Interspeech 2024
1-5 September 2024, Kos, Greece

RASU: Retrieval Augmented Speech Understanding through Generative Modeling

Hao Yang[1], *Min Zhang*[1], *Minghan Wang*[2], *Jiaxin Guo*[1]

[1]2012 Labs Huawei CO., LTD Beijing, China
[2]Monash University, Melbourne,Australia

yanghao30@huawei.com, zhangmin186@huawei.com, minghan.wang@monash.edu, guojiaxin@huawei.com

Abstract

Large language models have benefited from retrieval augmented generation (RAG) techniques, which allow relevant knowledge to be retrieved and provided as prompts to enhance natural language understanding capabilities. Extending this promising approach to spoken language understanding (SLU) tasks represents an important area of study. This paper introduces a novel RAG framework tailored for SLU called Retrieval Augmented Speech Understanding(RASU). The proposed model first employs the encoder from a pre-trained automatic speech recognition (ASR) model to retrieve relevant speech segments and transcripts from the training data given a new spoken utterance. The retrieved text transcripts and their corresponding intent labels are then formulated as prompts to conditionally guide the SLU decoder during generation. Additionally, a prompt attention mechanism is incorporated to strengthen the interaction between the generated outputs and the retrieved prompts. Empirical evaluations demonstrate that RASU substantially outperforms conventional end-to-end and cascaded SLU models on intent prediction from speech data. These results highlight the efficacy of leveraging retrieval-based prompting and external knowledge sources to markedly improve spoken language understanding performance. The RASU approach presents a promising direction for advancing SLU capabilities by bridging speech retrieval and generative language modeling.

Index Terms: spoken language understanding, retrieval augmented generation, large language models

1. Introduction

Large language models (LLMs) [1] [2]have emerged as the latest groundbreaking advances in artificial intelligence, substantially enhancing capabilities in natural language understanding (NLU) tasks [3]. Retrieval augmented generation (RAG) [4] [5] has proven to be an effective technique to further boost LLM performance. RAG systems retrieve relevant sentences or external knowledge sources to construct informative prompts that provide additional contextual information. These prompts are then leveraged to conditionally guide and inform the LLM's predictions [6] [7]. Recent studies have demonstrated that RAG can significantly improve LLM performance on various NLU benchmarks by equipping the models with useful retrieved knowledge through conditional prompting.

Extending this promising RAG paradigm to the spoken language domain represents an important research direction. Retrieval Augmented Speech Understanding (RASU) aims to enhance spoken language understanding capabilities of LLMs by retrieving and incorporating relevant speech data and transcripts as prompts during the generation process. RASU holds the potential to unlock further gains by allowing LLMs to effectively leverage external spoken knowledge sources and data when interpreting spoken utterances. As such, RASU remains a promising area for advancing the state-of-the-art in spoken language processing while fully capitalizing on the remarkable capabilities of large language models.

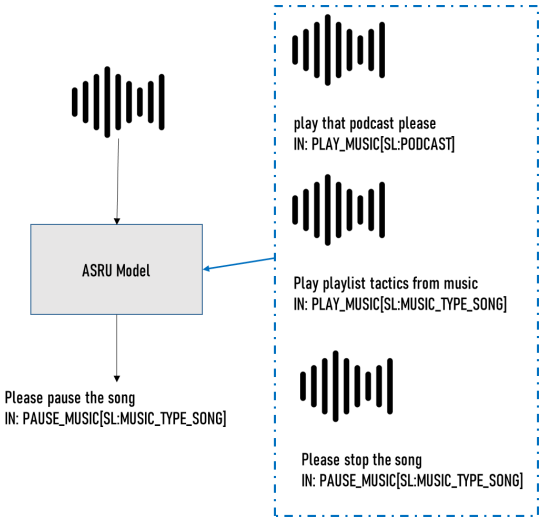

Figure 1: *Retrieval Augmented Speech Understanding.*

The application of large language models (LLMs) to spoken language understanding (SLU) tasks [8] [9] is an emerging research area with significant potential, but faces key challenges: 1) effectively incorporating external knowledge into LLMs for SLU, 2) designing suitable prompts conditioned on relevant knowledge sources, and 3) properly leveraging LLM capabilities to enhance SLU performance. Limited prior work has explored tailored prompting techniques to inject knowledge into LLMs for improving SLU, and the knowledge retrieval and prompt formulation methodologies designed specifically for the SLU domain remain understudied compared to other NLP tasks. Ample research opportunities exist in developing novel retrieval augmented speech understanding (RASU) approaches to equip LLMs with external knowledge sources for SLU, as well as investigating prompt-based transfer learning techniques to further boost spoken language understanding capabilities using LLMs, which could unlock their potential for state-of-the-art SLU performance.

Whisper [10] [11] is a pre-trained automatic speech recognition (ASR) model based on a convolutional and encoder-decoder architecture, and its convolutional or transformer encoder can be leveraged for text retrieval over the training corpus in RASU framework. The retrieved text transcripts and corre-

10.21437/Interspeech.2024-148

sponding intent labels can then be formulated as informative prompts to conditionally augment the intent prediction process. Akin to machine translation, a prompt attention mechanism can be explored to effectively incorporate relevant prompts for enhancing intent generation. This paradigm suggests the investigation of prompt-based transfer learning techniques like prompt tuning and prompt training to further adapt and fine-tune Whisper for intent detection tasks. Overall, capitalizing on Whisper's text retrieval capabilities and exploring prompt-based learning presents promising research directions for advancing the state-of-the-art in spoken language understanding, as evidenced by promising initial results.

In general, the key innovations presented in this article include:

1. A retrieval model is constructed by fine-tuning the encoder of a pre-trained automatic speech recognition (ASR) model.
2. For a given speech input, relevant text transcripts and their corresponding intent labels are retrieved from the training set to form informative prompt examples.
3. A prompt attention mechanism is then introduced to effectively incorporate the retrieved prompts for guiding the spoken language understanding (SLU) generator. This allows explicitly conditioning the intent prediction on relevant external knowledge retrieved from the corpus.

Comprehensive experiments on SLU benchmarks demonstrate that the proposed RASU approach achieves significant performance gains over both cascaded and end-to-end SLU baseline models. These results empirically validate that RASU can effectively improve spoken language understanding capabilities by leveraging retrieval-based knowledge augmentation and conditional prompting from relevant external data sources.

2. Related Work

2.1. SLU Model

Speech language understanding (SLU) [12] [13] systems aim to extract structured semantic representations from spoken natural language inputs by converting audio waveforms into representations of user intents and semantic arguments. This process involves automatic speech recognition to transcribe the audio, natural language understanding to extract meaning from the transcripts, and dialogue management systems to formulate meaningful responses. Key modules in SLU pipelines include automatic speech recognition, named entity recognition, intent detection, slot filling, and dialogue state tracking. Recent advances in deep learning and conversational AI have substantially improved the accuracy of modern SLU systems, enabling applications ranging from virtual personal assistants to customer service chatbots. However, significant challenges persist in effectively modeling implicit knowledge and handling the diversity present in human speech for SLU tasks. In this context, retrieval augmented speech understanding (RASU) presents a promising research direction to enhance SLU capabilities.

SLU [14] [15] aims to map sequence speech input $x_1, ..., x_m$ to semantic representations comprising intent classification Y_{intent} and slot filling Y_{slot}:

$$\begin{aligned} X &= [x_1, x_2, ..., x_m] \\ Y_{intent} &= f(x_1, x_2, ..., x_m) \\ Y_{slot} &= g(x_1, x_2, ..., x_m) \end{aligned} \tag{1}$$

where f and g denote the intent prediction and slot filling models respectively.

Key challenges for spoken language understanding (SLU) include the speech-text mismatch problem and the lack of sufficient labeled speech data for training robust SLU models. To mitigate these issues, recent research has explored transfer learning approaches that leverage pre-trained language models like BERT [16] and pre-trained speech representation models such as Wav2Vec 2.0 [17] [18]. However, advanced adaptation and robust learning techniques are required to effectively bridge the gap across speech and language modalities for improved SLU performance.

2.2. Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is a framework that leverages retrieved knowledge to enhance text generation capabilities [19] [19] [20]. Given an input context X, RAG first retrieves relevant knowledge k from a corpus or database. The retrieved knowledge is then utilized as a prompt to conditionally guide the generation process [21].

$$\begin{aligned} X &= [x_1, x_2, ..., x_m] \\ Y &= [y_1, y_2, ..., y_n] \\ K &= Retrieval(X) \\ P(Y|X) &= \prod_h P(y_t|X, K, y_1, ..., y_{t-1}) \end{aligned} \tag{2}$$

where y is the generated text. The prompt K guides the decoder to incorporate external knowledge. RAG has shown significant gains across various generation tasks including summarization, translation, dialog by retrieving task-specific examples and facts.

2.3. Retrieval Augmented Speech Understanding

Spoken language understanding (SLU) [17] [22] aims to map speech input x to a semantic intent representation y. However, the lack of sufficient labeled speech data poses significant challenges for training robust SLU models. To mitigate this issue, we propose a retrieval augmented speech understanding (RASU) approach. Given an input speech utterance x, the RASU framework first employs a retrieval model to identify similar speech examples and their corresponding transcripts and intent labels from a corpus:

$[Relevantspeechsegments, transcripts, intentlabels] = Retrieval(x, corpus)$

The retrieved [speech, text, label] tuples then serve as informative prompts to conditionally augment the SLU model's intent prediction:

$$Y_{pred} = SLU_Model(x, [Prompts]) \tag{3}$$

This RASU approach provides several key advantages: It enables leveraging external speech/text data as prompts to enhance the SLU model, alleviating labeled data scarcity. Prompting allows adapting large pre-trained language models for SLU via techniques like prompt tuning. Multimodal retrieval and prompting facilitate jointly conditioning on speech and text inputs. The retrieval model can be efficiently constructed by fine-tuning pretrained automatic speech recognition encoders.

3. Proposed Approach

The proposed retrieval augmented speech understanding (RASU) framework, inspired by the retrieval augmented generation (RAG) paradigm, comprises two key components: a retriever and a generator. The retriever performs a search over a

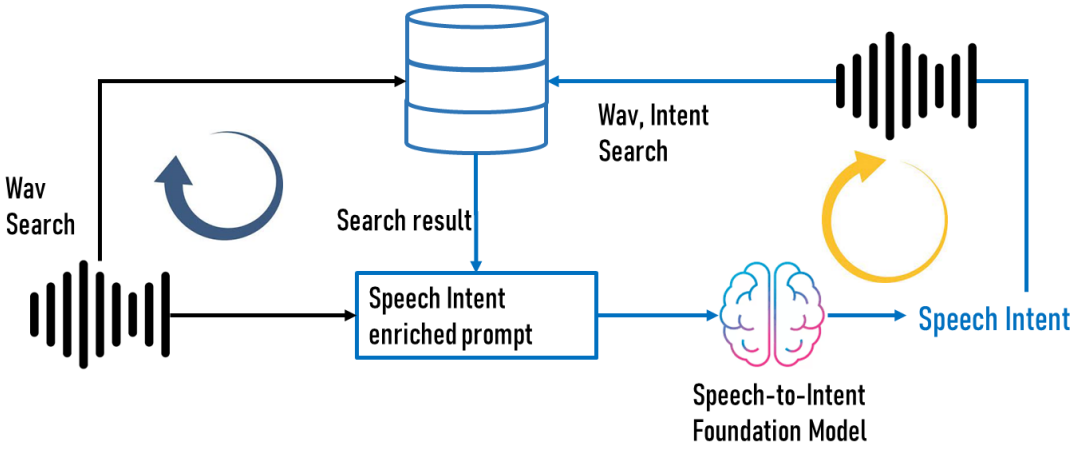

Figure 2: *RASU architecture for multi-task prompt fine-tuning based on Whisper*

knowledge repository containing speech data, transcripts, and intent labels to obtain the most relevant prompts for a given input speech utterance. Specifically, the top matching speech segments are retrieved along with their corresponding text transcripts and intent labels. These retrieved [speech, text, intent] tuples are then collated to form informative prompt examples. The generator takes the speech representation and the retrieved prompts as input to conduct auto-regressive decoding for intent prediction. The incorporation of high-quality prompts from the retrieval stage guides the model to generate accurate intent representations by explicitly conditioning on the relevant external knowledge retrieved from the training corpus.

3.1. RASU Retriever

The proposed retrieval augmented speech understanding (RASU) framework employs a retriever component that serves three key functions:(1)Encoding the input speech utterance into a dense embedding vector representation using a speech encoder model.(2)Indexing the speech embedding vectors of the entire knowledge corpus containing speech recordings, transcripts, and intent labels. This creates a searchable index over the speech data.(3)Conducting an efficient vector search over the indexed speech embeddings to retrieve the top-k most similar speech instances to the input, along with their aligned text transcripts and intent labels.

3.2. RASU Generator

The generator component of the proposed retrieval augmented speech understanding (RASU) framework is an autoregressive model that takes two key inputs:

1. The speech representation encoded from the input audio utterance.
2. The prompt representation encoded from the retrieved text transcripts and intent labels.

Based on these inputs, the generator's decoder conducts left-to-right autoregressive decoding to generate the predicted intent representation. Specifically, the RASU framework adopts the Whisper decoder as the generator model. The cross-attention sublayer attends to the encoded speech representation to model the acoustic context. Additionally, a prompt attention module is introduced that attends to the encoded prompt representations retrieved from the corpus. This allows the generator to explicitly model and incorporate both the speech input context as well as the informative retrieval knowledge from relevant examples.

The prompt attention mechanism is implemented using standard scaled dot-product attention for simplicity. As shown in Figure 2, the overall generator architecture contains dual attention sublayers over the speech encoding and the retrieved prompts within a unified autoregressive decoder.

This dual attention approach allows the RASU generator to transparently incorporate external knowledge from the retrieved prompts to enhance the spoken language understanding process. By conditioning the intent prediction on both the input speech and relevant retrieved exemplars, the generator can leverage informative contextual information to better interpret the speech utterance and predict the accurate intent representation.

4. Experiments

4.1. Experiments Setup

The proposed retrieval augmented speech understanding (RASU) approach is rigorously evaluated against state-of-the-art baselines on spoken language understanding benchmarks, including:

- Cascaded models: Wav2Vec2.0 + NLU and Hubert + NLU pipelines
- End-to-end model: WhiSLU

For the RASU model, three different configurations based on the Whisper architecture are explored - Whisper-base, Whisper-medium, and Whisper-large variants. Hyperparameter and training details are as follows:

- For the base and medium RASU models, all parameters are fine-tuned in an end-to-end fashion.
- For the large model, the encoder weights are frozen and only the decoder is updated during fine-tuning.
- A learning rate of 1e-4 is used for optimization.
- The per-GPU batch size is set to 8, with gradient accumulation steps adjusted for each model size to fit the available GPU memory.

4.2. Datasets and Metrics

The proposed retrieval augmented speech understanding approach is rigorously evaluated on the STOP dataset, which is a large-scale spoken language understanding benchmark. The STOP dataset comprises:

- 885 speakers
- 218 hours of transcribed speech data
- 236,477 aligned examples of [speech, text transcript, intent] tuples

Exact Match (EM) accuracy: Measures the percentage of predicted SLU output sequences (intent + slot values) that exactly match the ground truth sequences. This evaluates the overall accuracy including correctly predicting both the intent and all slot values. Tree EM: Evaluates the structural accuracy by ignoring the slot leaf node values during matching. It focuses specifically on correctly predicting the intent along with the hierarchical slot structures/types, without considering the slot value accuracy.

Using these EM and Tree EM metrics provides a comprehensive evaluation of the RASU model's performance on the spoken language understanding task. EM assesses the overall output quality, while Tree EM isolates the model's capability in identifying the correct intent and slot structures regardless of slot value errors.

4.3. Performance Comparison

The experimental results in Table 1 demonstrate that the proposed retrieval augmented speech understanding (RASU) approach, referred to as RASU in the paper, significantly outperforms cascaded ASR+NLU systems on the spoken language understanding benchmarks. Moreover, RASU surpasses the performance of the strong end-to-end WhiSLU-medium baseline by achieving absolute gains of 0.9% in Exact Match (EM) accuracy and 0.42% in Tree EM accuracy. These improvements highlight the key benefits of incorporating external knowledge through the retrieval-based prompting mechanism employed in RASU for enhancing speech understanding capabilities. Notably, while the word error rate (WER) of RASU is slightly lower compared to some baselines, the substantial gains in semantic parsing metrics like EM and Tree EM over robust models like WhiSLU validate that the performance enhancements stem from better linguistic comprehension rather than just literal transcription accuracy alone.

Table 1: *Overall SLU comparison result, including cascade models, E2E models and RASU.*

Model	EM	EM-tree	Wer
cascade	72.36	82.78	
wav2vec2	68.7	82.78	4.45
HuBERT	69.23	82.87	**4.26**
WhiSLU-medium	74.13	85.46	10.85
WhiSLU-large	76.68	86.37	6.14
RASU-large	**76.79**	**86.47**	4.81

4.4. Modle Size Comparison

The results in Table 2 provide an analysis of the impact of model size on the performance of the WhiSLU and RASU approaches. Several important observations can be made:

Model Scaling Benefits: For both WhiSLU and RASU, larger model sizes (large ¿ medium ¿ base) achieve superior performance in terms of both Exact Match (EM) and Tree EM metrics. This demonstrates the general benefits of increasing model capacity through scaling for spoken language understanding tasks. RASU Outperforms Across Scales: Notably, the RASU framework consistently outperforms the corresponding WhiSLU model across all three size configurations (base, medium, large). This highlights that the retrieval augmented speech understanding paradigm provides complementary performance gains over the strong end-to-end WhiSLU baseline, regardless of the underlying model scale.

Table 2: *Pre-trained model Analysis, large model is always better.*

Model	#Trainable Params	EM	EM-Tree
WhiSLU-base	74M	68.32	81.57
WhiSLU-medium	769M	74.13	85.46
WhiSLU-large	1550M	76.68	86.37
RASU-base	74M	68.52	82.34
RASU-medium	769M	74.13	85.46
RASU-large	1550M	76.79	86.47

4.5. Prompt Evaluation

The results presented in Table 3 provide insightful comparisons on the impact and quality of different prompt sources for the proposed RASU framework:(1) RASU prompts vs. No prompts: The RASU model utilizing retrieved speech/text prompts significantly outperforms the no prompt baseline. This validates the efficacy of incorporating external knowledge via conditional prompting for enhancing spoken language understanding. (2) Gap to ground-truth optimal prompts: However, there remains a sizable performance gap between RASU and using ground-truth optimal prompts extracted from the reference data. This implies there is considerable room for improvement in the current prompt retrieval approach based solely on acoustic speech similarity matching.

Table 3: *Prompt comparsion, Ground Truth(GT) prompt is best, Overall SLU comparison result, including cascade models, E2E models and RASU.*

	EM	EM-Tree	Wer
no Prompt	73.3	85.17	10.85
GT Prompt	85.25	87.85	3.34
RASU Prompt	74.13	85.46	4.81

5. Conclusion

This paper introduces a novel retrieval augmented speech understanding (RASU) framework that aims to enhance semantic comprehension in spoken language tasks. The key innovation is to complement the core speech recognition model with external knowledge retrieved from a database to better interpret the meaning behind spoken inputs, instead of solely relying on acoustic and textual signals. The proposed RASU approach employs a retrieval model to identify relevant speech segments, transcripts, and intent labels from a corpus based on similarity to the input utterance. These retrieved exemplars

are then formulated as prompts and incorporated into the spoken language understanding model through a tailored retrieval augmented generation (RAG) architecture. Comprehensive experiments across multiple SLU benchmarks demonstrate that augmenting speech models with this external retrieved knowledge via the RAG framework and conditional prompting leads to significant performance improvements on key language understanding metrics like intent prediction accuracy.

6. References

[1] E. L. Hill-Yardin, M. R. Hutchinson, R. Laycock, and S. J. Spencer, "A chat (gpt) about the future of scientific publishing," *Brain Behav Immun*, vol. 110, pp. 152–154, 2023.

[2] H. Yang, M. Zhang, S. Tao, M. Wang, D. Wei, and Y. Jiang, "Knowledge-prompted estimator: A novel approach to explainable machine translation assessment," *arXiv preprint arXiv:2306.07486*, 2023.

[3] G.-T. Lin, C.-J. Hsu, D.-R. Liu, H.-Y. Lee, and Y. Tsao, "Analyzing the robustness of unsupervised speech recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8202–8206.

[4] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang, "P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 61–68.

[5] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision*. Springer, 2022, pp. 709–727.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[7] Y. Hao, C. Junliang, M. Xiangwu, and Q. Bingyu, "Dynamically traveling web service clustering based on spatial and temporal aspects." Springer, 2007, pp. 348–357.

[8] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," *arXiv preprint arXiv:2011.13205*, 2020.

[9] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," *arXiv preprint arXiv:1909.02188*, 2019.

[10] 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

[11] H. Yang, M. Zhang, S. Tao, M. Ma, and Y. Qin, "Chinese asr and ner improvement based on whisper fine-tuning," in *2023 25th International Conference on Advanced Communication Technology (ICACT)*. IEEE, 2023, pp. 213–217.

[12] Z. Li, Z. Wu, Z. Rao, X. YuHao, G. JiaXin, D. Wei, H. Shang, W. Minghan, X. Chen, Z. Yu *et al.*, "Hw-tsc at iwslt2023: Break the quality ceiling of offline track via pre-training and domain adaptation," in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, 2023, pp. 187–193.

[13] M. Wang, Y. Li, J. Guo, X. Qiao, Z. Li, H. Shang, D. Wei, S. Tao, M. Zhang, and H. Yang, "Whislu: End-to-end spoken language understanding with whisper."

[14] C.-I. Lai, Y.-S. Chuang, H.-Y. Lee, S.-W. Li, and J. Glass, "Semi-supervised spoken language understanding via self-supervised speech and language model pretraining," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7468–7472.

[15] J. Shi, C.-J. Hsu, H. Chung, D. Gao, P. Garcia, S. Watanabe, A. Lee, and H.-y. Lee, "Bridging speech and textual pre-trained models with unsupervised asr," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] M. Wang, J. Guo, Y. Li, X. Qiao, Y. Wang, Z. Li, C. Su, Y. Chen, M. Zhang, S. Tao, H. Yang, and Y. Qin, "The HW-TSC's Simultaneous Speech Translation System for IWSLT 2022 Evaluation," in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Dublin, Ireland (in-person and online): Association for Computational Linguistics, 2022, pp. 247–254. [Online]. Available: https://aclanthology.org/2022.iwslt-1.21

[18] B. van Niekerk, M.-A. Carbonneau, J. Zaidi, M. Baas, H. Seute, and H. Kamper, "A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Singapore, Singapore: IEEE, May 2022, pp. 6562–6566. [Online]. Available: https://ieeexplore.ieee.org/document/9746484/

[19] K.-W. Chang, Y.-K. Wang, H. Shen, I.-t. Kang, W.-C. Tseng, S.-W. Li, and H.-y. Lee, "Speechprompt v2: Prompt tuning for speech classification tasks," *arXiv preprint arXiv:2303.00733*, 2023.

[20] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," *arXiv preprint arXiv:2305.11095*, 2023.

[21] H. Yang, Z. Wu, Z. Yu, X. Chen, D. Wei, Z. Li, H. Shang, M. Wang, J. Guo, L. Lei *et al.*, "Hw-tsc's submissions to the wmt21 biomedical translation task," in *Proceedings of the Sixth Conference on Machine Translation*, 2021, pp. 879–884.

[22] H. Yang, S. Tao, M. Wang, M. Zhang, D. Wei, S. Zhao, M. Ma, and Y. Qin, "CCDC: A Chinese-Centric Cross Domain Contrastive Learning Framework," in *Knowledge Science, Engineering and Management*, G. Memmi, B. Yang, L. Kong, T. Zhang, and M. Qiu, Eds. Cham: Springer International Publishing, 2022, pp. 225–236.