations from massive amounts of speech data, thus greatly reducing the reliance on manual feature engineering. Existing deep speech recognition methods can be further divided into: CNN-based (LeCun et al., 2015) speech recognition methods and transformer-based (Vaswani et al., 2017) speech recognition methods. CNN-based (LeCun et al., 2015) speech recognition methods mainly include CNN-HMM (Abdel-Hamid et al., 2012), CNN-RBM-ASAT (Song, 2020), CBN (Veselỳ et al., 2011), etc. Such methods can effectively extract local features from speech signals through structures such as convolutional layers and pooling layers. Specifically, their convolution kernels can slide over speech spectrograms to capture short-term features in speech, such as energy changes in specific frequency bands and formant features. In speech recognition tasks, these local features play a crucial role in distinguishing different phonemes and words. However, CNN-based methods also have certain limitations: their ability to model long-range dependencies is relatively weak. When processing speech sequences with strong coherence, they struggle to fully utilise contextual information, which affects the accurate understanding of the entire speech segment. Because transformer can capture the dependencies between various elements in a sequence on a global scale through the attention mechanism, transformer-based speech recognition methods have received increasing attention in recent years, with classic approaches such as Speech-Transformer (Dong et al., 2018) and Conformer (Gulati et al., 2020) having emerged. These methods utilise the self-attention mechanism to focus on information

such as semantic connections and grammatical structures between words in a sentence, thereby improving the accuracy of recognition. However, transformer is not perfect. It has high computational complexity, especially when processing long sequences. The computational load of the self-attention mechanism increases quadratically with the length of the sequence, which demands enormous computing resources and results in low efficiency in model training and inference. Moreover, although transformer has strong global modelling capabilities, it still has drawbacks in terms of local feature perception.

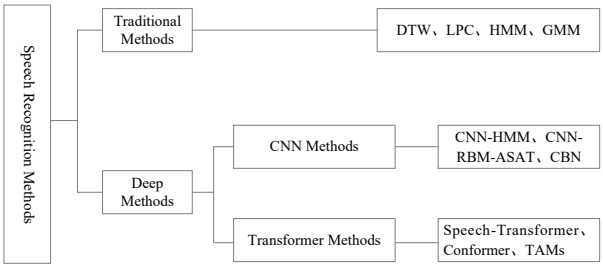
To fully leverage the advantages of CNN and transformer while overcoming their respective limitations, researchers have begun to explore hybrid network architectures that combine CNN and transformer. Such hybrid network architecture aims to utilise CNN's powerful local feature extraction capability and transformer's excellent global dependency modelling capability to achieve complementary advantages and improve the overall performance of English speech recognition systems. For example, in some existing studies, CNN is first used to perform preliminary feature extraction on speech signals to obtain local feature representations of speech, and then these features are input into the transformer module, where transformer is used to model and infer global information (Gu et al., 2023). This combination has improved the accuracy of speech recognition to a certain extent, but existing hybrid network models often have excessive parameters and high computational complexity. In practical applications, especially on some resource-constrained devices (such as mobile devices and embedded devices), their deployment and operation face huge challenges. In addition, existing hybrid models tend to add a transformer structure after the CNN model. This architecture ignores the interaction between local and global features within the intermediate layers of the network. Such an oversight prevents the model from fully utilising the interaction between global and local spatial information in the intermediate layers, thereby limiting its ability to model complex spatiotemporal relationships. In addition, these methods also significantly increase computational overhead.

To address the aforementioned issues, this paper designs a lightweight CNN-transformer hybrid network (LwCTHNet) for English speech recognition tasks, which requires only 0.9M parameters. LwCTHNet alternately learns local and global features, and introduces wavelet transform to supplement local details. Specifically, LwCTHNet consists of multiple cross-stacked 3×3 convolution layers, wavelet-enhanced convolution (WEC) modules, and lightweight transformer (LwT) modules. Among them, the 3×3 convolution layer is mainly used to extract local features, the WEC module is employed for downsampling and extracting frequency-domain details, and the LwT module is primarily responsible for capturing global feature relationships. In addition, during the network learning process, skip connections are employed to enable multi-scale feature extraction. By using this lightweight network, it is possible to significantly reduce the number of model parameters and computational load while ensuring model performance, thereby lowering the demand for computing resources. This enables efficient speech recognition and meets the application requirements in real-time and resource-constrained environments.

2 Related works

Existing speech recognition methods can generally be divided into two categories: traditional speech recognition methods and deep speech recognition methods (see Figure 1). Next, we will introduce these two types of methods separately.

Figure 1 The tree diagram of the speech recognition methods



2.1 Traditional speech recognition methods

In the early stage, speech recognition technology mainly included methods based on template matching and statistical models. In 1952, Bell Laboratories first realised a speech recognition system for isolated words using a simple template matching method, and created the first speech (digit) recogniser Audrey (Davis et al., 1952), which represented the initial stage of speech recognition technology. In 1959, a research team from the University of London adopted statistical principles to collect data information of speech sequences. This not only significantly improved the accuracy of speech recognition, but also realised a system suitable for multi-speech recognition. During this period, the template matching technology had emerged and become the dominant technology in the field of speech recognition. However, although the template matching-based methods can achieve good results in some specific scenarios, its inherent defects limit its further development and application in practical speech recognition. For instance, it requires pre-recording a large number of speech templates, and these templates need to cover various speech variations and features. Nevertheless, in practical applications, it is extremely difficult to pre-record all possible speech templates in advance.

Subsequently, with the advancement of computer technology, researchers began to explore how to introduce pattern recognition theory (Gao et al., 2024) into the field of speech recognition to improve the accuracy and reliability of systems. For example, Vintsyuk (1968) proposed the dynamic time warping (DTW) algorithm, which enables effective alignment of two time series, even if they have different lengths or different speeds. In speech recognition field, this capability is particularly important, as factors such as speaking rate and intonation among different speakers can lead to variations in the pronunciation of the same words. Anjum et al. (2020) proposed the linear predictive coding (LPC) method for isolated word recognition. Subsequently, the research focus of speech recognition shifted from template matching technology to statistical model methods, and the hidden Markov model (HMM) (Rabiner, 2002) gradually rose in technical status in the field of speech recognition. For example, Lee (1988) used

HMM to model speech temporal states, and further used Gaussian mixture model (GMM) for modelling, and finally developed the SPHINX speech recognition system. Subsequently, a large number of HMM-based speech recognition methods emerged. However, HMM-based methods still face many challenges. Firstly, the training process for HMM-based models is both cumbersome and difficult to optimise. Especially when dealing with large-scale speech data, a significant amount of computing resources and time are required for training and optimising the model, which increases the difficulty of model optimisation. Secondly, the HMM model is based on the assumption of conditional independence, meaning that given the hidden states, the observed states are independent of each other. This assumption does not always hold in practical situations, especially when processing complex speech signals, where there may be correlations and dependencies between multiple observed states. Therefore, over-reliance on the conditional independence assumption may lead to inaccurate modelling of data by the model, thereby affecting the performance of speech recognition. Despite the limitations of traditional models, they have provided an important foundation and reference for the development of the speech recognition field. The challenges faced during this period have also driven researchers to continuously seek innovations.

2.2 Deep speech recognition methods

With the rapid development of deep learning technology, it has also triggered a revolution in the field of speech recognition. A large number of deep learning-based speech recognition methods have emerged, which can be further divided into: CNN-based speech recognition methods and transformer-based speech recognition methods. Convolutional neural networks (CNN) (Taher and Abdulazeez, 2021) first became a key research area for researchers due to their efficiency in learning local feature information, translation invariance, and the characteristic of sharing convolution kernels. Furthermore, CNNs can reduce the number of model parameters by using weight sharing strategies, which not only endows the model with strong modelling performance but also significantly reduces computational complexity. Abdel-Hamid et al. (2012) first proposed the CNN-HMM method, which used local filtering and max pooling in the frequency domain to normalise speaker differences, thereby achieving higher performance in multi-speaker speech recognition. Song proposed the CNN-RBM-ASAT method (Song, 2020). This method first adopts a deep neural network supervised learning approach to extract high-level speech features, selects the output of a fixed hidden layer as new speech features for the newly generated network, and uses these new speech features to train a GMM-HMM acoustic model. Secondly, for various speech attributes, it trains a speech attribute extractor based on a deep neural network, and then classifies the extracted speech attributes into phonemes through the deep neural network. Finally, based on a linear feature fusion algorithm, the speech features and speech attribute features are fused into the same CNN framework through the neural network. Veselỳ et al. (2011) proposed the convolutive bottleneck network (CBN) method, which is an extension of the current state-of-the-art universal context network. Despite achieving certain results, CNN-based speech recognition methods have weak capabilities in modelling long-range

dependencies. When processing speech sequences with strong coherence, they struggle to fully utilise contextual information, which hinders the accurate understanding of entire speech segments.

Given that transformer can capture dependencies between various elements in a sequence on a global scale by virtue of the attention mechanism, transformer-based speech recognition methods have received increasing attention in recent years. Dong et al. (2018) proposed the Speech-Transformer method, which is a non-recursive sequence-to-sequence model. It relies entirely on the attention mechanism to learn positional dependencies and can be trained faster and more efficiently. In addition, the authors also proposed a two-dimensional attention mechanism that can jointly focus on the time and frequency axes of two-dimensional speech inputs, thereby providing more expressive representations for Speech-Transformer. Subsequently, Google launched a new speech recognition model called Conformer (Gulati et al., 2020), which is implemented by incorporating convolutional layers and a macaron structure into the transformer. It can not only capture global information but also learn local feature information, fully integrating the advantages of convolutional neural networks and transformer. This model successfully models both the global and local information of audio sequences and further set a new record in the speech recognition error rate on the English dataset LibriSpeech. Wang et al. (2020) proposed transformer-based acoustic models (TAMs) for hybrid speech recognition. This method discusses several modelling options, including various positional embedding methods and iterative loss, to enable the training of deep transformers. It also conducts a preliminary study on the use of limited-context in transformer models, which makes streaming applications possible. Despite achieving favourable results, transformer-based speech recognition methods still suffer from drawbacks such as a large number of parameters and high computational complexity, leading to low efficiency in model training and inference. Furthermore, although transformers possess strong global modelling capabilities, they still have limitations in terms of local feature perception.

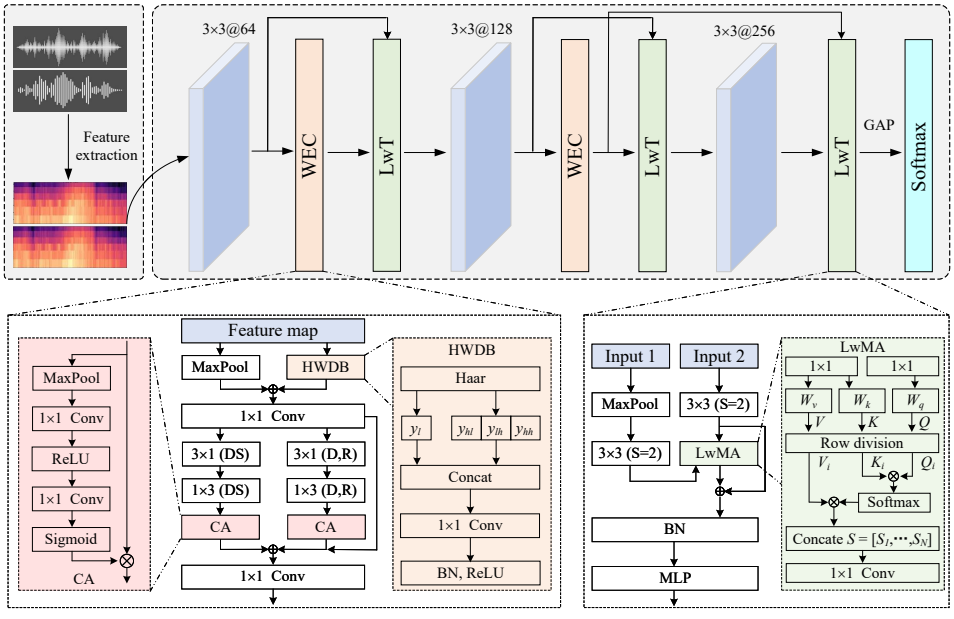
3 Lightweight CNN-transformer hybrid network

As shown in Figure 2, the proposed LwCTHNet method consists of multiple cross-stacked 3 × 3 convolution layers, wavelet-enhanced convolution (WEC) modules, and lightweight transformer (LwT) modules. Among them, the 3 × 3 convolution layer is mainly used to extract local detailed features, the WEC module is used for downsampling and extracting frequency-domain details, and the LwT module is mainly used to extract long-range dependencies.

More specifically, the input $X \in R^{H\times W\times C}$ (In this paper, X is the preprocessed Mel spectrogram, and we set $H = W = 224, C = 3$) is first passed through a 3 × 3 convolution layer, and becomes to $X^{(1)} \in R^{H\times W\times C_1}$. Next, through the first WEC module, the downsampled feature map $X^{(2)} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$ is output, which obtains various frequency-domain details of the input via wavelet transform. Then, $X^{(1)}$ and $X^{(2)}$ are jointly fed into the first LwT module to obtain $X^{(3)} \in R^{\frac{H}{4}\times\frac{W}{4}\times C_1}$. This module extracts long-range dependencies of the input data through an interactive attention mechanism while considering multi-scale feature information. Repeating the above process, $X^{(4)} \in R^{\frac{H}{4}\times\frac{W}{4}\times C_2}$, $X^{(5)} \in R^{\frac{H}{8}\times\frac{W}{8}\times C_2}$ and $X^{(6)} \in R^{\frac{H}{16}\times\frac{W}{16}\times C_2}$ are obtained respectively. Then, through the last 3 × 3 convolution layer to increase the

number of channels, $X^{(7)} \in R^{\frac{H}{16} \times \frac{W}{16} \times C_3}$ is obtained. After that, through the last LwT module, we obtain $X^{(8)} \in R^{\frac{H}{32} \times \frac{W}{32} \times C_3}$. The features after global average pooling are fed into the classifier layer for classification. The number of input neurons in the classifier layer is C_3, and the number of output neurons is equal to the number of classes.

Figure 2 A schematic illustration of our proposed LwCTHNet (see online version for colours)



Note: The input Mel spectrogram first passes through 3 × 3 convolutional layers and the WEC module to extract initial local features, and performing downsampling and frequency-domain detail extraction. Next, the outputs of these two components are jointly fed into the LwT module to capture long-range dependencies. The process is repeated several times to complete the speech recognition task.

Next, we will provide a detailed introduction to the two components of LwCTHNet: the WEC module and the LwT module.

3.1 WEC module

Traditional convolution operation can extract local information from input feature maps, but with the stacking of multiple convolution layers, using traditional pooling for downsampling will lose detailed information such as edge and texture. Even if multi-scale information is utilised through skip connections, these already lost details cannot be recovered. In the field of speech recognition, the detailed information plays a crucial role in improving recognition performance. Given that wavelet transform can extract local information in both the time and frequency domains while effectively removing noise, we have designed the WEC module in this section. This module

leverages the respective advantages of traditional convolution and Haar wavelet transform to accomplish the downsampling of feature maps. In addition, the WEC module also uses dilated convolution to expand the receptive field of traditional convolution. This not only extracts local information from the feature map but also utilises contextual information, helping the model better recognise objects. The structure of WEC is shown in the lower left corner of Figure 2.

More specifically, taking the first WEC module as an example, two types of downsampling are performed on the input feature map $X^{(1)} \in R^{H\times W\times C_1}$. In the left branch, 2 × 2 max pooling (MaxPool) with a stride of 2 is used to halve the size of the input feature map, and it outputs

$$X^{(11,L)} = MaxPool(X^{(1)}) \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}, \tag{1}$$

where L represents left branch. In the right branch, to better extract frequency-domain details and reduce the information loss caused by downsampling via traditional convolution, a Haar wavelet decomposition block (HWDB) is used for downsampling. Specifically, the Haar wavelet transform is employed to decompose the input into 1 low-frequency component $y_l \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$ and 3 high-frequency components $y_{hl} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$, $y_{lh} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$ and $y_{hh} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$ (representing high-frequency features in the horizontal, vertical, and diagonal directions, respectively), i.e.,

$$(y_l, y_{lh}, y_{hl}, y_{hh}) = Haar(X^{(1)}), \tag{2}$$

where $Haar(\cdot)$ denotes HWDB.

Subsequently, the low-frequency and high-frequency components are concatenated along the channel dimension to obtain features containing multi-scale contextual information

$$z = Concat(y_l, y_{lh}, y_{hl}, y_{hh}) \in R^{\frac{H}{2}\times\frac{W}{2}\times 4C_1}. \tag{3}$$

If the channel dimension is too large, it will affect the computational efficiency, so a 1 × 1 convolution is used to reduce the channel dimension, then a batch normalisation (BN) layer and ReLU function are used to further enhance the feature discriminative ability. Finally, the output of the right branch is

$$X^{(11,R)} = ReLU(BN(Conv_{1\times1}(z))) \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}, \tag{4}$$

where R represents right branch.

The features $X^{(11,L)}$, $X^{(11,R)}$ obtained from the two branches are concatenated along the channel dimension, and a 1 × 1 convolution is used to keep the channel dimension unchanged. Then, drawing on the ideas of Inception and MobileNet networks, we continue to learn multi-scale features. The left branch uses successive 3 × 1 and 1 × 3 depth-wise separable convolutions, which increase the network depth while effectively controlling the number of network parameters. The right branch uses successive 3 × 1 and 1 × 3 dilated convolutions to expand the receptive field and extract richer contextual information. The outputs of the two branches are respectively: $X^{(12,L)} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$, $X^{(12,R)} \in R^{\frac{H}{2}\times\frac{W}{2}\times C_1}$. Furthermore, to focus on valuable channel features and suppress irrelevant or redundant channel information,

thereby enhancing the model's feature expression ability, a channel attention (CA) operation is introduced at the end of each branch:

$$X^{(13,L)} = CA(X^{(12,L)}) \in R^{\frac{H}{2} \times \frac{W}{2} \times C_1}, \tag{5}$$

$$X^{(13,R)} = CA(X^{(12,R)}) \in R^{\frac{H}{2} \times \frac{W}{2} \times C_1}. \tag{6}$$

Then, the features obtained from the two branches are concatenated along the channel dimension, and a 1×1 convolution is used to keep the channel dimension unchanged, thereby obtaining the final output of this module:

$$X^{(13)} = Concat(X^{(13,L)}, X^{(13,R)}) \in R^{\frac{H}{2} \times \frac{W}{2} \times 2C_1}, \tag{7}$$

$$X^{(2)} = Conv_{1\times1}(X^{(13)}) \in R^{\frac{H}{2} \times \frac{W}{2} \times C_1}. \tag{8}$$

The WEC module enhances the model's ability to capture frequency-domain detailed features. Meanwhile, the use of depth-wise separable convolutions and dilated convolutions enables the model to flexibly capture features in horizontal and vertical directions while minimising the number of model parameters.

3.2 LwT module

Traditional convolutional neural networks have limited receptive fields and struggle to capture long-range dependencies. In contrast, transformer, leveraging self-attention mechanisms, demonstrates strong capabilities in acquiring long-range dependencies. However, due to the massive computational complexity and parameter count of standard transformers, they are difficult to deploy on mobile edge devices. To enhance the model's recognition performance while enabling broader application scenarios, the LwT is designed to improve recognition performance while balancing computational efficiency. The network structure of LwT is shown in the lower right corner of Figure 2, which includes a downsampling layer, a lightweight interactive multi-head attention layer, a BN layer, and a multi-layer perceptron (MLP) layer.

More specifically, taking the first LwT module as an example, for multi-scale inputs $X^{(1)} \in R^{H\times W\times C_1}$ and $X^{(2)} \in R^{H\times W\times C_1}$, different dimension reduction strategies are performed to them respectively. Specifically, $X^{(1)}$ goes through the left branch, where it undergoes 2×2 max pooling with a stride of 2 followed by a 3×3 convolution, converting the input into

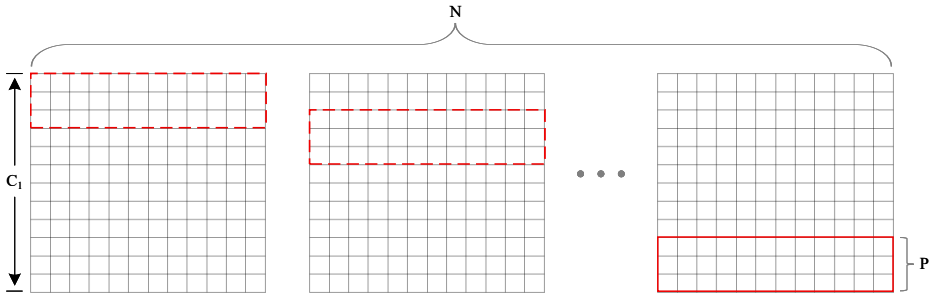
$$X^{(3,1)} = Conv(MaxPool(X^{(1)})) \in R^{\frac{H}{4} \times \frac{W}{4} \times C_1}. \tag{9}$$

$X^{(2)}$ goes through the right branch, where it undergoes a 3×3 convolution with a stride of 2, converting the input into

$$X^{(3,2)} = Conv(X^{(2)}) \in R^{\frac{H}{4} \times \frac{W}{4} \times C_1}. \tag{10}$$

Through this dual-branch structure, richer semantic information can be obtained. Then, $X^{(3,1)}$ and $X^{(3,2)}$ are jointly fed into the lightweight multi-head attention (LwMA) block.

Figure 3 A schematic illustration of row division strategy (see online version for colours)



In the standard multi-head attention mechanism, the input matrix is mapped to different Q, K, V spaces through multiple different linear transformations. For instance, if the input matrix has a size of $\frac{HW}{16} \times C_1$, the number of parameters for a single attention head is the sum of the parameters of matrices $W_q \in R^{D_c \times C_1}$, $W_k \in R^{D_c \times C_1}$ and $W_v \in R^{D_c \times C_1}$, which equals $3D_c \times C_1$. When using M attention heads, the total number of parameters amounts to $M \times 3D_c \times C_1$. To reduce the number of parameters and enable contextual interaction between multi-scale information, this section designs a novel attention mechanism: LwMA. First, 1 × 1 convolutions are used to reduce the number of channels of features from different levels to half of the input, obtaining $\hat{X}^{(3,1)} = Conv_{1\times1}(X^{(3,1)}) \in R^{\frac{H}{4} \times \frac{W}{4} \times \frac{C_1}{2}}$ and $\hat{X}^{(3,2)} = Conv_{1\times1}(X^{(3,2)}) \in R^{\frac{H}{4} \times \frac{W}{4} \times \frac{C_1}{2}}$. Then, the linear transformations $W_q \in R^{C_1 \times \frac{C_1}{2}}$, $W_k \in R^{C_1 \times \frac{C_1}{2}}$, $W_v \in R^{C_1 \times \frac{C_1}{2}}$ are applied to project the two inputs into:

$$\begin{aligned} Q &= W_q \times (\hat{X}^{(3,1)})^T \in R^{C_1 \times \frac{HW}{16}}, \\ K &= W_k \times (\hat{X}^{(3,2)})^T \in R^{C_1 \times \frac{HW}{16}}, \\ V &= W_v \times (\hat{X}^{(3,2)})^T \in R^{C_1 \times \frac{HW}{16}}. \end{aligned} \tag{11}$$

Next, we perform row division on the Q, K, V matrices, dividing them into N groups of submatrices to simulate the multi-head attention mechanism (see Figure 3). Assuming the size of each group of submatrices is $Q_i \in R^{P \times \frac{HW}{16}}$, $K_i \in R^{P \times \frac{HW}{16}}$, $V_i \in R^{P \times \frac{HW}{16}}$ respectively, then the self-attention is computed as follows:

$$S_i = V_i \times Softmax(\frac{K_i^T Q_i}{\sqrt{C_1}}) \in R^{P \times \frac{HW}{16}}, \ i = 1, \cdots, N. \tag{12}$$

Through the above process, we obtain N groups of results. These results are then restored to a 3D structure and concatenated along the channel dimension to simulate the multi-head attention mechanism:

$$S = Concat(S_1, \cdots, S_N) \in R^{\frac{H}{4} \times \frac{W}{4} \times NP}. \tag{13}$$

Subsequently, a 1 × 1 convolution is employed to reduce the NP channels back to C_1 channels:

$$X^{(3)} = Conv_{1\times1}(S) \in R^{\frac{H}{4} \times \frac{W}{4} \times C_1}. \tag{14}$$

Through the above calculations, the number of parameters can be reduced to $1/M$ of the standard multi-head attention mechanism (typically $M = 16, 32$, etc.), and the computational load will also be significantly reduced.

3.3 Loss function

In the experiments, we use the cross-entropy (CE) loss function to train the entire network, and its formula is as follows:

$$L_1 = -\frac{1}{B}\sum_{i=1}^{B} y_i \log(\hat{y}_i), \tag{15}$$

where B is the batch size, y_i is the real label, and $\hat{y}_i$ is the predicted label.

Furthermore, to enhance the discriminative ability of features, before training the entire network, we use contrastive loss (CL) to pre-train the backbone network. This ensures that in the feature space, samples of the same class are close to each other, while samples of different classes are separated from each other. The formula of the contrastive loss is as follows:

$$L_2 = y_{ij}||X_i^{(8)} - X_j^{(8)}||_2^2 + (1 - y_{ij})\max(0, t - ||X_i^{(8)} - X_j^{(8)}||_2^2), \tag{16}$$

where y_{ij} denotes the indicator function. if $X_i^{(8)}$ and $X_j^{(8)}$ belong to the same category, then $y_{ij} = 1$; otherwise $y_{ij} = 0$. Here, t is a hyperparameter.

Finally, the total loss function is:

$$L = L_1 + \lambda L_2, \tag{17}$$

where λ is the balance parameter.

4 Experimental results and analysis

To verify the effectiveness of the proposed method, extensive experiments will be conducted on three English speech recognition datasets. The word error rate (WER) is adopted as the evaluation metric, which is a crucial indicator for measuring the accuracy of speech recognition systems. WER is defined as the ratio of the number of incorrectly recognised words to the total number of words. This evaluation metric can intuitively reflect the model's performance in speech recognition tasks, and its calculation formula is as follows:

$$WER = \frac{S + D + I}{n} \times 100\%,$$

where S denotes the number of substituted words, D represents the number of deleted words, I stands for the number of newly inserted words, and n is the total number of words in the sentence.

4.1 Datasets

In this paper, we use LibriSpeech, GigaSpeech and LRS2 three English speech recognition datasets to verify the effectiveness of the proposed method.

The LibriSpeech dataset (Panayotov et al., 2015) is a large-scale English speech dataset containing approximately 1,000 hours of read English speech with a sampling rate of 16 kHz. Derived from audiobooks of the LibriVox project, the dataset encompasses diverse accents and reading styles, and has undergone careful segmentation and alignment. In this paper, 80% of the dataset (800 hours) is used as training data, 10% as the validation set, and the remaining 10% as the testing set. The average performance across ten random experiments is reported.

The GigaSpeech dataset (Chen et al., 2021), developed by the Speech and Audio Technology Laboratory of Tsinghua University, is an evolving multi-domain English speech recognition dataset, totalling 33,000 hours of audio, among which 10,000 hours are high-quality annotated data. In the experiments of this paper, the 10,000-hour high-quality annotated subset is used as the training set, 12 hours of the dataset as the validation set, and 40 hours as the test set.

LRS2 (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html) is an English speech dataset consisting of short clips extracted from BBC TV programs, with a total duration of over 200 hours. It covers daily conversations, news broadcasts, and other scenarios, and the speech in the dataset is closer to real-world speech situations. The LRS2 dataset includes three parts: training (train), evaluation (dev), and testing (test). All audio is extracted from video clips, involving factors such as complex background noise, different speaking speeds, and pronunciation habits, which make the dataset more challenging.

4.2 Implementation details

Our experiments were implemented using the TensorFlow framework on a 3080 GPU, with the Adam optimiser employed. The number of training epochs was set to 200, and the batch size was 64. The initial learning rate was set to 0.003, which gradually increased with the number of training epochs until reaching 0.03, after which it was gradually decreased.

For the input audio signal, we first perform preprocessing to reduce noise interference. Subsequently, we conduct framing, windowing, and short-time Fourier transform (STFT), and apply a Mel filter bank to generate a Mel spectrogram, which is a two-dimensional vector feature representation. Before feeding the Mel spectrogram into the network, it needs to be resized to a standard size of $224 \times 224 \times 3$.

4.3 Comparison methods

This paper will compare with state-of-the-art speech recognition methods, including: CNN-HMM, Speech-Transformer, Vanilla Transformer, LRT, Deep Speech, Conformer, CLSC-Transformer, etc. In addition, we also compare with several lightweight models, including: Fully Adaptive Transformer (FAT-B0) (Fan et al., 2023), Lightweight Transformer Image Feature Extraction (LTIFE) network (Zheng et al., 2024), MobileViT-XS (Mehta and Rastegari, 2021), etc.

4.4 Experiments on LibriSpeech dataset

We first conducted experiments on the LibriSpeech dataset, and the detail experimental results are shown in Table 1.

Table 1 Comparison results on the LibriSpeech dataset

Methods	*Params (M)*	*FLOPs (G)*	*WER*
CNN-HMM	31.3	2.3	15.41 ± 3.21
Speech-Transformer	26.2	2.1	13.24 ± 2.33
Vanilla Transformer	25.1	2.1	14.41 ± 2.19
LRT	12.7	1.9	13.33 ± 2.57
Deep Speech	40.8	2.5	12.56 ± 3.44
Conformer	10.5	1.8	13.47 ± 3.11
CLSC-Transformer	10.3	1.7	12.14 ± 2.17
FAT-B0	4.5	0.6	15.31 ± 2.14
LTIFE	2.2	2.3	15.35 ± 2.44
MobileViT-XS	2.3	1.2	14.47 ± 2.28
LwCTHNet	0.9	0.5	12.71 ± 2.13

From this table, we can observe that:

1 LwCTHNet exhibits significant lightweight characteristics in terms of parameter size: its parameter is only 0.9M, far lower than all comparison methods. Specifically, traditional methods such as CNN-HMM have parameters as high as 31.3M, while classic transformer-based methods such as Speech-Transformer and Vanilla Transformer have parameters exceeding 25M. Even models specifically designed for lightweighting, such as FAT-B0, LTIFE, and MobileViT XS, have parameters of 4.5M, 2.2M, and 2.3M, respectively, all of which are more than three times that of LwCTHNet. This result verifies that LwCTHNet has successfully achieved a significant reduction in model parameters through the structural design of cross stacking 3 × 3 convolution layers, WEC modules, and LwT modules, meeting the deployment requirements of resource constrained scenarios such as mobile devices.

2 In terms of computational complexity (FLOPs), LwCTHNet has the highest computational efficiency among all methods with a value of 0.5G. By comparison, the FLOPs of the traditional method CNN-HMM are 2.3G, while the transformer-based method Conformer is 1.8G. Even the FAT-B0 (0.6G), which has lower computational complexity in lightweight models, still has slightly higher computational complexity than LwCTHNet. This advantage stems from the application of lightweight strategies such as depthwise separable convolution and wavelet transform downsampling in LwCTHNet, which effectively reduces the computational load of the model during runtime and provides efficiency guarantees for real-time speech recognition scenarios.

3 On the core performance indicator WER, LwCTHNet has achieved a competitive level: its WER is 12.71, only slightly higher than the current optimal CLSC-Transformer (12.14), but the difference is controlled within 0.57% and significantly better than most comparison methods. Specifically, it outperforms traditional methods (such as CNN-HMM's 15.41) and early transformer models (such as Vanilla Transformer's 14.41). In addition, LwCTHNet is significantly ahead of other lightweight models (such as FAT-B0's 15.31 and LTIFE's 15.35), with a gap of over 2.6%. Compared with the Conformer method (13.47) which has a larger parameter scale, our method's WER decreased by 0.76%, reflecting an effective balance between 'lightweight' and 'high-performance'.

In summary, the experimental results in Table 1 indicate that the proposed LwCTHNet achieves the optimal balance between parameter size, computational complexity, and recognition accuracy on the LibriSpeech dataset, it reduces parameters by 97% (compared to CLSC-Transformer) and computational complexity by 68% (compared to Conformer), and it still maintains recognition accuracy close to the current optimal method. This validates the effectiveness of its design approach, providing a new solution for the application of English speech recognition in resource constrained scenarios.

4.5 Experiments on GigaSpeech and LRS2 datasets

We also conduct experiments on the GigaSpeech and LRS2 dataset, and the experimental results are reported in Table 2.

Table 2 The word error rate (WER) of different methods on the GigaSpeech and LRS2 datasets

Methods	*GigaSpeech*	*LRS2*
CNN-HMM	14.59	22.95
Speech-Transformer	14.47	18.57
Vanilla Transformer	13.25	17.44
LRT	13.74	17.21
Deep Speech	12.11	17.14
Conformer	13.54	18.05
CLSC-Transformer	12.17	17.03
FAT-B0	14.56	19.44
LTIFE	13.78	20.31
MobileViT-XS	13.67	18.78
LwCTHNet	12.74	17.03

From Table 2, it can be seen that the proposed LwCTHNet method achieved approximately optimal results on both datasets. Specifically, on the GigaSpeech dataset, WCTHNet achieved a WER of 12.74, ranking third among all comparison methods, only slightly higher than the non-lightweight models such as Deep Speech (12.11) and CLSC-Transformer (12.17), with gaps of 0.63% and 0.57%, respectively. However, it should be noted that the parameter size of our LwCTHNet is much lower than those of the two models. Compared to traditional methods (such as CNN-HMM's 14.59),

our method reduces WER by 1.85%; compared to classic transformer models such as Speech-Transformer (14.47) and Vanilla Transformer (13.25), the WER has decreased by 1.73% and 0.51%, respectively. Compared with other lightweight models, our LwCTHNet has particularly significant advantages. For example, the WER of FAT-B0 is 14.56, while LwCTHNet reduces it by 1.82%; The WER of MobileViT XS is 13.67, and LwCTHNet decreased by 0.93%. The above results indicate that even on large-scale and multi domain speech data, LwCTHNet can still maintain high accuracy, verifying its effectiveness in complex scenarios.

On the LRS2 dataset, our LwCTHNet achieved the optimal WER (17.03). Compared to traditional methods such as CNN-HMM (22.95), LwCTHNet reduces the WER by 5.92%. Compared to Speech-Transformer and Conformer, LwCTHNet reduces the WER by 1.54% and 1.02%, respectively. Compared to the lightweight models, the advantages of the proposed method are more prominent, it reduces the WER of LTIFE and FAT-B0 by 3.28% and 2.41%, respectively. All these results indicate that LwCTHNet can still maintain high performance in noisy and non-ideal real speech scenes. Its cross stacked local feature extraction and global dependency modelling mechanism, as well as the multi-scale feature fusion ability brought by skip connections, effectively enhance the robustness of the model to interference factors.

In summary, the experimental results in Table 2 further validate the generalisation ability and practicality of LwCTHNet. Whether in large-scale standardised data or real complex scenarios, the model can achieve or approach the recognition accuracy of the current optimal method with extremely low parameter size and computational complexity. This performance is attributed to its innovative hybrid architecture design, which supplements frequency domain details through WEC modules, efficiently captures global dependencies through LwT modules, and combines multi-scale feature fusion strategies, making it significantly applicable in resource constrained scenarios such as mobile devices and real-time voice interaction.

4.6 Ablation study

In this subsection, we will conduct ablation studies to verify the effects of different components and loss functions of the proposed LwCTHNet. All experiments are conducted on the LRS2 dataset.

4.6.1 Ablation on different components

We first conduct ablation study to verify the influences of different components, and the experimental results are shown in Table 3. It can be seen from the table that when only a single component is retained, the performance of the model decreases significantly, which confirms that a single module cannot support high-precision speech recognition. Specifically, comparing 'order' 1, 4, and 6, it can be seen that when only the 3×3 convolution layer is retained, the WER is 40.88, when only the WEC module is retained, the WER is 35.43, and when only the LwT module is retained, the WER is 25.47, indicating that the LwT model is more important.

When two components are retained, although there is an improvement in performance, it still does not reach the optimal level, indicating that the collaboration of the three components is irreplaceable. Specifically, if '3×3 convolution + WEC'

(30.43) is retained simultaneously, WER is significantly reduced compared to a single component, but still higher than the complete model. The reason is the lack of LwT module, which makes it impossible to model long-range dependencies. If both '3 × 3 convolution + LwT' (23.74) are retained, the WER will be lower than '3 × 3 convolution + WEC', indicating that LwT is more important than WEC. If 'WEC + LwT' (20.32) is kept, the WER is the lowest, indicating that WEC is more important than 3 × 3 convolution.

Table 3 Ablation experiments of different components on LRS2 dataset

Order	*3 × 3 Conv*	*WEC*	*LwT*	*WER*
1	✓			40.88
2	✓	✓		30.43
3	✓		✓	23.74
4		✓		35.43
5		✓	✓	20.32
6			✓	25.47
7	✓	✓	✓	17.03

When all three components (3 × 3 convolution, WEC, LwT) are retained, the model performance reaches its optimal level and significantly outperforms all ablation combinations. This result validates the rationality of LwCTHNet's design.

4.6.2 Ablation on different loss functions

In this subsection, we will conduct ablation studies to verify the effects of different loss functions, and the experimental results are shown in Table 4.

Table 4 Ablation experiments of different loss function on LRS2 dataset

Order	*CE loss*	*CL loss*	*WER*
1	✓		19.57
2		✓	18.79
3	✓	✓	17.03

From this table, we can observe that:

1 When using only cross entropy (CE) loss, a WER of 19.57 was achieved, which is significantly higher than the complete model (17.03), with a difference of 2.54%. This is because the CE loss focuses more on whether the samples are correctly classified, while ignoring the aggregation of similar samples and the separation of heterogeneous samples in the feature space – in speech recognition, features of similar phonemes may be confused due to insufficient discriminability, especially in noisy scenes of the LRS2 dataset, where features are easily disturbed, and a single CE loss is difficult to ensure the robustness of features.

2 When using only contrastive loss (CL), our LwCTHNet achieves a WER of 18.79. Compared to the single CE loss, its WER decreased by 0.78%, indicating that feature space optimisation has a positive effect on speech recognition. However, when relying solely on CL loss, the model lacks direct classification objective constraints, which may lead to the problem of 'good feature separation but fuzzy classification boundaries', resulting in suboptimal performance.

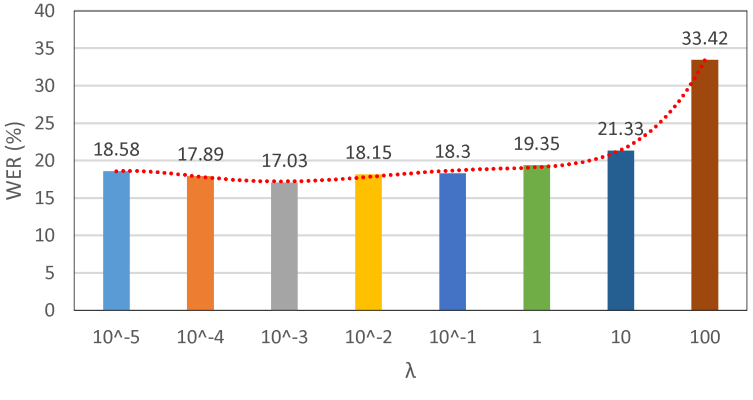
3 When both CE loss and CL loss are used, the model achieves optimal performance. These results fully demonstrate that combining CE loss and CL loss can effectively improve speech recognition performance.

4.7 Parameter analysis

4.7.1 Analysis of parameter λ

In the proposed method, λ is an important parameter that can control the balance between CE loss and CL loss. Moreover, different λ values can lead to different experimental results. Therefore, in this section, we will verify the impact of λ on the LRS2 dataset. Figure 4 shows the impact of the balance parameter λ on model performance. The results show that when λ is within a reasonable range (such as $10^{-4} \sim 10^{-3}$), the WER of the model is the lowest (17.03). If λ is too large (such as $\lambda \geq 1$), the CL loss weight may be too high, which may lead to the model focusing too much on feature separation and neglecting classification accuracy, resulting in an increase in WER. If λ is too small (such as $\lambda \leq 10^{-4}$), the effect of CL loss is weak, and the model performance is close to that of using only CE loss. This result further confirms the importance of balancing feature discriminability and classification accuracy in the mixed loss function. In other words, for speech recognition tasks, it is necessary to moderately enhance the structuring of the feature space, but not deviate from the classification objective.

Figure 4 The influence of different λ on LRS2 dataset (see online version for colours)

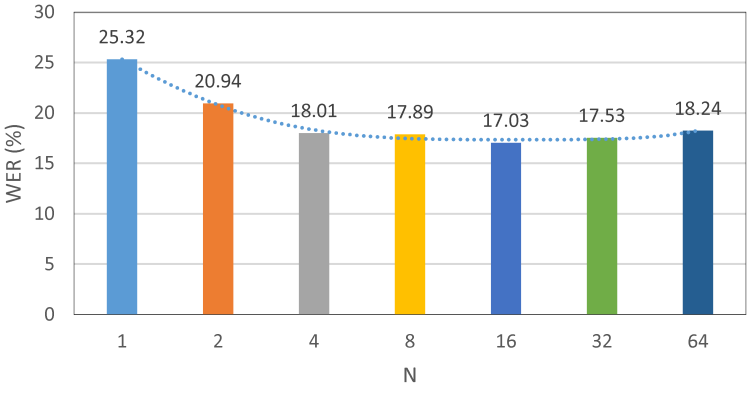


4.7.2 Analysis of parameter N

In this paper, parameter N is also an important parameter, which directly affects the granularity of the multi-head attention mechanism and further impacts the model's

ability to capture global dependencies and computational efficiency. To explore the optimal value of N, we conducted control experiments on the LRS2 dataset by setting N to 1, 2, 4, 8, 16, 32, and 64 while keeping other hyper-parameters unchanged. As seen in Figure 5, the model's WER shows a 'U-shaped' trend with the change of N: when N= 1 (single-head attention), the WER is as high as 25.32% due to the inability to capture multi-dimensional contextual information; as N increases to 16, the WER decreases to the lowest 17.03% because appropriate multi-group division enables parallel attention calculation from multiple granularities, fully capturing global semantic relationships; when N exceeds 16 and continues to increase to 32 and 64, the WER rises to 17.53% and 18.24% respectively – this is because excessive grouping leads to 'over-fragmentation' of Q, K, V matrices (each submatrix contains limited semantic information, making it difficult to model long-range dependencies between words) and increased computational overhead, which impairs model efficiency and recognition accuracy. In conclusion, the optimal value of N for LwCTHNet on the LRS2 dataset is 16, which balances the model's global dependency capture capability and computational complexity effectively.

Figure 5 The influence of different N on LRS2 dataset (see online version for colours)



4.7.3 Analysis of batch size

In the proposed method, batch size is a critical hyper-parameter that affects both the stability of model training and the final recognition performance. A larger batch size can enhance the generalisation ability of the model by utilising more sample information for gradient calculation, while an excessively large batch size may increase memory consumption and reduce training efficiency. To determine the optimal batch size, the comparison experiments on the LibriSpeech, GigaSpeech, and LRS2 datasets are performed by setting the batch size as {16, 32, 64, 128, 256}. The experimental results are shown in Table 5.

From this table, we observed that as the batch size increases from 16 to 64, the WER of LwCTHNet decreases significantly on all three datasets, this is because a moderate increase in batch size enables the model to learn more comprehensive data distribution characteristics, thereby improving feature discriminability. However, when the batch size exceeds 64 and continues to increase to 128 and 256, the WER reduction becomes negligible, this phenomenon is due to the fact that after the batch size reaches a certain

threshold (i.e., 64), the model has already captured the main data distribution, and continuing to increase the batch size can only bring marginal performance gains while significantly increasing GPU memory usage. So, the optimal batch size for LwCTHNet is 64, which can balance training efficiency, memory consumption, and recognition performance effectively.

Table 5 The influence of different batch sizes on three datasets

Datasets	*16*	*32*	*64*	*128*	*256*
LibriSpeech	13.95	13.21	12.71	12.19	12.03
GigaSpeech	14.14	13.45	12.74	12.38	12.11
LRS2	18.89	17.55	17.03	17.01	16.83

4.8 Running time analysis

To further verify the efficiency advantage of LwCTHNet in practical inference scenarios, Table 6 compares the running time of LwCTHNet with several representative methods when processing a single test sample on the LRS2 dataset. As can be seen from this table that LwCTHNet achieves the shortest running time and it is significantly faster than all comparison methods. Specifically, our method only needs 0.4 seconds, while the non-lightweight model Conformer needs 2.3 seconds, the lightweight model LTIFE needs 1.9 seconds, FAT-B0 needs 1.5 seconds, and MobileViT-XS – the previously most efficient lightweight model, needs 0.5 seconds. This result fully demonstrates that the lightweight design of LwCTHNet (such as the application of depthwise separable convolution, wavelet transform downsampling, and parameter-reduced LwT module) not only reduces model parameters and computational complexity but also directly translates to faster inference speed. This advantage makes LwCTHNet highly suitable for real-time speech recognition scenarios with strict latency requirements, such as mobile devices and edge computing platforms.

Table 6 The comparison of running time (s) on LRS2 dataset when processing one testing sample

Datasets	*FAT-B0*	*LTIFE*	*MobileViT-XS*	*Conformer*	*LwCTHNet*
LRS2	1.5	1.9	0.5	2.3	0.4

5 Conclusions

In this paper, we propose a lightweight CNN-Transformer hybrid network LwCTHNet for English speech recognition, which combines the advantages of both CNN and transformer to balance recognition accuracy, computational complexity, and model size. By introducing a 3×3 convolution layer, WEC module, and LwT module, LwCTHNet not only accurately captures local details but also effectively extracts global semantic information, achieving accurate recognition of English speech. Experiments conducted on the LibriSpeech, GigaSpeech, and LRS2 datasets demonstrate that LwCTHNet achieves a word error rate comparable to state-of-the-art methods, while outperforming

other lightweight models by a significant margin. Notably, this performance is attained with a substantially reduced parameter size and enhanced inference speed.

However, LwCTHNet still has certain limitations in extreme scenarios: for example, when dealing with English speech with extremely low signal-to-noise ratios or speech with strong accent variations, its WER increases more significantly compared to standard scenarios. In future work, we will further optimise the WEC module by introducing adaptive noise reduction mechanisms and accent-aware feature extraction strategies to enhance the model's ability to understand context in extreme scenarios, thereby expanding its applicability in more complex real-world environments. Besides, we also want to extend LwCTHNet to business speech analysis tasks with practical value, such as real-time extraction of key demand information from customer service speech logs and automatic identification of risk-related content in financial business communications, to explore its application potential in business intelligence scenarios. □

Declarations

All authors declare that they have no conflicts of interest.

References

Abdel-Hamid, O., Mohamed, A-r., Jiang, H. and Penn, G. (2012) 'Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition', *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, pp.4277–4280.

Anjum, M.F., Dasgupta, S., Mudumbai, R., Singh, A., Cavanagh, J.F. and Narayanan, N.S. (2020) 'Linear predictive coding distinguishes spectral EEG features of Parkinson's disease', *Parkinsonism & Related Disorders*, Vol. 79, pp.79–85.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J. et al. (2021) *GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio*, arXiv preprint arXiv:2106.06909.

Chen, X., Yi, B., Li, Q., Zhu, F., Nian, Y., Shankar, A., Nappi, M. and Tolba, A. (2025a) 'Deep customized network slicing and efficient routing for IoT applications in B5G-enabled edge computing networks', *IEEE Internet of Things Journal*, Vol. 12, No. 3, pp.2763–2774.

Chen, X., Zhu, F., Li, D., Li, Q., Anwar, M.S., Shan, G. and Jiang, J. (2025b) 'Towards clinically applicable large-model-based privacy-preserving polyp segmentation: a federated LoRA approach to colonoscopy', *IEEE Journal of Biomedical and Health Informatics*, pp.1–13.

Davis, K.H., Biddulph, R. and Balashek, S. (1952) 'Automatic recognition of spoken digits', *The Journal of the Acoustical Society of America*, Vol. 24, No. 6, pp.637–642.

Dong, L., Xu, S. and Xu, B. (2018) 'Speech-Transformer: a no-recurrence sequence-to-sequence model for speech recognition', *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Calgary, pp.5884–5888.

Fan, Q., Huang, H., Zhou, X. and He, R. (2023) 'Lightweight vision transformer with bidirectional interaction', *Advances in Neural Information Processing Systems*, Vol. 36, pp.15234–15251.

Gao, X., Niu, S., Wei, D., Liu, X., Wang, T., Zhu, F., Dong, J. and Sun, Q. (2024) 'Joint metric learning-based class-specific representation for image set classification', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 5, pp.6731–6745.

Gu, P., Zhang, Y., Wang, C. and Chen, D.Z. (2023) 'Convformer: combining CNN and transformer for medical image segmentation', *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Cartagena de Indias, Colombia, pp.1–5.

Gulati, A., Qin, J., Chiu, C-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. et al. (2020) *Conformer: Convolution-Augmented Transformer for Speech Recognition*, arXiv preprint arXiv:2005.08100.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, Vol. 521, No. 7553, pp.436–444.

Lee, K-F. (1988) *Automatic Speech Recognition: The Development of the SPHINX System*, Vol. 62, Springer Science & Business Media, New York.

Mehta, S. and Rastegari, M. (2021) *MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer*, arXiv preprint arXiv:2110.02178.

Panayotov, V., Chen, G., Povey, D. and Khudanpur, S. (2015) 'LibriSpeech: an ASR corpus based on public domain audio books', *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, South Brisbane, pp.5206–5210.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiát, M., Rastrow, A. et al. (2011) 'The subspace Gaussian mixture model – a structured model for speech recognition', *Computer Speech & Language*, Vol. 25, No. 2, pp.404–439.

Rabiner, L.R. (2002) 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE*, Vol. 77, No. 2, pp.257–286.

Song, Z. (2020) 'English speech recognition based on deep learning with multiple features', *Computing*, Vol. 102, No. 3, pp.663–682.

Taher, K.I. and Abdulazeez, A.M. (2021) 'Deep learning convolutional neural network for speech recognition: a review', *International Journal of Science and Business*, Vol. 5, No. 3, pp.1–14.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) 'Attention is all you need', *Advances in Neural Information Processing Systems*, Vol. 30, pp.5998–6008.

Veselỳ, K., Karafiát, M. and Grézl, F. (2011) 'Convolutive bottleneck network features for lvcsr', *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, IEEE, Hawaii, pp.42–47.

Vintsyuk, T.K. (1968) 'Speech discrimination by dynamic programming', *Cybernetics*, Vol. 4, No. 1, pp.52–57.

Wang, J., Gao, X., Zhu, F. and Chen, X. (2024) 'Exploring frontier technologies in video-based person re-identification: a survey on deep learning approach', *Computers, Materials & Continua*, Vol. 81, No. 1, pp.25–51.

Wang, Y., Mohamed, A., Le, D., Liu, C., Xiao, A., Mahadeokar, J., Huang, H., Tjandra, A., Zhang, X., Zhang, F., Fuegen, C., Zweig, G. and Seltzer, M.L. (2020) 'Transformer-based acoustic modeling for hybrid speech recognition', *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, pp.6874–6878.

Zhao, J. and Itti, L. (2018) 'shapeDTW: shape dynamic time warping', *Pattern Recognition*, Vol. 74, pp.171–184.

Zheng, W., Lu, S., Yang, Y., Yin, Z. and Yin, L. (2024) 'Lightweight transformer image feature extraction network', *PeerJ Computer Science*, Vol. 10, p.e1755.