An Effective Contextualized Automatic Speech Recognition Approach Leveraging Self-Supervised Phoneme Features

Li-Ting Pai[1], Yi-Cheng Wang[1], Bi-Cheng Yan[1], Hsin-Wei Wang[1], Jia-Liang Lu[1],
Chi-Han Lin[2], Juan-Wei Xu[2], Berlin Chen[1]
Department of Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
E.SUN Financial Holding Co., Ltd
E-mail: {61147095s, yichengwang, 80847001s, hsinweiwang, jialianglu, berlin}@ntnu.edu.tw;
E-mail: {finalspaceman-19590, weixu-22681}@esunbank.com

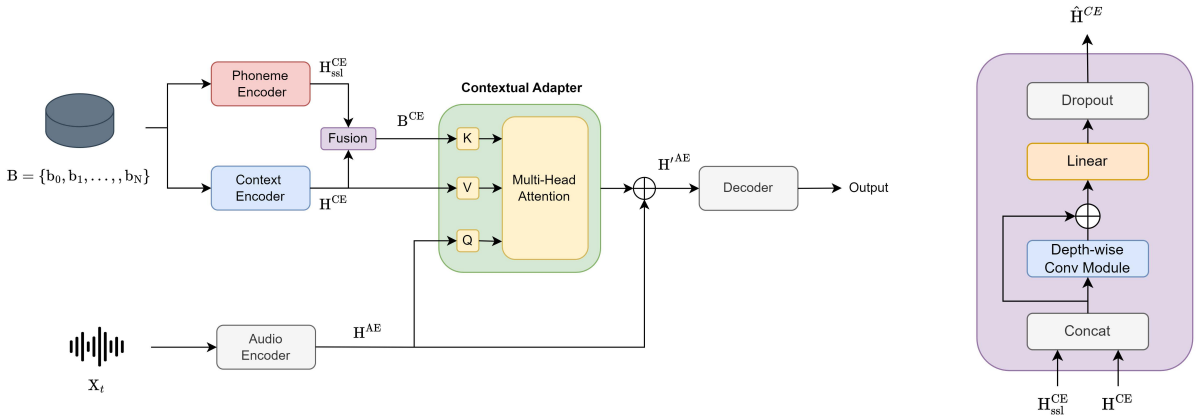
Abstract— **Years of scholarly efforts have led to extensive studies on end-to-end automatic speech recognition (E2E ASR), now demonstrating robust performance in everyday applications such as voice assistants, transcription services, and many others. However, E2E ASR struggles to recognize domain-specific phrases, such as keywords or name entities. To address this, contextualized ASR (CASR) has been developed to improve keyword recognition accuracy by incorporating specific contextual information, represented by a keyword list, into the ASR model. Despite their effectiveness, CASR systems still fall short in distinguishing between keywords with similar sounds, as well as generalizing to uncommon keyword pronunciations. Previous studies have focused primarily on enriching keyword representations by integrating keyword phoneme features derived from a simple sequence encoder with keyword grapheme features to overcome these obstacles. However, such phoneme representations are insufficient, as human pronunciation varies in different contexts, involving phenomena like linking and variations. In this paper, we argue that integration of more fine-grained phoneme features instrumental to accurate keyword recognition in CASR. To this end, we propose leveraging a self-supervised learning (SSL) phoneme encoder to provide more subtle phonemic details of keywords, effectively addressing these variations and alleviating the phonetic confusion between keywords. A series of experiments conducted on the SlideSpeech benchmark dataset demonstrates the effectiveness of our approach in alleviating keyword phonemic confusion and enhancing out-of-domain keyword recognition.**

I. INTRODUCTION

End-to-end (E2E) automatic speech recognition (ASR) [1] has garnered significant interest from both research communities and industrial applications in recent years due to its streamlined design and scalability compared to conventional hybrid DNN-HMM models. Although E2E ASR models demonstrate impressive recognition abilities for common words, they struggle with rare words, including keywords like personal names [2]. This difficulty arises from the imbalance of word distributions in the training set. Accurate keyword recognition is crucial for downstream tasks like natural language processing, making the improvement of keyword recognition accuracy essential for advancing E2E ASR models in real-world applications.

In contrast to machines, humans, with their rich sensory inputs, can easily recognize unfamiliar phrases by leveraging various clues from multiple modalities such as visual cues, body language, and environmental sounds. This ability enables them to adapt swiftly to different scenarios, including keynote speeches, financial discussions, and medical consultations. However, E2E ASR models rely solely on speech signals and lack these additional perceptual inputs, making it challenging for them to recognize domain-specific phrases. To address this limitation, researchers are developing Contextualized ASR (CASR) systems [3] that incorporate knowledge beyond speech signals. This approach enhances the understanding capabilities of ASR models and adapts them to diverse contexts, leading to more accurate and reliable speech recognition.

CASR research can be broadly categorized into three groups: shallow fusion, deep biasing, and prompting. 1) Shallow fusion [3, 4, 5] uses weighted finite state transducers (WFST) to combine the output of external *N*-gram language models with the output of the ASR model. While this can improve ASR performance, it requires careful tuning of the fusion weights for optimal results. 2) Deep biasing [6, 7, 8, 9] employs cross-attention [10] mechanisms to integrate keyword representations into the hidden layers of ASR models, effectively improving the recognition of domain-specific terms. 3) Prompting [11, 12, 13] involves providing contextual information as an input sequence to the ASR model. Among these CASR methods, the contextual adapter (CA) [14] stands out for its simplicity and effectiveness. It integrates a list of pre-defined keywords into the ASR model using a cross-attention mechanism. Notably, CA employs an adapter-style training method, necessitating the tuning of only about 3% of the model weights. This approach offers a



(a) Attention-based encoder-decoder (AED) model with contextual adapter (b) Fusion mechanism

Fig 1. (a) The architecture of phoneme-aware contextual adapter with attention-encoder-decoder architecture. (b) Fusion mechanism of the context feature and SSL feature.

significant advantage over other CASR methods. Despite their effectiveness, CA models still face challenges in distinguishing between keywords with similar sounds and generalizing to uncommon keyword pronunciations, especially in scenarios with limited resources.

Previous studies [15, 16, 17] have primarily focused on enriching keyword representations by integrating keyword phoneme features derived from a simple sequence encoder with keyword grapheme features to overcome these obstacles. However, such phoneme representations are insufficient due to limited keyword variation in the training corpus and the fact that human pronunciation varies in different contexts, involving linking and variations. For example, in English, the pronunciation of “black cab” may result in a linking sound, making it sound like “bla-cab,” while “red apple” may sound like “re-dapple.”

To address these challenges, recent advancements in self-supervised learning (SSL) offer promising solutions, particularly for scenarios with limited resources. SSL involves pre-training a model on large amounts of unlabeled data to obtain rich feature representations. This method does not rely on a large amount of manually labeled data but instead uses the intrinsic structure within the data to learn useful representations.

In the field of CASR, research on enhancing contextual adapters through phoneme-aware encoding using self-supervised learning is limited. In this paper, we argue that incorporating more fine-grained phoneme features is necessary to improve keyword recognition. We propose utilizing a self-supervised learning (SSL) phoneme encoder to provide detailed phoneme information, which can effectively address variations and reduce phonetic confusion between keywords. This approach aims to enhance the accuracy and reliability of CASR systems in recognizing domain-specific terms and uncommon keyword pronunciations.

In summary, our contributions are at least three-fold:

- **SSL Phoneme Encoder Integration:** We introduce a self-supervised learning (SSL) phoneme encoder to enhance keyword phoneme representation in CASR systems, capturing fine-grained phonemic details to improve recognition accuracy.
- **Phonetic Confusion Reduction:** Our approach significantly reduces phonetic confusion between keywords with similar sounds, leveraging detailed phonemic features for more precise keyword recognition.
- **Enhanced Out-of-Domain Generalization:** Experiments on the SlideSpeech benchmark dataset show the effectiveness of our methods in improving both in-domain and out-of-domain keyword recognition, demonstrating robustness and versatility.

II. Related Work

Recently, contextual adapter (CA) [14] has demonstrated considerable success; however, its performance often declines when phoneme confusion increases in rare word lists. To address this issue, a phoneme-based encoding method was introduced to enhance the recognition of words with irregular pronunciations [15, 16, 17]. Despite its benefits, this method still lacks context consideration when encoding phonemes. To overcome this limitation, we propose combining detailed phoneme-aware features generated by the self-supervised learning (SSL) model XPhoneBERT [18].

XPhoneBERT was originally developed to enhance text-to-speech (TTS) tasks by learning robust phoneme representations. We propose utilizing XPhoneBERT to CASR models. XPhoneBERT is a pioneering multilingual model with a BERT-base architecture [19], trained using the masked prediction objective on a dataset of 330 million phoneme-level sentences from nearly 100 languages. This extensive training enables it to

Table 1. In-domain testing: WER, K-WER and NK-WER results for History and Computer Science test datasets.

Domain	Model	WER	K-WER	NK-WER
Computer Science	Conformer	13.75	13.21	**13.90**
	+ CA	14.80	14.41	14.90
	+ SSL feature	**13.61**	**9.83**	14.64
History	Conformer	**13.90**	15.18	**13.66**
	+ CA	14.86	16.57	14.55
	+ SSL feature	14.80	**14.38**	14.86

generate high-quality phoneme representations across various languages. Key features of XPhoneBERT include multilingual capability and a focus on phoneme-level data, capturing fine-grained phonetic nuances crucial for both TTS and ASR tasks. We adapt XPhoneBERT for contextualized ASR models by incorporating it as a phoneme encoder.

III. Proposed Method

Before we delve into our proposed methods, we will first describe the architecture of our backbone E2E ASR model, which is an attention-based encoder-decoder (AED) network. Next, we introduce the contextual adapter (CA) methods. Finally, we elaborate on our proposed SSL phoneme features enriched CA approach.

A. Backbone E2E ASR Model

An AED model generally consists of an encoder network and a decoder network, as shown in Figure 1(a) in the gray part. Given an audio signal $\mathrm{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ of length T, represent the acoustic feature vectors, and $\mathbf{y} = (y_1, y_2, \dots, y_M)$ of length M be the corresponding subword sequence. The encoder network, $\mathrm{Enc_{AE}}(\cdot)$, processes a sequence of acoustic feature vectors, X, and produces a sequence of high-level acoustic representations, $\mathrm{H^{AE}} = \mathrm{Enc_{AE}}(\mathbf{x}_t)$. The decoder network, $\mathrm{Dec}(\cdot)$, then fuses these acoustic embeddings $\mathrm{H^{AE}}$ with the previously decoded text tokens $\mathbf{y}_{1:m} = (y_1, y_2, \dots, y_m)$ using a cross-attention mechanism. The output probability of a possible upcoming token $\mathbf{y}_{m+1}$ can be derived by

$$\mathbf{y}_{m+1} = \mathrm{Dec}(\mathrm{H^{AE}}, \mathbf{y}_{1:m}). \tag{1}$$

B. Contextual Adapter

Contextual adaptation (CA) is a mechanism designed to enhance the recognition accuracy of ASR models for rare or context-specific words by incorporating additional contextual information into the ASR model. CA introduces two additional components to the original ASR model: a context encoder and an attention-based adapter.

The Context Encoder, denoted as $\mathrm{Enc_{CE}}(\cdot)$, processes a keyword list, consist of N keywords $\mathrm{B} = \{\mathbf{b}_0, \mathbf{b}_1, \dots, \mathbf{b}_N\}$ into the contextual representations $\mathrm{H^{CE}} = \mathrm{Enc_{CE}}(\mathrm{B})$.

The attention-based adapter then integrates these contextual embeddings $\mathrm{H^{CE}}$ into the ASR model using a multi-head cross-attention mechanism. The attention process can be derived as:

$$\mathrm{B^{CE}} = \mathrm{softmax}\left(\frac{\mathrm{H^{AE}W^{Q}(H^{CE}W^{K})^{T}}}{\sqrt{d}}\right)\mathrm{H^{CE}W^{V}}, \tag{2}$$

where $\mathrm{W^{Q}}, \mathrm{W^{K}}, \mathrm{W^{V}}$ are trainable weight matrices, and d is the dimensionality of the embeddings. This biasing matrix $\mathrm{B^{CE}}$ is then used to contextualize the original acoustic embeddings, resulting in $\acute{\mathrm{H}}^{\mathrm{AE}} = \mathrm{H^{AE}} + \mathrm{B^{CE}}$. By incorporating the contextual information in this manner, the ASR model can more accurately recognize and transcribe rare words that are crucial for understanding the input audio sequence.

C. SSL Phoneme Features Enriched Contextual Adapter

We leverage phoneme features extracted from the SSL phoneme encoder, XPhoneBERT, alongside the context embedding $\mathrm{H^{CE}}$ generated by the context encoder to produce phoneme-aware context embeddings, as illustrated in Fig. 1(a).

The phoneme encoder, denoted as $\mathrm{Enc_{pho}}(\cdot)$, extracts SSL features represented as $\mathrm{H^{CE}_{ssl}} = \mathrm{Enc_{pho}}(\mathrm{B})$. We then concatenate these phoneme features $\mathrm{H^{CE}_{ssl}}$ with the context features $\mathrm{H^{CE}}$ to form a combined feature representation $\mathrm{H^{CE}_{cat}} = \mathrm{Concat}(\mathrm{H^{CE}_{ssl}}, \mathrm{H^{CE}})$.

Next, we apply a depth-wise convolution to this combined representation to capture information from adjacent dimensions, as investigate in [20], resulting in $\mathrm{H^{CE}_{dw}} = \mathrm{DepthwiseConv}(\mathrm{H^{CE}_{cat}})$. The output of this convolution is summed with the original concatenated features, followed by a linear projection to produce the final phoneme-aware context embedding $\widehat{\mathrm{H}}^{\mathrm{CE}} = (\mathrm{H^{CE}_{dw}} + \mathrm{H^{CE}_{cat}})\mathrm{W}$, as shown in Figure 1(b). These phoneme-aware context embeddings are subsequently used to generate the key embedding of the attention-based adapter, which is jointly optimized with the ASR model.

IV. Experimental Setup

A. Dataset and Evaluation Metrics

The SlideSpeech corpus [21] is a comprehensive audio-visual dataset containing over 1,000 hours of slide presentations. It includes real-time synchronized slides, pre-processed Optical

Table 2. Out of domain generalization: WER, K-WER and NK-WER results for History and Computer Science test datasets.

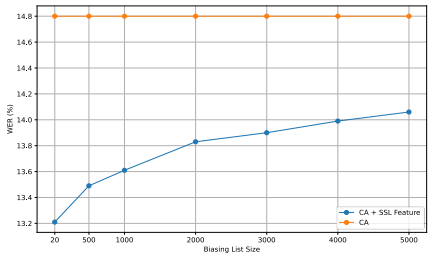
Training Domain	Testing Domain	Model	WER	K-WER	NK-WER
History	Computer Science	Conformer	16.59	16.15	**16.65**
		+ CA	17.48	16.95	17.57
		+ SSL feature	**16.45**	**13.79**	16.87
Computer Science	History	Conformer	**21.06**	**19.75**	**21.28**
		+ CA	22.42	21.32	22.60
		+ SSL feature	22.22	21.92	22.27

Character Recognition (OCR) results, and extracted keywords corresponding to the slides, making it an ideal dataset for evaluating the CASR task. SlideSpeech spans 22 diverse domain categories, with Computer Science and History being the largest. Therefore, our experiments focused on these two domains, as they are well-represented in the corpus. For the Computer Science dataset, we divided it into a training set with 109 hours, a development set with 20 hours, and a test set with 32 hours. Similarly, for the History dataset, the training set contained 173 hours, the development set 20 hours, and the test set 47 hours.

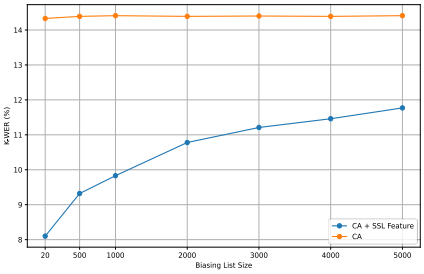
During training, we dynamically generated a keyword list that included keywords randomly selected from multiple utterances within the batch as distractors, with the total number of distractors set to 200 for all experiments. We adapted the keywords provided by the SlideSpeech corpus. Since training with the entire keyword list is memory-intensive, we dynamically generated a subset of the keyword list during training. This subset included keywords randomly selected from the full list, the reference keywords in the utterance, and a special token <OOV>, which serves as a fallback when no relevant context words are available. During testing, the keyword lists consist of keywords from the target utterance along with distractors, with the total number of distractors set to 1000. We evaluate contextual biasing performance using word error rate (WER), keyword word error rate (K-WER), and non-keyword word error rate (NK-WER). K-WER is calculated as the number of incorrect keywords in the keyword list divided by the total number of keywords in the test set. NK-WER is the number of incorrect non-keywords in the keyword list divided by the total number of non-keywords in the test set.

B. Baseline Models

We used a Conformer-Transformer model as the backbone for long-tailed speech recognition experiments. The network consists of a Conformer encoder [22] and a Transformer decoder [23] (denoted by Conformer for short). The Conformer encoder consists of 12 blocks, each with 2,048 hidden units and 8 attention heads. The Trans-former decoder network includes 6 blocks, each also with 2,048 hidden units. Our proposed method is compared against the iconic contextual adapter (CA) method, which serves as a strong baseline in this study. The contextual adapter is equipped with a cross-attention layer having 8 heads, each with a size of 64. The input features are audio features from the Slidepeech dataset, with each audio segment converted into 80-dimensional Mel spectrograms and normalized at the frame level. During training, we use the Adam optimizer with an initial learning rate set to 0.001, employing a learning rate decay strategy. The batch size is set to 64, and the model is trained for 35 epochs.



(a) WER

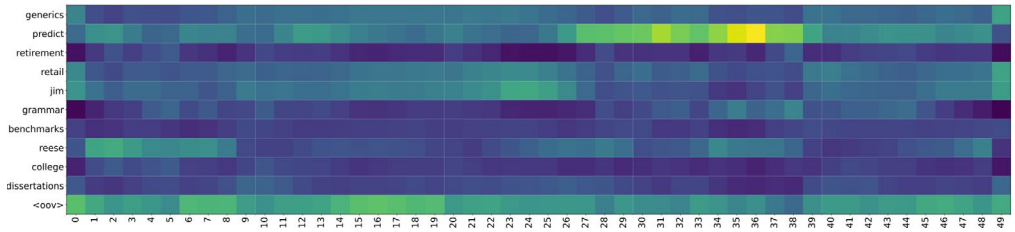


(b) K-WER

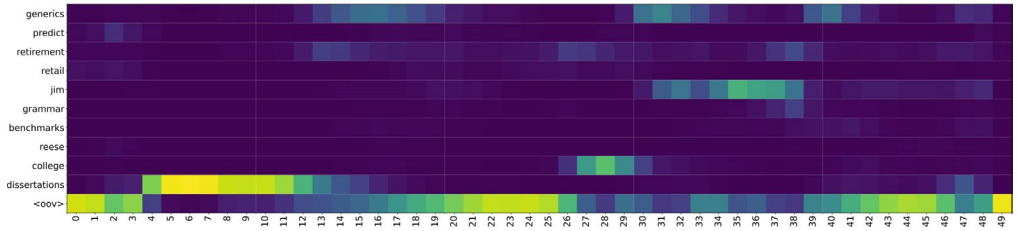
Fig 2. The impact of WER and K-WER under different keyword list sizes.

The components of the Contextual Adapter are as follows: The phoneme encoder uses XPhoneBERT[1], which converts phoneme sequences into 768-dimensional phonetic representations. The encoder includes an embedding layer that maps input tokens to 512-dimensional embeddings, an OOV embedding

[1] https://huggingface.co/vinai/xphonebert-base



(a) Without SSL feature



(b) With SSL features

Fig 3. Visualization of the attention scores within the adapter, "dissertations" and "college" are the target keywords, in the exemplar utterance: "dissertations and college and alumni magazines." In (a), the contextual adapter is easily confused by phonetically similar keywords. In contrast, (b) demonstrates that our SSL feature-enriched contextual adapter effectively distinguishes keywords from other distractors.

layer that converts 512-dimensional embeddings to a single dimension to handle unknown tokens, and a padding embedding layer similar to the OOV embedding but used for padding tokens. The RNN encoder is composed of a bidirectional LSTM with two layers, each with 512 hidden units, where the final LSTM output is projected to 512 dimensions. A dropout layer with a rate of 0.1 is applied to prevent overfitting, and layer normalization is applied to intermediate features. Depth-wise convolution fusion is performed with a 1D convolutional layer with 1280 groups for channel feature fusion, followed by a projection layer that projects the fused features to 512 dimensions. The adapter includes a custom multi-headed attention mechanism, which consists of linear layers for query, key, and value transformations, each outputting 512 dimensions. A dropout rate of 0.1 is applied to the attention mechanism, and a projection layer projects the attention outputs back to 512 dimensions.

To encourage further research, we will make our code publicly available for benchmarking and replication experiments.

V. Experimental Results

We begin our experiments by evaluating the performance on the in-domain dataset, comparing baseline Conformer models with and without contextual adapter (CA) and Self-Supervised Learning (SSL) features. Next, we assess the out-of-domain generalization adaptability of the CASR model by training and testing on different domain datasets. We then analyze how variations in the size of the keyword list affect the WER and K-WER. Finally, we visualize attention maps to understand the alignment between audio features and keywords, both with and without SSL features.

A. Main Results

Table 1 presents the in-domain testing results for Computer Science and History subset. The baseline Conformer model shows decent performance, but incorporating the CA results in a slight increase in error rates. In contrast, adding SSL features significantly improves performance, particularly in reducing the K-WER. These results indicate that while CA alone may not reduce errors, combining CA with SSL features enhances performance, especially in recognizing keywords, thus improving the overall accuracy of the ASR model in the Computer Science and History domain.

B. Cross-domain Generalizations

Table 3 highlights significant differences in performance across cross-domain datasets with various model configurations. When training on History domain dataset and testing on Computer Science domain dataset, the baseline model shows higher error rates, which increase further with CA. However, adding SSL features reduces these rates. Similarly, when training on Computer Science data and testing on History domain dataset, the baseline model shows high error rates, which increase with CA and decrease slightly with SSL features. These results suggest that while contextual adapters and SSL features can improve performance on cross-domain datasets, they may also lead to higher error rates in some cases.

C. Impact of Different Size of the Keyword list

Figure 2(a) shows that the WER decreases significantly as the size of the keyword list decreases. For instance, the WER is

lower with a smaller keyword list, highlighting the effectiveness of our method in accurately handling rare words. Figure 2(b) illustrates that the K-WER increases as the size of the keyword list grows. This suggests that while larger keyword lists provide more context, they also introduce complexity, negatively impacting keyword recognition performance. Therefore, a balanced keyword list size is crucial for optimizing the trade-off between context and complexity, enhancing the accuracy of specific keywords.

D. Qualitative Analysis

Figure 3 visually depicts the corresponding results. On the X-axis stands for time stamps while the Y-axis displays keywords, with "dissertations," "collect" being the target keywords. Brighter pixels in the attention map signify higher attention weights assigned by the model to each rare word at a given time stamp. The figure illustrates the attention relationship between audio features and keywords, demonstrating the effectiveness of our method in integrating phoneme and contextual information. Without SSL features, the bias word does not effectively align with the frames. In contrast, with SSL features, the bias word aligns well with the frames, indicating better model convergence.

VI. CONCLUSIONS

In this paper, we presented a novel method to enhance the performance of ASR systems by integrating fine-grained phoneme SSL features into a contextual adapter. Our approach leverages both phoneme and contextual information to improve the accuracy of recognizing context-specific words. By incorporating SSL models to extract phoneme representations from the keyword list, we provided dual representations that include both phonetic and grapheme information. This method significantly improves keyword recognition accuracy by comprehensively utilizing external knowledge and contextual information. Furthermore, by employing SSL models to extract phoneme features, we enhanced the accuracy of recognizing keywords. In future work, we will explore applying our approach to a wider range of datasets, investigate real-time ASR applications.

REFERENCES

[1] J. K. Chorowski et al., "Attention-based models for speech recognition," in *Proc. Neural Information Processing Systems*, vol. 28, 2015.

[2] D. Bahdanau et al., "End-to-end attention-based large vocabulary speech recognition," in *Proc. ICASSP. IEEE*, 2016, pp. 4945-4949.

[3] I. Williams et al., "Contextual speech recognition in end-to-end neural network systems using beam search," in *Proc. Interspeech*, 2018, pp. 2227-2231.

[4] D. Zhao et al., "Shallow-fusion end-to-end contextual biasing," in *Proc. Interspeech*, 2019, pp. 1418-1422.

[5] S. Kim et al., "Improved neural language model fusion for streaming recurrent neural network transducer." in *Proc. ICASSP*, 2021, pp. 7333-7337.

[6] G. Pundak et al., "Deep context: end-to-end contextual speech recognition," in *Proc. IEEE SLT*, 2018, pp. 418-425.

[7] M. Han et al., "Improving end-to-end contextual speech recognition with fine-grained contextual knowledge selection," in *Proc. ICASSP*, 2022.

[8] D. Le et al., "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," *arXiv preprint arXiv:2104.02194*, 2021.

[9] F. J. Chang et al., "Context-aware transformer transducer for speech recognition," in *Proc. IEEE ASRU*, 2021, pp. 503-510.

[10] A. Vaswani et al., "Attention is all you need." in *Proc. NeurIPS*, 2017.

[11] X. Yang et al., "PromptASR for contextualized ASR with controllable style," in *Proc. ICASSP. IEEE*, 2024, pp. 10536-10540.

[12] G. Yang et al., "MaLa-ASR: Multimedia-Assisted LLM-Based ASR," *arXiv preprint arXiv:2406.05839*, 2024.

[13] Li, Yuang, et al. "Using Large Language Model for End-to-End Chinese ASR and NER." *arXiv preprint arxiv:2401.11382*, 2024.

[14] K. M. Sathyendra et al., "Contextual adapters for personalized speech recognition in neural transducers," in *Proc. ICASSP. IEEE*, 2022, pp. 8537-8541.

[15] R. Pandey et al., "PROCTER: pronunciation-aware contextual adapter for personalized speech recognition in neural transducers," in *Proc. ICASSP. IEEE*, 2023, pp. 1-5.

[16] Z. Chen et al., "Joint grapheme and phoneme embeddings for contextual end-to-end ASR," in *Proc. Interspeech*, 2019, pp. 3490-3494.

[17] A. Bruguier et al., "Phoebe: Pronunciation-aware contextualization for end-to-end speech recognition," in *Proc. ICASSP*, 2019, pp. 6171-6175.

[18] L. T. Nguyen et al., "XPhoneBERT: A pre-trained multilingual model for phoneme representations for text-to-speech," in *Proc. Interspeech*, 2023.

[19] J. Devlin et al., "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171-4186.

[20] C. Si et al., "Inception Transformer," in *Proc. NeurIPS*, vol. 30, 2022, pp. 23495-23509.

[21] H. Wang et al., "SlideSpeech: A large scale slide-enriched audio-visual corpus," in *Proc. ICASSP. IEEE*, 2024, pp. 11076-11080.

[22] A. Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, vol. 30, 2017.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.