IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2024, accepted 26 June 2024, date of publication 5 July 2024, date of current version 15 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3424260

RESEARCH ARTICLE

Contextual Biasing for End-to-End Chinese ASR

KAI ZHANG[1], **QIUXIA ZHANG**[2], **CHUNG-CHE WANG**[2],
AND JYH-SHING ROGER JANG[2], **(Member, IEEE)**
[1]Yiwu Industrial and Commercial College, Yiwu, Zhejiang 322000, China
[2]Department of Computer Science & Information Engineering, National Taiwan University, Taipei 106, Taiwan

Corresponding author: Kai Zhang (randomjerry@foxmail.com)

This work was supported by Jinhua Social Science Federation Research Project under Grant ZD2024032.

ABSTRACT The end-to-end speech recognition approach exhibits higher robustness compared to conventional methods, enhancing recognition accuracy across diverse contexts. However, due to the absence of an independent language model, it struggles to identify vocabulary beyond the training data, thus impacting the recognition of certain specific terms. Adapting to various scenarios necessitates a pivot towards specific domains. This study, based on the CATSLU dataset, constructed two tasks for Chinese contextual biasing, targeting both proper nouns and mixed-domain sentences. Additionally, it explored four methods of contextual biasing at different stages within the speech recognition process: pre-recognition, within the model, decoding, and post-processing stages. Experimental results indicate that all biasing methods to some extent improved the recognition efficacy of the speech recognition model within specific domains.

INDEX TERMS Automatic speech recognition, context biasing, intent classification, hotwords.

I. INTRODUCTION

This section delineates the challenges present in speech recognition in daily life, elucidating the research motivation. It outlines the subsequent chapters' content and the contributions of this study.

A. RESEARCH MOTIVATION

Speech recognition technology finds extensive application in various everyday scenarios, spanning from in-car assistants to music streaming and academic conferences [1]. Within different contexts, commonly used vocal commands and vocabulary exhibit notable variations. For instance, in vehicular systems, prevalent recognition terms mainly revolve around destination details and navigation instructions. Meanwhile, music streaming applications concentrate on song titles, artists, and genres. Academic conferences involve the usage of specialized terms and jargon [2]. However, certain domains lack sufficient relevant speech data to effectively train end-to-end speech recognition systems, or the available data is not substantial enough to support a robust system.

Hence, there's a necessity to utilize general speech recognition data to train models, filling the void in these domains. However, general speech data often lacks domain-specific vocabulary or doesn't align with the usage frequency of common terms within specific contexts. Consequently, models trained solely on general speech data struggle to accurately recognize domain-specific common terms and specialized vocabulary in these contexts.

When it comes to speech recognition methods represented by Kaldi [3], it stands as a relatively mature approach in the field. It encompasses independent acoustic models, pronunciation dictionaries, and language models. For phonetic recognition tasks, it can employ GMM-HMM [4] or DNN-HMM [5] acoustic models, followed by n-gram language models or other neural network language models to generate text outputs. During language model training, adjusting the weights of common contextual words in the language model training data enhances the recognition rate of frequently used terms according to different contexts. Additionally, by modifying the WFST (Weighted Finite State Transducer) method, specific proprietary terms can be assigned initial scores to improve the recognition rate of domain-specific terms [6].

Present-day end-to-end speech recognition models, compared to traditional hybrid models, offer higher recognition rates and stronger adaptability to various environments [7]. However, due to the integration of

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

acoustic, pronunciation, and language models without an independent language model, improving the accuracy of recognizing context-specific words and proprietary terms through language model training and modification is not feasible [8].

Enhancing the accuracy of recognizing context-specific sentences when only relevant textual corpora or proprietary terms are available becomes a crucial functionality in commercial-grade end-to-end speech recognition. In practical scenarios, user utterances often involve uncertainty about the domain. For instance, when using a smart speaker, a user might want to request a video or inquire about the weather, two domains that often have distinct common words and proprietary terms.

Determining the context of user utterances is essential in addressing contextual biasings. Sometimes, discerning the context is necessary to prevent interference from proprietary terms with similar pronunciations across different domains. For example, in requesting a song, the name of the singer might be “Wang Lin,” while in requesting a movie, the name of the film might be “Wang Ling.”

If it were possible to determine the intent of a statement or which domain it belongs to before speech recognition, biasing known proprietary terms and common words in established domains could result in improved recognition outcomes.

B. RESEARCH CONTRIBUTION

The contributions of this paper encompass several key aspects:

1) Pioneering Chinese contextual biasing Tasks: Drawing from prior English research [9], leveraging the open-source dataset CATSLU [10], and introducing two open-source contextual biasing task methods in Chinese, thereby filling a void in this field.
2) Demonstrated the superiority of end-to-end intent identification methods: Utilizing pre-trained self-supervised learning models for context detection. Empirical evidence showcases that in terms of computational demands and accuracy, adopting end-to-end intent recognition methods surpasses traditional context detection methods. Additionally, it confirms that using end-to-end context detection effectively reduces the occurrence of misalignments.
3) Proposed Chinese alternative word prediction model: Proposing a Chinese alternative word prediction model to generate phonetically similar and commonly confused words for less frequently encountered Chinese proprietary terms. Its effectiveness has been empirically validated when applied in the post-processing stage of sentences.
4) Applied previous methods on the proposed context biasing task: Validating the effectiveness of replicating prior research methodologies on the proposed two contextual biasing tasks, specifically in model training and sentence decoding stages, and confirming their effectiveness through empirical validation.

C. SECTION OVERVIEW

The paper is structured into six chapters:

1) Introduction (Chapter 1): Presents the research motivation and outlines the content of this paper.
2) Literature Review (Chapter 2): Explores the field of context biasing based on previous studies, discussing its primary goals, the current landscape, various directions in context biasing research in end-to-end speech recognition, and introduces classic approaches in context biasing from prominent papers.
3) Dataset and Task Introduction (Chapter 3): Introduces the context biasing open-source task proposed in this paper based on the Chinese open dataset CATSLU [10], along with the datasets used in subsequent chapters.
4) Research Methodology (Chapter 4): Details the methods employed in this paper for context biasing across four stages of speech recognition.
5) Experimental Design and Results (Chapter 5): Provides a detailed account of experiments designed in four directions, encompassing experiment objectives, model designs, parameter configurations, dataset selections, and comparative analysis of experimental outcomes.
6) Conclusion and Future Work (Chapter 6): Summarizes conclusions drawn from the thesis experiments and proposes potential avenues for future research endeavors.

II. RELATED WORK

Because end-to-end speech recognition lacks an independent language model, it cannot be biased towards a specific domain solely by adjusting the language model. However, a mature commercial ASR system requires the capability to adjust the probabilities of different words appearing based on various contexts. Consequently, following the emergence of end-to-end speech recognition, there has been considerable research exploring contextual biasings in different directions to adapt speech recognition systems. This section will cover the current contextual biasing methods applied in end-to-end speech recognition systems, including approaches across multiple stages of speech recognition. It will also provide a brief overview of the development of each method and conclude with an assessment based on literature.

A. CLEAN TRAINING DATA

In a study conducted by Google [11], it was mentioned that data cleansing can facilitate contextual biasings, incorporating textual training data from text searches into speech recognition systems for voice searches. This paper introduced three straightforward data selection strategies aimed at reducing the corpus size required for language model training while enhancing the recognition quality of rare vocabulary, all without compromising overall performance. These strategies encompass:

1) Gradually downsampling high-frequency sentences using a logarithmic function to alleviate the “head effect” in the corpus.

2) Explicitly filtering sentences within the acoustic data that contain rare words to augment their representation during training.
3) Employing contrastive data selection based on perplexity, filtering solely those queries that closely resemble the target domain (voice search).

As depicted in Figure 1, conducted data selection on a portion of anonymized Google search traffic, yielding a subset 53 times smaller than the original corpus. Notably, this subset exhibited improvements in both the head and tail ends of the data distribution. Through shallow fusion, the language model refined via data selection achieved up to a 24% relative decrease in Word Error Rate (WER) for rare words compared to the language model trained on the original corpus, while maintaining the overall WER unchanged.

B. SHALLOW FUSION

In end-to-end speech recognition, the method involving the language model participating in scoring during the decoding stage is referred to as shallow fusion. Initially proposed in a study dating back to 2018, introduced a technique known as shallow fusion [12]. It involves using logarithmic linear interpolation at each search step to combine the probability distributions of an external language model and the internal language model (i.e., the decoder). This approach enhances the fluency and accuracy of the output sequence. Compared to deep fusion techniques, shallow fusion offers the following advantages:

1) There's no need to modify the original sequence-to-sequence model structure or retrain it.
2) The weights between the external and internal language models can be flexibly adjusted.
3) It's convenient to interchange external language models of different types or sources seamlessly.

Compared the effectiveness of shallow fusion across various types of language models (neural network and n-gram), different decoding units (words, characters, and subwords), and diverse tasks (Google Voice Search and Switchboard). In the case of Google Voice Search, employing a neural network language model and subwords as an external language model, shallow fusion led to a 9.1% relative reduction in WER compared to the baseline sequence-to-sequence model, eliminating the need for a second rescoring pass. However, the method proposed in this study performed shallow fusion at the word unit level. As a result, words not present in the training data were pruned prematurely before beam search, hindering the possibility of biasing them.

In a study conducted in 2019 [13], the approach involved executing shallow fusion at the subword unit level to prevent the premature pruning of rare words during beam search. Additionally, to mitigate the impact on text that doesn't require biasing, the study experimented with a prefix-triggering method (e.g., "call," "text"), performing the bias only when such prefixes appeared. This enhanced version of shallow fusion, post-improvements, gained widespread adoption in subsequent end-to-end speech recognition systems. For instance, it was integrated into open-source tools like wenet2.0 [14].

C. ALTERNATIVE SPELLING PREDICTIONS

In the post-decoding correction phase, there's a method for contextual biasings as well. It involves gathering incorrectly recognized words from ASR outputs and training an alternative spelling prediction model. This model aids in generating alternative spellings for the target biased words [9].

This study utilized an end-to-end speech recognition model based on CTC and attention mechanisms. It amalgamated two distinct decoding algorithms, namely CTC prefix beam search and WFST decoding. During the decoding process, the authors employed a shallow fusion approach. They compiled a user-defined contextual bias word list into WFST representation and combined it with the speech recognition model to augment the weightage of bias words.

The authors proposed an alternative spelling prediction model to enhance the identification of rare and out-of-vocabulary (OOV) words. This model leverages the speech recognition model and training data to generate alternative spellings for biased words, subsequently adding them to the biased WFST. This method eliminates the need for additional pronunciation dictionaries or speech synthesis systems. The approach significantly improved the accuracy of rare and OOV words, presenting a novel avenue for contextual biasing methodologies.

D. CONTEXTUAL LISTEN, ATTEND AND SPELL (CLAS) CONTEXT BIASING

In the CLAS study, an attention-based context biasing method was proposed [15]. This system possesses the ability during inference to handle contextual phrases that might contain out-of-vocabulary (OOV) terms unseen during training. This approach eliminates the need for specific context information during training or meticulous adjustments of rescoring weights, yet it retains the capability to incorporate OOV terms. Across multiple tasks, the CLAS system exhibited a 68% improvement in WER compared to the baseline method, demonstrating the advantage of joint optimization over individually trained components. One key advantage of this method is its dynamic inclusion of contextual information during recognition, without the necessity for on-the-fly rescoring using externally trained language models during inference.

However, experimental results indicate that this method has limitations when dealing with a substantial number of biased terms, as depicted in Figure 2, leading to a noticeable decrease in the bias effectiveness. In real-world applications, there's a necessity for biasing a large number of potentially existing specialized terms.

E. WHISPER CONTEXT BIASING

In the current Whisper open-source model known for its robustness, there exists a context biasing method as well

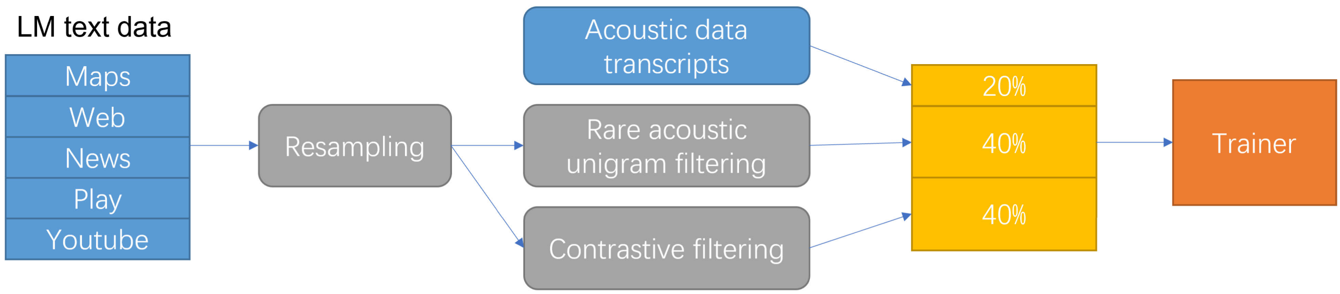


FIGURE 1. Schematic diagram of the method for cleaning training data [11].

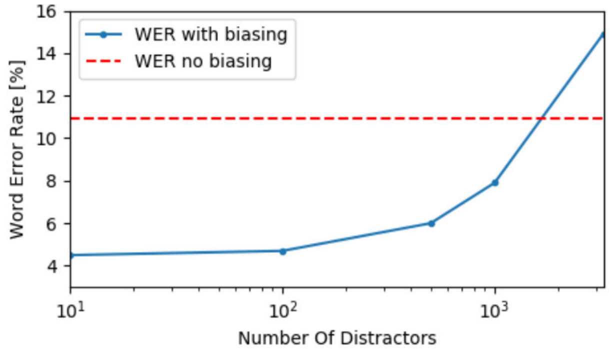


FIGURE 2. The impact of the number of CLAS bias keywords on WER [15].

TABLE 1. Detailed information on each area of the CATSLU dataset.

Domain	Number of Speakers	Training set (Number of statements)	Validation set (Number of statements)	Test set (Number of statements)
Map	1788	5093	921	1578
Music	268	2189	381	676
Weather	276	341	378	2660
Video	227	205	195	1641

[16]. This method remains based on attention mechanisms. By providing contextual information as a prompt input to the Whisper model right before inputting the audio, we can create a preceding context. During decoding, the model refers to the specialized terms, context, language, and grammar structures present in this preceding context. If we can predefine the domain of the input speech, this method allows inputting potential preceding contexts for that domain to enhance the decoding accuracy.

However, this approach has its limitations. It's challenging to specify a vast array of specialized terms for biasing, and designing the prompt might necessitate manual intervention. The content of the prompt could significantly impact the effectiveness of the subsequent context biasing.

F. CONTEXT BIASING TASK

There are diverse directions and methods for conducting context biasing, yet evaluating them consistently remains a challenge as each study often follows its own approach. For instance, in Amazon's research, comparisons were made using different methods on internal datasets [17]. Similarly, Google's research relied on its own text and speech search datasets for evaluations [11].

The usage of non-open-source data hinders reproducibility in other research endeavors, which is not the desired approach. Exploring methods to design contextual biasing tasks on an open-source dataset, a notable study presented an effective approach [9]. This research utilized the Earnings21 dataset [18], derived from public company phone meetings, rich in executive and company names. The study marked proprietary terms using spaCy [1] as the targeted keywords for biasing. They also introduced interference, words not present in the test set but existing in the target keyword list.

By leveraging an open-source ASR model, they applied various contextual biasing methods to decode the test set. The decoding outcomes were categorized into four types: words, phrases, rare words, and OOV words. The assessment of these categories' decoding performance served as a metric to evaluate the effectiveness of the contextual biasing methods. This approach was instrumental in guiding our subsequent research efforts to design open-source contextual biasing tasks in Chinese, ensuring replicability for future researchers.

III. DATASET AND TASK

A. DATASET

1) CATSLU DATASET

The CATSLU dataset serves as an oral comprehension corpus comprising both audio and textual information, encompassing four conversational domains: map navigation, music search, weather forecasting, and video (movies and TV) search. These data are sourced from real dialogue systems, inclusive of speech signals from diverse users, automatically transcribed speech-to-text, and manually transcribed text. The aim of this dataset is to explore leveraging multimodal information and domain knowledge to enhance the performance and robustness of oral comprehension. It's divided into two scenarios: single-domain oral comprehension and cross-domain adaptive oral comprehension [10]

The CATSLU dataset comprises 521 users, over 40,000 dialogues, and more than 100,000 slot-value pairs. The total duration of this dataset is approximately 30 hours, with an average length of 2.7 seconds per record. CATSLU is the pioneering publicly available Chinese oral comprehension dataset and the first to include both audio and textual features for oral comprehension. For detailed information, refer to Table 1.

2) WENETSPEECH

WenetSpeech constitutes a large-scale, multi-domain Chinese speech recognition dataset, encompassing the following components [19]:

1) The dataset comprises over 10,000 hours of meticulously annotated speech data, spanning across 10 domains. These domains encompass audiobooks, reviews, documentaries, dramas, interviews, readings, conversations, variety shows, and others. Sourced from YouTube and Podcasts, these datasets exhibit diverse speaking styles, settings, topics, and noise conditions.
2) More than 2,400 hours of lightly annotated speech data primarily sourced from Podcasts. An advanced automatic speech recognition system is utilized for initial transcription, followed by an end-to-end label error detection technique for filtering and refinement.
3) Over 10,000 hours of unlabeled speech data, suitable for semi-supervised or unsupervised speech representation learning purposes.
4) Three manually annotated high-quality test sets are available for evaluating the performance of the speech recognition system. The "Dev" set is used for cross-validation during the training process. The "Test Net" set is gathered from the internet and serves as a matching test set. Lastly, the "Test Meeting" set, recorded from actual meetings, is a more challenging non-matching test set.

The WenetSpeech dataset stands as one of the most extensive open-source Chinese speech recognition resources available. Leveraging optical character recognition and automatic speech recognition techniques, it generates candidate audio/text pairs. Furthermore, it introduces an innovative end-to-end label error detection method to further verify and filter these candidate pairs. This method utilizes a pre-trained wav2vec 2.0 model as a feature extractor, incorporating a binary classifier based on CTC loss function and edit distance threshold policy. This approach efficiently detects low-quality or mismatched candidate pairs while retaining high-quality or matching ones.

Designed to offer the research community a resource with diverse domains and significant scale, the WenetSpeech dataset aims to develop more universal and robust automatic speech recognition systems. It's released under the CC-BY 4.0 license, providing benchmark systems built on three popular toolkits: Kaldi, ESPnet, and WeNet. These benchmark systems have achieved satisfactory or even surpassed the best results on three test sets compared to existing open-source Chinese datasets.

TABLE 2. WenetSpeech statement duration distribution in various fields.

Domain	Youtube	Podcast	Total
audiobook	0	250.9	250.9
commentary	112.6	135.7	248.3
documentary	386.7	90.5	477.2
drama	4338.2	0	4338.2
interview	324.2	614	938.2
news	0	868	868
reading	0	1110.2	1110.2
talk	204	90.7	294.7
variety	603.3	224.5	827.8
others	144	507.5	651.5
Total	6113	3892	10005

3) AISHELL-1

The AISHELL-1 is an open-source Chinese Mandarin speech corpus released by Beijing Shell-Tech Co., Ltd. It's primarily utilized for research and system construction in the field of speech recognition.

AISHELL-1 comprises approximately 178 hours (approximately 1.5 million words) of content recorded by 400 individuals from various Chinese accent regions, including Mandarin, Cantonese, Shanghainese, among others. These recordings were captured in quiet indoor environments using high-fidelity microphones. The content covers 11 different domains, such as smart homes, autonomous driving, and industrial production. Each recording is manually transcribed and annotated. The data has a sampling rate of 16kHz and boasts transcription accuracy of over 97%. The dataset is divided into three sections: a training set with 120,098 sentences, a development set with 14,326 sentences, and a test set with 7,176 sentences. Additionally, the dataset includes dictionary files corresponding to the transcriptions [20].

AISHELL-1, being the inaugural release in the AISHELL series, is among the most widely used versions. It has been referenced in numerous research papers and employed as a benchmark or comparative dataset in various evaluation tasks. For instance, in the 12th International Chinese Spoken Language Processing Competition (CCLASR 2018) held in 2018, AISHELL-1 was used as one of the training datasets provided to participants. Similarly, in the 14th International Chinese Spoken Language Processing Competition (CCLASR 2020) held in 2020, AISHELL-1 was utilized as one of the test datasets for evaluating participating systems.

4) AISHELL-2

AISHELL-2 dataset, an open-source extensive speech corpus primarily tailored for Mandarin speech recognition research, was released by the ShellBe Foundation. This dataset encompasses 1,000 hours of high-quality Mandarin speech data, recorded through iOS devices, and is made freely available to the academic research community.

The recordings involved 1991 speakers, including 845 males and 1,146 females, spanning ages from 11 years old to over 40 years old. The speakers were instructed to articulate all recorded content in standard Mandarin,

TABLE 3. Detailed information on each area of the CATSLU dataset.

Domain	Statements	Keywords	Distractor
Movie Titles	785	555	381
Artist Names	504	390	210
City Names	601	930	690

albeit minor accent differences exist. Based on accent characteristics, 1,293 speakers used a northern accent, 678 speakers used a southern accent, and 20 speakers had other accents. Among the participants, 1347 speakers were recorded in studio conditions, while the rest were recorded in acoustically uncontrolled living room settings. The recorded content encompasses eight primary themes, including voice commands (such as IoT device control and numerical sequence input), tourist attractions, entertainment, finance, technology, sports, English spelling, and free speech without a specific theme [21]

AISHELL-1 is actually a subset of AISHELL-2.

B. TASKS

As per the previous study [9] that introduced context biasing tasks evaluation criteria on the English dataset Earnings21, we utilized the Chinese dataset CATSLU for the context biasing tasks. This dataset contains a variety of proprietary terms and a keyword list, coupled with 16kHz audio files and transcriptions. Additionally, based on different application scenarios, I proposed the following two context biasing tasks.

1) CATSLU PROPRIETARY TERM CONTEXT BIASING TASK

In this task, sentences only contain occurrences of specific proprietary terms, segmented into three subtasks based on distinct proprietary terms.

1) Contextual biasing within sentences containing movie titles.
2) Contextual biasing within sentences containing artist names.
3) Contextual biasing within sentences containing city names.

Each task entails its unique set of keywords, acting as target terms for context biasing. However, not all keywords listed are guaranteed to appear in the sentences. Those keywords that do not feature in the sentences are termed “interference items.” The distribution of sentences and keywords across the three subtasks is illustrated in Table 3.

This task operates under zero-shot learning conditions, meaning no training data is provided. It requires testing using a standardized open-source model. The chosen open-source model is based on the default Conformer configuration in Wenet [22], trained on the heavily annotated WenetSpeech dataset encompassing 10,000 hours of data. The aim of this task is to assess the contextual biasing capability of methods specifically regarding specialized terms.

TABLE 4. CATSLU multi-domain mixed corpus context bias data set division.

Info	Training set	Test set
Statements	4191	1050
Time Length (Hours)	3.6	0.9

2) CATSLU MULTI-DOMAIN MIXED CORPUS CONTEXT BIASING

In this task, all sentences from three domains—video search, music search, and weather inquiry—are incorporated. They are divided into training and testing sets using a class-balanced approach. The training set accounts for 80%, while the testing set represents 20% of the data. You can find detailed information in Table 4.

In this task, leveraging the open-source model trained on the 10,000-hour WenetSpeech corpus, context biasing can be applied using the training set data and a keyword list. The keywords represent commonly used terms in various domains, like “play” or “navigation,” but they might not specifically be proper nouns. During testing, the data set is mixed, without explicit domain labels. This setup mirrors real-world usage scenarios, aiming to assess the impact of context detection methods on context biasing and evaluate the performance of domain-specific context biasing methods.

C. ASSESSMENT METHOD

1) CHARACTER ERROR RATE

By computing the CER, which involves comparing all decoded sentences to the standard answers after decoding, we can assess the overall impact of the context shift method. This helps us determine whether over-Biasing might lead to a high error rate for words outside the keywords. A lower CER value indicates fewer errors in the corpus. Let’s denote S as the number of substituted characters, D as the number of deleted characters, I as the number of inserted characters, and N as the total number of characters in the reference sequence. The formula for calculating CER is as follows:

$$CER = \frac{S + D + I}{N} \tag{1}$$

2) KEYWORDS ERROR RATE

To evaluate the effectiveness of different context biasing methods in speech recognition tasks, we calculated the error rate of all keywords present in the corpus within the identification results. The keyword error rate signifies the proportion of keywords originally present in correct sentences but inaccurately predicted, which is equivalent to 1 minus the recall rate.

IV. RESEARCH METHODS

In this study, we aim to reduce the Keyword Error Rate (KER) associated with target keywords and minimize the Character Error Rate (CER) in specific domains through context biasing across four stages of speech recognition: pre-recognition, recognition model, decoding, and post-decoding correction.

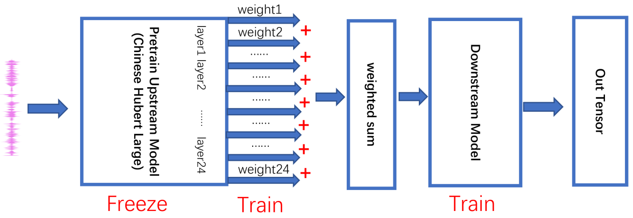


FIGURE 3. Schematic diagram of end-to-end intent identification process.

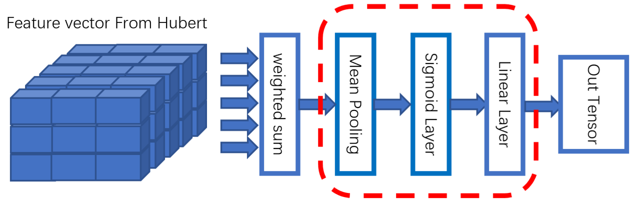


FIGURE 4. Schematic diagram of linear downstream model.

Prior to speech recognition, we employed an end-to-end intent recognition method for context detection to ensure the accuracy of the bias direction. For the recognition model, we utilized fine-tuning of pre-trained models to adapt to domain and environmental features. During decoding, we applied a shallow fusion technique in Mandarin, assigning higher reward scores to target keywords. Post-decoding, we used the trained model to generate a list of candidate words for the target keywords and conducted context biasing for subsequent correction.

In this section, we'll provide a detailed explanation of these four methods and demonstrate their efficacy through experiments in the following sections.

A. END-TO-END INTENT RECOGNITION

Given the diverse application scenarios of speech recognition systems, the domain of incoming speech may entail uncertainties. In situations where the domain of incoming speech is unknown, we cannot be certain about the method of biasing. We attempted to use intent recognition as a means of context detection. The aim was to determine the domain to which the incoming speech belongs, thereby guiding the subsequent methods of biasing in the right direction.

1) OVERALL PROCESS

We adopted an end-to-end approach for intent recognition, which differs from conventional methods. Traditional approaches often involve using ASR systems for speech recognition first, followed by inputting the recognized text into a language model for intent recognition. However, studies have shown that an end-to-end approach requires less training data and yields higher accuracy [23]. This is particularly valuable in context-biasing problems, where relevant context training data is often lacking. Conventional ASR models may not offer optimal decoding results. Hence, we opted for an end-to-end method for intent recognition to determine the context requiring biasing.

As shown in Figure 3, we initially extract speech features using the pre-trained upstream model HuBERT [24] from the first 5 seconds of audio files longer than 5 seconds. Then, these feature vectors from each layer of HuBERT are inputted into the downstream model in the form of a weighted sum, where the weights are learned along with the downstream model, ultimately directly outputting the intent.

2) HUBERT FEATURE EXTRACTION

HuBERT [24], a self-supervised learning model, employs a clustering method to classify and label audio data. Trained on WenetSpeech [19], the Chinese version of HuBERT effectively extracts features from input sentences. Unlike conventional deep learning models, self-supervised learning doesn't require extensive labeled data, which makes it adept at handling noise and variations in speech datasets.

Using a masking technique similar to BERT during training, HuBERT predicts partially masked labels, allowing the model to learn audio features and patterns. The clustering method in HuBERT automatically annotates audio data, enabling the model to naturally capture relevant features during training.

To leverage HuBERT-extracted features optimally, this study refers to SUPERB's [25] approach, summing the outputs of each HuBERT layer with weighted values for downstream model training. This method maximizes the utilization of HuBERT's features, enhancing the performance of the downstream model. Weight values are trained alongside the downstream model, enabling automatic selection of specific HuBERT layer information. This approach effectively addresses overfitting and information propagation issues in deep neural networks, boosting the downstream model's generalization capabilities.

3) INTENT RECOGNITION DOWNSTREAM MODEL

The downstream model for intent classification functions by determining which intent class a given feature vector belongs to. It takes feature vectors extracted from audio files by the HuBERT upstream model as input and outputs the probability for each category. When performing intent classification, the choice of downstream model architecture can depend on the complexity of the task.

With smaller datasets, opting for a Linear model as the downstream model architecture is a viable option. After extracting features through mean pooling, the features are directly fed into a Linear layer for classification prediction, as in Figure 4

When there's an adequate amount of data available, selecting a Bi-LSTM architecture as the downstream model tends to yield better results. To reduce the parameter count, the last-dimensional features are reduced to 256 using a Linear Layer. To consider continuous contextual features, the output from the linear model is input into a Bi-directional Long Short-Term Memory (Bi-LSTM) layer with a hidden

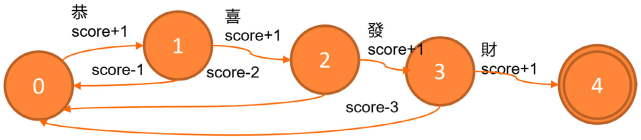


FIGURE 5. Schematic diagram of Chinese shallow fusion WFST.

size of 256. To further extract intent features, average pooling with a kernel size of (3,3) is applied, followed by a linear layer for the final output.

This approach offers the advantage of directly extracting intent information from the input speech without the need for ASR transcription into text. ASR transcription might be influenced by the ASR model's recognition outcomes, which could impact downstream tasks.

B. FINE-TUNING A PRE-TRAINED MODEL

If there's a limited amount of data in the target domain, fine-tuning the initial ASR model can be a beneficial strategy.

With a small dataset, training a complete ASR system may be challenging. Therefore, leveraging a pre-trained end-to-end ASR model that already possesses certain robustness as initial parameters and adjusting the training steps and learning rate helps gauge the impact of domain-specific data on the original model. Typically, to maximize domain adaptation effects, an iterative process is followed, iterating through training while selecting the model that performs best on a validation set for fine-tuning usage.

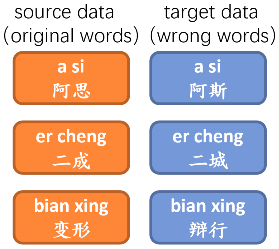
When there's an abundance of training data, achieving optimal results often requires a sufficient number of training steps. In contrast, with limited training data, achieving fine-tuning might necessitate only a few steps to adapt the original model.

C. CHINESE SHALLOW FUSION

In numerous contextual biasing tasks, training data is not provided, featuring only select keywords to bias. It's impractical to train specific speech recognition models for each small domain. Hence, we opt for the shallow fusion method during the decoding stage.

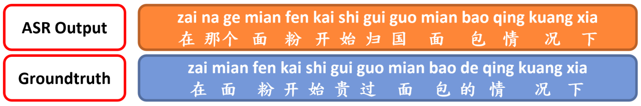
In Chinese, the smallest unit in speech recognition corresponds to characters, aligning well with phonemes in English. Thus, we utilize the Wenet tool [14] to construct the target keyword needing biasing into WFST. As illustrated in Figure 5, "Congratulations" is compiled into WFST [26]. Each decoded character in sequence fetches a corresponding reward score until the entire word is completely decoded.

During decoding, the system concurrently explores the WFST of the target keyword, considering both the CTC loss and the bonus assigned by the keyword. If it fails to decode the target keyword correctly, it will deduct the previously added scores. The assigned bonus scores are customizable, where higher scores correspond to greater biasing weight. These scores' configuration can be tailored based on different datasets and specific requirements.



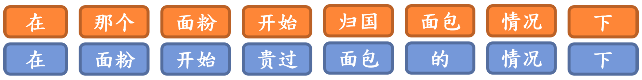
source data (original words)	target data (wrong words)
a si 阿思	a si 阿斯
er cheng 二成	er cheng 二城
bian xing 变形	bian xing 辩行

FIGURE 6. Alternative word prediction model training data example.



ASR Output	zai na ge mian fen kai shi gui guo mian bao qing kuang xia 在 那个 面 粉 开 始 归 国 面 包 情 况 下
Groundtruth	zai mian fen kai shi gui guo mian bao de qing kuang xia 在 面 粉 开 始 贵 过 面 包 的 情 况 下

FIGURE 7. ASR model output and groundtruth examples.



在	那个	面粉	开始	归国	面包	情况	下
在	面粉	开始	贵过	面包	的	情况	下

FIGURE 8. Example of input and output text jieba word segmentation.

D. ALTERNATIVE WORD PREDICTION (AWP)

To further enhance the recognition accuracy of proper nouns in the post-correction stage, we explored an approach for predicting alternative terms in Chinese. This method involves training a Transformer specifically designed to generate common misrecognized terms for proper nouns. Using this trained Transformer, a list of alternative terms for the proper nouns is generated, containing "n" commonly mistaken variations. The "n" value can be adjusted as needed. Post decoding, a search for these common mistakes is conducted, replacing them with potential proper nouns.

1) GENERATE TRAINING DATA

Each training data point for the alternative term prediction model consists of a pair of words. One word is the ground truth from the corpus, while the other corresponds to the decoded mistaken word, as illustrated in Figure 6.

For a more precise model adaptation, we need to use the target-shifted model to generate training data for the alternative word prediction model. By preparing a substantial amount of speech data, open-source corpora such as AISHELL can be utilized. It is preferable to choose corpora related to the target-shifted domain if available. The prepared data is input into the ASR model that needs shifting. This process yields the results of speech after decoding by the ASR model, as illustrated in Figure 7.

In order to acquire word pairs for training data, identifying pairs of words associated with decoding errors is essential. However, the current output units of end-to-end ASR models mostly consist of characters. Therefore, initial segmentation of input and output sentences is required. We've opted for employing the Chinese segmentation tool, jieba [27],

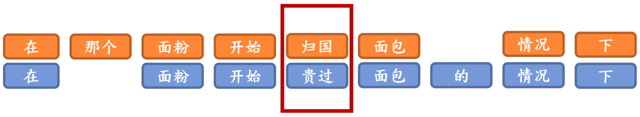


FIGURE 9. Complete sentence examples after alignment.

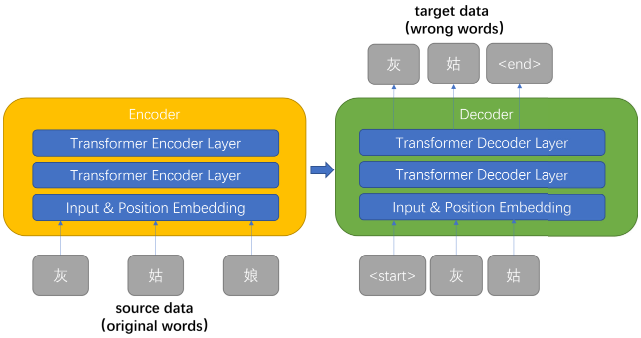

target data
(wrong words)
灰
姑
<end>
Encoder
Transformer Encoder Layer
Transformer Encoder Layer
Input & Position Embedding
Decoder
Transformer Decoder Layer
Transformer Decoder Layer
Input & Position Embedding
灰
姑
娘
source data
(original words)
<start>
灰
姑

FIGURE 10. Alternative word prediction model architecture diagram.

to segment both input and output sentences. Once segmented, the sentences appear as illustrated in Figure 8.

We also require identification of the words undergoing substitution from the segmented input and output text to obtain training pairs. This alignment is achieved through the algorithm for computing the shortest edit distance, expressed by the following formula:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} \end{cases} \tag{2}$$

After the alignment is completed, the sentence style shown in Figure 9 can be obtained

In this manner, we can identify the words that have undergone substitution. The uncertainty in the length of the output from the ASR model often results in one word overlapping into another. To address issues arising from ASR output errors causing different word lengths after alignment, we filtered out data where the word lengths differed. Only the data with the same word lengths were retained for training purposes.

2) ALTERNATIVE WORD PREDICTION MODEL ARCHITECTURE

We opted for a Transformer architecture for the candidate word prediction model, comprising two encoders and two decoders with a hidden layer size of 256, employing 8 multi-head settings and setting the feed-forward hidden size to 2048. Regarding the handling of the training data, we treated Chinese characters as the smallest units, considering each word as a sequence composed of multiple characters. The length of the output word is determined by the Transformer, extending until it generates an end symbol. The model architecture is illustrated in Figure 10.

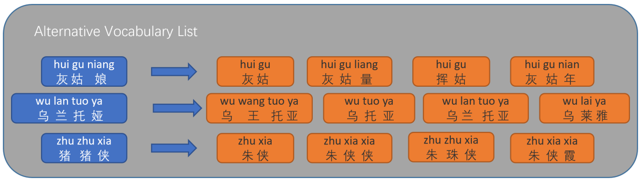


FIGURE 11. Example image of alternative vocabulary list.

This model's specific implementation was derived from the network architecture and data processing methods used in end-to-end machine translation. We utilized the OpenNMT [28] open-source tool to execute this implementation.

3) GENERATION AND USE OF ALTERNATIVE VOCABULARY LISTS

The model utilizes the collected training data to train and subsequently generate alternative words for specialized terms. By inputting the target keywords for biasing into the predictive model, we can configure it to produce an output of N best alternative words while providing the similarity of each word with the input. Setting the size of N allows us to control the number of alternative words outputted by the model. Additionally, establishing a similarity threshold helps filter out words with low similarity, ensuring the reasonableness of subsequent replacements. These parameters enable us to manage the quantity of alternative words for each keyword in the generated table, as illustrated in Figure 11.

We can employ the alternative word table for conducting post-corrections on the keywords. We'll start by examining the input text, initially determining if it contains sentences that potentially include specialized terms. If it likely contains specialized terms, we proceed with the post-correction steps. Next, we search the sentence to check for the presence of keywords. If the keywords are already present, we halt the correction. If they're not, we begin searching for the alternative words from the alternative word table. The priority is given to alternative words with more characters. If an alternative word from the table is found in the sentence needing correction, we replace it with the keyword and terminate the search. In practical applications, searching for keywords and alternative words can be implemented using WFST [26].

4) COMMON WORD LIST

To prevent the frequent occurrence of generated alternative words in everyday conversations, which may lead to erroneous replacements, we employ a common word table. We conduct a 4-gram count on the text corpus of an open-source dataset. This involves counting the occurrences of each four-character combination around each word. Naturally, the counting includes three-character, two-character, and one-character combinations as well. Let us illustrate the statistical method using the Chinese phrase "gong xi fa cai" as an example, as shown in Figure 12.

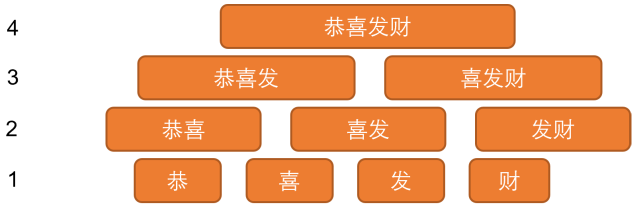


FIGURE 12. Schematic diagram of common word counting method.

When a character combination exceeds a count of ‘‘n’’ appearances, that combination is included in the common word table. The value of ‘‘n’’ can be adjusted according to the specific scenario to avoid erroneous replacements.

Once the common word table is established, it can be directly applied during the correction phase. When alternative words are identified during post-correction, an additional check is performed to verify whether these words exist in the common word table. If they are present, no replacement is made. Alternatively, the common word table can be used during the generation of the alternative word table. If any words generated are found in the common word table, they are not included in the list of alternative words.

V. EXPERIMENTAL DESIGN AND DISCUSSION OF RESULTS

In this section, we focus on contextual displacement tasks and conduct experiments using four distinct methods across four stages to validate their effectiveness. We introduce the experimental workflow, environmental parameters, and basic information. For each stage, we've designed comparative experiments with different models and parameters to demonstrate the effectiveness of our methods.

A. EXPERIMENTAL PROCESS AND SETTINGS

This subsection outlines the experiment's procedures and the hardware environment used for conducting the experiments.

1) EXPERIMENTAL PROCESS

In Section V-B, we designed experiments related to context detection, consisting of three parts. These segments contrasted various downstream models, different intent recognition methods, and observed their impact on overall context biasing when applied in the context biasing scenario.

Section V-E covered experiments conducted during model training for contextual biasing methods. By comparing the outcomes of using related-domain corpora to fine-tune models, we aimed to demonstrate the effectiveness of the fine-tuning approach.

Section V-F comprised experiments focusing on decoding and post-processing stages of context biasing. This section included four sub-experiments: the first involved adjusting reward scores to observe their impact on the Chinese shallow fusion method. The second compared different alternative word prediction model architectures. The third attempted to prevent erroneous substitutions by utilizing common word lists. Lastly, the fourth experiment applied the Chinese shallow fusion method and alternative word prediction in the context biasing task, contrasting it with using only the shallow fusion approach, demonstrating the method's advantages.

TABLE 5. Experimental environment.

item	parameter
CPU	Intel ® Core (TM) i7-12700KF
GPU	NVIDIA GeForce RTX 3080ti
RAM	64GB
OS	Ubuntu 20.04 LTS

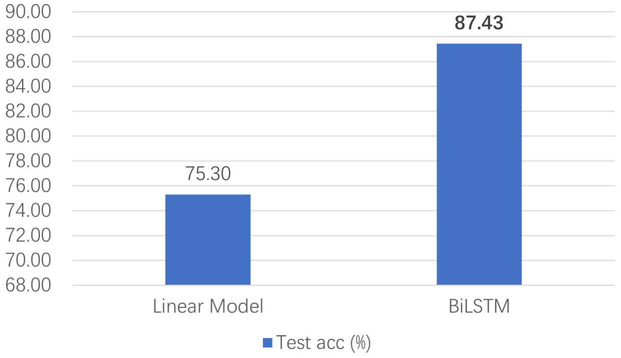


FIGURE 13. Compare experimental results using the accuracy of different downstream models.

The experimental environment for this study is detailed in Table 5.

B. END-TO-END INTENT RECOGNITION MODEL FOR AUTOMATIC CONTEXT DETECTION

In this set of experiments, we utilized the CATSLU multi-domain mixed corpus context biasing task as our context biasing task. The aim was to test the effectiveness of the method when faced with uncertain domain input sentences.

1) THE IMPACT OF DIFFERENT DOWNSTREAM MODELS ON INTENT RECOGNITION

In this experiment, our aim was to compare the accuracy of various downstream models. We initiated the process by utilizing the upstream model for feature extraction from the audio data. Subsequently, we applied different downstream models for intent recognition. The intent recognition models were trained using the training dataset and evaluated on the test dataset. For the comparison of downstream models, we specifically chose the Linear model and the Bi-LSTM model. Our configurations included a batch size of 32, a learning rate set to 1e-4, employing the Adam optimizer [29], and implementing a warm-up strategy, conducting training over 10 epochs. The experimental findings are illustrated in Figure 13.

The experimental outcomes demonstrate that our enhanced Bi-LSTM downstream model exhibited superior accuracy in this task. We attribute this performance to the Bi-LSTM

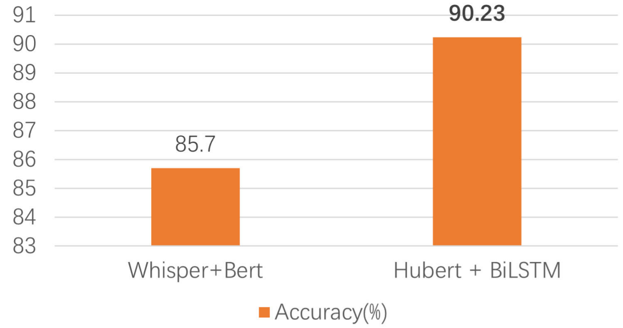


FIGURE 14. Comparison of experimental results on the accuracy of different methods.

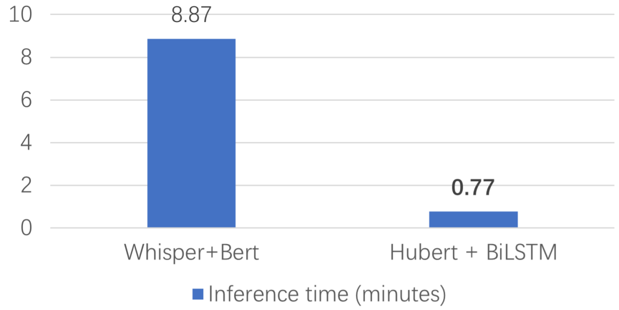


FIGURE 15. Experimental results comparing the efficiency of different methods.

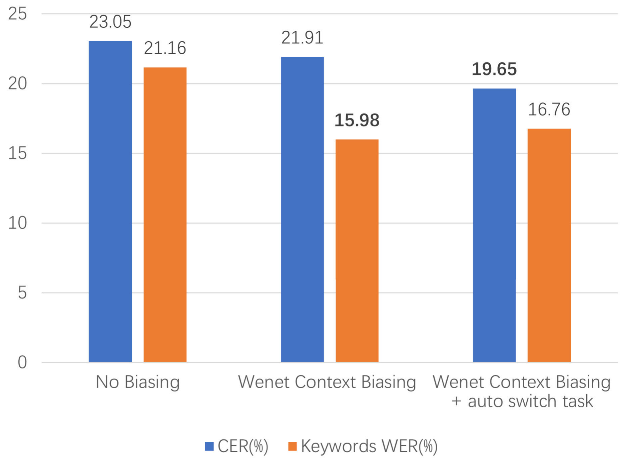


FIGURE 16. Experimental results of context detection applied in context biasing.

model's ability to capture contextual variations within continuous feature vectors more effectively. This model achieved an intent recognition accuracy of 87.43% on input sentences, validating its potential application in real-world context detection.

C. COMPARISON OF COMPUTING POWER AND ACCURACY OF DIFFERENT METHODS FOR INTENT RECOGNITION

The aim of this experiment was to compare the accuracy and computational requirements between end-to-end intent recognition and traditional intent recognition methods. To achieve this, we employed two different approaches for intent recognition. The first is the conventional method, involving the use of a speech recognition model to convert speech into text, followed by language model analysis of the text. The second approach is the end-to-end method, which directly analyzes the speech using a single model without any intermediary conversion.

In the traditional method, we opted for the Whisper model, currently considered one of the most advanced open-source speech recognition models [16], along with the widely used BERT language model [30]. For the Whisper model, we chose the Large version known for its superior recognition performance, while for BERT, we utilized the Chinese version suitable for Mandarin. During training, we kept the Whisper model fixed and solely trained the downstream classification layer of the BERT model, which is a single Linear layer. We set the batch size to 16, the learning rate to 5e-4, employed the AdamW optimizer [31], and used 10 epochs as the stopping condition for training.

In the end-to-end method, the Bi-LSTM model was selected as the downstream model due to its better performance in the previous experiment. Keeping the training parameters consistent with the traditional method, we conducted a comparison of the accuracy in intent recognition on the test set. The experimental results are depicted in Figure 14.

Additionally, we conducted efficiency tests on both methods in the experimental environment described in this section. We individually measured the time required for traditional intent recognition and end-to-end intent recognition to process one hour of speech data. The results are visualized in Figure 15.

The end-to-end intent recognition method not only outperformed the traditional approach in accuracy but also required less computational resources. This suggests that the end-to-end method might be more efficient and accurate for intent recognition tasks in your experiment.

D. APPLYING CONTEXT DETECTION METHOD BEFORE CONTEXT BIASING

The aim of this experiment was to assess the performance of end-to-end intent recognition in dealing with context biasing in unknown domain sentences. Initially, we employed the end-to-end intent recognition method for contextual detection and executed the relevant context biasing based on the detected context. We employed the Chinese shallow fusion method to effectuate context biasing, setting the biasing score at 3.0. We compared three distinct approaches in terms of accuracy: no biasing, shallow fusion applied to all keywords, and shallow fusion based on contextual detection results. Simultaneously, we established a control group with perfect contextual detection results. The experimental outcomes are illustrated in Figure 16.

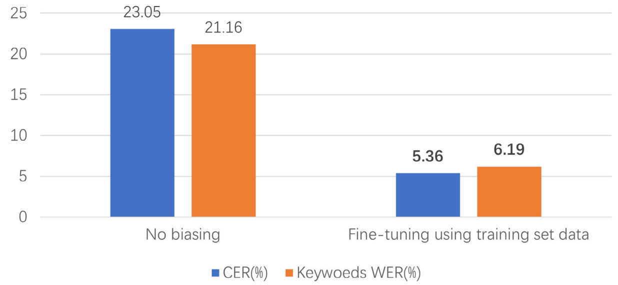


FIGURE 17. **Fine-tuning model experimental results.**

The data in the chart indicates that employing the context detection method effectively reduces the overall CER of the sentences, suggesting that this method can rectify certain erroneous bias directions. However, we've observed a slight increase in keyword error rates, which we suspect might be due to inherent inaccuracies in the context detection itself. Moreover, owing to the limited number of categorized keywords in the current test data, the method of using context detection doesn't outperform the non-biasing approach concerning keyword accuracy. Nevertheless, in practical applications involving multiple domains and extensive keyword lists, we anticipate that employing context detection would yield superior results.

We simulated the best-case scenario of context detection by using the original task labels, and found it achieved optimal contextual biasing effects. Surprisingly, contrary to our expectations, the CER for automated context detection, where all contexts were correctly detected, was higher than when there were errors in the context detection. We hypothesize this discrepancy might be attributed to certain classification errors in parts of the end-to-end intent recognition that involve non-specific terms, leading to confusion. The automated context detection effectively categorizes these terms into more probable classes, increasing the likelihood of correctly identifying the non-specific parts. This indirectly demonstrates the advantages of automated context detection in aiding ambiguous sentences to find their most fitting categories and execute context biasings.

E. FINE-TUNING THE MODEL WITH A SMALL AMOUNT OF CORPUS TO ACHIEVE CONTEXT BIASING

The primary objective of this study is to validate whether pre-trained models, after fine-tuning on limited data, can effectively adapt to various contextual Biasing scenarios.

For this purpose, we devised an experiment employing fine-tuning on a limited dataset, using the CATSLU mixed-corpus contextual biasing task as the evaluation benchmark. This task offers a portion of training data for fine-tuning our model. We opted for a pre-trained Conformer model trained on the open-source WenetSpeech [19] corpus (approximately 10,000 hours), initializing our fine-tuning process with its parameters. Throughout fine-tuning, we utilized 80-dimensional FBank features as input, with hyperparameters set at gradient accumulation of 16, batch size of 64, Adam optimizer, and a learning rate of 2e-3, conducting a total of 640 epochs.

Figure 17 illustrates the performance contrast between the fine-tuned and original models. This comparison demonstrates that fine-tuning the original ASR model using domain-specific data yields impressive results. This approach allows training distinct ASR models for different domains, and when used in conjunction with context detection methods, can achieve notably superior performance. However, it's crucial to note that this method requires access to relevant domain-specific speech data.

F. CONTEXTUAL BIASING IN DECODING AND POST-PROCESSING STAGES

In this experimental setup, we employed the CATSLU specialized noun contextual biasing task to assess the contextual biasing methods in the decoding and post-processing stages concerning specialized nouns.

1) THE IMPACT OF BIASING SCORES ON CHINESE SHALLOW FUSION

The objective of this experiment is to investigate the impact of different bias reward scores on the contextual biasing effect, using CER (Character Error Rate) and KER (Keyword Error Rate) as evaluation metrics. The aim is to identify the most suitable biasing score setting. For this purpose, we selected a challenging subtask within the CATSLU specialized noun contextual biasing task, specifically, video title recognition, as the experimental scenario.

In the experimental setup, we utilized a consistent open-source model, specifically the Conformer model trained on the WenetSpeech corpus [19] (approximately 10,000 hours). During the decoding phase, we implemented the Chinese shallow fusion technique to enhance the model's ability to recognize video titles. We experimented with different shallow fusion reward scores, ranging from 0 to 9 as integer values, where 0 indicates no utilization of the shallow fusion method. The experimental results are depicted in a line chart format, as shown in Figure 18.

The experimental results have shown some patterns that align with our expectations.

Initially, a moderate increase in the reward score when it's low effectively enhances both the overall sentence and keyword recognition rates. This suggests that the shallow fusion method assists in boosting the model's capability to handle biases. However, once the reward score surpasses a certain threshold, it leads to a rise in the overall CER of the sentences. This might occur due to excessively high reward scores causing the model to overly prioritize keywords at the expense of other parts of the sentence. Eventually, when the reward score reaches an extremely high value, it can even have a negative impact on KER, potentially leading to erroneous or repetitive keywords. Therefore, for practical applications, determining the optimal reward score

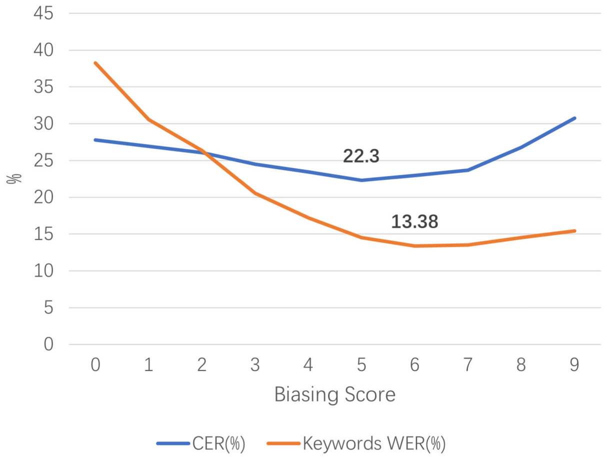


FIGURE 18. Biasing score adjustment experiment results.

values requires experiments tailored to specific use cases and requirements.

2) AWP MODEL SELECTION

The purpose of this experiment is to investigate how different types of AWP model architectures impact the performance of the AWP model in predicting ASR errors.

The evaluation criterion in this experiment is accuracy. Accuracy measures how well the AWP model predicts the same error output as the target biased ASR model when the ground truth is used as input. A well-performing AWP model should effectively predict how the ASR model misrecognizes the target keywords.

Both the training and testing datasets used in this study are derived from the AISHELL-2 corpus and feature extraction is carried out using the method described in this chapter. The testing dataset comprises 53,817 word pairs, accounting for approximately 10% of all data, while the training dataset consists of 538,169 word pairs, representing roughly 90% of all the data.

For this experiment, we selected a model based on the Transformer architecture, comprising two self-attention encoding layers and two self-attention decoding layers. We contrasted this model with a pretrained language model, mengzi-t5-base-mt, which is based on the t5 model and trained on a large-scale Chinese corpus.

During the training of the Transformer model, we tokenized the Chinese text at the character level as input units and employed the Adam optimizer for parameter updates. We set the batch size to 32, the learning rate to 2, conducted learning rate warm-up for the first 8000 steps, and kept the learning rate unchanged for subsequent steps until training reached 100,000 steps. When fine-tuning the T5 model, we used the same input units and batch size but switched to the Adafactor optimizer for parameter updates. The learning rate was set to 1e-3, and the entire training dataset underwent 10 iterations to ensure model convergence.

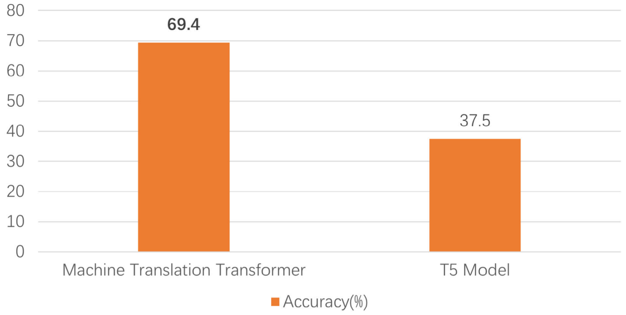


FIGURE 19. Experimental results of different AWP model architectures.

Upon concluding the training, we conducted an experimental assessment of the two models and showcased the outcomes in Figure 19.

The Transformer model displayed significantly superior accuracy in the task of predicting alternative words compared to utilizing the pre-trained language model T5. This suggests that the Transformer model adeptly generates suitable alternative words, whereas the pre-training of the T5 language model doesn't contribute significantly to this task. Consequently, in subsequent experiments, we'll employ the Transformer architecture as the foundational structure for the AWP model and further enhance and optimize it based on this.

3) COMMON WORD LIST

The experiment aimed to assess the efficacy of employing a common word list within the AWP method. We chose the CATSLU proprietary noun context biasing task, particularly focusing on the context biasing of movie titles as a sub-task for the experiment. The ASR model used for decoding was based on the Conformer model trained with WenetSpeech (approximately 10,000 hours).

To create the common word list, we performed a 4-gram analysis on AISHELL-2, a large-scale Chinese speech dataset, and extracted 4-grams appearing more than 500 times as common words. In the AWP method, during the candidate word search phase, we excluded all common words to reduce the chance of erroneously replacing them. We compared the AWP method with and without the common word list and contrasted them with the baseline method that didn't use the AWP method. All methods involved a post-correction step after employing the Chinese shallow fusion technique during the decoding stage. Figure 20 illustrates the experimental results across various metrics for each method.

Based on the data illustrated in the chart, it can be discerned that employing a common word list method, as opposed to the AWP method which does not utilize a common word list, slightly reduces the CER metric. This indicates that utilizing the common word list method effectively avoids erroneous substitutions of candidate words, thereby enhancing the accuracy of speech recognition. Consequently, in subsequent experiments, we'll incorporate the common word list method

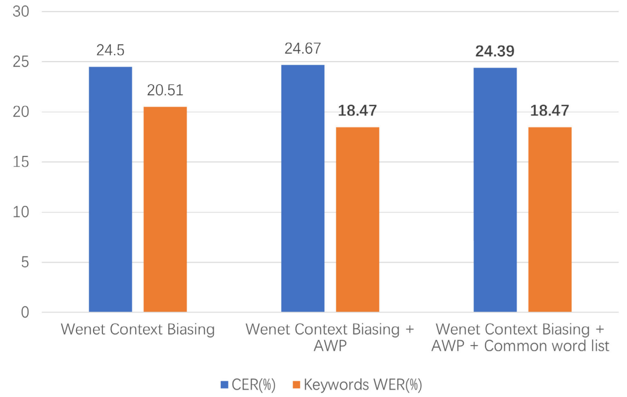


FIGURE 20. Common word list comparison experiment results.

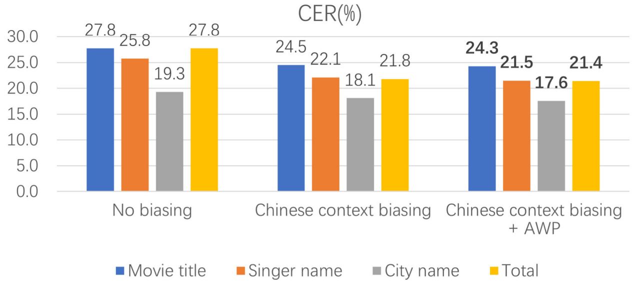


FIGURE 21. CATSLU proper noun context bias all subtasks experimental sentences overall CER.

as a crucial component of the AWP approach and filter candidate words during the candidate word search phase.

G. APPLICATION OF CHINESE SHALLOW FUSION AND AWP METHODS IN MULTI-DOMAIN TASKS

To validate the universality and effectiveness of the AWP method across various subtasks within the CATSLU specialized term context biasing task, experiments were conducted.

We employed three approaches for decoding: no biasing, solely using the Chinese shallow fusion method, and combining the Chinese shallow fusion method with the AWP method. Comparisons were made among these approaches based on various metrics across all subtasks. The ASR model utilized in this experiment was a Conformer model trained on the extensive WenetSpeech Chinese speech dataset, which demonstrates robust generalization and recognition capabilities. The experimental outcomes for different subtasks are displayed in Figures 21 and 22.

Based on the data presented in the graphs, it's evident that the method combining Chinese shallow fusion with AWP outperforms both no-biasing and sole Chinese shallow fusion across all subtasks in terms of CER and KER metrics. This underscores the efficacy of our proposed Chinese alternative word prediction method in rectifying errors associated with specialized terms, further enhancing the identification capacity of specialized terms based on the Chinese shallow fusion approach.

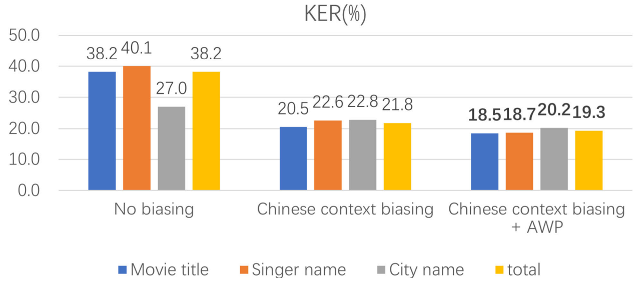


FIGURE 22. CATSLU proper noun context biasing all subtask experimental keyword error rate KER.

VI. CONCLUSION AND FUTURE WORK

This section is divided into two subsections, presenting the conclusion and future work respectively.

A. CONCLUSION

This paper defines two Chinese context biasing tasks based on the open-source CATSLU dataset: one targeting proper nouns and the other focusing on context biasing issues in mixed-domain corpora. Four different stages of context biasing methods were employed to bias the end-to-end ASR model. Beyond the existing shallow fusion method, this paper introduces a Chinese alternative word prediction model to generate a list of alternatives and perform post-processing error correction, thus further enhancing the effectiveness of context biasing.

By adapting previous research's methods of defining context biasing tasks on English open-source datasets, this study defines two Chinese context biasing tasks using the CATSLU dataset. One is the CATSLU proper noun context biasing task, evaluating the effectiveness of context biasing methods targeting proper nouns. The other is the CATSLU mixed-domain context biasing task, assessing how context biasing methods enhance the performance of end-to-end ASR models across various domains.

This paper delves into four different stages of context biasing methods, encompassing pre-recognition, modeling, decoding, and post-processing stages. In the pre-recognition stage, a self-supervised pre-trained model and a bidirectional LSTM-based end-to-end context recognition model were employed to automatically detect the domain category of the input sentence. During the modeling stage, a small amount of target domain data was used to fine-tune the end-to-end model, enhancing its performance in a specific domain. In the decoding stage, a Chinese shallow fusion method using keyword reward mechanisms prioritized decoding for specific domain keywords. In the post-processing stage, a Transformer-based alternative word prediction model was utilized to generate alternative word lists and perform error correction.

From the experimental results, it's evident that the context detection method reduced the error rate (ERR) in CER by 10.31%. Fine-tuning the original model with corpora reduced the CER error rate by 76.75%. The use of Chinese shallow fusion resulted in a 21.58% reduction in CER errors and a 42.93% reduction in KER errors. Furthermore, employing the alternative word prediction method on top of the Chinese shallow fusion achieved a 23.02% reduction in CER errors and a 50.52% reduction in KER errors. These results indicate the effectiveness of context biasing methods in all four directions.

B. FUTURE WORK

The primary challenge faced in this study is the lack of adequately large-scale open-source datasets for performing context biasing tasks across different domains and scenarios comprehensively. Due to limited coverage in open-source datasets, it's challenging to conduct a thorough evaluation and analysis of various possible context biasing scenarios. Therefore, future work will focus on leveraging large-scale open-source datasets like WenetSpeech to establish representative and challenging context biasing tasks. This entails adopting reasonable and effective selection criteria to choose sentences that meet specific contextual requirements, aiming for an objective and comprehensive assessment and comparison of different types of context biasing methods.

Regarding the alternative word prediction method, there's room for further improvement and expansion in this research. For instance, considering incorporating the predicted alternative words into the shallow fusion strategy during the generation process of the biased word list to enhance its diversity and richness. Additionally, exploring the impact of different domain corpora on the training effectiveness of alternative word prediction models and comparing the performance of different domain models in various scenarios could be beneficial.

In terms of alternative word search methods, there are possibilities for enhancement and optimization in this study. For example, transforming the alternative word list into a Finite State Transducer (FST) and utilizing FST to match whether the sentence contains the target alternative words. As real-world scenarios might involve numerous keywords and sentences, there's a need to find more efficient and accurate search and replacement algorithms.

For context biasing methods combined with end-to-end intent recognition, an avenue to explore could be integrating Prompt input via whisper. This involves providing contextual cues while recognizing the input speech, aiding the whisper model in better discernment by incorporating potential contextual information.

REFERENCES

[1] R. Graham, L. Aldridge, K. Carter, and T. C. Lansdow, "The design of in-car speech recognition interfaces for usability and user acceptance," in *Proc. 2nd Int. Conf. Eng. Psychol. Cogn. Ergonom.*, 1998.

[2] P. N. Garner, J. Dines, T. Hain, A. E. Hannani, M. Karafiát, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang, "Real-time ASR from meetings," in *Proc. Interspeech*, Sep. 2009, pp. 1–13.

[3] M. Ravanelli, T. Parcollet, and Y. Bengio, "The Pytorch–Kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6465–6469.

[4] P. Swietojanski, A. Ghoshal, and S. Renals, "Revisiting hybrid and GMM-HMM system combination techniques," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6744–6748.

[5] P. Dighe, A. Asaei, and H. Bourlard, "On quantifying the quality of acoustic models in hybrid DNN-HMM ASR," *Speech Commun.*, vol. 119, pp. 24–35, May 2020.

[6] N. Jung, G. Kim, and J. S. Chung, "Spell my name: Keyword boosted speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6642–6646.

[7] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," 2018, *arXiv:1804.00015*.

[8] S. Husnjak, D. Perakovic, and I. Jovovic, "Possibilities of using speech recognition systems of smart terminal devices in traffic environment," *Proc. Eng.*, vol. 69, no. 1, pp. 778–787, Jun. 2014.

[9] J. Drexler Fox and N. Delworth, "Improving contextual recognition of rare words with an alternate spelling prediction model," 2022, *arXiv:2209.01250*.

[10] S. Zhu, Z. Zhao, T. Zhao, C. Zong, and K. Yu, "CATSLU: The 1st Chinese audio-textual spoken language understanding challenge," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 521–525.

[11] W. Ronny Huang, C. Peyser, T. N. Sainath, R. Pang, T. Strohman, and S. Kumar, "Sentence-select: Large-scale language model data selection for rare-word speech recognition," 2022, *arXiv:2203.05008*.

[12] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1–5828.

[13] D. Zhao, T. N. Sainath, D. Rybach, P. Rondon, D. Bhatia, B. Li, and R. Pang, "Shallow-fusion end-to-end contextual biasing," in *Proc. Interspeech*, Sep. 2019, pp. 1418–1422.

[14] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu, "WeNet 2.0: More productive end-to-end speech recognition toolkit," 2022, *arXiv:2203.15455*.

[15] G. Pundak, T. N. Sainath, R. Prabhavalkar, A. Kannan, and D. Zhao, "Deep context: End-to-end contextual speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 418–425.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1–12.

[17] A. Gourav, L. Liu, A. Gandhe, Y. Gu, G. Lan, X. Huang, S. Kalmane, G. Tiwari, D. Filimonov, A. Rastrow, A. Stolcke, and I. Bulyko, "Personalization strategies for end-to-end speech recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7348–7352.

[18] M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Zelasko, and M. Jette, "Earnings-21: A practical benchmark for ASR in the wild," 2021, *arXiv:2104.11348*.

[19] B. Zhang, H. Lv, P. Guo, Q. Shao, C. Yang, L. Xie, X. Xu, H. Bu, X. Chen, C. Zeng, D. Wu, and Z. Peng, "WENETSPEECH: A 10000+ hours multi-domain Mandarin corpus for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6182–6186.

[20] 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. 20th Conf. Oriental Chapter Int. Coordinating Committee Speech Databases Speech I/O Syst. Assessment*, Nov. 2017, pp. 1–5.

[21] J. Du, X. Na, X. Liu, and H. Bu, "AISHELL-2: Transforming Mandarin ASR research into industrial scale," 2018, *arXiv:1808.10583*.

[22] Z. Yao, D. Wu, X. Wang, B. Zhang, F. Yu, C. Yang, Z. Peng, X. Chen, L. Xie, and X. Lei, "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," 2021, *arXiv:2102.01547*.

[23] L. Borgholt, J. D. Havtorn, M. Abdou, J. Edin, L. Maaløe, A. Søgaard, and C. Igel, "Do we still need automatic speech recognition for spoken language understanding?" 2021, *arXiv:2111.14842*.

[24] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, no. 1, pp. 3451–3460, Jul. 2021.

[25] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-T. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech processing universal performance benchmark," 2021, *arXiv:2105.01051*.

[26] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Comput. Speech Lang.*, vol. 16, no. 1, pp. 69–88, Jan. 2002.

[27] Fxsjy. (2020). *Jieba*. [Online]. Available: https://github.com/fxsjy/jieba

[28] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," 2017, *arXiv:1701.02810*.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017, *arXiv:1412.6980*.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019, *arXiv:1810.04805*.

[31] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.

KAI ZHANG received the bachelor's degree from Providence University, in 2021, and the master's degree from National Taiwan University, in 2023. His research interests include speech recognition, intent recognition, large language models, and large speech models.

QIUXIA ZHANG received the B.S. degree in computer science and information engineering from the National Taiwan University of Science and Technology, in 2021, and the M.S. degree in computer science and information engineering from National Taiwan University, in 2023. From 2021 to 2023, she has published three articles. Her research interests include few-shot learning, slot filling, named entity recognition, and intent recognition.

CHUNG-CHE WANG received the Ph.D. degree from the CS Department, National Tsing Hua University, Hsinchu, Taiwan, in 2017. His research interests include audio retrieval and audio processing.

JYH-SHING ROGER JANG (Member, IEEE) received the Ph.D. degree in electrical engineering and computer sciences from the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, CA, USA, in 1992. He studied fuzzy logic and artificial neural networks with Prof. Lotfi Zadeh, the father of fuzzy logic. After the Ph.D. degree, he joined MathWorks to co-author the Fuzzy Logic Toolbox (for MATLAB). He has since cultivated a keen interest in implementing industrial software for machine learning. From 1995 to 2012, he was a Professor with the Computer Science Department, National Tsing Hua University, Hsinchu, Taiwan. Since August 2012, he has been a Professor with the Department of Computer Science and Information Engineering, National Taiwan University (NTU), Taipei, Taiwan. He was the IT Director of NTU Hospital, from 2017 to 2019, and the Director of the FinTech Center, NTU, from 2018 to 2022. He is currently the CTO of E.SUN Financial Holding Company, Taipei. He has authored or co-authored one book titled *Neuro-Fuzzy and Soft Computing* (Prentice Hall, 1997). He has also maintained toolboxes for machine learning and speech/audio processing. His research interests include machine learning in practice, with wide applications to speech recognition/assessment/synthesis, music analysis/retrieval, image classification, medical/healthcare data analytics, and FinTech. As of November 2022, Google Scholar shows more than 19 000 citations for his seminal article on adaptive neuro-fuzzy inference systems (ANFIS) published in 1993. He was the General Chair of the International Society for Music Information Retrieval (ISMIR) Conference, Taipei, 2014, and the General Co-Chair of the ISMIR Conference, Suzhou, 2017.