frontiers | Frontiers in Communications and Networks

TYPE Original Research
PUBLISHED 15 October 2025
DOI 10.3389/frcmn.2025.1662788

OPEN ACCESS

EDITED BY
Utkal Mehta,
University of the South Pacific, Fiji

REVIEWED BY
Anil Gavade,
KLS Gogte Institute of Technology, India
Bo Wang,
University of Minho, Portugal

*CORRESPONDENCE
Xu Li,
lixu@xjnu.edu.cn

RECEIVED 09 July 2025
ACCEPTED 23 September 2025
PUBLISHED 15 October 2025

CITATION
Li H, Tang C, Yue X and Li X (2025) Sentence-level consistency of conformer based pre-training distillation for Chinese speech recognition.
Front. Commun. Netw. 6:1662788.
doi: 10.3389/frcmn.2025.1662788

Sentence-level consistency of conformer based pre-training distillation for Chinese speech recognition

Haifang Li, Chao Tang, Xin Yue and Xu Li*

School of Computer Science and Technology, Xinjiang Normal University, Xinjiang, China

Introduction: We address robustness and efficiency in Chinese automatic speech recognition (ASR), focusing on long-form broadcast speech where sentence-level semantic consistency is often lost.
Methods: We propose a Conformer-based framework that integrates sentence-level consistency with pre-training knowledge distillation. We also construct CH Broadcast ASR, a domain-specific Chinese corpus for the broadcast and television domain, and evaluate on AISHELL-1, AISHELL-3, and CH Broadcast ASR.
Results: The proposed model consistently outperforms strong baselines (TDNN, DFSMN-T, TCN-Transformer), achieving CER = 3.3% on AISHELL-1, 3.7% on AISHELL-3, and 3.9% on CH Broadcast ASR, while reducing model size by >10%.
Discussion: Enforcing sentence-level semantic alignment together with distillation improves robustness for long-form broadcast speech and enhances efficiency for real-time deployment.

KEYWORDS

speech recognition, model distillation, con-former, pre-trained mode, sentence-level consistency

1 Introduction

Speech recognition, a key technology in human–computer interaction, aims to convert raw speech signals into readable text through computational models. The fundamental pipeline of a- = n automatic speech recognition (ASR) system generally consists of three main stages: First, the acoustic signal is processed and transformed into spectral features such as Mel-frequency cepstral coefficients (MFCC) or filter banks (Davis and Mermelstein, 1980) Next, these features are passed into sequence modeling modules, where neural architectures such as Transformer or Conformer learn to map them into meaningful linguistic representations (Gulati et al., 2020; Zhang et al., 2022). Finally, the decoding stage integrates the predicted representations with language models or discriminative sequence criteria, such as lattice-free MMI, to generate the final transcription (Hadian et al., 2018). Figure 1 illustrates the overall AI pipeline used in this study, from speech input to recognized text, highlighting the integration of sentence-level consistency and knowledge distillation within the Conformer backbone.

Despite remarkable progress enabled by deep learning (Prabhavalkar et al., 2023), existing ASR approaches still face challenges. Traditional ASR models often rely on local or word-level objectives and therefore struggle to maintain consistency across long utterances. Recent advances such as chunk-wise attention and selective state space mechanisms have been proposed to improve robustness in streaming and long-form recognition scenarios

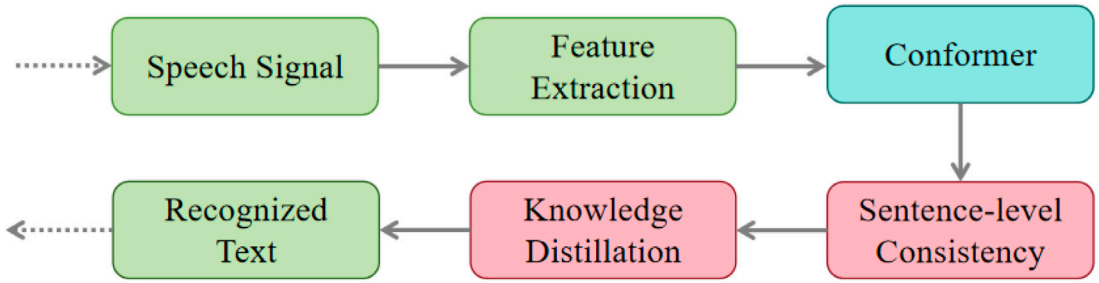

FIGURE 1
Conformer-based ASR pipeline integrating sentence-level consistency and knowledge distillation.

(Mimura et al., 2025a). In parallel, studies on long-form multimodal narration highlight the importance of maintaining semantic coherence across extended contexts (Zhang et al., 2024a), reinforcing the need for sentence-level objectives in speech recognition. In addition, large pre-trained models such as Transformer and Conformer deliver high accuracy (Gulati et al., 2020; Zhang et al., 2022) but suffer from heavy computational costs, making real-time deployment difficult (Fan et al., 2024; Zhao et al., 2022).

To address these limitations, this paper introduces a novel framework, that combines sentence-level consistency with model distillation (Tian et al., 2022), aiming to enhance both robustness and efficiency in Chinese broadcast-domain speech recognition. We further contribute a new domain-specific dataset, CH Broadcast ASR, to support evaluation.

The remainder of this paper is organized as follows: Section 2 reviews related work in speech recognition, model distillation, and pre-training. Section 3 introduces the CH Broadcast ASR corpus. Section 4 details the proposed methodology, including the sentence-level consistency module and the distillation strategy. Section 5 presents experimental setups and results. Section 6 discusses conclusions and future work.

2 Related work

2.1 Speech recognition

In the 21st century, deep learning techniques have undergone significant advancements. In 2011, Deng et al. from Microsoft Research introduced a breakthrough in speech recognition with the proposal of the deep neural network-Hidden Markov model (DNN-HMM) approach. The DNN model replaces the conventional Gaussian mixture model (GMM) in each state, leading to a remarkable reduction in the error rate. Nevertheless, the performance of DNN-HMM speech recognition models remains constrained by challenges such as forced segmentation, alignment, and the independent training of multiple modules inherent to HMM -based systems (Hadian et al., 2018).

Transformer-based architectures have subsequently brought substantial improvements to ASR by leveraging self-attention for long-range context modeling. Building on this foundation, recent studies have revisited convolution-free Transformer variants to further enhance both accuracy and computational efficiency (Vaswani et al., 2017; Hou et al., 2024). These advancements have effectively addressed several limitations of traditional speech recognition techniques.

Beyond architectural innovations, pre-training strategies for encoder–decoder models have attracted increasing attention. For example, Wang et al. (2024) proposed an encoder–decoder pre-training paradigm tailored to minority language ASR, demonstrating significant gains under scarce-resource conditions. Wu et al. (2023) introduced Wav2Seq, a self-supervised framework leveraging pseudo-languages to pre-train both encoder and decoder. More recently, researchers have explored multimodal and unified pre-training frameworks. Wang et al. (2024) presented a paralinguistics-aware speech-empowered large language model that leverages multimodal pre-training for conversational ASR, while Kim E. et al. (2024) introduced a joint end-to-end framework for spoken language understanding and ASR, based on unified speech-to-text pre-training. In addition, Baevski and Mohamed (2020) demonstrated the effectiveness of self-supervised pre-training for ASR, highlighting the potential of pre-trained representations to improve robustness in low-resource or domain-specific conditions.

Beyond scaling and pre-training, another line of research has targeted long-form and streaming ASR, which requires maintaining consistency across extended utterances. Mimura et al. (2025b) developed chunk-wise attention and trans-chunk selective state space mechanisms to improve robustness in streaming scenarios, while Zhang et al. (2024b) emphasized long-context modeling in multimodal narration tasks, underscoring the importance of semantic consistency. These insights motivate our focus on integrating sentence-level consistency into broadcast-domain ASR, where utterances are often long and context-rich.

2.2 Model distillation in speech recognition

While the Conformer model has become a strong backbone for speech recognition by effectively combining self-attention with convolutional modules, the challenge now lies less in designing new architectures and more in adapting large-scale pre-trained models for efficient deployment. To this end, knowledge distillation (KD) has emerged as a promising technique for transferring knowledge from large teacher models to smaller

TABLE 1 Topics in the mainstream Chinese corpus.

Datasets	Hours	Topics
Speech ocean	10	Daily conversation, etc.
THCHS30	30	Verses, books, etc.
Prime words	100	Voice chat and intelligent voice control
ST-CMDS	122	History, books, etc.
MAGICDATA	755	Interactive quiz, music search, etc.
Aishell_1	178	Technology, sports, entertainment, etc.
Aishell_2	1,000	Smart home, driverless, etc.
Aidatatang_200zh	200	General Telephone Corpus

student models, thereby reducing model size while maintaining accuracy (Zhao et al., 2022; Tian et al., 2022).

Several KD strategies have been explored in speech-related tasks. Zhao et al. (2022) proposed a module-replacing distillation strategy for RNN-Transducer models, demonstrating that structured KD can achieve substantial compression without significant loss of accuracy. Lee et al. (2024) further applied KD to the training of speech enhancement modules, showing improved robustness of ASR under noisy conditions. In addition, Sun et al. (2024) introduced a joint-embedding predictive KD framework for visual speech recognition, while Fan et al. (2025) nvestigated cross-modal KD with multi-stage adaptive feature fusion for speech separation. These works illustrate the versatility of KD across ASR, enhancement, and multimodal speech tasks.

In contrast to these prior efforts, which mainly emphasize frame-level or modality-specific objectives, our work focuses on sentence-level semantic consistency within the KD framework. Specifically, we integrate sentence-level alignment between speech and text embeddings into the distillation process. Moreover, we jointly optimize conventional token-level losses with sentence-level consistency, striking a balance between reducing computational overhead and improving robustness in long-form recognition. By doing so, our method not only compresses Conformer-based ASR models but also explicitly bridges the semantic gap between speech and text—an aspect insufficiently addressed in prior KD approaches.

3 Speech recognition corpus for broadcast and television

3.1 CH broadcast ASR

Table 1 provides an overview of existing open-source Chinese speech datasets, which primarily cater to general-purpose speech recognition tasks (Bu et al., 2017a; Li et al., 2019). How-ever, it is important to note that datasets specific to the broadcast domain are often not freely available or open source due to copyright restrictions. Therefore, this paper addresses this gap by introducing the construction of CH Broadcast ASR, a Chinese speech dataset specifically designed for the broadcast and television domain.

The data for the CH Broadcast ASR dataset primarily originates from FM radio and television broad-cast programs. The corpus construction process consists of the following steps:

1. Audio segmentation and text recognition: The audio files are segmented using the PyAnnote.audio toolkit (Bredin et al., 2020) and Baidu and Tencent's speech recognition interfaces are employed to transcribe the audio. The accuracy of the transcriptions from both interfaces is evaluated using the Bilingual Evaluation Understudy (BLEU) metric (Papineni et al., 2002) allowing for the filtering of unreliable text.
2. Audio auditing: This step involves verifying the integrity of the filtered audio, ensuring that sentence boundaries coincide with natural pauses in speech rather than interrupting speech segments.
3. Proofreading: The audio content is carefully cross-checked with the corresponding text to ensure alignment and accuracy between the two.

By following these steps, the CH Broadcast ASR dataset is constructed, providing a reliable and accurately transcribed speech corpus for the broadcast and television domain. Since the original audio contains the voices of multiple speakers, it is necessary to differentiate between speakers and segment the corpus into shorter audio files. To achieve this, the PyAnnote.audio toolkit is employed to split the audio into segments of natural durations.

Audio segmentation is a crucial step in the analysis of audio data as it involves dividing the continuous audio signal into uniform or "similar" segments. The process can be outlined as follows: First, Voice Activity Detection (VAD) is employed to ensure that the noise level in silent portions remains below 20dB, and both the pre-audio and post-audio silent activities last for at least 0.2 s. Additionally, the duration of each split audio segment should exceed 2 s. These requirements aid deep learning models in audio signal feature extraction. Based on previous data collation experiences, it was found that sections such as interviews, openings, and closings are often excluded. The resulting audio files are then subjected to identification using Baidu and Tencent interfaces to complete the annotation of the audio files. Subsequently, BLEU scores are calculated and segments with BLEU values less than 0.5 are selected for manual audio auditing.

During the audio segmentation process, complete accuracy is not guaranteed. Therefore, it becomes necessary to conduct manual reviews to assess the completeness of the audio, the presence of a single speaker, and the overall acoustic environment cleanliness. This meticulous evaluation ensures the quality of the audio. Only audio that satisfies all criteria is considered qualified and included in the corpus as usable data.

The transcribed text obtained after the screening undergoes further adjustments based on the actual content of the audio. Inappropriate content involving sensitive political issues, user privacy, pornography, violence, and so on, is removed. Additionally, numbers, dates, percentages, and other numerical expressions are converted into their corresponding Chinese readings. Furthermore, symbols such as '[', ']', '/', etc., are eliminated from the transcriptions.

Similar to most Chinese speech datasets, the finalized compliant corpus files are saved in a 16 kHz sample rate. Each file has a frame

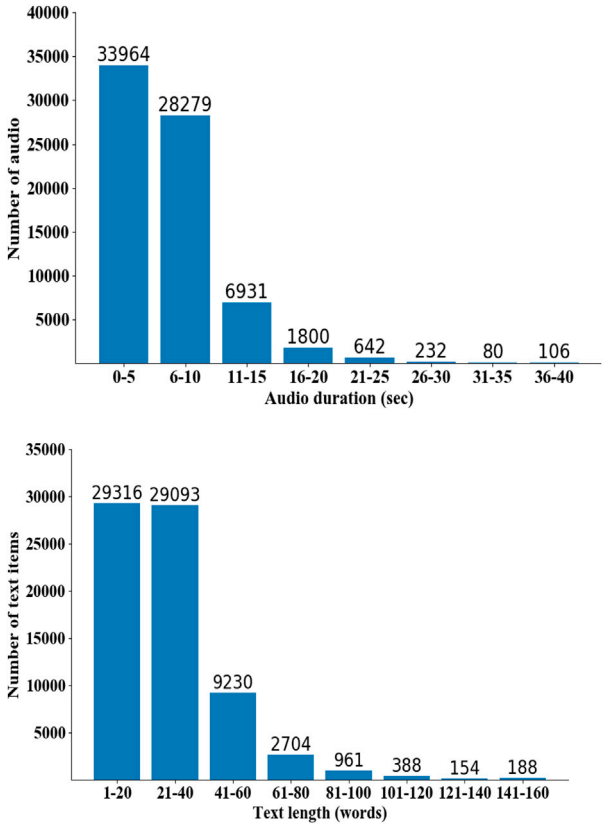

FIGURE 2
Corpus distribution information.

length of 25 ms and a frame shift length of 10 ms. Additionally, the files are in 16-bit monaural WAV format.

3.2 Corpus statistics

The CH Broadcast ASR dataset comprises 72,034 Chinese speech recognition data specifically collected from the broadcast and television domain. It encompasses a total duration of 127 h and features recordings from 23 speakers, including 12 males and 11 females. The training set consists of 56,891 sentences, equivalent to 100 h of speech, contributed by 7 male speakers and 6 female speakers. The validation includes 8,312 sentences, totaling 15 h of speech, and involves 2 male speakers and 2 female speakers. Finally, the test set comprises 6,831 sentences, accounting for 12 h of speech, featuring 3 male speakers and 3 female speakers. The distribution of the dataset is visually depicted in Figure 2.

On average, the audio duration in the dataset is around 6 s, while the average text length spans approximately 30 characters. Most of the texts fall within the range of 10–80 characters, with only a small portion extending beyond 100 characters. During the data processing phase, it was observed that the speech rate of the announcers tends to be faster compared to everyday spoken communication. As a result, broadcast audio contains more characters within the same time duration compared to other types of cor-puses. To streamline the model's training time and reduce the parameter count, content exceeding 160 characters in text length or 40 s in audio duration was excluded from the dataset.

4 Proposed methodology

The aforementioned challenges motivate our study. Current speech recognition approaches still face two critical problems:

Lack of sentence-level modeling: Existing models primarily optimize at the word or sub-word level, failing to capture semantic consistency across entire sentences. This limitation becomes especially problematic in domains such as broadcast and television, where utterances are typically long and context-rich.

High computational cost of large models: Although pre-trained Conformer and Transformer architectures achieve high accuracy, their massive parameter counts restrict scalability and real-time deployment.

To address these problems, we propose a Conformer-based approach that integrates sentence-level consistency modeling with pre-training distillation, aiming to simultaneously improve recognition accuracy and reduce model size.

4.1 Incorporating sentence-level consistency

In contrast to traditional cascaded speech recognition models, end-to-end automatic speech recognition (ASR) employs attention mechanisms to directly align output text with input audio by estimating probability distributions over acoustic sequences for each target token (Bu et al., 2017b; Karita et al., 2019). However, since audio and text lie in heterogeneous representation spaces, this mapping process can be complex and may lead to semantic inconsistencies or recognition errors.

The main objective of end-to-end ASR training is to minimize the loss between recognized words and their ground-truth transcriptions. While this approach establishes direct correspondences between audio and text, prior studies have predominantly optimized word-level or token-level relationships, often neglecting global sentence-level semantic consistency.

Although Transformer-based architectures and their recent convolution-free variants have significantly advanced ASR accuracy and efficiency (Hou et al., 2024), they still primarily capture local or phrase-level dependencies. By contrast, sentence-level similarity modeling has been effectively applied in natural language processing tasks, such as abstractive summarization using contrastive learning (Huang et al., 2024) and coherence modeling for sentence ordering (Logeswaran et al., 2018). These approaches highlight the importance of preserving global semantic consistency beyond local word alignments.

Motivated by these insights, we introduce a novel sentence-level consistency module into the Conformer-based ASR framework. The goal is to minimize discrepancies between source speech embeddings and target text embeddings at the sentence level, thereby complementing token-level supervision with global semantic alignment. In parallel, recent advances in long-form speech recognition have underscored the importance of maintaining consistency across extended sequences (Gong et al.,

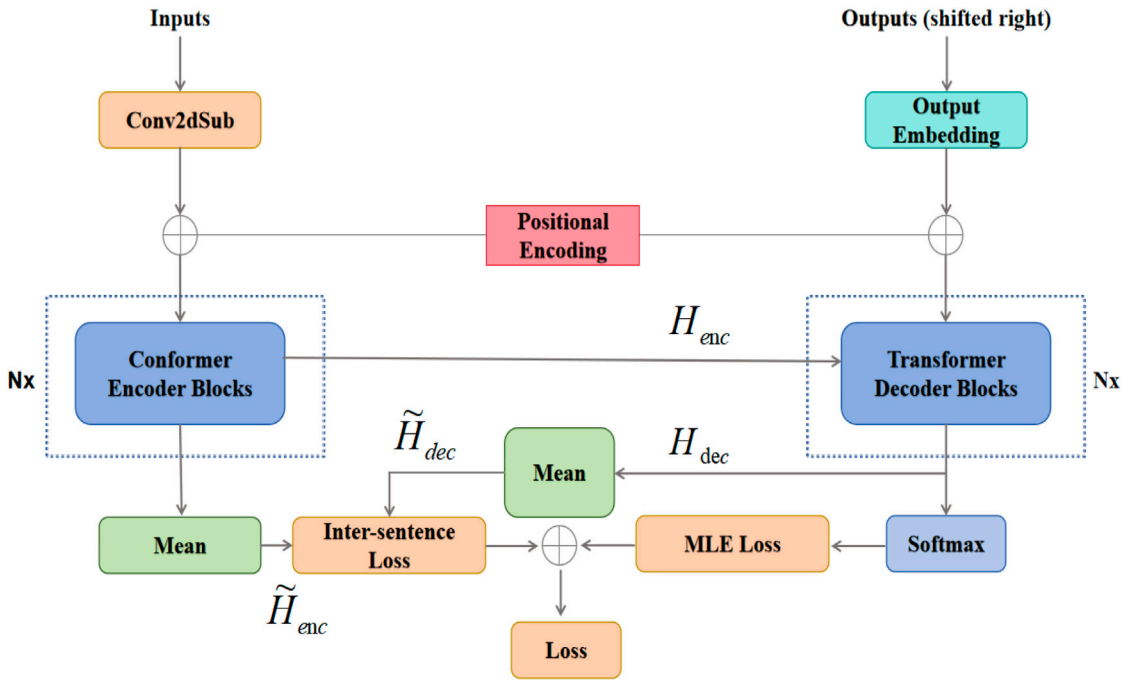

FIGURE 3
Conformer model combining sentence-level consistency.

2024), further motivating the integration of sentence-level objectives in broadcast-domain ASR. The structure of the proposed model is illustrated in Figure 3.

It has been shown that averaging operations provide an effective way to construct sentence-level representations by emphasizing global semantics while reducing local alignment noise. Specifically, we obtain the speech-level representation by averaging the encoder outputs, as in Equation 1.

$$h_s = \frac{1}{T}\sum_{i=1}^{T} h_i^{(enc)} \tag{1}$$

where T is the number of encoder frames and $h_i^{(\mathrm{enc})}$ denotes the hidden vector at frame i. Similarly, the text-level representation is obtained by averaging the decoder input embeddings, as in Equation 2.

$$h_t = \frac{1}{N}\sum_{j=1}^{N} h_j^{(dec)} \tag{2}$$

where N is the number of tokens and $h_j^{(dec)}$ is the embedding of token j.

While conventional ASR training objectives such as CTC or attention-based loss operate at the token or alignment level, they cannot guarantee that h_s and h_t are close in the semantic space. This often leads to unstable predictions, especially for long utterances. To address this limitation, we define a sentence-level consistency loss that directly minimizes the distance between these two representations, see Equation 3.

$$L_{sent} = \|h_s - h_t\|_2^2 \tag{3}$$

From an information-theoretic perspective, this loss reduces the conditional entropy $H(h_t \mid h_s)$, thereby enhancing the mutual information between speech and text modalities. In practice, averaging across tokens enables the model to capture global semantic meaning, while avoiding error accumulation caused by imperfect local alignments.

To fully exploit the advantages of both fine-grained and global supervision, we jointly optimize the original Conformer loss L_{conf} with the sentence-level consistency loss. The overall training objective is formulated as in Equation 4.

$$L = L_{conf} + \lambda L_{sent} \tag{4}$$

where λ is a weighting coefficient that balances local token-level learning and global sentence-level alignment. This hybrid objective guides the model to achieve accurate word-level predictions while maintaining semantic coherence at the sentence level, which is particularly beneficial in broadcast and television speech recognition tasks. We optimize the model using the embeddings in Equations 1, 2 and the losses in Equations 3, 4.

4.2 Pre-training-based model distillation method

The pre-trained Transformer architecture has achieved remarkable success in various NLP tasks such as text summarization, machine translation, question answering systems, and information extraction. However, these models often have an extensive parameter count, reaching hundreds of millions or even billions, which poses challenges in terms of computational and memory requirements. Consequently,

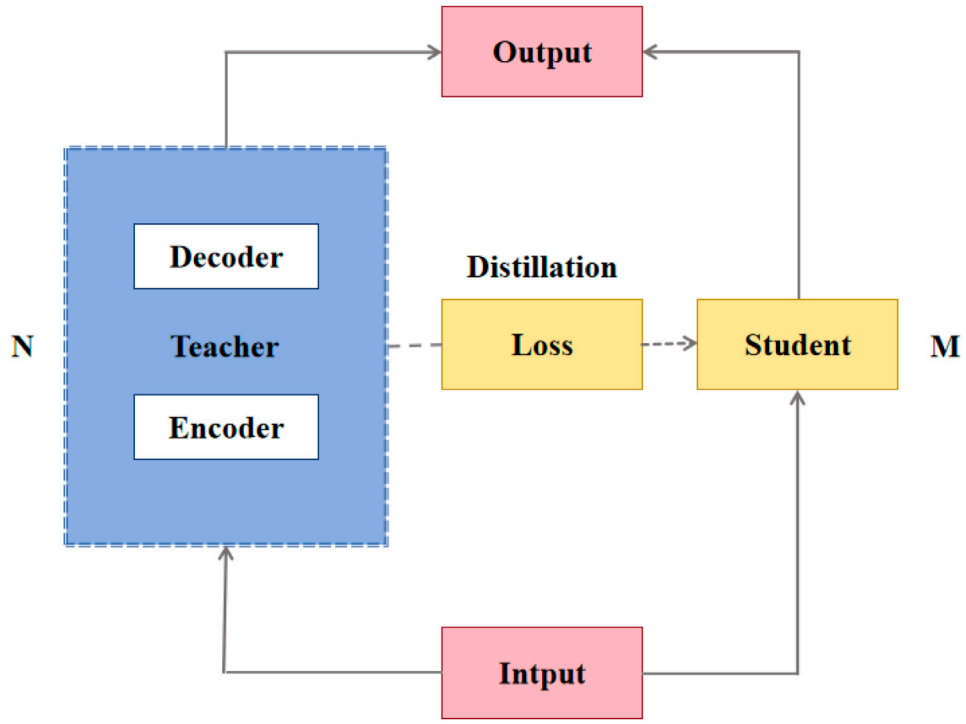

FIGURE 4
Structure of conformer distillation.

deploying such models in real-world scenarios, especially those with real-time and resource-constrained constraints, becomes challenging. To address these issues, research efforts have focused on model compression techniques, aiming to create faster and smaller models while maintaining comparable performance to the original model.

Model distillation is a technique that enables knowledge transfer from a large-scale teacher model to a small-scale student model. The goal is to reduce the model size, improve inference speed, and achieve performance that is essentially equivalent to the original model (Pham et al., 2021; Yoon et al., 2021). The general idea is to train smaller student models to emulate the performance of larger teacher models, including probabilistic outputs of downstream tasks, attention layers, and hidden layer representations. A common practice is to initialize the student model by copying weights from selected layers of the teacher model. For example, when training a 3-layer student model with a 6-layer teacher, layers 0, 3, and 5 of the teacher can be used to initialize layers 0, 1, and 2 of the student, respectively, thereby ensuring effective information transfer. Empirical evidence further shows that using the teacher's final layer tends to yield better results; thus, when the student consists of only a single layer, it should be initialized from the teacher's last layer rather than the first. Following this principle, we constructed the distillation structure as shown in Figure 4.

The loss function L for distillation includes the cross-entropy loss function L_{pred} for the downstream task and the distillation loss L_{dist}. The distillation loss comprises three components, as implemented in modern ASR toolkits such as PyTorch-Kaldi (Ravanelli et al., 2019), all of which are mean squared errors (MSE) (Lee et al., 2020) between the teacher and student models under different conditions: output probability L_{embd}, attention layer output L_{attn} (including errors in the encoder and decoder's respective attention layers as well as cross attention) and hidden layer output L_{hidn} (including errors in the encoder and decoder's respective hidden layer state). The loss function L is defined as shown in Equation 5.

$$L = \begin{Bmatrix} L_{embd} & m = 0 \\ L_{hidn} + L_{attn}, & 0 \le m \le M \\ L_{pred} & m = M + 1 \end{Bmatrix} \quad (5)$$

4.3 Synergy of sentence-level consistency and model distillation

Although sentence-level consistency and model distillation are effective independently, their integration in a joint training framework produces complementary benefits. The sentence-level loss enforces global semantic alignment between source speech and target text, reducing recognition errors that arise from long or complex utterances. In contrast, the distillation module primarily improves computational efficiency and stabilizes the optimization of smaller student models by transferring knowledge from a large teacher model.

When combined, sentence-level consistency acts as a semantic regularizer that guides both the teacher and student networks toward preserving holistic meaning, while distillation ensures that this information is efficiently compressed into the student model. This synergy prevents the student model from merely imitating local distributions and instead encourages it to capture sentence-level coherence.

Ablation results (see Section *Experiments*) confirm this complementary relationship: the consistency loss alone reduces sentence error rate significantly, while distillation alone mainly compresses the model without large accuracy gains. However, the

joint framework achieves both lower error rates and smaller model size, demonstrating that the two modules are not piecemeal but mutually reinforcing components of a unified training strategy.

4.4 Implementation details

To ensure reproducibility, we provide key implementation details and parameter settings for both the sentence-level consistency module and the distillation process.

4.4.1 Sentence-level consistency module

1. Sentence embeddings are obtained by averaging the encoder output vectors for speech and the decoder input embeddings for text.
2. The similarity loss is measured using mean squared error (MSE) between the two sentence-level embeddings.
3. A weighting coefficient $\lambda = 0.3$ is applied to balance the sentence-level consistency loss with the original CTC/attention loss during joint training.
4. Optimization is performed using the Adam optimizer with an initial learning rate of 1e-4, warmup steps of 25k, and gradient clipping set to 5.

4.4.2 Distillation process

1. The teacher model is a 12-layer Conformer encoder-decoder with 512 hidden units, while the student model has 6 layers and 256 hidden units.
2. For layer-wise initialization, every second layer of the teacher is used to initialize the student model.
3. The distillation loss includes three components: probability distribution alignment, attention map alignment, and hidden state alignment, all implemented with MSE.
4. The weights for these three components are set to 0.5 : 0.25: 0.25, respectively.
5. Temperature scaling with $T = 2.0$ is used for soft targets in the probability distillation.

4.4.3 Training configuration

1. All models are trained on 4 NVIDIA V100 GPUs with a batch size of 32 per GPU.
2. Early stopping with patience of 10 epochs is used to prevent overfitting.
3. These details ensure that our method can be reliably reproduced and evaluated by the research community.

5 Experiments and results

5.1 Baseline

We compared our method with various representative studies, including recurrent neural network-based models and recent studies based on self-supervised learning.

- TDNN (Waibel et al., 2013): Time-delay neural networks, which utilize phoneme-recognition-based to independently discover acoustic-phonetic features and temporal relationships, regardless of their position in time.
- DFSMN-T (Li et al., 2019): A Lightweight speech recognition system consisting of an acoustic model DF-SMN and language model transformer with fast deco-ding speed.
- TCN-Transformer (Xie et al., 2022): Transformer-based a fusion of temporal convolutio-nal neural networks and connected temporal classification.

TABLE 2 Information about datasets.

Datasets	—	Sentences	Hours
Aishell_1	Training	120,098	150
	Validation	144,326	18
	Testing	7,176	10
Aishell_3	Training	70,620	68
	Validation	8,803	9
	Testing	8,612	8
CH Broadcast ASR	Training	56,891	100
	Validation	8,312	15
	Testing	6,831	12

5.2 Datasets

The dataset information is presented in Table 2.

The dataset was divided into training, testing, and validation sets without crossover, In all experiments, we used 80-dimensional Fbank features. The audio data have a sample rate of 16 kHz, frame length of 25 ms, frame shift length of 10 ms, and are stored in 16-bit monaural WAV format.

5.3 Performance comparison

We show the performance of the baseline model and the proposed model on three different datasets in Table 3.

The results clearly demonstrate the superiority of the proposed model over the baseline model across all three datasets. This demonstrates that incorporating sentence-level consistency and model distillation effectively reduces both sentence and character error rates, while enhancing the model's robustness and reducing its size.

Notably, the performance of the proposed model on the CH Broadcast ASR dataset significantly out-performs that of the Transformer-based speech recognition model. This can be attributed to the dataset's longer average sentence length. The inclusion of sentence-level consistency improves alignment between sentences, thereby enhancing the model's recognition performance.

5.4 Ablation study

To assess the usability of the CH Broadcast ASR dataset, we conducted experiments utilizing four models and varying CTC weights ranging from 0.1 to 0.9. The results of these experiments are presented in Table 4.

TABLE 3 Performance of the proposed model and the baseline model on the three datasets.

Datasets	Models	W.corr (%)	SER (%)	CER (%)	Model size (MB)
Aishell_1	TDNN	95.3	39.6	3.8	158.76
	DFSMN-T	95.8	38.2	3.6	159.23
	TCN-Transformer	96.5	37.1	3.4	180.44
	Proposed model	97.0	36.6	3.3	164.03
Aishell_3	TDNN	94.8	38.8	4.2	141.86
	DFSMN-T	95.2	37.8	3.8	142.30
	TCN-Transformer	96.0	36.6	3.9	168.29
	Proposed model	96.3	36.7	3.7	147.99
CH Broadcast ASR	TDNN	95.8	40.4	4.8	160.11
	DFSMN-T	96.5	38.1	4.2	160.78
	TCN-Transformer	96.1	39.9	4.1	179.45
	Proposed model	97.1	36.5	3.9	163.13

TABLE 4 Performance of different neural networks on CH Broadcast ASR.

Model	Metric (%)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
RNN	W.corr	91.9	92.3	92.5	92.6	92.8	93.1	92.7	92.6	93.8
	SER	60.4	58.6	55.3	56.3	54.2	53.4	53.8	56.4	58.4
	CER	8.4	8.2	8.0	7.9	7.6	7.3	7.8	7.9	8.1
LSTM	W.corr	92.5	92.8	93.0	93.3	92.9	92.7	92.4	92.1	91.8
	SER	60.1	57.8	54.5	53.2	54.1	54.9	55.5	56.6	57.5
	CER	8.1	7.8	7.3	7.0	9.4	9.4	9.5	9.5	9.5
Transformer	W.corr	93.4	93.6	94.2	92.4	92.2	92.3	92.1	92	92.1
	SER	46.5	46.1	44.5	46.6	48.1	47.8	48.5	48.8	48.6
	CER	7.4	7.4	6.5	9.3	9.4	9.4	9.5	9.5	9.5
Conformer	W.corr	94.7	92.5	95.2	95.2	95.5	94.8	94.6	94.2	92.2
	SER	42.5	45.8	41.3	41.5	41.3	41.9	42.4	43.5	44.8
	CER	5.9	8.1	5.5	5.5	5.1	5.6	5.9	6.2	6.5

The results demonstrate that the recurrent neural network, self-attentive mechanism, and Conformer with convolutional neural network all exhibit excellent performance on the CH Broadcast ASR dataset. In the subsequent analysis, we augment the dataset through speed perturbation techniques and evaluate the impact on model performance.

To facilitate a more efficient comprehension of the conclusions and to visually analyse the impact of the fused sentence-level consistency and model distillation method proposed in this study on model performance improvement, we conducted ablation studies on the Aishell_1 and CH Broadcast ASR datasets. Figure 5 presents the results of these studies, where "Base" represents the Conformer model with CTC loss and speed perturbation augmentation after pre-training, "+Loss" indicates the end-to-end speech recognition model incorporating sentence-level consistency, and "+Distil" denotes the model after distillation.

Incorporating sentence-level consistency into the Base model leads to a significant reduction in the sentence error rate. Additionally, the character error rate is also notably reduced, particularly in the CH Broadcast ASR dataset, which comprises long sentences. Conversely, when the model distillation method is employed independently, the error rate does not exhibit a significant decrease. However, when both methods are combined, the error rate is further reduced, indicating that model distillation also contributes to improved model performance.

Figure 6 illustrates the performance variation on the CH Broadcast ASR dataset after applying speed perturbation. The results show that this augmentation substantially improves robustness, with the Transformer model exhibiting the most pronounced reduction in

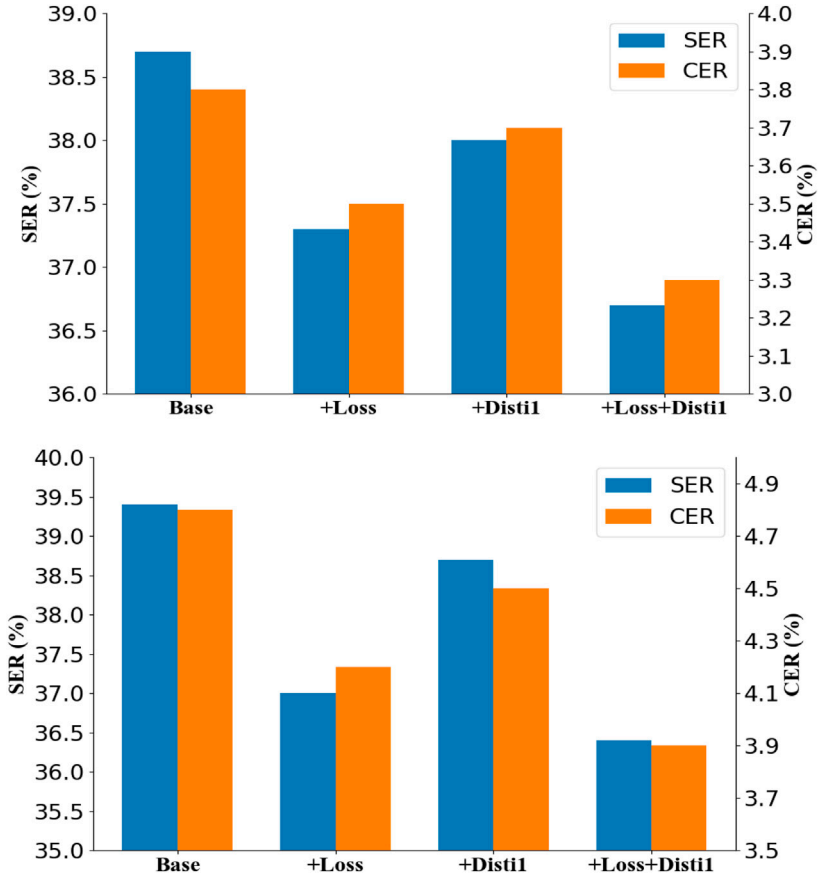

FIGURE 5
Ablation experiments on the Aishell_1 and CH Broadcast ASR datasets.

CER and increase in W. corr. This indicates that speed perturbation is particularly effective in alleviating the instability of self-attention mechanisms when handling long and complex sequences. In comparison, the Conformer maintains consistently superior performance, reflecting the advantages of its hybrid architecture that integrates convolution and self-attention to adapt to temporal variations more effectively.

Furthermore, the figure highlights the benefits of combining speed perturbation with CTC training. Unlike phonetic perturbation, CTC leads to faster convergence and more stable training dynamics due to its many-to-one alignment property, which reduces the impact of local temporal distortions. These observations not only validate the quality and reliability of the CH Broadcast ASR dataset but also confirm that appropriate data augmentation strategies can enhance both accuracy and generalization. This provides strong evidence for the practical applicability of the proposed approach in broadcast-domain speech recognition.

6 Conclusion

In this paper, we have presented a Chinese corpus specifically designed for speech recognition tasks in the broadcast and TV-based domain. We have also pro- posed a novel approach that combines model distillation and an end-to-end speech recognition model with sentence-level consistency to improve algorithmic ac- curacy. Our objective was to address the challenges associated with end-to-end model, which can have a significant impact on the accuracy and robustness of speech recognition systems.

The experimental results demonstrate the effectiveness of our method in improving the accuracy and robustness of Chinese speech recognition across different datasets. Specifically, our approach outperformed state-of-the-art methods in terms of multiple evaluation metrics, including CER and SER. These results validate the effectiveness of our approach in tackling challenging scenarios.

Experimental results demonstrate that our proposed sentence-level consistency distillation framework outperforms standard Transformer and Conformer baselines, achieving a relative CER reduction of 7.5% on CH Broadcast ASR and 5.8% on AISHELL-1. Furthermore, compared with conventional knowledge distillation methods, our approach delivers superior robustness in long-form speech recognition while maintaining real-time efficiency through significant model compression. These results highlight the practical value of our method for broadcast-domain ASR, where both accuracy and deployment efficiency are critical.

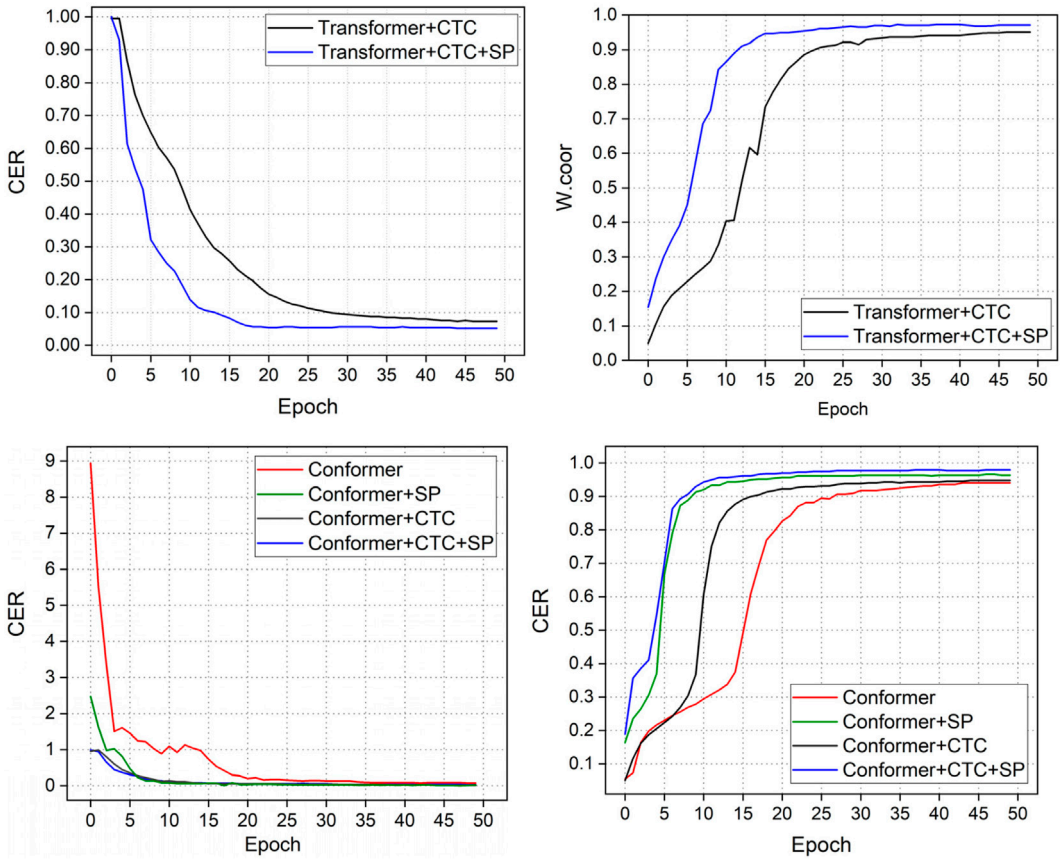

FIGURE 6
Model performance on CH Broadcast ASR combined with data augmentation.

Data availability statement

The dataset supporting this study is publicly available at: https://github.com/zsdflash/CH_Broadcast_ASR.

Author contributions

HL: Funding acquisition, Writing – review and editing, Writing – original draft, Resources. CT: Data curation, Validation, Writing – review and editing, Project administration, Supervision. XY: Validation, Writing – review and editing, Project administration, Data curation, Supervision. XL: Supervision, Validation, Project administration, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The Doctoral Research Foundation of Xinjiang Normal University (no.XJNUZBS2411); the Xinjiang Tianchi Youth Fund Project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baevski, A., and Mohamed, A. (2020). "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 7694–7698.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., and Gill, M. P. (2020). "Pyannote. Audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 7124–7128.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (2017a). "Aishell-1: an open-source mandarin speech corpus and a speech recognition base-line," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)* (IEEE), 1–5.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (2017b). "Aishell-1: an open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)* (IEEE), 1–5.

Davis, S., and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. speech, signal Process.* 28 (4), 357–366. doi:10.1109/tassp.1980.1163420

Fan, Z., Zhang, X., Huang, M., and Bu, Z. (2024). Sampleformer: an efficient conformer-based neural network for automatic speech recognition. *Intell. Data Anal.* 28 (1), 1647–1659. doi:10.3233/ida-230612

Fan, C., Xiang, W., Tao, J., Yi, J., and Lv, Z. (2025). Cross-Modal knowledge distillation with multi-stage adaptive feature fusion for speech separation. *IEEE Trans. Audio, Speech Lang. Process.* 33, 935–948. doi:10.1109/taslpro.2025.3533359

Gong, X., Wu, Y., Li, J., Liu, S., Zhao, R., Chen, X., et al. (2024). Advanced long-content speech recognition with factorized neural transducer. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 32, 1803–1815. doi:10.1109/taslp.2024.3350893

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., et al. (2020). Conformer: convolution-augmented transformer for speech recognition. *Proc. Interspeech*, 5036–5040. doi:10.21437/interspeech.2020-3015

Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. *Proc. Interspeech*, 12–16. doi:10.21437/interspeech.2018-1423

Hou, Z., Huybrechts, G., Bhatia, A., Garcia-Romero, D., Han, K., and Kirchhoff, K. (2024). Revisiting convolution-free transformer for speech recognition. 4568, 4572. doi:10.21437/interspeech.2024-588

Huang, Y., Li, Z., Chen, Z., Zhang, C., and Ma, H. (2024). Sentence salience contrastive learning for abstractive text summarization. *Neurocomputing* 593, 127808. doi:10.1016/j.neucom.2024.127808

Karita, S., Chen, N., Hayashi, T., Hori, T., and Zhang, W. (2019). "A comparative study on transformer vs rnn in speech applications," in *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (IEEE), 449–456.

Kim, E., Tang, Y., Ki, T., Neelagiri, D., and Apsingek, V. R. (2024). "Joint end-to-end spoken language understanding and automatic speech recognition training based on unified speech-to-text pre-training," in *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 10971–10975.

Lee, W., Lee, J., Kim, D., and Ham, B. (2020). "Learning with privileged information for efficient image super-resolution," in *Computer Vision–ECCV 2020: 16Th european conference, Glasgow, UK, August 23–28, 2020, proceedings, Part XXIV 16* (Springer), 465–482.

Lee, G. W., Kim, H. K., and Kong, D. J. (2024). Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition. *IEEE Access* 12, 72707–72720. doi:10.1109/access.2024.3403761

Li, J., Wang, X., and Li, Y. (2019). "The speechtransformer for large-scale mandarin chinese speech recognition," in *ICASSP 2019 - IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 7095–7099. doi:10.1109/icassp.2019.8682586

Logeswaran, L., Lee, H., and Radev, D. (2018). in Sentence ordering and coherence modeling using recurrent neural networks. In Proceedings of the AAAI Conference on Artificial Intelligence.32

Mimura, M., Moriya, T., and Matsuura, K. (2025a). "Advancing streaming ASR with chunk-wise attention and trans-chunk selective state spaces," in *ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 1–5.

Mimura, M., Moriya, T., and Matsuura, K. (2025b). Advancing streaming ASR with chunk-wise attention and trans-chunk selective state spaces. *Proc. ICASSP*, 1–5. doi:10.1109/icassp49660.2025.10889802

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002). "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311–318.

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. (2021). "Meta pseudo labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11557–11568.

Prabhavalkar, R., Hori, T., Sainath, T. N., Schlüter, R., and Watanabe, S. (2023). End-to-end speech recognition: a survey. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 32, 325–351. doi:10.1109/taslp.2023.3328283

Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (IEEE), 6465–6469.

Sun, C., Qin, B., and Yang, H. (2024). Jep-kd: joint-embedding predictive architecture based knowledge distillation for visual speech recognition. *IEEE Open J. Signal Process.* 5, 1147–1152. doi:10.1109/ojsp.2024.3496819

Tian, S., Deng, K., Li, Z., Ye, L., Cheng, G., Li, T., et al. (2022). Knowledge distillation for CTC-based speech recognition *via* consistent acoustic representation learning. *Proc. Interspeech*, 2022 243–247. doi:10.21437/interspeech.2022-775

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 5998–6008. doi:10.1007/978-3-031-84300-6_13

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (2013). Phoneme recognition using time-delay neural networks, In Editors: D. E. Rumelhart, G. E. Hinton, and R. J. Williams *Backpropagation: Theory, Architectures, and Applications* 35–61. (New York, NY, USA: Psychology Press).

Wang, B., Jin, X., Yu, M., Wang, G., and Chen, J. (2024). "Pre-training encoder-decoder for minority language speech recognition," in *2024 international joint conference on neural networks (IJCNN)* (IEEE), 1–8.

Wang, C., Liao, M., Huang, Z., Wu, J., Zong, C., and Zhang, J. (2024). BLSP-Emo: Towards empathetic large speech-language models. *Proceedings of EMNLP 2024*. doi:10.18653/v1/2024.emnlp-main.1070

Wu, Y., Kim, K., Watanabe, S., Han, K. J., McDonald, R., Weinberger, K. Q., et al. (2023). Wav2Seq: pre-Training speech-to-text encoder–decoder models using pseudo languages. *Proc. ICASSP*, 1–5. doi:10.1109/icassp49357.2023.10096988

Xie, X., Chen, G., Ssun, J., and Chen, Q. (2022). Tcn-transformer-ctc for end-to-end speech recognition. *Appl. Res. Computers/Jisuanji Yingyong Yanjiu* 39 (3). doi:10.19734/j.issn.1001-3695.2021.08.0323

Yoon, J. W., Lee, H., Kim, H. Y., Cho, W. I., and Kim, N. S. (2021). Tutornet: towards flexible knowledge distillation for end-to-end speech recognition. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* 29, 1626–1638. doi:10.1109/taslp.2021.3071662

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., et al. (2022). BigSSL: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE J. Sel. Top. Signal Process.* 16 (6), 1519–1532. doi:10.1109/jstsp.2022.3182537

Zhang, C., Lin, K., Yang, Z., Wang, J., and Wang, L. (2024a). "Mm-narrator: narrating long-form videos with multimodal in-context learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13647–13657.

Zhang, C., Lin, K., Yang, Z., Wang, J., Li, L., Lin, C. C., et al. (2024b). MM-Narrator: narrating long-form videos with multimodal in-context learning. *Proc. IEEE/CVF CVPR*, 13647–13657. doi:10.1109/cvpr52733.2024.01295

Zhao, K., Nguyen, H. D., Jain, A., Susanj, N., Mouchtaris, A., Gupta, L., et al. (2022). "Knowledge distillation *via* module replacing for automatic speech recognition with recurrent neural network transducer," in *23rd interspeech conference*. doi:10.21437/interspeech.2022-500