

# INTEGRATION OF PRE-TRAINED NETWORKS WITH CONTINUOUS TOKEN INTERFACE FOR END-TO-END SPOKEN LANGUAGE UNDERSTANDING

Seunghyun Seo<sup>1,2\*</sup>, Donghyun Kwak<sup>1\*</sup>, Bowon Lee<sup>2</sup>

<sup>1</sup>Clova AI, NAVER Corp. <sup>2</sup>Inha University

{real.seunghyun.seo, donghyun.kwak}@navercorp.com, bowon.lee@inha.ac.kr

## ABSTRACT

Most End-to-End (E2E) Spoken Language Understanding (SLU) networks leverage the pre-trained Automatic Speech Recognition (ASR) networks but still lack the capability to understand the semantics of utterances, crucial for the SLU task. To solve this, recently proposed studies use pre-trained Natural Language Understanding (NLU) networks. However, it is not trivial to fully utilize both pre-trained networks; many solutions were proposed, such as Knowledge Distillation (KD), cross-modal shared embedding and network integration with Interface. We propose a simple and robust integration method for the E2E SLU network with a novel Interface, Continuous Token Interface (CTI). CTI is a junctional representation of the ASR and NLU networks when both networks are pre-trained with the same vocabulary. Thus, we can train our SLU network in an E2E manner without additional modules, such as Gumbel-Softmax. We evaluate our model using SLURP, a challenging SLU dataset and achieve state-of-the-art scores on intent classification and slot filling tasks. We also verify that the NLU network, pre-trained with Masked Language Model (MLM), can utilize a noisy textual representation of CTI. Moreover, we train our model with extra data, SLURP-Synth, and get better results.

**Index Terms**— end-to-end spoken language understanding, interface of networks, intent classification, slot filling

## 1. INTRODUCTION

Spoken Language Understanding (SLU), a front-end of many spoken dialogue systems, is a task of extracting semantic information such as intents, slots, or emotions from speech. Conventional SLU pipelines consist of the Automatic Speech Recognition (ASR) model and the Natural Language Understanding (NLU) model, where the ASR model converts speech to text, and the NLU model predicts task-specific information from the text [1]. However, these pipelines propagate errors from the ASR model to the NLU model, degrading the system's overall performance. To address this problem, recent studies focused on End-to-End (E2E) SLU networks that directly extract semantic information from

speech [2, 3]. Most E2E SLU networks leverage the encoder part of pre-trained ASR networks but cannot fully understand the semantics of utterances, which is crucial for the SLU task.

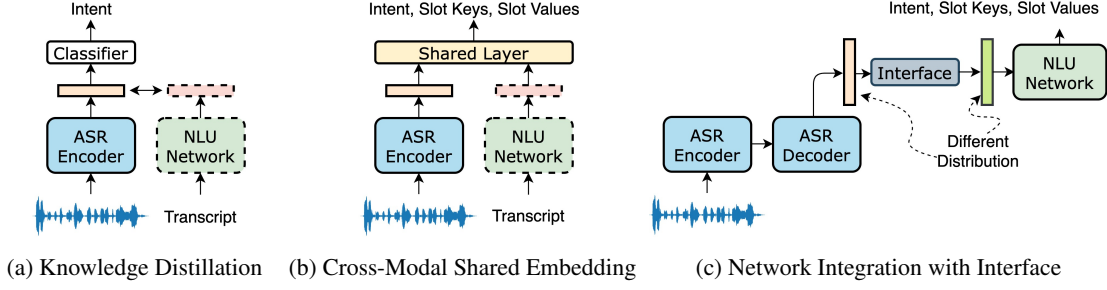
To solve this, recently proposed studies use pre-trained NLU networks to leverage the powerful language representation of the Pre-trained Language Model (PLM), such as BERT [4]. Some approaches use Knowledge Distillation (KD) from PLM to SLU network [5, 6] or introduce cross-modal shared embedding space between acoustic and textual features [7, 8, 9], and others try to integrate the ASR and NLU networks with appropriate Interface [10, 11].

However, these methods may lose the powerful language representation of BERT in the process of KD or lose important features while embedding cross-modal inputs in a shared space. In the case of network integration with Interface, it depends on the gap between the two representations. The Interface is a junctional representation of two pre-trained networks to be combined [11]. To make the gradient flow via Interface, it is usually necessary to employ additional modules to match the representations.

For example, If we integrate the networks with Discrete Token Interface (DTI), one-hot encoding of ASR hypothesis, the model needs additional modules, such as the Gumbel-Max trick and Gumbel-Softmax distribution for forward and backward passes to match the distribution [12]. This learning process is complicated, as the pre-trained networks could miss their acoustic or linguistic representations.

In this paper, we propose a novel Interface, Continuous Token Interface (CTI), for a simple and robust network integration. The CTI is naturally derived by pre-training both ASR and NLU networks with the same vocabulary. With our interface, there is a small representation gap between the networks, and no additional modules are needed. Because the networks have the same vocabulary, the softmax probability distribution of the ASR network can be considered as a noisy textual representation from the point of view of the NLU network. We assume it only differs in the noise pattern; our Interface has a neglectable gap of representations, unlike other Interfaces. Therefore we directly feed the ASR network's output to the NLU network, which makes our SLU network trainable in an E2E manner without additional modules, such as Gumbel-Softmax.

\* Equal contribution.



**Fig. 1:** Various methods for leveraging the capacity of the pre-trained NLU network. From left to right, they are methodologies in the order described in Related Work. Note that the blocks with dotted lines are only used in the training procedure.

Finally, we evaluate our model on the SLURP dataset [13], a recently proposed challenging SLU dataset, and achieve state-of-the-art performance in both Intent Classification (IC) and Slot Filling (SF). We also conduct ablation studies to verify that the NLU network, pre-trained with Masked Language Modeling (MLM), can utilize a noisy textual representation of CTI. Moreover, we train our model with extra data, SLURP-Synth, and get better results.

Our main contributions can be summarized as follows:

1. We integrate the pre-trained networks by a novel Interface, CTI, in an E2E manner without any modules.
2. Our CTI integration allows each part of the model to be trained independently based on data type.
3. We evaluate our model on the SLURP dataset and achieve state-of-the-art scores in both IC and SF.

## 2. RELATED WORKS

Recently, the E2E SLU approaches are designed to understand contextual semantic information of speech by utilizing PLM [14, 15]. These works were studied in various ways, including KD, learning cross-modal latent space, and network integration via Interface. Fig. 1 shows these works.

### 2.1. Knowledge Distillation

Inspired by [16], KD-based E2E SLU models were proposed to leverage the representation power of PLM [5, 6]. The SLU network can learn high-level semantic representations from the NLU network by distilling the knowledge.

### 2.2. Cross-Modal Shared Embedding

These approaches were designed to learn a cross-modal latent space by inducing the paired speech and transcript to become closer to each other in the common embedding space [7, 9]. The SLU network then has a cross-modal shared embedding space derived from both representations of ASR and NLU networks.

### 2.3. Network Integration with Interface

These approaches were proposed to combine ASR and NLU networks and train the models in an end-to-end manner. They extracted acoustic information from the ASR network, including logits, hidden features, or text, and feed them into the

NLU network via Interface [17]. If the Interface is not differentiable, additional modules such as Gumbel-softmax or attention gate are needed [11, 17, 18].

## 3. PROPOSED METHOD

### 3.1. Integration Method and Model Components

The SLU network we proposed consists of two main components: ASR and NLU networks. For simple and robust integration via CTI, both networks are pre-trained with the same vocabulary. Given speech input  $S$ , the ASR network produces tokens with a continuous probability distribution by *softmax* function and delivers the output vectors,  $Z$  to the NLU network directly as in Eq. (1). The NLU network finally outputs  $Y$  (intent, slot key, and slot value) as in Eq. (2).

$$Z = \text{softmax}(\text{ASR}_{Dec}(\text{ASR}_{Enc}(S))) \quad (1)$$

$$Y = \text{NLU}(Z) \quad (2)$$

We call this  $Z$  (the output of the ASR network) as CTI, which are considered few-hot encoded vectors from the point of view of the NLU network. We hypothesize that it is noisy but more informative than one-hot encoded vectors generated by the conventionally used *argmax* function.

If there is a confusing case between two tokens, the conventional ASR network makes only true/false binary predictions. In our proposed model, however, both pieces of information can be processed at the same time even with some noise. We may think of this as a richer feature that contains information about how the token sounds, rather than just noise.

As a result, the entire SLU network can be trained in an end-to-end manner, with the intermediate noisy text representation. Fig. 2 shows our proposed model.

#### 3.1.1. ASR Network : Wav2Vec 2.0 based Seq2Seq Model

Our ASR network is a Sequence-to-Sequence (Seq2Seq) model consisting of Wav2Vec 2.0 Encoder [19] and Transformer Decoder [20], where the Wav2Vec 2.0 Encoder is known to have remarkable representation power by contrastive learning task.

#### 3.1.2. NLU Network : RoBERTa

We use RoBERTa [21], one of the popular PLMs as our NLU network. We added an attention pooling layer, softmax classi-



**Table 1:** The comparison of the IC accuracy between our baseline and other models on the FSC dataset.

Model (E2E SLU)	Input	Dev	Test
Lugosh et al. [3]	Speech	-	98.8
Kim et al. [6]	Speech	97.8	99.7
Qian et al. [26]	Speech	-	99.7
Wav2Vec2.0-Classifer (Ours)	Speech	<b>98.9</b>	<b>99.7</b>

We use pre-trained RoBERTa as the NLU network, and finally, we connect two components via our proposed method (CTI), then fine-tune the model on the SLURP dataset with multi-task losses. Note that, unlike other E2E SLU methods, each part of our model can behave like a stand-alone model. Therefore, we can train the NLU network with text-to-label data and the ASR network with speech-to-text data at will. We leave 46 audio files out of 50,628 training audio files during the fine-tuning because there are mismatches between speech and labels.

### 4.3. Results and Analysis

We evaluate the performance of SLU models with the metric of IC accuracy and SLU-F1 score, where SLU-F1 is a new metric for slot filling tasks proposed with the SLURP dataset [13]. The scores are shown in Table. 2. There are three types of models: NLU, SLU (Inference Only), and SLU (E2E Train). The score of our NLU network is 87.73, which is higher than 84.84 of the NLU network from the SLURP paper in IC. To experiment with SLU (Inference Only) case, these stand-alone NLU networks need to be fine-tuned on the SLURP dataset. On the other hand, in SLU (E2E Train), the NLU networks are not fine-tuned independently on the SLURP dataset but fine-tuned in an end-to-end manner. Wav2Vec2.0-Classifer only shows 76.6 IC accuracy because it does not explicitly understand the language information. Among all SLU experiments, our model with all multi-task losses achieves state-of-the-art scores, 82.93 (IC) and 71.12 (SLU-F1).

#### 4.3.1. Effectiveness of CTI

We hypothesize that CTI makes NLU networks take richer information about the words that the ASR network confuses; the confusing words are represented as an interpolation of embedding vectors in the NLU network. We compare SLU models with DTI, Gumbel-Interface [11], and CTI. For SLU (Inference Only), CTI shows meaningful improvement compared to DTI by 0.8 (IC), and for SLU (E2E Train), CTI outperforms Gumbel-Interface by 0.83 (IC) and 0.57 (SLU-F1) under the same conditions except for the Interface. We assume that the comparison between Gumbel-Interface and CTI could be different from [17] according to the domain of dataset. Moreover, we introduce more multi-task losses, such as  $L_{NLU}$  compared to [17] for training with each interface.

#### 4.3.2. NLU Loss

We train our model with and without  $L_{NLU}$  because CTI allows NLU networks to be trained independently on text-only

**Table 2:** The comparison of IC accuracy and SLU-F1 score between proposed models and baselines. ASR→NLU means inference with DTI and ASR⇒NLU means inference with CTI. In SLU (E2E Train), A, S, and N means ASR loss, SLU loss, and NLU loss for each.

Model Type	Model	Intent	SLU-F1
NLU	NLU [13]	84.84	-
	NLU (Ours)	87.73	84.34
SLU (Inference Only)	ASR→NLU [13]	78.33	70.84
	ASR→NLU (Ours)	80.37	70.23
	ASR⇒NLU (Ours)	81.17	70.20
SLU (E2E Train)	Wav2Vec2.0-Classifer (Ours)	76.6	-
	Gumbel-Interface (A+S+N) [11]	82.10	70.55
	CTI (A+S)	82.39	70.61
	CTI (A+S+N)	<b>82.93</b>	<b>71.12</b>
	CTI (A+S+N) + Extra data (Text-only)	84.34	71.08
	CTI (A+S+N) + Extra data (All)	<b>86.92</b>	<b>74.66</b>

data. Adding  $L_{NLU}$  increases the model score by 0.54 (IC) and 0.51 (SLU-F1). This shows that CTI can preserve the NLU network’s powerful text representation, which is learned in the pre-training procedure.

#### 4.3.3. Extra Data : SLURP-Synth

We conduct an additional experiment to check if using extra data improves the model performance as studied in [27]. The extra data is SLURP-Synth where it consist of text, labels and synthesized speech by Google’s Text-to-Speech system [13]. Here we leave 24,365 files that have out-of-distribution labels among the 69,253 files.

We first train the entire network with additional Text-only training data. To do this, we extract the Transcript and Intent pairs from SLURP-synth. The results show that using additional Text-only data increases the IC by 2.5. Second, we train the entire network with SLURP-Synth data containing all text, labels, and synthesized speech. This increase the score by 2.58 (IC) and 3.58 (SLU-F1) than Extra data (Text-only) case, and we finally achieved 86.92 (IC) and 74.66 (SLU-F1) which is the state-of-the-art score. This means that critical information still remains in speech-to-label training that has not been fully discovered in text-to-label training.

## 5. CONCLUSIONS AND FUTURE WORK

We propose a simple and robust method for integrating two pre-trained networks via novel Interface, CTI. We achieve state-of-the-art scores of 82.93 (IC), 71.12 (SLU-F1) on the SLURP dataset, and 86.92 (IC), 74.66 (SLU-F1) when we add the SLURP-Synth dataset. In the future, we plan to investigate a new pre-training strategy in which the NLU recovers some tokens corrupted by acoustic noise, such as phoneme-level ASR error distributions. Then the noise pattern of the NLU network inputs will match the ASR network outputs; the E2E SLU model will be less susceptible to the problem of representation gap.

## 6. REFERENCES

- [1] G. Tur and R. De Mori, *Spoken language understanding: Systems for extracting semantic information from speech*, John Wiley & Sons, 2011.
- [2] D. Serdyuk, Y. Wang, C. Fuegen, A. Kumar, B. Liu, and Y. Bengio, “Towards end-to-end spoken language understanding,” in *ICASSP*, 2018, pp. 5754–5758.
- [3] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, “Speech Model Pre-Training for End-to-End Spoken Language Understanding,” in *Proc. Interspeech*, 2019, pp. 814–818.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019, pp. 4171–4186.
- [5] W. I. Cho, D. Kwak, J. W. Yoon, and N. S. Kim, “Speech to Text Adaptation: Towards an Efficient Cross-Modal Distillation,” in *Proc. Interspeech*, 2020, pp. 896–900.
- [6] S. Kim, G. Kim, S. Shin, and S. Lee, “Two-stage textual knowledge distillation to speech encoder for spoken language understanding,” in *ICASSP*, 2021, pp. 7463–7467.
- [7] B. Agrawal, M. Müller, M. Radfar, S. Choudhary, A. Mouchtaris, and S. Kunzmann, “Tie your embeddings down: Cross-modal latent spaces for end-to-end spoken language understanding,” *arXiv preprint arXiv:2011.09044*, 2020.
- [8] Y. Huang, H.-K. Kuo, S. Thomas, Z. Kons, K. Audhkhasi, B. Kingsbury, R. Hoory, and M. Picheny, “Leveraging unpaired text data for training end-to-end speech-to-intent systems,” in *ICASSP*, 2020, pp. 7984–7988.
- [9] M. Kim, G. Kim, S.-W. Lee, and J.-W. Ha, “St-bert: Cross-modal language model pre-training for end-to-end spoken language understanding,” in *ICASSP*, 2021, pp. 7478–7482.
- [10] C.-I. Lai, J. Cao, S. Bodapati, and S.-W. Li, “Towards semi-supervised semantics understanding from speech,” in *Workshop on Self-Supervised Learning for Speech and Audio Processing, NeurIPS*, 2020.
- [11] M. Rao, P. Dheram, G. Tiwari, A. Raju, J. Droppo, A. Rastrow, and A. Stolcke, “Do as i mean, not as i say: Sequence loss training for spoken language understanding,” in *ICASSP*, 2021, pp. 7473–7477.
- [12] E. Jang, S. Gu, and B. Poole, “Categorical reparametrization with gumble-softmax,” in *ICLR*, 2017.
- [13] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, “Slurp: A spoken language understanding resource package,” in *EMNLP*, 2020, pp. 7252–7262.
- [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 pre-training,” in *ICASSP*, 2021.
- [15] Y.-A. Chung, C. Zhu, and M. Zeng, “Splat: Speech-language joint pre-training for spoken language understanding,” in *NAACL*, 2021, pp. 1897–1907.
- [16] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Workshop on Deep Learning and Representation Learning, NeurIPS*, 2015.
- [17] M. Saxon, S. Choudhary, J. P. McKenna, and A. Mouchtaris, “End-to-end spoken language understanding for generalized voice assistants,” in *Proc. Interspeech*, 2021.
- [18] Y. Chen, W. Lu, A. Mottini, L. E. Li, J. Droppo, Z. Du, and B. Zeng, “Top-down attention in end-to-end spoken language understanding,” in *ICASSP*, 2021, pp. 6199–6203.
- [19] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *NeurIPS*, 2020.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.
- [21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [22] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tur, X. He, L. Heck, G. Tur, D. Yu, et al., “Using recurrent neural networks for slot filling in spoken language understanding,” *TASLP*, vol. 23, no. 3, pp. 530–539, 2014.
- [23] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *ICASSP*, 2016, pp. 4960–4964.
- [24] J. P. McKenna, S. Choudhary, M. Saxon, G. P. Strimel, and A. Mouchtaris, “Semantic complexity in end-to-end spoken language understanding,” in *Proc. Interspeech*, 2020.
- [25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*, 2015, pp. 5206–5210.
- [26] Y. Qian, X. Bianv, Y. Shi, N. Kanda, L. Shen, Z. Xiao, and M. Zeng, “Speech-language pre-training for end-to-end spoken language understanding,” in *ICASSP*, 2021, pp. 7458–7462.
- [27] L. Lugosch, B. H. Meyer, D. Nowrouzezahrai, and M. Ravanelli, “Using speech synthesis to train end-to-end spoken language understanding models,” in *ICASSP*, 2020, pp. 8499–8503.