



GLCLAP: A Novel Contrastive Learning Pre-trained Model for Contextual Biasing in ASR

Yuxiang Kong*, Fan Cui*, Liyong Guo, Heinrich Dinkel, Lichun Fan, Junbo Zhang, Jian Luan

MiLM Plus, Xiaomi Inc., China

kongyuxiang1, cuifan, guoliyong, dinkelheinrich, fanlichun1, zhangjunbo1,
luanjian@xiaomi.com

Abstract

Recently, Automatic Speech Recognition (ASR) that supports prompts has shown remarkable versatility. For contextual biasing with these systems, a pivotal factor lies in obtaining well-matched prompts. To address this issue, Contrastive Language-Audio Pre-training is exploited to retrieve matched entities from a user-specified list. Instead of only confining contrastive learning to the sentence level, we propose the Global-Local Contrastive Language-Audio Pre-trained model (GLCLAP). On the global scale, semantic information is extracted from audio and text, enabling a holistic understanding of the input. On the local scale, detailed local word information of individual segments is focused. This multi-scale information has led to a remarkable improvement in bias word retrieval accuracy. By using the GLCLAP bias word retrieval system as the prompts generation component, the accuracy of the final ASR decoding result is significantly improved without finetuning.

Index Terms: Speech Recognition, Contrastive Learning, Contextual biasing

1. Introduction

In recent years, with the help of large-scale datasets, speech recognition models such as Whisper have manifested significant capabilities. Beyond high recognition accuracy, they also support multiple languages and prompt functionality [1, 2]. During the evolution of ASR, the demand for personalized recognition has become particularly pressing. Usually, personalized speech recognition is specifically known as contextual biasing, which entails integrating contextual knowledge into the ASR system [3, 4, 5, 6, 7, 8, 9] to improve recognition accuracy for domain-specific vocabulary. This approach can quickly and effectively enhance recognition performance and meet the needs of specific scenarios.

In traditional contextual biasing ASR solutions, two primary paradigms exist. The first one relies on pronunciation dictionaries, such as the Weighted Finite-State Transducer(WFST) based methods [10, 11, 12, 13]. These systems leverage the pre-defined pronunciation information to improve the accuracy of recognizing specific terms.

The second one involves directly incorporating the biasing mechanism into the architecture of the ASR model. This is achieved by jointly training the biasing mechanism with the ASR model [14], such as SeAco-Paraformer [15] and Contextual Phrase Prediction Network [16].

Neither of these two systems is conducive to handling biasing words in the context of contemporary prompt supported ASR scenarios. For WFST-based system, it is difficult to obtain

pronunciation dictionaries for minority languages or dialects. While, end-to-end contextual biasing typically requires modifications to the ASR model and joint training process, which is not flexible for updating and iterating within the prompt supported model paradigm. It is cumbersome since training large models demands substantial time and computational resources.

The integration of the prompt mechanism in large language models (LLMs) with Retrieval-Augmented Generation (RAG) [17] offers great hints. RAG obtains the desired output by optimizing the prompts without modifying the network structure of the LLM or conducting finetuning. Inspired by this paradigm, biasing prompts generation can be decoupled as an independent module from the recognition process. In this way, the model does not need to rely on pronunciation dictionaries, nor does it need to depend on ASR model during training. This approach aligns seamlessly with the current large-model framework, leveraging the RAG approach for large-scale contextual biasing enhancement.

Contrastive Language-Audio Pre-training (CLAP) [18] is a multimodal pre-trained model based on contrastive learning, designed to learn the joint representation of audio and language through pre-training on an extensive corpus of audio-text pair data. It is capable of extracting embeddings that contain semantic information from both text and audio, laying the foundation for the subsequent calculation of the similarity between audio and text. By obtaining the most matched audio-text pairs, methods can be used to provide bias prompts for subsequent ASR systems. However, CLAP emphasizes the complete matching of the entire audio and text sentences, while biasing words often only correspond to a segment within the audio. Aiming at this issue, we propose Global-Local Contrastive Language-Audio pre-trained model (GLCLAP), which models both global and local information simultaneously, thus augmenting CLAP's applicability to the biasing prompts generation task. Our key contributions are as follows:

- Using audio-language pre-training for user-defined biasing prompt generation.
- Proposing the Global-Local Contrastive Language-Audio pre-trained model (GLCLAP). It extracts audio information at different scales, which can significantly boost the biasing prompt accuracy within the sentences.
- Integrating a GLCLAP based biasing prompt generation component into the ASR model. This corrects decoding results without finetuning.

2. Proposed Method

*These authors contributed equally.

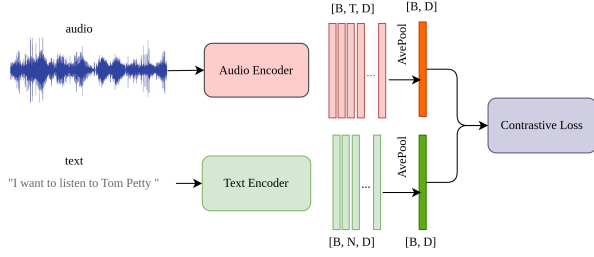


Figure 1: The architecture of the baseline CLAP model.

2.1. Local Subtext Extraction for CLAP

Our main objective is to align the audio embeddings with the embeddings generated from the user-defined biasing list. By calculating the similarity between these embeddings, we can identify the bias that provides the best match. As shown in figure 1, the original CLAP model is designed to capture the semantic information of entire audio and text inputs. However, it is not fully compatible with biasing word retrieval tasks, because biasing word is often a part of the whole sentence. To overcome this limitation, we modify the training process. Specifically, we randomly extract sub-text from the original text annotations. This approach helps to enhance the model’s representation ability for short contexts within the sentences.

For instance, if the original text transcription is "Have you ever heard Taylor Swift’s songs", a randomly selected subtext could be "Taylor Swift". In this way, the model can better learn to handle and retrieve short biasing words during the training process.

2.2. GLCLAP

To enhance the representations from both local and global perspectives, we design the training architecture of the Global-Local Contrastive learning model for the data prepared in Section 2.1, as shown in figure 2.

For the processing of text, besides the original processing approach (referred to as a global branch), we have added a local branch to process the subtext. Let $f_t(\cdot)$ represent the text encoder. The local and global branches share the same weights and are followed by an average pooling layer $p(\cdot)$ to reduce the word dimension. The global branch captures embeddings E^t from the complete texts $X^t \in \mathbb{R}^{B \times N}$, while the local branch focuses on extracting embeddings E_t^t for subtexts $X^{t'} \in \mathbb{R}^{B \times N'}$, where N represents the number of text tokens $N' \leq N$.

$$\begin{aligned} E^t &= p(f_t(X^t)) \\ E_t^t &= p(f_t(X^{t'})) \end{aligned} \quad (1)$$

Both branches produce embeddings E^t, E_t^t of size $[B, D]$, where B denotes the batch size and D represents the embedding dimension.

Similarly, in the audio branch, $f_a(\cdot)$ represents the audio encoder with audio mel spectrogram as input $X^a \in \mathbb{R}^{B \times T \times F}$ where F are the number of Mel bins and T are the number of frames.

When dealing with the audio branch, we perform contrastive learning both before and after average pooling. The reason is to reduce local temporal information loss. The local

audio embedding $E^{a'}$ has a size of $[B, T//4, D]$, where $T//4$ indicates that the audio encoder has performed a 4 times down-sampling. While the global audio embeddings E^a has a size of $[B, D]$, obtained by applying average pooling over the time dimension T on the local audio embedding. The formulas are:

$$\begin{aligned} E^{a'} &= f_a(X^a) \\ E^a &= p(E^{a'}) \end{aligned} \quad (2)$$

Finally, we compute the contrastive loss for both local and global representations of text and audio separately. The global contrastive loss \mathcal{L}_g between the audio and text embeddings:

$$\mathcal{L}_g = l(E^t \cdot E^{aT}) + l(E^a \cdot E^{tT}) \quad (3)$$

And the local max-pooling contrastive loss \mathcal{L}_l is:

$$\mathcal{L}_l = l(\max_t(E_t^t \cdot E^{a'T})) + l(\max_t(E^{a'} \cdot E_t^{t'T})) \quad (4)$$

where \max_t denotes to get the maximum value taken along the time dimension. $l(\cdot) = \frac{1}{B} \sum \log(\text{diag}(\text{softmax}(\cdot)))$, where diag means taking the diagonal elements of the matrix after applying the softmax function. This function is used to measure the similarity between the predicted and target distributions. For training LCLAP, only the \mathcal{L}_l loss is used. The total loss for training GLCLAP is:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_l \quad (5)$$

2.3. GLCLAP for Contextual Biasing ASR

The GLCLAP model is capable of retrieving the best matching biasing words as the prompts to help ASR model accurately recognize rare words that are frequently misidentified. As depicted in Figure 3, the user-defined biasing list $X_1^t, X_2^t \dots X_K^t$ (K is the size of the list) feed into the text encoder to generate the associated text embeddings $E_{1 \dots K}^t$ with a shape of $[K, C]$. For each audio input X_i^a , it is sent to the audio branch without average pooling. In this way, a temporal embedding sequence containing local information can be obtained, marked as $E_i^{a'}$, with a shape of $[T, D]$. Subsequently, a similarity matrix is constructed by calculating the similarity between the audio and text embeddings:

$$\text{Sim} = E_{1 \dots K}^t \cdot E_i^{a'T} \quad (6)$$

The similarity matrix has a shape of $[K, T]$. Then, we perform max-pooling in the temporal domain to obtain a vector of shape $[K]$, which corresponds to the similarity between each text embedding and audio embedding. Ultimately, if there are words whose similarity scores exceed the preset threshold, they will be selected as part of the prompt. These selected words, together with the original audio, are fed into the ASR model to obtain the final recognition results. Our GLCLAP-based retrieval process provides the ASR model with well-matched biasing words, and enhances the overall performance of the ASR system in handling rare and misidentified words.

3. Experiments

3.1. Dataset

Training Datasets To train the GLCLAP model, we utilize four datasets comprising two Chinese and two English datasets.

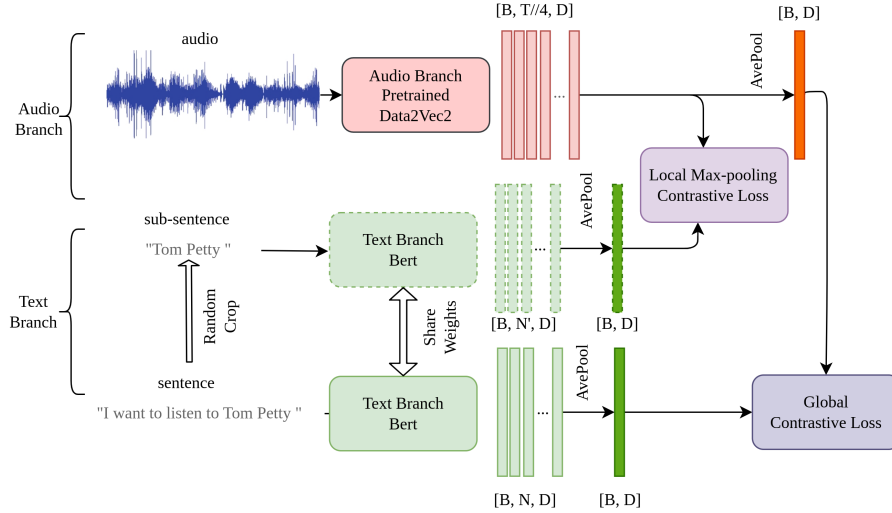


Figure 2: The proposed GLCLAP model.

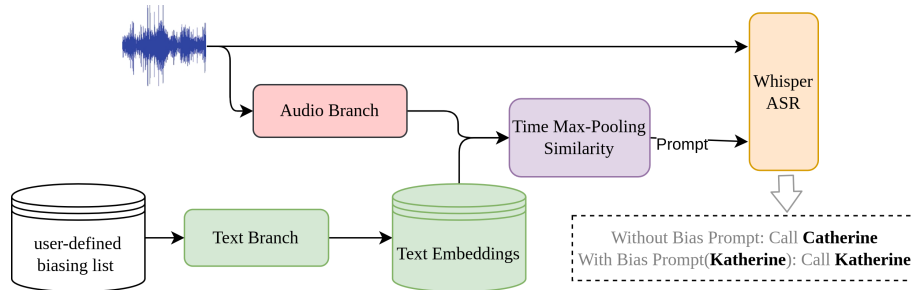


Figure 3: The proposed GLCLAP based context biasing ASR.

Specifically, the Chinese datasets included WenetSpeech [19] and Aishell [20], while the English datasets consisted of Gigaspeech [21] and Librispeech [22].

Test Datasets We utilized the four datasets to evaluate the performance of our contextual biasing ASR system. Specifically, the following test sets were used:

- **Aishell-1 test NT [15]:** This subset contains approximately 808 examples, which are typically used to assess the performance of Chinese named entity retrieval.
- **PhoneCall Dataset:** A private chinese dataset containing queries designed to simulate real-world phone call scenarios is used, with each query including a person’s name to evaluate the effectiveness of personalized speech recognition in user authentication.
- **STOP1:** This dataset is extracted from the STOP [23] test and evaluation datasets. It is filtered to include queries that contain person name information.
- **STOP2¹:** Similar to the STOP1 dataset, this one is filtered to include queries that contain location information.

3.2. Experimental Details

Hyperparameters The training process involves tuning several hyperparameters to achieve optimal performance. The learning

rate is set to $5e-4$, and the batch size is set to 64. The model is trained for 100 epochs, with early stopping activated to prevent overfitting.

Evaluation Metrics For the named entity retrieval, we utilize Top-1 recall and F1 score to assess the hit rate of biasing words. To evaluate the overall improvement of the ASR system, the word error rate (WER) is used as the primary metric.

3.3. Models

The model architecture consists of multiple layers designed to capture both local and global information from the speech and text data. The CLAP and GLCLAP models in the experiment are initialized with the same audio and text encoders:

Audio Encoder We utilize the same structure and pre-training method as Data2Vec2.0-large [24]. Specifically, the Data2VecAudioModel is employed, which is a transformer-based architecture designed for self-supervised learning of speech representations. It is pre-trained with a private dataset, including both English and Chinese data.

Text Encoder The text encoder is initialized with bert-base-multilingual-uncased². It consists of 12 layers of transformer layers, which enables the model to capture contextual information effectively [25].

¹The STOP1 and STOP2 datasets are available at: <https://github.com/GLCLAP/GLCLAP-stop1-stop2-dataset>

²<https://huggingface.co/bert-base-multilingual-uncased>

3.4. Results

Biasing Word Retrieval Evaluation We assess the top-1 recall of the proposed methods on the PhoneCall, STOP1, and STOP2 datasets. The evaluation results are presented in Table 1. The top-1 recall of the base ASR is calculated by matching the biasing list with ASR recognition results. The ASR model is based on the Conformer architecture [26] and has approximately 130M parameters. It is trained on the WenetSpeech, Gigaspeech, Librispeech, Aishell datasets, as well as the 20kh private dataset. Notably, the performance of the base CLAP is comparable to that of the base ASR, and neither of them performs satisfactorily. Significantly, we observe that both the SubText and Global-Local components achieve substantially higher recall values compared to the base model across data of different entity types and with different language. Specifically, the SubText CLAP outperforms the base CLAP by 24.83%, 16.59%, and 30.14%, respectively. The LCLAP shows the most significant improvement, with an increase of 40.85%, 23.29%, and 36.23% on various test sets. Meanwhile, incorporating the Global components further boosts the recall to 97.35%, 88.01%, and 88.81%. Additionally, as shown in Table 2, we compared the F1 scores of LCLAP and GLCLAP with SeACo-Paraformer on the Aishell-1 test NT dataset. LCLAP achieved comparable results, while GLCLAP improved the F1 score by 0.96%.

Table 1: Comparison of the proposed biasing word retrieval methods.

top-1 recall (%)	PhoneCall	STOP1	STOP2
Base ASR	46.52	39.87	37.0
Base CLAP	30.02	45.44	19.4
+ Subtext	54.85	62.03	49.54
+ LCLAP	95.70	85.32	85.77
+ GLCLAP	97.35	88.01	88.81

Table 2: Comparison of LCLAP, GLCLAP, and SeACo-Paraformer.

F1 score (%)	SeACo-Paraformer	LCLAP	GLCLAP
test Aishell-1 NT	96	96	96.96

Multi-Modal Alignment Evaluation To further analyze the relationship between text and audio embeddings, a cosine similarity matrix is calculated from different scales between the two modalities. As shown in Figure 4, the matrix shows higher similarity at the positions where the temporal audio features and the text features correspond. Furthermore, we compared the similarity matrices at different textual granularities such as word and phrase levels. Accurate alignment of the text-audio was achieved at both granularities. It allows the model to better match audio segments with their corresponding named entities. **Contextual biasing ASR evaluation** Finally, the results of the ASR evaluation are summarized in table 3. Compared to the original whisper small model, the application of GLCLAP-based contextual biasing achieves an 16.08% 0.91% 3.31% absolute WER reductions on the PhoneCall, STOP1, STOP2 datasets. These improvements are comparable to the results obtained by providing the Whisper model with ideal contextual biasing words. The relative WER reductions indicate that the integration of GLCLAP biasing words retrieval significantly

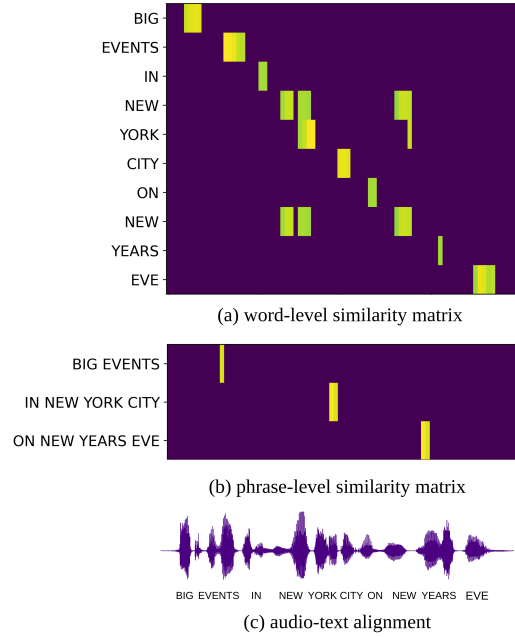


Figure 4: Text-audio similarity matrix at different scales. The similarity values that are 0.5 less than the top-1 similarity value are masked out.

Table 3: Comparison of the contextual biasing ASR systems.

WER(%)	PhoneCall	STOP1	STOP2
Whisper Small	19.96	7.51	11.21
+ideal	3.48	6.21	7.63
+Base CLAP	19.14	7.23	9.93
+Subtext CLAP	10.82	7.01	9.53
+LCLAP	3.96	6.6	8.04
+GLCLAP	3.88	6.6	7.90

enhances the ASR system’s ability to recognize rare words and named entities.

4. Conclusion

In this paper, experiments are conducted to investigate a new contrastive learning pre-trained model for contextual biasing ASR. Compared with traditional methods, contrastive language-audio pre-training is considered to recognize user-defined named entities. Rather than calculating the audio-text similarity at the sentence level, a Global-Local Contrastive Language-Audio pre-trained model is proposed to extract audio semantic information from different scales, which has shown a significant improvement in retrieval accuracy. Additionally, integration of the GLCLAP model into the ASR system provides a dynamic and effective way to enhance recognition accuracy without modifying the original model parameters. The results show that the GLCLAP model is a valuable addition to the ASR system, providing significant improvements in both named entity retrieval and overall ASR performance.

5. References

- [1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [2] G. Yang, Z. Ma, F. Yu, Z. Gao, S. Zhang, and X. Chen, "Ctc-assisted llm-based contextual asr," 2024. [Online]. Available: <https://arxiv.org/abs/2411.06437>
- [3] T. Xu, Z. Yang, K. Huang, P. Guo, A. Zhang, B. Li, C. Chen, C. Li, and L. Xie, "Adaptive contextual biasing for transducer based streaming speech recognition," in *Proc. Interspeech*, 2023, pp. 1668–1672.
- [4] K. M. Sathyendra, T. Muniyappa, F.-J. Chang, J. Liu, J. Su, G. P. Strimel, A. Mouchtaris, and S. Kunzmann, "Contextual adapters for personalized speech recognition in neural transducers," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8537–8541.
- [5] D. Kulshreshtha, S. Dingliwal, B. Houston, and S. Bodapati, "Multilingual contextual adapters to improve custom word recognition in low-resource languages," in *Proc. Interspeech*, 2023, pp. 3302–3306.
- [6] M. A. B. Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, "Investigating the effect of asr tuning on named entity recognition," in *Proc. Interspeech*, 2017, pp. 2490–2494, interspeech 2017. [Online]. Available: https://www.isca-archive.org/interspeech_2017/jannet17_interspeech.pdf
- [7] Z. Meng, Z. Wu, R. Prabhavalkar, C. Peyser, W. Wang, N. Chen, T. Sainath, and B. Ramabhadran, "Text injection for neural contextual biasing," in *Proc. Interspeech*, 2024, pp. 2985–2989, arXiv:2406.02921. [Online]. Available: <https://arxiv.org/abs/2406.02921>
- [8] Y. Li, Y. Li, M. Zhang, C. Su, J. Yu, M. Piao, X. Qiao, M. Ma, Y. Zhao, and H. Yang, "CB-Whisper: Contextual Biasing Whisper Using Open-Vocabulary Keyword-Spotting," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, 2024, pp. 2941–2946. [Online]. Available: <https://aclanthology.org/2024.lrec-main.262/>
- [9] X. Gong, A. Lv, Z. Wang, and Y. Qian, "Contextual biasing speech recognition in speech-enhanced large language model," in *Interspeech*. Graz, Austria: ISCA, 2024, pp. 1–5. [Online]. Available: https://www.isca-archive.org/interspeech_2024/gong24b_interspeech.pdf
- [10] J. D. Fox and N. Delworth, "Improving contextual recognition of rare words with an alternate spelling prediction model," in *Proc. Interspeech*, 2022, arXiv:2209.01250. [Online]. Available: <https://arxiv.org/pdf/2209.01250>
- [11] P. R. Dixon, C. Hori, and H. Kashioka, "A specialized wfst approach for class models and dynamic vocabulary," in *Proc. Interspeech*, 2012. [Online]. Available: <http://www.interspeech2020.org/uploadfile/pdf/Thu-2-8-3.pdf>
- [12] Y. Li *et al.*, "Minimising biasing word errors for contextual asr with the wfst framework," in *Proc. Interspeech*, 2022, arXiv:2205.09058. [Online]. Available: <https://arxiv.org/pdf/2205.09058>
- [13] P. Aleksic, C. Allauzen, D. Elson, A. Kracun, D. M. Casado, and P. J. Moreno, "Improved recognition of contact names in voice commands," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5172–5175.
- [14] J. Tang, K. Kim, S. Shon, F. Wu, and P. Sridhar, "Improving ASR contextual biasing with guided attention," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Seoul, Republic of Korea: IEEE, April 2024, pp. 12 096–12 100. [Online]. Available: <https://ieeexplore.ieee.org/document/10447438>
- [15] X. Shi, Y. Yang, Z. Li, Y. Chen, Z. Gao, and S. Zhang, "Seaco-parafomer: A non-autoregressive asr system with flexible and effective hotword customization ability," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 346–10 350.
- [16] K. Huang, A. Zhang, Z. Yang, P. Guo, B. Mu, T. Xu, and L. Xie, "Contextualized end-to-end speech recognition with contextual phrase prediction network," *arXiv preprint arXiv:2305.12493*, 2023.
- [17] Y. Gao, "Retrieval-augmented generation for large language models: A survey," 2023, arXiv:2312.10997v5. [Online]. Available: <https://arxiv.org/abs/2312.10997>
- [18] Y. Gao, Y. Wang, Y. Wang, and Y. Wang, "Clap: Contrastive language-audio pretraining," *arXiv preprint arXiv:2206.04769*, 2022, accessed: 2024-12-11. [Online]. Available: <https://arxiv.org/pdf/2206.04769>
- [19] W. Team, "Wenetspeech: A large-scale chinese speech corpus," *arXiv preprint arXiv:2101.08605*, 2021, accessed: 2024-12-11. [Online]. Available: <https://arxiv.org/abs/2101.08605>
- [20] J. Bu, X. Wu, Y. Zhang, L. Zhang, J. Zhang, and Y. Yan, "Aishell: An easier chinese mandarin speech recognition dataset," *arXiv preprint arXiv:1709.05522*, 2017, accessed: 2024-12-11. [Online]. Available: <https://arxiv.org/abs/1709.05522>
- [21] G. Team, "Gigaspeech: A large-scale chinese speech corpus," *arXiv preprint arXiv:2101.08605*, 2021, accessed: 2024-12-11. [Online]. Available: <https://arxiv.org/abs/2101.08605>
- [22] D. Paul and J. Baker, "The LibriSpeech asr corpus," *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015, accessed: 2024-12-11. [Online]. Available: <https://www.tensorflow.org/datasets/catalog/librispeech>
- [23] P. Tomaseo, A. Shrivastava, D. Lazar, P.-C. Hsu, D. Le, A. Sagar, A. Elkahky, J. Copet, W.-N. Hsu, Y. Adi, R. Algayres, T. A. Nguyen, E. Dupoux, L. Zettlemoyer, and A. Mohamed, "Stop: A dataset for spoken task oriented semantic parsing," 2022, arXiv:2207.10643v3. [Online]. Available: <https://arxiv.org/abs/2207.10643>
- [24] A. Baevski, A. Babu, W.-N. Hsu, and M. Auli, "Efficient self-supervised learning with contextualized target representations for vision, speech and language," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1416–1429.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, accessed: 2024-12-11. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [26] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.