Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.046746

ARTICLE

Audio-Text Multimodal Speech Recognition via Dual-Tower Architecture for Mandarin Air Traffic Control Communications

Shuting Ge[1,2], Jin Ren[2,3,*], Yihua Shi[4], Yujun Zhang[1], Shunzhi Yang[2] and Jinfeng Yang[2]

[1]School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China

[2]Institute of Applied Artificial Intelligence of the Guangdong-Hong Kong-Macao Greater Bay Area, Shenzhen Polytechnic University, Shenzhen, 518055, China

[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055, China

[4]Industrial Training Centre, Shenzhen Polytechnic University, Shenzhen, 518055, China

*Corresponding Author: Jin Ren. Email: renjin666@szpu.edu.cn

Received: 13 October 2023 Accepted: 18 December 2023 Published: 26 March 2024

ABSTRACT

In air traffic control communications (ATCC), misunderstandings between pilots and controllers could result in fatal aviation accidents. Fortunately, advanced automatic speech recognition technology has emerged as a promising means of preventing miscommunications and enhancing aviation safety. However, most existing speech recognition methods merely incorporate external language models on the decoder side, leading to insufficient semantic alignment between speech and text modalities during the encoding phase. Furthermore, it is challenging to model acoustic context dependencies over long distances due to the longer speech sequences than text, especially for the extended ATCC data. To address these issues, we propose a speech-text multimodal dual-tower architecture for speech recognition. It employs cross-modal interactions to achieve close semantic alignment during the encoding stage and strengthen its capabilities in modeling auditory long-distance context dependencies. In addition, a two-stage training strategy is elaborately devised to derive semantics-aware acoustic representations effectively. The first stage focuses on pre-training the speech-text multimodal encoding module to enhance inter-modal semantic alignment and aural long-distance context dependencies. The second stage fine-tunes the entire network to bridge the input modality variation gap between the training and inference phases and boost generalization performance. Extensive experiments demonstrate the effectiveness of the proposed speech-text multimodal speech recognition method on the ATCC and AISHELL-1 datasets. It reduces the character error rate to 6.54% and 8.73%, respectively, and exhibits substantial performance gains of 28.76% and 23.82% compared with the best baseline model. The case studies indicate that the obtained semantics-aware acoustic representations aid in accurately recognizing terms with similar pronunciations but distinctive semantics. The research provides a novel modeling paradigm for semantics-aware speech recognition in air traffic control communications, which could contribute to the advancement of intelligent and efficient aviation safety management.

KEYWORDS

Speech-text multimodal; automatic speech recognition; semantic alignment; air traffic control communications; dual-tower architecture

1 Introduction

Air traffic controllers (ATCOs) and pilots interact via radio to confirm vital flight information, such as flight phase, flight status, traffic situation, and weather conditions, to maintain the safety and reliability of aircraft flight in a double-check manner [1], as illustrated in Fig. 1. Nonetheless, misunderstandings may occur by accident due to background noise disturbance, lack of concentration, tiredness, and overwhelming pressure, leading to catastrophic aviation accidents [2–4]. In particular, as air traffic volume rises, the high-pressure workload ATCOs face further drives up the risk of miscommunications [5]. Fortunately, recent speech and language processing advances have shed light on the automation of air traffic control communications (ATCC) [6,7]. Automatic speech recognition (ASR) technologies make it practical to transcribe utterances to text and verify the semantic consistency between the ATCO's instructions and the pilot's readbacks, thereby reducing the workload of ATCOs and enhancing aviation safety [8,9].

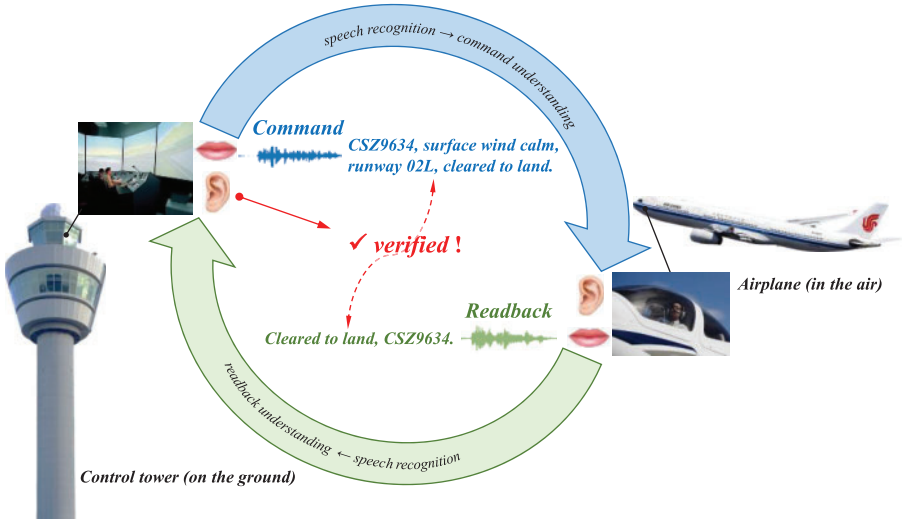

Figure 1: The double-check interaction procedure in air traffic control communication between air traffic controllers (ATCOs) and pilots. The whole process comprises three steps: (1) the ATCO issues an instruction to the pilot; (2) the pilot reads back the perceived instruction to the ATCO; and (3) the ATCO ensures the flight safety by verifying the semantic consistency between his/her issued instruction and the readback received from the pilot

ASR technology has evolved dramatically in recent decades. As its representative, end-to-end (E2E) ASR approaches directly model the mapping from raw voice signals to text outputs, simplifying architecture design and training processes of conventional ASR systems and attracting research interest [10–13]. ASR models developed for general domains face limitations in directly addressing the unique challenges of air traffic control communications (ATCC), including scarce data, prevalent background noise, multilingualism, and challenging accents. Numerous research efforts have emerged in recent years, attempting to design ASR models more suitable for the ATCC domain. The works [14,15] addressed the challenge of data scarcity in ATCC and developed a robust ASR system for ATCC on a limited dataset. Lin et al. explored a series of ASR solutions for ATCC in the context of multilingual scenarios [16–18]. To address environmental background noise, solutions like adversarial generation [19], multi-task learning [20,21], and self-supervised learning [22,23] offer practical strategies for building more robust speech recognition models for ATCC. However, a

significant challenge lies in semantic information, which is the focus of this study as one of the crucial directions for advancing spoken language understanding tasks. Extensive research in various domains has demonstrated its importance [24,25], making it equally vital for enhancing performance in speech recognition models for ATCC. However, most existing methods [26,27] merely incorporate external language models on the decoder side and capture the interactions between speech and text relatively separately without explicit semantic alignment guidance, leading to insufficient semantics-aware acoustic modeling during the encoding phase.

Inspired by how the brain's auditory cortex and language center work together to help humans comprehend spoken language [28,29], speech-text multimodal collaborative processing for speech recognition is a promising solution to alleviate the concerns above. Nevertheless, multimodal joint modeling must address the inherent heterogeneity challenge between speech and text modalities [30,31]. Text is a sequence of discrete symbols that convey human linguistic syntax, semantics, and pragmatics. In contrast, speech is a continuous-wave signal that can be discretized into a protracted frame sequence with substantial redundancy between adjacent frames. Specifically, as depicted in Fig. 2, the speech-text heterogeneity is reflected in the following four aspects:

- **Information components.** The text conveys primarily semantic information, whereas speech encompasses diverse components, such as semantics, speaker characteristics, emotions, accents, and surrounding noises.
- **Information density.** The text is more informative intensive than speech; thus, arbitrary masking of a few textual tokens can considerably influence the semantics, while masking random frames is unlikely to corrupt the speech notably.
- **Sequence length.** Speech sequences are much longer than their corresponding text sequences.
- **Context-dependent distance.** Text sequences can model longer context dependencies, while speech sequences typically only model adjacent local frames.

Speech modality inherently contains more abundant components than text modality, which makes it hard for semantic-focused speech recognition models in ATCC to extract semantically relevant information from the intricate parts of the speech signals. Furthermore, speech signals are generally processed at the frame level, leading to longer acoustic sequences than text, especially for the ATCC data characterized by long durations. However, speech models can only effectively capture short-range dependencies between locally adjacent frames, which further limits the modeling of acoustic long-distance context dependencies and affects recognition performance.

Fortunately, from the opposite perspective, the heterogeneity between speech and text modalities could also serve as an inspiration for their collaborative modeling, rather than just a challenge. Text modality with more intensive semantics and longer context dependencies than speech could tackle the mentioned difficulties. Ideally, by incorporating these beneficial textual semantics into speech encoding for ASR in a cross-modal cooperative fashion, the derived semantics-aware acoustic representation will achieve a close semantic alignment between speech and text, thereby enhancing the performance of ASR.

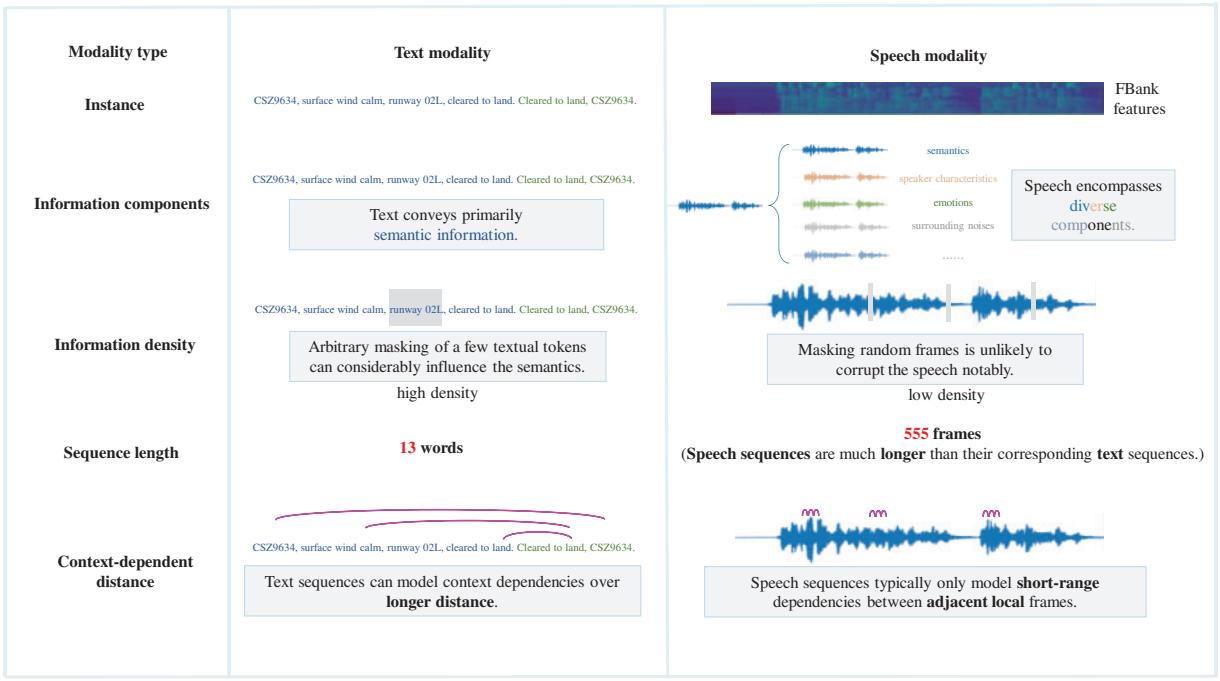

Figure 2: The illustration of the inherent heterogeneity between speech and text modalities regarding information components, information density, sequence length, and context-dependent distance

Along this line of thought, a speech-text multimodal dual-tower architecture is proposed in this work for speech recognition of Mandarin Air Traffic Control Communications, which employs cross-modal interactions to align semantic information across modalities. In addition, a two-stage training strategy is elaborately devised to derive semantics-aware acoustic representations effectively from paired speech and text data. In the first phase, the speech-text multimodal encoding module is pre-trained to enhance inter-modal semantic alignment and acoustic long-distance context dependencies with the help of masked language modeling (MLM) and cross-modal masked acoustic modeling (CMAM) strategies. In the second stage, two methods are adopted to fine-tune the entire network, including deactivating the text encoder or employing all <mask> tokens as text-side input. Therefore, it bridges the input modality variation gap between the training and inference phases and boosts the generalization performance thanks to better adaptation to speech-only inputs. The main contributions of this paper can be summarized as follows:

- We propose a speech-text multimodal dual-tower framework for speech recognition of Mandarin air traffic control communications. By incorporating multimodal information via intra-modal and inter-modal interactions, the ASR model can achieve a close semantic alignment between speech and text modalities during the encoding stage and superior recognition performance while preserving efficiency.
- We elaborately devise a two-stage training strategy to derive semantics-aware acoustic representations effectively. The first stage focuses on pre-training the speech-text multimodal encoding module to enhance inter-modal semantic alignment and acoustic long-distance context dependencies. The second stage fine-tunes the entire network to bridge the input modality variation gap between the training and inference phases and boost the generalization performance.

- We demonstrate the effectiveness of the proposed multimodal speech recognition method on the ATCC and public AISHELL-1 datasets through extensive experiments, including comparative experiments, ablation studies, and case studies.

The remaining sections of the paper are organized as follows. Section 2 discusses the related work. Section 3 comprehensively describes the model architecture and training strategy for speech-text multimodal speech recognition in Mandarin ATCC. Section 4 introduces the datasets and experimental setup and presents the analysis of the experimental results. Lastly, Section 5 provides a summary of the research.

2 Related Work

2.1 Speech-Text Multimodal Structure

In recent years, inspired by how humans perceive the world through multiple sources, employing multimodal information interactions instead of unimodal data for effective representation learning has gained popularity [30,32]. As crucial information mediums for human communication, the distinct modalities of speech and text mutually complement each other, which is also the focus of this study. Numerous studies leveraging the multimodal interaction between speech and text have demonstrated superior performance over unimodal approaches in various tasks, including automatic speech recognition [33,34], speech emotion recognition [35,36], and spoken language understanding [37,38].

From the perspective of model architecture, speech-text multimodal structures can be broadly divided into three categories: single-tower networks [39,40], dual-tower architectures [41,42], and encoder-decoder frameworks [43–45], as presented in Table 1. In the single-tower network, the speech and text tokens are embedded and concatenated as a unified input sequence for a shared multimodal encoder, enabling the model to learn their representations with semantic associations. SpeechBERT [40] takes aligned speech-text pairs as joint input for the shared encoder and constructs an end-to-end model for spoken question answering. However, speech and text are lengthy sequences, leading to high computational complexity and inevitable information loss when concatenated as input. In the dual-tower architecture, the two-modal information flows separately into their respective encoding branches, enabling them to learn their correlation collaboratively via interaction in the middle or later stages. The work [42] employed cross-attention and self-attention modules to explore the inter-modal and intra-modal interactions between acoustic and textual features. Encoder-decoder framework aims to acquire shared semantics by jointly training the encoder and decoder. SpeechT5 [43] converts diverse speech processing tasks into speech/text-to-speech/text problems, allowing cooperative learning of representations for both modalities to enhance cross-modal modeling capabilities. While the encoder-decoder framework has advantages for generative tasks, differences in optimization objectives between the encoder and decoder may lead to interference when handling various tasks. Alternatively, a discrete processing module can bridge the semantic gap between speech and text in an intermediate manner. SpeechUT [45] demonstrates that hidden units can effectively convey the relationships between speech and text modalities by decomposing the speech-to-text model into a speech-to-unit model and a unit-to-text model. CodeBERT [44] transforms speech into discrete code form, and the model learns speech-text multimodal representations via self-supervised tasks involving code prediction. The discrete code for speech and text requires careful design; otherwise, it may lead to information loss and pose challenges during model training.

Table 1: Overview of the relevant literature on speech-text multimodal structure

References		Main challenges concern	Dataset	Supervision	Multilingual	Evaluation criteria	Tasks
Single-tower	[39]	1. Limited label data; 2. Cross-modal representation discrepancy.	MuST-C; MT Dataset	Supervised	✓	BLEU	Speech translation
	[40]	1. Cross-modal representation discrepancy; 2. Contextual semantic comprehension.	SQuAD	Supervised	×	F1 score; Exact Match Score	Spoken question answering
Dual-tower	[41]	1. Speech and text representation discrepancy; 2. Catastrophic forgetting.	IARPA BABEL; AISHELL-1	Supervised	✓	CER	Speech recognition
	[42]	Effectively integrate textual and acoustic information.	IEMOCAP; MELD	Supervised	×	Overall accuracy; Average accuracy	Speech emotion Recognition
Encoder-Decoder	[43]	1. Ignored text data and modality transformation; 2. Absence of pre-trained sequence Decoders.	LibriSpeech; LibriTTS; 3. MuST-C	Self-supervised	✓	WER; MOS/CMOS; BLEU	Speech recognition; Text to speech; Speech translation
	[44]	Alignment of unpaired speech and text representations.	LibriSpeech; CoVoT	Self-supervised	✓	WER; BLEU	Speech recognition; Speech translation
	[45]	The mapping of speech and text.	LibriSpeech; MuST-C	Self-supervised	✓	WER; BLEU	Speech recognition; Speech translation

In this paper, we design a dual-tower structure to jointly model speech and text modalities during the encoding phase for enhanced efficiency and flexibility. Each modality branch is scheduled with distinct optimization objectives to learn the intrinsic information within each modality. In addition, the model learns connections between modalities through a flexible cross-modal interaction mechanism, thereby acquiring richer contextual semantic information.

2.2 Speech-Text Multimodal Speech Recognition

Conventional hybrid-based and end-to-end (E2E) [12] acoustic modeling are the two significant research categories for automatic speech recognition (ASR). The hybrid-based ASR method comprises several independently optimized model components. In the early period, most techniques relied on Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) modeling. With the advancement of deep learning technology, deep neural networks (DNNs) have replaced GMM to estimate the probability of HMM states, leading to the development of the DNN-HMM [46] framework. However, submodules are cascaded in the hybrid-based approach after being individually trained and optimized for their targets. This results in error accumulation, propagation, and potential inconsistency between optimal local and global solutions. Furthermore, the hybrid model heavily relies on the strict alignment of annotation information. End-to-end modeling mitigates these issues by directly integrating all modules into a unified system for joint optimization. The widespread use of connectionist temporal classification (CTC) [47] enables automatic variable-length mapping from speech frame features to output label sequences, creating a new paradigm for end-to-end speech recognition. Additionally, attention-based ASR methods are also frequently employed. The attention-based encoder-decoder [48,49] does not require preliminary alignment information. Instead, it concentrates dynamically and flexibly on various portions of the input sequence based on the input acoustic features and previous context information.

Even though these unimodal works [48] play a dominant role in ASR applications, more modalities, such as audio-visual [50,51] or speech-text [52,53], have been employed to enhance acoustic representation with precise semantic alignment [54], as presented in Table 2. The paper [55] demonstrated that multimodal models incorporating heterogeneous features are preferable to unimodal models. A type of work explores incorporating the pre-trained BERT model into the ASR model to inject text knowledge [56]; however, it cannot effectively merge multimodal context information in a shared space. To alleviate the above problem, Zheng et al. [41] integrated BERT and wav2vec 2.0 as an end-to-end unified framework, which uses the attention and representation aggregation modules to facilitate the cooperative learning of the speech-text pre-training model.

Table 2: Overview of the relevant literature on speech-text multimodal speech recognition

References		Main challenges concern	Dataset	Supervision	Multilingual	Evaluation criteria	Tasks
Hybrid-based	[46]	1. Statistical inefficiency and limited capacity of GMMs; 2. Computational efficiency	TIMIT	Supervised	×	WER	Speech recognition
End-to-end	[47]	1. Limitations of pre-segmentation and post-processing; 2. Aligning variable-length sequences	TIMIT	Supervised	×	LER (Label Error Rate)	Speech recognition
	[48]	1. Explore large-scale Mandarin ASR; 2. Potential limitations of the long sequences; 3. Computational efficiency	AISHELL-1; LiveShow; Voice Comment	Supervised	×	CER	Speech recognition
	[49]	1. Handling long and similar speech segments; 2. Handling noise and unstructured input	TIMIT	Supervised	×	PER (Phoneme Error Rate)	Speech recognition
BERT-integrated	[56]	1. Limited label data; 2. Speech-text length discrepancy	CALLHOME; HKUST	Self-supervised	✓	CER; WER	Speech recognition
Self-supervised learning	[57]	Align non-matching speech-text representations.	VoxPopuli; CoVoST 2	Self-supervised	✓	WER; BLEU	Speech recognition; Speech translation
	[58]	Limited label data	FLUERS	Unsupervised	✓	CER	Speech recognition

(Continued)

Table 2 (continued)

References		Main challenges concern	Dataset	Supervision	Multilingual	Evaluation criteria	Tasks
Multi-task learning	[52]	1. Character-sound mapping gap; 2. Underutilization of textual data	AISHELL-1; AISHELL-2; WenetSpeech	Self-supervised + Supervised	×	WER	Speech recognition
	[59]	1. Limited label data; 2. Representation discrepancy between speech and text	LibriSpeech; MuST-C	Supervised	✓	WER; BLEU	Speech recognition; Speech translation
	[60]	Limited label data	LibriSpeech; MuST-C	Self-supervised + Supervised	✓	WER; BLEU	Speech recognition; Speech translation
	[61]	Transfer interference and capacity dilution in multimodal pre-trained model	LibriSpeech; SpeechStew; CoVoST 2	Unsupervised	✓	WER; BLEU	Speech recognition; Speech translation

Another research category seeks to reduce reliance on massive amounts of paired speech and text data by simultaneously learning their representations. Chen et al. [57] presented a self-supervised method to cooperatively learn a unified representation from both modalities via pretext objectives, including sequence alignment, duration prediction, and aligned masked language modeling tasks. With this jointly learned speech-text expression, they developed a multi-lingual ASR model with only unlabeled speech and text in the target language [58]. Several studies introduced the multi-task learning framework to effectively employ limited paired speech-text data as supervision, and abundant unlabeled speech and text data as unsupervised supplements. The paper [59] utilized denoising coding and machine translation tasks as auxiliary training objectives to enhance the performance of ASR. Meta AI [60] integrates four self-supervised and supervised sub-tasks to facilitate cross-modal learning between speech and text. The study [61] used additional supervised speech-text multimodal tasks to align speech and text representations. The paper [52] leveraged five training strategies to capture modality-invariant information between Mandarin speech and text. They include self-supervised phoneme-to-text, speech-to-pseudocodes, masked speech prediction tasks, and supervised phoneme prediction and speech-to-text tasks.

In this paper, we investigate a dual-tower structure to achieve multimodal modeling of speech and text. With the help of intra-modal and inter-modal interactions during the encoding stage, we further design a two-stage training strategy to effectively derive semantics-aware acoustic representations from paired speech and text data. Considering the practical speech recognition tasks that involve only speech input, we need to deactivate the text encoder, which remains active during training in the multimodal framework, to facilitate regular inference. To address the input inconsistency between training and inference, we explore solutions for missing modalities to bridge the gap effectively.

2.3 Automatic Speech Recognition for Air Traffic Control communications

The ASR system in air traffic control communications could substantially decrease the workload of air traffic controllers and boost their efficiency [1,9]. While extensive research has been conducted on speech recognition for ATCC over the past few decades, as presented in Table 3, it is essential to acknowledge that ATCC presents unique challenges, including data scarcity, high environmental noise interference, multilingual, accent, unstable speech rates, and a lack of semantic contextual information [1,62]. These challenges call for more in-depth explorations and innovative solutions to achieve robust and reliable speech recognition in ATCC.

Table 3: Overview of the relevant literature on speech recognition for air traffic control communications

Main challenges concerned	References	Modality	Dataset	Supervision	Evaluation criteria
Limited label data	[15]	Speech	Real-world ATC dataset (unavailable)	Unsupervised	CER
	[14]	Speech	Real-world ATC dataset (unavailable)	Self-supervised	LER (Label Error Rate)

(Continued)

Table 3 (continued)

Main challenges concerned	References	Modality	Dataset	Supervision	Evaluation criteria
	[63]	Speech	NATS, ISAVIA (unavailable); LiveATC, ATCO2, LDC-ATCC, UWB-ATCC and ATCOSIM (public)	Self-supervised	WER
Multilingual (accent)	[17]	Speech	ATCSpeech (application)	Supervised	CER
	[64]	Speech	ATCSpeech (application)	Self-supervised	LER
	[16]	Speech	ATCSpeech (application)	Supervised	LER
	[18]	Speech	ATCSpeech (application)	Supervised	CER
High environmental and noise interference	[65]	Speech	ATC Corpus (unavailable)	Supervised	CER; SER (Sentence Error Rate)
Insufficient contextual information	[66]	Speech	AISHELL-2, (application) ATC Corpus (unavailable)	Supervised	CER; RTF (Real Time Factor); RT (Real Time)
	[67]	Speech	Atcosim, UWB ATCC, LDC ATCC, MALORCA, AIRBUS and LiveATC (public)	Semi-supervised	WER
	[68]	Speech	LiveATC and MALORCA (public)	Semi-supervised	WER; CallWER (Call-Sign WER); Accuracy
	[69]	Speech-text	Surveillance database of OpenSky (public)	Supervised	WER; CSA (Call-Sign Recognition Accuracy)
	[70] Speaker role identification Task	Speech-text	ATCSpeech (application)	Supervised	Accuracy; Precision; Recall; F1 score

Annotating ATCC speech requires the expertise of specialists and a deep understanding of air traffic control, incurring high labeling costs. The paper [15] proposed a practical approach to address the issue of small training samples, which employs unsupervised pre-training to learn speech representations from unannotated speech samples, followed by supervised transfer learning for sub-domain adaptation. The study [14] used the self-supervised model wav2vec 2.0 to learn the general acoustic representation. It combines multi-task learning to fine-tune the model to promote the performance of low-resource speech recognition for ATCC. By investigating self-supervised methods and rapidly fine-tuning ASR models with limited labeled data, Zuluaga-Gomez et al. [63] yielded comparable performance. Experiments on a multilingual dataset showed a slight yet noticeable boost in ASR performance for ATCC compared with monolingual approaches. Lin et al. concentrated on the end-to-end framework for multilingual and accented automatic speech recognition systems in ATCC and design several effective models [16–18,64]. However, the majority of these are tailored for English accents. The intelligibility of ATCC speech may decrease due to background noise interference. ASR encounters variable environmental noise from factors like speaker switches, transmission equipment, and electromagnetic interference in the ATCC domain. Zhou et al. [65] improved a hybrid CTC-attention end-to-end system in ATCC, addressing the impact of noise and enhancing the model's robustness.

Insufficient contextual information may lead to erroneous command interpretations and interactive misunderstandings, which is one of the critical factors affecting the performance of ASR for ATCC and the focus of this study. The paper [66] used a combination of deep residual convolution and gated attention units to improve the performance of Chinese radiotelephony speech recognition by capturing local correlations and long-distance dependencies. The papers [67–69] integrated contextual information from air surveillance data into respective strategies to enhance call-sign recognition in the ASR system. The research [1] illustrated multimodal inputs to facilitate the extraction of contextual information, thereby improving the performance of spoken instruction understanding. Closest to the idea presented in this paper is the work [70], which proposes a modal fusion module to integrate acoustic and textual knowledge for ATCC-related speaker role recognition tasks. However, applications of multimodal information in ATCC are still in the exploratory phase. Unlike previous work, we explore integrating text information with rich semantics and long-range dependencies as auxiliary inputs, constructing a dual-tower multimodal speech recognition framework for ATCC. Furthermore, a two-stage strategy is designed to facilitate the interaction modeling of multimodal semantic information, enhancing the model's acoustic long-distance context modeling capability and improving ATCC speech recognition accuracy.

3 Method

3.1 The Overall Model Architecture

This section overviews the proposed speech-text multimodal speech recognition method for Mandarin air traffic control communications. Fig. 3 illustrates the overall dual-tower architecture, which consists of three components: 1) Two input representation modules are used to preprocess raw speech and text data into model input representations. 2) A speech-text multimodal encoding module, comprising a text encoder and a cross-modal speech encoder, contributes to cross-modal interaction through collaborative learning with contextual information from both modalities. 3) An attention-based decoder generates target sequences by focusing on different parts of the encoder output. The following sections will discuss the details of each module and the training strategies for the overall model.

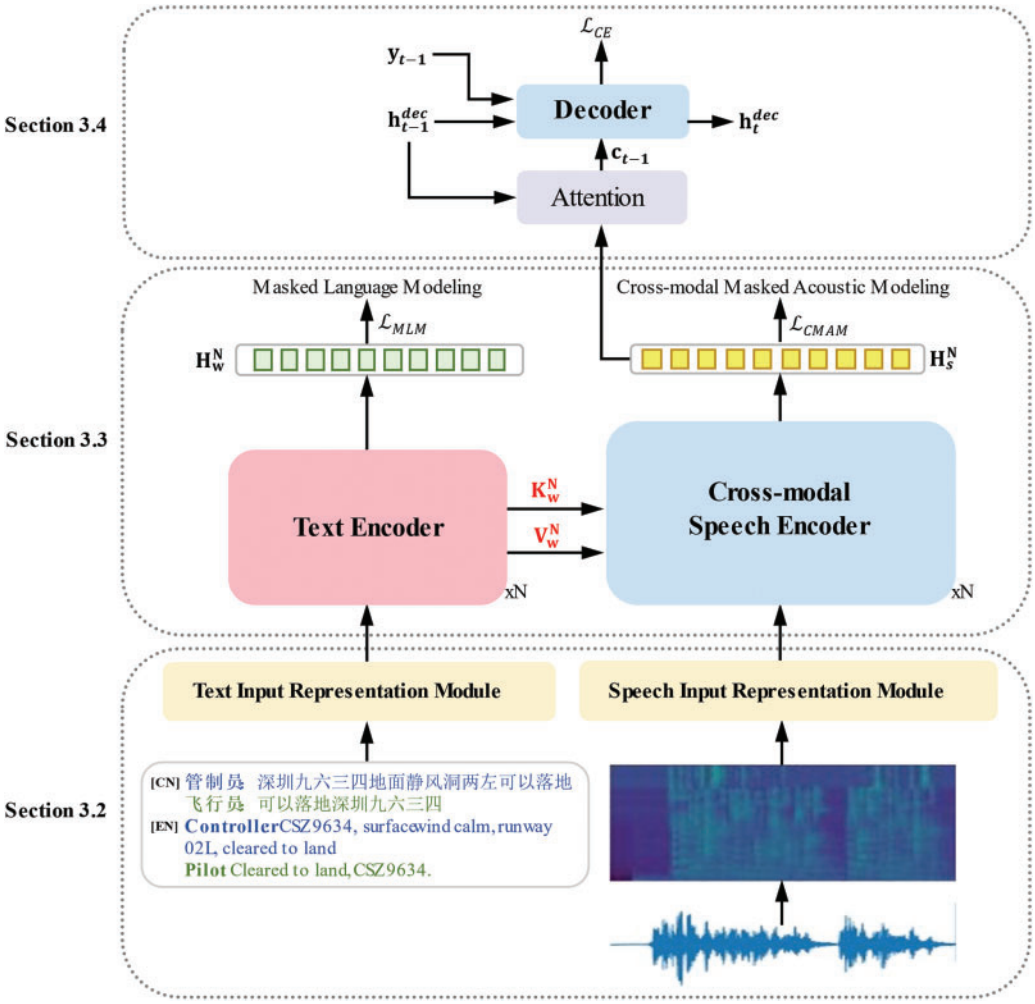

Figure 3: The overall architecture of the speech-text multimodal dual-tower framework for speech recognition in Mandarin air traffic control communications. "$\mathbf{K}_w^N$" and "$\mathbf{V}_w^N$" denote the final (deep) key matrix and value matrix of the text encoder at the last layer, respectively. "$\mathbf{H}_w^N$" is the final text representation and "$\mathbf{H}_s^N$" is the final speech representation. "$\mathbf{h}_{t-1}^{dec}$" and "$\mathbf{h}_t^{dec}$" represent the decoder's hidden state at step $t-1$ and step t. "$\mathbf{c}_{t-1}$" is the context vector for the decoder at step $t-1$. "$\mathcal{L}_{MLM}$" and "$\mathcal{L}_{CMAM}$" denote the loss of MLM task and CMAM task, respectively. "$\mathcal{L}_{CE}$" is the cross-entropy loss

3.2 Text/Speech Input Representation Module

For the text modality, the RoBERTa-wwm tokenizer[1] is adopted to tokenize and encode input texts with a vocabulary size of 21128 units. In addition, the special tokens <s> and </s> are introduced to indicate the start and end identifiers. Then the encoded token embedding is added with the corresponding positional embedding as the final textual input representation. We represent it as $\mathbf{E}_w = \{\mathbf{e}_w^1, \mathbf{e}_w^2, \ldots, \mathbf{e}_w^{T_w}\} \in \mathbb{R}^{T_w \times d_w}$, where T_w is the sequence length of textual tokens and d_w denotes the hidden size of the text representation.

[1]The tokenizer can be accessed from the website https://huggingface.co/hfl/chinese-roberta-wwm-ext.

For the speech modality, the input audio signal is initially divided into frames with a duration of 50 ms and a step size of 12.5 ms. Next, the corresponding Mel-spectrograms are computed to extract 80-dimensional filter bank (FBank) features from each frame with the Librosa toolkit. To comprehensively capture the speech signal's temporal and spectral characteristics, the speech features are concatenated with their first-order derivatives, which expands the feature dimension to 160. Finally, the processed acoustic features are projected through a dense layer and combined with the positional embedding to obtain the input representation $\mathbf{E}_s = \{\mathbf{e}_s^1, \mathbf{e}_s^2, \ldots, \mathbf{e}_s^{T_s}\} \in \mathbb{R}^{T_s \times d_s}$ for the speech encoder, where T_s is the total number of acoustic frames, and d_s denotes the hidden size of the audio representation.

3.3 Speech-Text Multimodal Encoding Module

The multimodal encoding module comprises a text encoder and a cross-modal speech encoder in a dual-tower structure. Text and speech inputs undergo comprehensive feature extraction in their respective encoders. The cross-modal attention module, resembling a bridge connecting the two encoders, is a crucial component for facilitating semantic interactive learning between them. Treating the text as auxiliary guidance, the last hidden state of the text encoder serves as one of the inputs to the cross-modal attention module. Cross-modal attention interaction occurs between the high-level textual and all-level acoustic representations, allowing the model to flexibly adjust its focus on different parts of the speech based on the contextual and semantic information from the text. Through interactive correlations between speech and text, it facilitates the modeling of acoustic long-distance context dependencies and achieves close semantic alignment between speech and text, thereby enabling a better understanding of heterogeneous yet complementary cross-modal information.

3.3.1 Text Encoder

The text encoder adopts the vanilla Transformer encoder structure stacked with N layers. Each layer mainly consists of a multi-head self-attention sublayer, a position-wise feed-forward sublayer, and two add-and-norm modules. For the i-th layer, linear projection first transforms the text feature sequence into query, key, and value matrices, mapping the feature sequence to distinct representation spaces with various weight matrices. Then, multi-head self-attention [71] parallelly performs single-head attention several times and connects each head's outputs. In this way, the model captures semantic associations and significance in the text input sequence by weighted aggregating contextual information at different locations, as the following formulas:

$$\mathbf{Q}_w^j = \mathbf{H}_w^{j-1}\mathbf{W}_w^Q, \tag{1}$$

$$\mathbf{K}_{\mathrm{w}}^j = \mathbf{H}_w^{j-1}\mathbf{W}_{\mathrm{w}}^K, \tag{2}$$

$$\mathbf{V}_w^j = \mathbf{H}_w^{j-1}\mathbf{W}_w^V, \tag{3}$$

$$\Delta\mathbf{H}_w^i = \mathit{MultiHeadAttention}\left(\mathbf{Q}_w^i, \mathbf{K}_w^i, \mathbf{V}_w^i\right), \tag{4}$$

where $\mathbf{Q}_w^i, \mathbf{K}_w^i, \mathbf{V}_w^i \in \mathbb{R}^{T_w \times d_w}$ are the query, key, and value matrices of the textual feature sequence. In the attention mechanism, the ***query*** is used to pose questions, the ***key*** represents potential correlations with the ***query***, and the ***value*** contains the information to be attended to. $\mathbf{W}_w^Q, \mathbf{W}_w^K, \mathbf{W}_w^V \in \mathbb{R}^{d_w \times d_w}$ are linear transformation matrices concatenated along the columns corresponding to all attention heads. Notably, the output of (i-1)-th layer serves as the input to the i-th layer, where the very first representation $\mathbf{H}_w^0$ is initialized with the text embedding $\mathbf{E}_w$.

In addition, the position-wise feed-forward sublayer introduces non-linear activation functions to enhance the fitting capability of the model, and the add-and-norm modules are employed to facilitate deeper information propagation and accelerate model convergence. Thus, it derives the output text representation $\mathbf{H}_w^i \in \mathbb{R}^{T_w \times d_w}$ through the following formulas:

$$\hat{\mathbf{H}}_w^{\mathrm{i}} = \mathrm{LayerNorm}\left(\Delta\mathbf{H}_w^i + \mathbf{H}_w^{i-1}\right), \tag{5}$$

$$\mathbf{H}_w^i = \mathit{LayerNorm}\left(\hat{\mathbf{H}}_w^i + \mathit{FFN}\left(\hat{\mathbf{H}}_w^i\right)\right). \tag{6}$$

3.3.2 Cross-Modal Speech Encoder

The cross-modal speech encoder distinguishes from the original Transformer in two attention modules: the self-attention module and the cross-modal attention module, as depicted in Fig. 4. The self-attention module in the cross-modal speech encoder is similar to that in the text encoder. However, unlike the vanilla Transformer decoder, it eliminates the look-ahead mask operation for future positions. Instead, it employs a bidirectional self-attention module to learn the intra-modal information of the speech modality thoroughly. Moreover, it is a crucial component for achieving cross-modal semantic alignment.

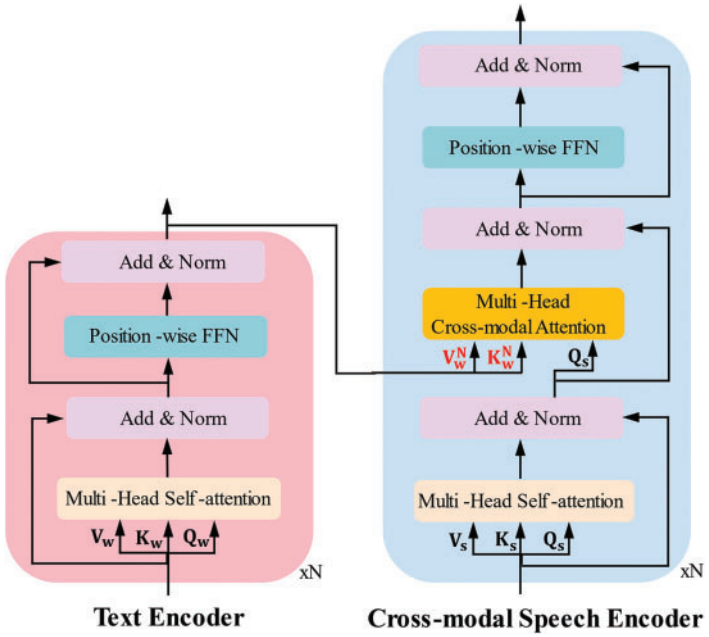

Figure 4: The detailed dual-tower structure of speech-text multimodal encoder. "V_w, K_w and Q_w" are the value matrix, key matrix, and query matrix of the text encoder, respectively. "V_s, K_s and Q_s" are the value matrix, key matrix, and query matrix of the speech encoder, respectively. "$\mathbf{V}_w^N$" and "$\mathbf{K}_w^N$" denote the value matrix, key matrix of the text encoder at the last layer. "$\mathcal{L}_{MLM}$" and "$\mathcal{L}_{CMAM}$" denote the loss of MLM task and CMAM task, respectively

Firstly, by applying linear transformations with different weight matrices to the hidden states of the previous layer of the speech modality, we obtain query, key, and value matrices. It facilitates capturing crucial information in the speech sequence more effectively during attention computations, which can be calculated as follows:

$$\mathbf{Q}_s^j = \mathbf{H}_s^{j-1}\mathbf{W}_s^Q, \tag{7}$$

$$\mathbf{K}_s^j = \mathbf{H}_s^{j-1}\mathbf{W}_s^K, \tag{8}$$

$$\mathbf{V}_s^j = \mathbf{H}_s^{j-1}\mathbf{W}_s^V, \tag{9}$$

where $\mathbf{Q}_s^j, \mathbf{K}_s^j, \mathbf{V}_s^j \in \mathbb{R}^{T_s \times d_s}$ are the query, key, and value matrices of the acoustic feature sequence, $\mathbf{W}_s^Q, \mathbf{W}_s^K, \mathbf{W}_s^V \in \mathbb{R}^{d_s \times d_s}$ are linear transformation matrices concatenated along the columns corresponding to all attention heads. Next, parallel attention calculations are performed on the speech's query, key, and value using the multi-head attention mechanism to learn representations of different focus aspects, enhancing the model's expressiveness and generalization capabilities, which can be expressed as the following formula:

$$\Delta\mathbf{H}_s^j = MultiHeadAttention\left(\mathbf{Q}_s^j, \mathbf{K}_s^j, \mathbf{V}_s^j\right), \tag{10}$$

where $\Delta\mathbf{H}_s^j \in \mathbb{R}^{T_s \times d_s}$ is the propagated information within speech modality. Finally, the representation obtained through the attention module is added to the previous layer's representation, and layer normalization is applied to alleviate the gradient vanishing problem during model training. The whole process can be summarized as follows:

$$\hat{\mathbf{H}}_s^j = LayerNorm\left(\Delta\mathbf{H}_s^j + \mathbf{H}_s^{j-1}\right). \tag{11}$$

Notably, the output of (j-1)-th layer serves as the input to the j-th layer, where the first representation $\mathbf{H}_s^0$ is initialized with the acoustic embedding $\mathbf{E}_s$.

The other is the cross-modal attention module, which aims to establish a strong link and semantic alignment between text and speech. The interaction between the two modalities is determined by computing the dot product between the speech queries and the text keys, resulting in cross-modal attention weights that capture their interrelationships. Then, a weighted sum of the textual value items is conducted in each feature sequence via the attention weights to obtain cross-modal interaction information. The process captures the interdependency and semantic alignment between the speech and text modalities, enhancing their mutual understanding through attention-weighted information fusion, which can be computed as follows:

$$\Delta\boldsymbol{H}_{w\to s}^j = MultiHeadAttention\left(\mathbf{Q}_s^j, \mathbf{K}_w^N, \mathbf{V}_w^N\right), \tag{12}$$

where $\Delta\boldsymbol{H}_{w\to s}^j \in \mathbb{R}^{T_s \times d_s}$ denotes the propagated cross-modal information from text to speech, and d_s is equal to d_w for unified modality. Notably, $\mathbf{K}_w^N$ and $\mathbf{V}_w^N$ are the final (deep) representations of the text encoder at the last layer.

Subsequently, a fully connected feed-forward layer is passed to enhance the representation ability further. Finally, it derives the cross-modal semantics-aware acoustic representation $\mathbf{H}_s^j \in \mathbb{R}^{T_s \times d_s}$ via the following formulas:

$$\tilde{\boldsymbol{H}}_s^j = LayerNorm\left(\Delta\boldsymbol{H}_{w\to s}^j + \hat{\mathbf{H}}_s^j\right), \tag{13}$$

$$\mathbf{H}_s^j = LayerNorm\left(\tilde{\boldsymbol{H}}_s^j + FFN(\tilde{\boldsymbol{H}}_s^j)\right). \tag{14}$$

3.4 Attention-Based Decoder

With the cross-modal speech encoder output as input, an attention-based decoder focuses dynamically on various portions of the encoder output to generate each target sequence element. The context vector $\mathbf{c}_{t-1}$ for the decoder at step $t-1$ is derived by performing a weighted summation of the encoder's context vectors with the attention weights as follows:

$$\mathbf{c}_{t-1} = \sum_{k=1}^{T_s} \alpha_{t-1,k} \mathbf{h}_k^{enc}, \tag{15}$$

where $\alpha_{t-1,k}$ is the attention weight between the corresponding decoder's hidden state $\mathbf{h}_{t-1}^{dec}$ and encoder's hidden state $\mathbf{h}_k^{enc}$ via the position-aware attention mechanism [72]. Specifically, it can be computed as follows:

$$\alpha_{t-1,k} = \frac{exp\left(s_{t-1,k}\right)}{\sum_{m=1}^{T_s} exp\left(s_{t-1,m}\right)}, \tag{16}$$

$$s_{t-1,k} = score\left(\mathbf{h}_{t-1}^{dec}, \mathbf{h}_k^{enc}\right), \tag{17}$$

where $s_{t-1,k}$ is the attention score measuring the similarity between the decoder's hidden state $\mathbf{h}_{t-1}^{dec}$ and encoder's hidden state $\mathbf{h}_k^{enc}$, and the hidden state $\mathbf{h}_k^{enc}$ is the k-th row vector of the final speech representation $\mathbf{H}_s^N$.

Furthermore, a long short-term memory (LSTM) network utilizes the current hidden state, decoder output, and context vector to compute a new hidden state along the time series autoregressively as follows:

$$\mathbf{h}_t^{dec} = LSTM\left(\mathbf{h}_{t-1}^{dec}, \boldsymbol{y}_{t-1}, \boldsymbol{c}_{t-1}\right). \tag{18}$$

Lastly, the final decoder's hidden states pass through a dense layer and a softmax function to obtain the predicted probability distribution vectors. The cross-entropy loss is calculated to measure the discrepancy between the predicted distribution and the ground truth labels as follows:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \sum_{u=1}^{V} \mathrm{y_u} \log(\hat{\mathrm{y}}_\mathrm{u}), \tag{19}$$

where V represents the vocabulary size, T denotes the sequence length of predicted text, and $\hat{y}_u$ and y_u represent the u-th element of the predicted distribution vector $\hat{\mathbf{y}}$ and the true label vector $\mathbf{y}$, respectively.

3.5 Two-Stage Training Strategy

The speech-text multimodal framework introduces a new branch of text modality during the encoding phase. Consequently, a significant gap emerges between the training and inference phases when the model is evaluated with only pure speech as input. To this end, we devised a two-stage training strategy, as shown in Fig. 5. It consists of the first-stage pre-training approach for the speech-text multimodal encoding module and the second-stage fine-tuning method for the entire end-to-end speech-text multimodal speech recognition network. In the first stage, the speech-text multimodal encoding module undergoes masked language modeling and cross-modal acoustic modeling tasks, which help the model comprehensively understand the intra-modal context and interactively integrate speech and text information. The well-initialized foundation from the first pre-training stage allows the model to adapt to speech recognition features rapidly. In the second stage, two approaches are designed to narrow the gap between pre-training and fine-tuning for speech-only input. Leveraging the established initialization, the model focuses on understanding speech features, optimizing for the specific speech recognition task, and improving accuracy.

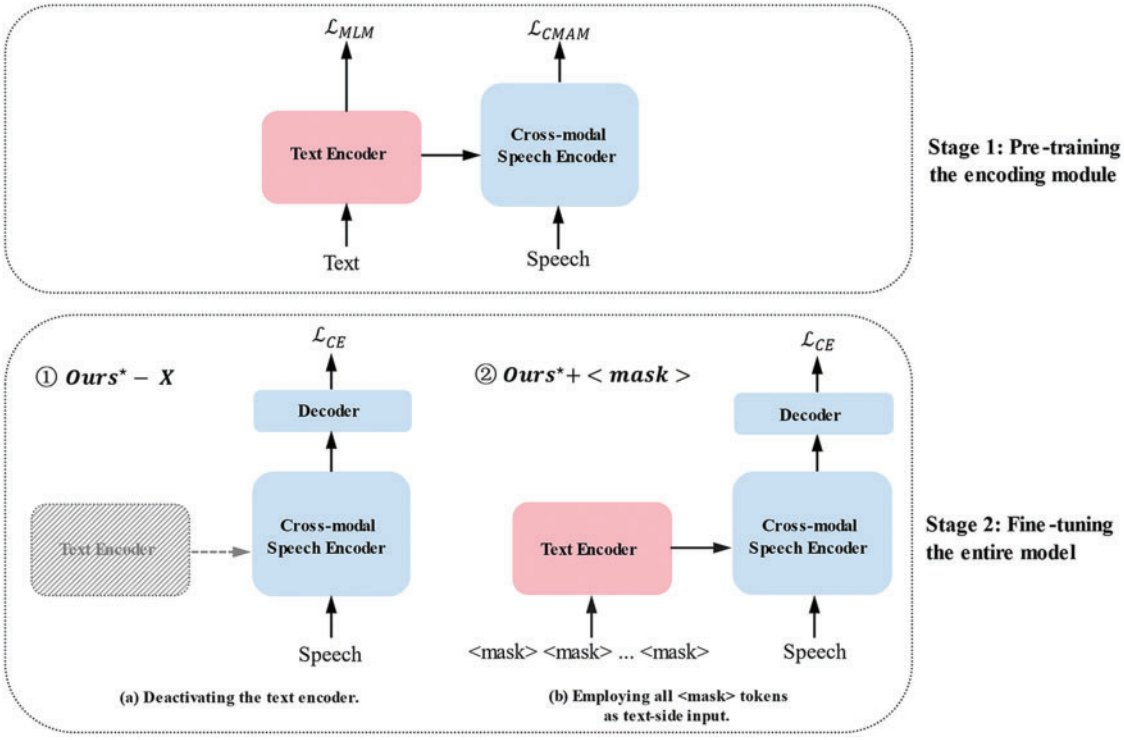

Figure 5: Two-stage Training Strategy. "$\mathcal{L}_{MLM}$" and "$\mathcal{L}_{CMAM}$" denote the loss of MLM task and CMAM task, respectively. The "<mask>" token is a special marker to represent a blank space that needs to be filled or predicted based on context. "*Ours** $-$ X" and "*Ours**$+$ $<$ *mask* $>$" are the two second-stage fine-tuning strategies corresponding to the descriptions (a) and (b) in the figure

3.5.1 First-Stage Pre-Training Approach

In the first stage, with coupled speech-text inputs, we pre-trained the speech-text multimodal encoding module to achieve close semantic alignment between speech and text modalities. The text encoder employs a cloze-like masked language modeling (MLM) strategy to assist the model in learning contextual relationships and semantic information within the textual modality. The objective is to mask certain tokens in the input text and then predict them depending on their contexts. Following the RoBERTa [73] configuration, we dynamically mask input tokens with a probability of 15%, where 80% of the masked positions are replaced with <mask>, 10% with random tokens, and the remaining 10% remain unchanged. Lastly, the cross-entropy loss serves as the objective function to optimize the model as follows:

$$\mathcal{L}_{MLM} = -\sum_{\hat{w}\in m(w)} log\left(P\left(\hat{w}|w_{\setminus m(w)}\right)\right), \tag{20}$$

where $m\,(\boldsymbol{w})$ and $\boldsymbol{w}_{\setminus \boldsymbol{m}(\boldsymbol{w})}$ denote the masked characters from the textual sequence $\boldsymbol{w}$ and the rest context characters, respectively.

The cross-modal speech encoder adopts a cross-modal masked acoustic modeling (CMAM) strategy to learn the speech representation by masking a portion of audio frames and guiding the model to predict the masked parts depending on their contexts. Thus, it captures inter-modal interactions with close semantic alignment by incorporating contextual information from both modalities. Specifically,

we first divide the audio into individual segments by the number of consecutive frames per segment and select some pieces with a probability of 15%, which masks them all to zero 80% of the time, replaces them with randomly selected audio frames 10% of the time, and leaves them unchanged for the remaining instances. In this way, the model seeks to reconstruct the masked acoustic features by optimizing the L1 loss, which reduces the distributional distinction between predicted representations and the ground truth concerning the masked positions as follows:

$$\mathcal{L}_{CMAM} = -\frac{1}{|T_{masked}|}\sum_{i\in T_{masked}} |s_i - \hat{s}_i|, \tag{21}$$

where T_{masked} represents the number of masked positions, s_i represents the original masked audio features, and $\hat{s}_i$ represents the predicted audio features.

3.5.2 Second-Stage Fine-Tuning Method

In the second stage, with unpaired speech-only input, we fine-tuned the entire end-to-end speech-text multimodal speech recognition network using cross-entropy loss. In particular, the pre-trained speech-text multimodal encoding module was employed to initialize the speech-text multimodal encoder thanks to its effectively aligned multimodal representations acquired in the first stage, simultaneously taking both speech and text inputs.

The objective of the second stage is to fine-tune the entire model to adapt to speech-only inputs. To mitigate the discrepancy between the inputs of the encoding modules in the first and second stages due to inconsistent modal inputs, inspired by the research [62], two methods were attempted: 1) Deactivating the text encoder and textual input. 2) Employing all <mask> tokens as text-side input. For the former, the text encoder and cross-modal attention are eliminated or deactivated since the dual-tower architecture independently captures two modalities and uses cross-modal attention for interaction. For the latter, the original multimodal architecture remains in effect while utilizing a sequence of all <mask> tokens as text-side input. However, it is challenging for the model to convey sufficient meaningful information from the fixed text sequence input. Instead, it prioritizes fine-tuning the speech encoder and decoder to mitigate the domain shift. The two-stage training strategy offers a feasible means of establishing semantic connections between speech and text modalities, enabling the multimodal model to handle speech-only input better during inference and produce semantics-aware speech recognition results.

4 Experiment and Result Analysis

4.1 Dataset

In this study, we adopt a paraphrase dataset of Mandarin air traffic control communications (ATCC), recorded in a quiet environment under the supervision of seasoned air traffic control experts and professional course materials. The dataset comprises pairs of the air traffic controller's command and the pilot's readback, resulting in longer sentences than the average dataset. The speech has a subtle accent, and the speech rate is relatively stable. Each voice sample is stored in the WAV format with a sampling rate of 16 kHz, 16 bits, and mono encoding. The dataset consists of 10971 voice samples with a total duration of approximately 25.2 h and is relatively small-scale. We randomly shuffled the entire dataset and divided it into the train, validation, and test sets in an 8:1:1 ratio.

The study [1] provided summaries of datasets in the domains of common use and ATCC. To further verify the effectiveness of the proposed method, we also conducted experiments on the

AISHELL-1 [74] dataset[2], an open-source, high-quality Mandarin speech recognition corpus recorded in a quiet indoor environment. It is extensively used in the speech community and encompasses eleven domains, including smart home, autonomous driving, industrial production, etc. The total duration of the dataset is 178 h, indicating a dataset of moderate scale. It sampled at 16 kHz, mono, and 16 bits, and stored in WAV format. In addition, we divided the dataset into training, validation, and testing sets in an identical ratio of 8:1:1. Table 4 presents the detailed division of train, validation, and test sets for the two datasets.

Table 4: Division and size of the two datasets. The symbols #*Utterances* and #*Hours* denote the sample size and total duration of speech utterances, respectively

Dataset	Train		Dev		Test	
	#Utterances	#Hours	#Utterances	#Hours	#Utterances	#Hours
ATCC	8777	20.13	1097	2.52	1097	2.52
AISHELL-1	113280	142.4	14160	17.8	14160	17.8

4.2 Implementation Detail

The text encoder and cross-modal speech encoder each consist of six stacked layers of Transformer with a representation dimension of 768, 12 attention heads, and a feed-forward layer with a size of 3027. On the other hand, the decoder comprises a single layer of LSTM with a dimension of 768. Using an initial learning rate of 5e-5, we used the Adam [75] optimizer to optimize the trainable parameters during model training. In addition, we utilized a linear-decayed learning rate schedule with a warm-up period [76]. During the fine-tuning phase, we leveraged the AdamW [77] optimizer with an initial learning rate of 1e-5. Additionally, we employed a cosine annealing learning rate schedule [78] to achieve optimal performance. We used four NVIDIA A100-SXM4-40G GPUs for the experiment, with a batch size of eight.

We adopted the Character Error Rate (CER) as the evaluation metric since speech recognition aims to transcribe speech into text with character as the primary modeling unit. The CER directly corresponds to individual character errors within the text to ensure more accurate recognition precision, which can be calculated using the following formula:

$$CER = \frac{I + S + D}{N}, \tag{22}$$

where I, S, D represent the number of inserted, substituted, and deleted characters, respectively, and N represents the total number of characters in the ground truth labels. The CER can range from a minimum of 0 to a maximum greater than 1, with lower values indicating higher recognition accuracy. The Levenshtein distance algorithm[3] is the most commonly used implementation method for calculating the CER.

The proposed two-stage training strategy consists of the first-stage pre-training of the speech-text multimodal encoding module for 80 epochs and the second-stage fine-tuning of the entire E2E network for 30 epochs. In contrast, we introduced an additional experimental setting for comparison to demonstrate the effectiveness of the proposed strategy, as illustrated in Fig. 6b. In the first stage,

[2]The AISHELL-1 dataset can be downloaded from the website https://www.openslr.org/33/.
[3]The details of Levenshtein distance algorithm can be accessed from http://en.wikipedia.org/wiki/Levenshtein_distance.

the encoding and decoding modules are trained E2E with paired speech-text multimodal inputs, integrating the MLM loss, CMAM loss, and cross-entropy loss to optimize the entire network simultaneously. In the second stage, the whole E2E speech-text multimodal speech recognition network is also fine-tuned using solely speech data.

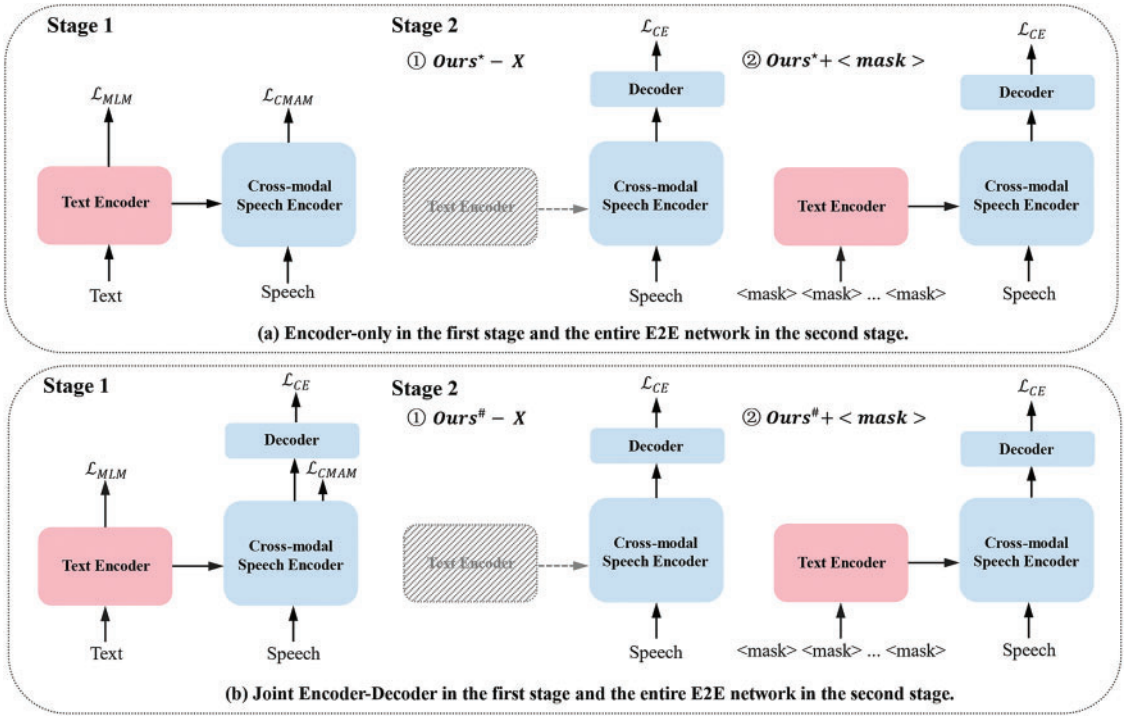

Figure 6: Two experimental setup variants for the two-stage training strategy. "$\mathcal{L}_{MLM}$" and "$\mathcal{L}_{CMAM}$" denote the loss of MLM task and CMAM task, respectively. "$\mathcal{L}_{CE}$" is the cross-entropy loss. The "<mask>" token is a special marker to represent a blank space that needs to be filled or predicted based on context. "$Ours^{\star} - X$" and "$Ours^{\star} + < mask >$" are the two second-stage fine-tuning strategies corresponding to the descriptions (a) in the figure. "$Ours^{\#} - X$" and "$Ours^{\#} + < mask >$" are the two second-stage fine-tuning strategies corresponding to the descriptions (b) in the figure

4.3 Comparative Experiments

Tables 5 and 6 display the recognition results under the same settings on the ATCC and AISHELL-1 datasets, respectively. The experimental results demonstrate that the proposed multimodal models outperform the unimodal baseline model by a substantial margin for both datasets, delivering performance improvements of 28.76% and 23.82% compared with the optimal baseline model. Remarkably, our multimodal method denoted as $Ours^{\star}$, achieves the best results on both datasets, demonstrating the effectiveness of the proposed two-stage training strategy. It indicates that the first-stage pre-training process concentrating solely on the multimodal encoder contributes to close semantic alignment between speech and text modalities, serving as a beneficial initialization for the entire training procedure. Besides, two second-stage fine-tuning strategies are correspondingly suitable for the two datasets.

Table 5: The performance comparison of different methods on the ATCC dataset. The symbols $\star$ and # denote the two experimental setup variants for the proposed two-stage training strategy in Figs. 6a and 6b, respectively. The symbols "$+ < \boldsymbol{mask} >$" and "$-\boldsymbol{X}$" represent the two second-stage fine-tuning strategies in Figs. 5b and 5a, respectively. Note that a lower Character Error Rate (CER) indicates better speech recognition performance. The bold values indicate the best recognition performance in the corresponding column regarding CER (%)

Model	Feature	CER (%)	
		Dev	Test
Speech-transformer [48]	FBank	23.78	24.82
CTC-based [47]	FBank	9.12	9.18
$Ours^{\star} + < mask >$	FBank	7.32	7.76
$Ours^{\star} - X$	FBank	**6.42**	**6.54**
$Ours^{\#} + < mask >$	FBank	16.26	16.31
$Ours^{\#} - X$	FBank	13.92	13.98

Table 6: The performance comparison of different methods on the AISHELL-1 dataset. The symbols $\star$ and # denote the two experimental setup variants for the proposed two-stage training strategy in Figs. 6a and 6b, respectively. The symbols "$+ < \boldsymbol{mask} >$" and "$-\boldsymbol{X}$" represent the two second-stage fine-tuning strategies in Figs. 5b and 5a, respectively. Note that a lower Character Error Rate (CER) indicates better speech recognition performance. The bold values indicate the best recognition performance in the corresponding column regarding CER (%)

Model	Feature	CER (%)	
		Dev	Test
Speech-transformer [48]	FBank	11.28	11.46
CTC-based [47]	FBank	15.82	15.95
$Ours^{\star} + < mask >$	FBank	**8.53**	**8.73**
$Ours^{\star} - X$	FBank	11.02	11.14
$Ours^{\#} + < mask >$	FBank	31.66	31.78
$Ours^{\#} - X$	FBank	15.10	15.14

For the ATCC dataset, the unimodal baseline model performs worse due to its limited modeling capability on small amounts of data and long sequences of the paraphrase dataset. In contrast, the proposed multimodal method efficiently captures and aligns cross-modal information with more substantial context modeling capability, thereby increasing accuracy and enhancing generalization capabilities. Hence, as presented in Table 5, the best results occur when pre-training only the multimodal encoder in the first stage and turning off the text encoder directly in the second stage. Inter-modal interaction in the first stage enhances acoustic long-distance context dependencies, and deactivating the text encoder in the second stage benefits the generalization performance of semantics-aware acoustic modeling.

For the AISHELL-1 dataset, the model performs optimally when employing all <mask> tokens as text-side input in the second stage, demonstrating the feasibility of this approach for fine-tuning and inference, as shown in Table 6. Following the ATCC dataset, the performance of jointly training the E2E encoder-decoder is inferior to training only the multimodal encoder module in the first stage. Ideally, after pre-training the multimodal encoder, the model focuses more on the transfer learning of semantics-aware acoustic feature distribution rather than feature extraction during the fine-tuning phase. Otherwise, the model devotes more attention to adjusting parameters for speech recognition tasks when training the encoder-decoder network following three objectives simultaneously in the first stage, leading to potential interference between subtasks.

4.4 Ablation Study

In this section, ablation experiments were conducted on the ATCC and AISHELL-1 datasets to validate the effectiveness of critical components in our model, as reported in Table 7. By comparing the results of setting (a) with our approach, we observed a performance degradation when modeling without the text modality, which indicates that incorporating semantically rich textual information as an auxiliary dramatically contributes to recognition performance. In configuration (b), the CER goes up when excluding the MLM task from the text encoder. It demonstrates the significance of learning high-level semantic representations for the speech recognition task in the text modality. For setting (c), we retained only the information flow between speech and text, removing the CMAM task from the cross-modal speech encoder. The dramatic drop in performance highlights the necessity of modeling both the intra-modal acoustic information and the inter-modal interaction between speech and text for comprehensive understanding and modeling of speech signals.

Table 7: The results of the ablation study conducted on critical components of the proposed speech-text multimodal approach. The indicator MAM denotes that the speech encoding module is pre-trained via a masked acoustic modeling objective. Note that a lower Character Error Rate (CER) indicates better speech recognition performance. The bold values indicate the best recognition performance in the corresponding column regarding CER (%)

Settings	Text Input	MLM	CMAM	ATCC		AISHELL-1	
				CER (%)		CER (%)	
				Dev	Test	Dev	Test
w/o text input			(MAM)	14.85	14.94	16.95	16.96
w/o MLM	✓		✓	13.19	13.33	14.38	14.41
w/o CMAM	✓	✓		16.39	16.41	17.08	17.11
Ours⋆	✓	✓	✓	**6.42**	**6.54**	**8.53**	**8.73**

4.5 Effect of Model Capacity on Recognition Performance

Hyperparameter selection experiments were conducted on the ATCC dataset to investigate the effect of the model capacity on recognition performance by varying the number of encoder layers for two modalities, as presented in Table 8. The recognition performance boosts as the number of encoder layers increases. The larger capacity enables the model to convey more semantic information, producing better results for semantics-related speech recognition tasks. Ultimately, the performance

peaks at six encoder layers; however, it then degrades since the model may struggle to learn and adapt due to insufficient data.

Table 8: Effect of model capacity on recognition performance on the ATCC dataset. Note that a lower Character Error Rate (CER) indicates better speech recognition performance. The bold values indicate the best recognition performance in the corresponding column regarding CER (%)

Settings	CER (%)	
	Dev	Test
N = 2	14.15	14.22
N = 4	8.38	8.45
N = 6	**6.42**	**6.54**
N = 12	18.22	18.34

4.6 Case Study

Examples of the proposed speech-text multimodal speech recognition model on the ATCC and AISHELL-1 datasets are presented in Tables 9 and 10, respectively. The tables provide the Mandarin pinyin pronunciation of keywords within the sentences and the corresponding English translation of the entire sentences, with the light gray background marking the incorrectly recognized characters. The proposed multimodal method outperforms the unimodal approach on both datasets regarding recognition accuracy. As Table 9 shows, the unimodal network suffers from significant insertion errors due to insufficient acoustic modeling capability for long sequences. Regarding semantic reliability, the results derived by the unimodal method are syntactically reasonable but contain several semantic errors, especially evident in the AISHELL-1 dataset.

Table 9: Examples of different speech recognition methods on the ATCC dataset, with the light gray background marking the incorrectly recognized characters. /../ denotes pinyin pronunciation for the Mandarin characters, where the number represents the character's tone. The symbol *** is employed to align character units for ease of observation regarding the insertion errors

	Reference	/guai3/ /liu4/ /san1/ 南方六拐 拐 六 地面风洞八洞洞 三 *** *** 洞两左可以起飞 /guai3/ /liu4/ 可以起飞洞两左南方六拐 拐 六 Controller: CSN6776, surface wind 080 degrees, 3 m/s, runway 02L, cleared for takeoff. Pilot: Cleared for takeoff, runway 02L, CSN6776.
1	Speech-transformer [48]	/ba1/ /wu3/ /liang3/ 南方六拐 八 五 地面风洞八洞洞 两 跑 道 洞两左可以起飞 eight five two runway /ba1/ /wu3/

(Continued)

Table 9 (continued)

		可以起飞洞两左南方六拐 八 五 eight five
	Ours⋆	/guai3/ /liu4/ /san1/ 南方六拐 拐 六 地面风洞八洞洞 三 *** *** 洞两左可以起飞 seven six three /guai3/ /liu4/ 可以起飞洞两左南方六拐 拐 六 seven six
	Reference	/jiu3/ 你好东方五两幺六航道阿尔法 九 头朝东跑道外等 /jiu3/ 阿尔法 九 头朝东跑道外等东方五两幺六 Controller: Hello, CES5216, runway A9, facing east, hold short of the runway. Pilot: Runway A9, facing east, holds short of runway, CES5216.
2	Speech-transformer [48]	/jiu4/ 你好东方五两幺六航道阿尔法 就 头朝东跑道外等 at once /jiu4/ 阿尔法 就 头朝东跑道外等东方五两幺六 at once
	Ours⋆	/jiu3/ 你好东方五两幺六航道阿尔法 九 头朝东跑道外等 nine /jiu3/ 阿尔法 九 头朝东跑道外等东方五两幺六 nine

In some cases, compared with ground truth, specific character recognition results have the same pronunciation but incorrect semantics. For example, in Case 1 of Table 10, the unimodal method misidentified "视作" (*shi4 zuo4*, *regard as*) as "试做" (*shi4 zuo4*, *try doing*) while the multimodal approach correctly recognized it, demonstrating that the multimodal method has superior representation capability with a closer semantic alignment between speech and text. Introducing textual information for multimodal interaction enhances contextual semantic information and certainty, and reduces sensitivity to phonemic overlap and confusion in acoustic modeling, thereby mitigating errors related to homophony. Nonetheless, as exhibited in Case 3 of Table 10, it is not ideal for the multimodal network to cope with rare proper nouns, which may benefit from more effective semantics-aware acoustic representations.

Table 10: Examples of different speech recognition methods on the AISHELL-1 dataset, with the light gray background marking the incorrectly recognized characters. /../ denotes pinyin pronunciation for the Mandarin characters, where the number represents the character's tone

		AISHELL-1
1	Reference	/shi4/ /zuo4/ /sha1/ /pai2/ 被 视 作 中国 沙 排 的明日之星 Regarded as the rising star of Chinese beach volleyball.
	Speech-transformer [48]	/shi4/ /zuo4/ /sha1/ /pai2/ 被 试 做 中国 杀 牌 的 明 日 之 星 try doing killing cards
	Ours⋆	/shi4/ /zuo4/ /sha1/ /pai2/ 被 视 作 中国 沙 排 的明日之星 regarded as beach volleyball
2	Reference	/yin1/ /qi2/ /li2/ /gang3/ 朱 茵 与好姐妹蔡少芬齐 齐 离 港 Yin Zhu (Athena Chu) and her close friend Shaofen Cai (Ada Choi) left Hong Kong together.
	Speech-transformer [48]	/yin1/ /qi1/ /li3/ /gang3/ 朱 因 与好姐妹蔡少芬齐 妻 李 岗 cause wife Lee post
	Ours⋆	/yin1/ /qi2/ /li2/ /gang3/ 朱 茵 与好姐妹蔡少芬齐 齐 离 港 Yin together left Hong Kong
3	Reference	/bao1/ /yi1/ /ji4/ /ju4/ /yi3/ /er4/ /chun2/ 其他上榜的还有 包 衣 剂 胶囊和 聚 乙 二 醇 Other items on the list include enteric-coated capsules and polyethylene glycol.
	Speech-transformer [48]	/bao4/ /yi4/ /ji4/ /ju4/ /yu4/ /er4/ /chun2/ 其他上榜的还有 暴 逸 记 胶囊和 巨 预 二 纯 violent escape memory giant preview pure
	Ours⋆	/bao4/ /yi1/ /ji4/ /ju4/ /yi3/ /er4/ /chun2/ 其他上榜的还有 暴 一 剂 胶囊和 巨 乙 二 纯 violent one giant pure

5 Conclusion

In this paper, we propose a speech-text multimodal speech recognition method via dual-tower architecture for Mandarin Air Traffic Control Communications (ATCC), which employs cross-modal interactions to achieve close semantic alignment across modalities during the encoding phase. Besides, we devise a two-stage training strategy to derive semantics-aware acoustic representations effectively. The first stage focuses on pre-training the speech-text multimodal encoding module with masked language modeling and cross-modal masked acoustic modeling strategies to enhance inter-modal semantic alignment and acoustic long-distance context dependencies. The second stage fine-tunes the entire network to bridge the input modality variation gap between the training and inference phases by deactivating the text encoder or employing all <mask> tokens as text-side input.

Extensive experiments demonstrate the effectiveness of the proposed speech-text multimodal speech recognition method on the ATCC and public AISHELL-1 datasets. It reduces the character error rate to 6.54% and 8.73%, respectively, and exhibits substantial performance gains of 28.76% and 23.82% over the best baseline model. Remarkably, the first-stage pre-training process concentrating solely on the multimodal encoder contributes to close semantic alignment between speech and text modalities, serving as a beneficial initialization for the second-stage training procedure. Even though the ATCC dataset has a small amount of data and long sequences, the first-stage inter-modal interactions make it easier to model acoustic long-distance context dependencies. Besides, two second-stage fine-tuning strategies are correspondingly suitable for the two datasets. Deactivating the text encoder works best for the ATCC dataset, while the AISHELL-1 dataset benefits most from employing all <mask> tokens as text-side input in the second stage.

In the future, we plan to investigate more speech-text multimodal frameworks and optimize training strategies further. In addition, we will also explore various efficient ways to lessen the dependence on multimodal paired data and tackle the issue of low-resource speech recognition.

Acknowledgement: The authors would like to thank all contributors to the AISHELL-1 corpus by Beijing Shell Shell Technology Co., Ltd.

Funding Statement: This research was funded by Shenzhen Science and Technology Program (Grant No. RCBS20221008093121051), the General Higher Education Project of Guangdong Provincial Education Department (Grant No. 2020ZDZX3085), China Postdoctoral Science Foundation (Grant No. 2021M703371) and the Post-Doctoral Foundation Project of Shenzhen Polytechnic (Grant No. 6021330002K).

Author Contributions: Study conception and design, S. Ge and J. Ren; methodology, S. Ge and J. Ren; software, S. Ge; validation, S. Ge and J. Ren; formal analysis, S. Ge and J. Ren; investigation, S. Ge and J. Ren; resources, J. Ren, S. Yang and J. Yang; data curation, Y. Shi and J. Yang; writing—original draft preparation, S. Ge and J. Ren; writing—review and editing, J. Ren, Y. Shi, Y. Zhang, S. Yang and J. Yang; visualization, S. Ge and J. Ren; supervision, Y. Zhang and J. Yang; project administration, Y. Shi and J. Yang; funding acquisition, J. Ren, Y. Shi and J. Yang. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Lin, "Spoken instruction understanding in air traffic control: Challenge, technique, and application," *Aerospace*, vol. 8, no. 3, pp. 65, 2021.

[2] O. Alharasees, A. Jazzar, U. Kale and D. Rohacs, "Aviation communication: The effect of critical factors on the rate of misunderstandings," *Aircr. Eng. Aerosp. Tec.*, vol. 95, no. 3, pp. 379–388, 2023.

[3] H. Yang, Y. Chang, and Y. Chou, "Subjective measures of communication errors between pilots and air traffic controllers," *J. Air. Transp. Manag.*, vol. 112, pp. 102461, 2023.

[4] M. Bongo and R. Seva, "Effect of fatigue in air traffic controllers' workload, situation awareness, and control strategy," *Int. J. Aerosp. Psychol.*, vol. 32, no. 1, pp. 1–23, 2022.

[5] H. An, N. King, and S. O. Hwang, "Issues and solutions in air-traffic infrastructure and flow management for sustainable aviation growth: A literature review," *World Rev. Intermodal Transp. Res.*, vol. 8, no. 4, pp. 293–319, 2019.

[6] J. Zuluaga-Gomez *et al.*, "BERTraffic: BERT-based joint speaker role and speaker change detection for air traffic control communications," in *Proc. 2022 IEEE Spok. Lang. Technol. Workshop (SLT)*, Doha, Qatar, 2023, pp. 633–640.

[7] O. Ohneiser *et al.*, "Prediction and extraction of tower controller commands for speech recognition applications," *J. Air. Transp. Manag.*, vol. 95, pp. 102089, 2021.

[8] H. Helmke *et al.*, "Automatic speech recognition and understanding for radar label maintenance support increases safety and reduces air traffic controllers' workload," in *Proc. Fifteenth USA/Europe Air Traffic Manag. Res. Dev. Semin. (ATM2023)*, Savannah, Georgia, USA, 2023.

[9] N. Ahrenhold *et al.*, "Validating automatic speech recognition and understanding for pre-filling radar labels—increasing safety while reducing air traffic controllers' workload," *Aerospace*, vol. 10, no. 6, pp. 538, 2023.

[10] Y. Yang, Y. Li, and B. Du, "Improving CTC-based ASR models with gated interlayer collaboration," in *Proc. ICASSP, 2023–2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.

[11] G. Cheng, H. Miao, R. Yang, K. Deng, and Y. Yan, "ETEH: Unified attention-based end-to-end ASR and KWS architecture," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1360–1373, 2022.

[12] J. Li, "Recent advances in end-to-end automatic speech recognition," *Apsipa Trans. Signal Inf. Process.*, vol. 11, no. 1, pp. e8, 2022.

[13] T. Moriya *et al.*, "Streaming end-to-end speech recognition for hybrid RNN-T/attention architecture," in *Proc. Interspeech 2021*, Brno, Czechia, 2021, pp. 1787–1791.

[14] D. Guo, Z. Zhang, B. Yang, J. Zhang, and Y. Lin, "Boosting low-resource speech recognition in air traffic communication via pretrained feature aggregation and multi-task learning," *IEEE Trans. Circuits Syst. II: Express Briefs*, vol. 70, no. 9, pp. 3714–3718, 2023.

[15] Y. Lin, Q. Li, B. Yang, Z. Yan, H. Tan, and Z. Chen, "Improving speech recognition models with small samples for air traffic control systems," *Neurocomputing*, vol. 445, pp. 287–297, 2021.

[16] Y. Lin, D. Guo, J. Zhang, Z. Chen, and B. Yang, "A unified framework for multilingual speech recognition in air traffic control systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3608–3620, 2020.

[17] Y. Lin, B. Yang, D. Guo, and P. Fan, "Towards multilingual end-to-end speech recognition for air traffic control," *IET Intell. Transp. Syst*, vol. 15, no. 9, pp. 1203–1214, 2021.

[18] P. Fan, D. Guo, J. Zhang, B. Yang, and Y. Lin, "Enhancing multilingual speech recognition in air traffic control by sentence-level language identification," arXiv preprint arXiv:2305.00170, 2023.

[19] C. Chen, N. Hou, Y. Hu, S. Shirol, and E. S. Chng, "Noise-robust speech recognition with 10 minutes unparalleled in-domain data," in *Proc. ICASSP 2022–2022 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, 2022, pp. 4298–4302.

[20] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Gradient remedy for multi-task learning in end-to-end noise-robust speech recognition," in *Proc. ICASSP 2023–2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.
[21] Y. Qiu, R. Wang, F. Hou, S. Singh, Z. Ma, and X. Jia, "Adversarial multi-task learning with inverse mapping for speech enhancement," *Appl. Soft Comput.*, vol. 120, pp. 108568, 2022.
[22] Q. Zhu *et al.*, "Robust data2vec: Noise-robust speech representation learning for ASR by combining regression and improved contrastive learning," in *Proc. ICASSP 2023–2023 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, 2023, pp. 1–5.
[23] Q. Zhu, J. Zhang, Z. Zhang, and L. Dai, "A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 1927–1939, 2023.
[24] L. Cheng *et al.*, "Improving speaker diarization using semantic information: Joint pairwise constraints propagation," arXiv preprint arXiv:2309.10456, 2023.
[25] J. Sevilla-Salcedo, E. Fernádez-Rodicio, L. Martín-Galván, A. Castro-González, J. Castillo, and M. A. Salichs, "Using large language models to shape social robots' speech," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 3, pp. 6–20, 2023.
[26] D. Guo, Z. Zhang, P. Fan, J. Zhang, and B. Yang, "A context-aware language model to improve the speech recognition in air traffic control," *Aerospace*, vol. 8, no. 11, pp. 348, 2021.
[27] R. Cabrera, X. Liu, M. Ghodsi, Z. Matteson, E. Weinstein, and A. Kannan, "Language model fusion for streaming end to end speech recognition," arXiv preprint arXiv:2104.04487, 2021.
[28] C. Wingfield *et al.*, "On the similarities of representations in artificial and brain neural networks for speech recognition," *Front. Comput. Neurosc.*, vol. 16, pp. 1057439, 2022.
[29] C. Caucheteux, A. Gramfort, and J. King, "Disentangling syntax and semantics in the brain with deep networks," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 1336–1348.
[30] W. Cao *et al.*, "A review on multimodal zero-shot learning," *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, vol. 13, no. 2, pp. e1488, 2023.
[31] A. Mohamed *et al.*, "Self-supervised speech representation learning: A review," *IEEE J. Sel. Top. Signal. Process.*, vol. 16, no. 6, pp. 1179–1210, 2022.
[32] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2022.
[33] T. N. Sainath *et al.*, "JOIST: A joint speech and text streaming model for ASR," in *Proc. 2022 IEEE Spok. Lang. Technol. Workshop (SLT)*, Doha, Qatar, 2023, pp. 52–59.
[34] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from BERT," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 1897–1911, 2021.
[35] Y. Fu *et al.*, "Context-and knowledge-aware graph convolutional network for multimodal emotion recognition," *IEEE Multimed.*, vol. 29, no. 3, pp. 91–100, 2022.
[36] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang Process.*, vol. 29, pp. 985–1000, 2021.
[37] Y. Chung, C. Zhu, and M. Zeng, "SPLAT: Speech-language joint pre-training for spoken language understanding," in *Proc. 2021 Conf. N. Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol. (NAACL)*, Online, 2021, pp. 1897–1907.
[38] Q. Chen, W. Wang, and Q. Zhang, "Pre-training for spoken language understanding with joint textual and phonetic representation learning," in *Proc. Interspeech 2021*, Brno, Czechia, 2021, pp. 1244–1248.
[39] Q. Fang, R. Ye, L. Li, Y. Feng, and M. Wang, "STEMM: Self-learning with speech-text manifold mixup for speech translation," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Dublin, Ireland, 2022, pp. 7050–7062.
[40] Y. Chuang, C. Liu, H. Lee, and L. Lee, "SpeechBERT: An audio-and-text jointly learned language model for end-to-end spoken question answering," in *Proc. Interspeech 2020*, Shanghai, China, 2020, pp. 4168–4172.

[41] G. Zheng, Y. Xiao, K. Gong, P. Zhou, X. Liang, and L. Lin, "Wav-BERT: Cooperative acoustic and linguistic representation learning for low-resource speech recognition," in *Findings of the Assoc. Comput. Linguist.: EMNLP 2021*, Punta Cana, Dominican Republic, 2021, pp. 2765–2777.
[42] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 4275–4279.
[43] J. Ao *et al.*, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Dublin, Ireland, 2022, pp. 5723–5738.
[44] C. Meng, J. Ao, T. Ko, M. Wang, and H. Li, "CoBERT: Self-supervised speech representation learning through code representation learning," arXiv preprint arXiv:2210.04062, 2022.
[45] Z. Zhang *et al.*, "SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training," in *Proc. 2022 Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 1663–1676.
[46] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Proc. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.
[47] A. Graves, S. Fern, A. Ndez, F. Gomez, and J. U. R. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2006, pp. 369–376.
[48] Y. Zhao, J. Li, X. Wang, and Y. Li, "The speechtransformer for large-scale Mandarin Chinese speech recognition," in *2019 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, UK, 2019, pp. 7095–7099.
[49] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. 28th Int Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, Canada, 2015, pp. 577–585.
[50] Y. Lin, Y. Sung, J. Lei, M. Bansal, and G. Bertasius, "Vision transformers are parameter-efficient audio-visual learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, Canada, 2023, pp. 2299–2309.
[51] H. Zhu, M. Luo, R. Wang, A. Zheng, and R. He, "Deep audio-visual learning: A survey," *Int. J. Autom. Comput.*, vol. 18, pp. 351–376, 2021.
[52] X. Zhou *et al.*, "MMSpeech: Multi-modal multi-task encoder-decoder pre-training for speech recognition," arXiv preprint arXiv:2212.00500, 2022.
[53] M. Kim, G. Kim, S. Lee, and J. Ha, "St-BERT: Cross-modal language model pre-training for end-to-end spoken language understanding," in *Proc. ICASSP, 2021–2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 7478–7482.
[54] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, 2023.
[55] Y. Huang *et al.*, "What makes multi-modal learning better than single (provably)," in *Proc. 35th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2021, pp. 10944–10956.
[56] C. Yi, S. Zhou, and B. Xu, "Efficiently fusing pretrained acoustic and linguistic encoders for low-resource speech recognition," *IEEE Signal Proc. Let.*, vol. 28, pp. 788–792, 2021.
[57] Z. Chen *et al.*, "MAESTRO: Matched speech text representations through modality matching," in *Proc. Interspeech 2022*, Incheon, Korea, 2022, pp. 4093–4097.
[58] Z. Chen *et al.*, "MAESTRO-U: Leveraging joint speech-text representation learning for zero supervised speech ASR," in *Proc. 2022 IEEE Spok. Lang. Technol. Workshop (SLT)*, Doha, Qatar, 2023, pp. 68–75.
[59] Y. Tang, J. Pino, C. Wang, X. Ma, and D. Genzel, "A general multi-task learning framework to leverage text data for speech to text tasks," in *Proc. ICASSP, 2021–2021 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, 2021, pp. 6209–6213.
[60] Y. Tang *et al.*, "Unified speech-text pre-training for speech translation and recognition," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (ACL)*, Dublin, Ireland, 2022, pp. 1488–1499.

[61] A. Bapna *et al.*, "SLAM: A unified encoder for speech and language modeling via speech-text joint pre-training," arXiv preprint arXiv:2110.10329, 2021.
[62] Z. Juan, P. Motlicek, Q. Zhan, R. Braun, and K. Vesely, "Automatic speech recognition benchmark for air-traffic communications," in *Proc. Interspeech 2020*, Shanghai, China, 2020, pp. 2297–2301.
[63] J. Zuluaga-Gomez *et al.*, "How does pre-trained wav2vec 2.0 perform on domain-shifted ASR? An extensive benchmark on air traffic control communications," in *Proc. 2022 IEEE Spok. Lang. Technol. Workshop (SLT)*, Doha, Qatar, 2023, pp. 205–212.
[64] Y. Lin *et al.*, "ATCSpeechNet: A multilingual end-to-end speech recognition framework for air traffic control systems," *Appl. Soft Comput.*, vol. 112, pp. 107847, 2021.
[65] K. Zhou, Q. Yang, X. Sun, S. Liu, and J. Lu, "Improved CTC-attention based end-to-end speech recognition on air traffic control," in *Proc. Intell. Sci. Big Data Eng. Big Data Mach. Learn.: 9th Int. Conf. (IScIDE)*, Nanjing, China, 2019, pp. 187–196.
[66] S. Zhang, J. Kong, C. Chen, Y. Li, and H. Liang, "Speech GAU: A single head attention for Mandarin speech recognition for air traffic control," *Aerospace*, vol. 9, no. 8, pp. 395, 2022.
[67] J. Zuluaga-Gomez *et al.*, "Contextual semi-supervised learning: An approach to leverage air-surveillance and untranscribed ATC data in ASR systems," in *Proc. Interspeech 2021*, Brno, Czechia, 2021, pp. 3296–3300.
[68] I. Nigmatulina, R. Braun, J. Zuluaga-Gomez, and P. Motlicek, "Improving callsign recognition with air-surveillance data in air-traffic communication," arXiv preprint arXiv:2108.12156, 2021.
[69] M. Kocour *et al.*, "Boosting of contextual information in ASR for air-traffic call-sign recognition," in *Proc. Interspeech 2021*, Brno, Czechia, 2021, pp. 3301–3305.
[70] D. Guo, J. Zhang, B. Yang, and Y. Lin, "A comparative study of speaker role identification in air traffic communication using deep learning approaches," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 4, pp. 1–17, 2023.
[71] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 6000–6010.
[72] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Repr. (ICLR)*, San Diego, CA, USA, 2015.
[73] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.
[74] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Proc. 2017 20th Conf. Orient. Chapter Int. Coord. Comm. Speech Databases Speech I/O Syst. Assess. (O-COCOSDA)*, Seoul, Korea (South), 2017, pp. 1–5.
[75] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations (ICLR)*, San Diego, CA, USA, 2015.
[76] J. D. M. C. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Technol. (NAACL)*, Minneapolis, Minnesota, 2019, pp. 4171–4186.
[77] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, 2018.
[78] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. 5th Int. Conf. Learn. Repr. (ICLR)*, Toulon, France, 2016.