Joint Automatic Speech Recognition And Structure Learning For Better Speech Understanding

Jiliang Hu[1], Zuchao Li[2,*], Mengjia Shen[3], Haojun Ai[1], Sheng Li[4], Jun Zhang[3]
[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education,
School of Cyber Science and Engineering, Wuhan University, Wuhan, China,
[2]School of Computer Science, Wuhan University, Wuhan, China,
[3]Wuhan Second Ship Design and Research Institute, Wuhan, China,
[4]National Institute of Information and Communications Technology, Japan.

***Abstract*—Spoken language understanding (SLU) is a structure prediction task in the field of speech. Recently, many works on SLU that treat it as a sequence-to-sequence task have achieved great success. However, This method is not suitable for simultaneous speech recognition and understanding. In this paper, we propose a joint speech recognition and structure learning framework (JSRSL), an end-to-end SLU model based on span, which can accurately transcribe speech and extract structured content simultaneously. We conduct experiments on name entity recognition and intent classification using the Chinese dataset AISHELL-NER and the English dataset SLURP. The results show that our proposed method not only outperforms the traditional sequence-to-sequence method in both transcription and extraction capabilities but also achieves state-of-the-art performance on the two datasets.**

***Index Terms*—Speech Recognition, Spoken Language Understanding, Information Extraction.**

I. Introduction

Automatic speech recognition (ASR) aims to convert human speech into text in the corresponding language [1]. On the other hand, SLU seeks to enhance machines' ability to comprehend and react to human language by extracting specific structured information from speech. SLU tasks encompass name entity recognition (NER), intent classification (IC), sentiment analysis (SA), among others [2]. SLU models can be categorized into two types: the pipeline model, which uses an ASR module first, followed by a natural language understanding (NLU) module in a sequential manner, and the end-to-end (E2E) model, which directly converts speech representations into structured information. The pipeline model encounters challenges related to error propagation due to the weak linkage between the two modules. Currently, there is a greater emphasis on the E2E model [3].

There has been significant progress in the field of SLU. However, few works consider how to improve or maintain the accuracy of transcription during spoken language understanding. For NER, most papers utilize the sequence-to-sequence method to achieve entity recognition. For instance, [4] use ”— *]*”, ”*$]*”, and ”{ *]*” to annotate *Person*, *Place*, and *Organization* in the transcribed text. Subsequently, the annotated text was used to train ASR model. However, [5] have shown that this method of labeling in the transcribed text can affect the recognition performance of the ASR model and increase transcription errors. For classification tasks such as SA and IC, some studies [6], [7] also achieve success by annotating classification types in the transcribed text, while others [8], [9] directly connect the output of the ASR model to a classification layer. The former will also reduce the performance of transcription, while the latter cannot transcribe the speech simultaneously. It is a significant challenge at the SLU to ensure accurate transcription by the ASR module and correct extraction by the SLU module simultaneously.

In this paper, we propose a joint speech recognition and structure learning framework (JSRSL), an E2E SLU model based on span, which can accurately recognize speech while extracting structured information. This model utilizes a refinement module to enhance text representations for structure prediction. Specifically, we adopt a parallel transformer for non-autoregressive end-to-end speech recognition as our framework and introduce the hidden layer output of the ASR decoder into a span for structure prediction. We additionally introduce a refinement module between the ASR model and the span to strengthen extraction performance. We conduct experiments on NER and IC using the AISHELL-NER and SLURP. The results indicate that the proposed idea has superior capabilities in speech transcription and speech understanding compared to the traditional sequence-to-sequence method. It also outperforms the current state-of-the-art (SOTA) on the two datasets.

II. Related Work

So far, NLU has been well developed, and many works have involved key tasks of natural language processing (NLP), such as natural language inference (NLI) [10], semantic role labeling (SRL) [11], [12], and SA [13], [14]. In recent years, a number of excellent SLU systems have emerged. [15] propose an acoustic model called TDT and they attempt to use the proposed model to do SLU tasks by jointly optimizing token

[*]Corresponding author.

This work was supported by the National Natural Science Foundation of China (No. 62306216), the Natural Science Foundation of Hubei Province of China (No. 2023AFB816), the Fundamental Research Funds for the Central Universities (No. 2042023kf0133).

Codes are available at https://github.com/193746/JSRSL prediction and temporal prediction. [8] introduce a novel method known as context-aware fine-tuning. They incorporate a c~~ontext module into the pre-trained~~ model to extract the context embedding vector, which is subsequently used as extra features for the ultimate prediction. [16] combine pre-trained self-supervised learning (SSL), ASR, language model (LM) and SLU models to explore the model combination that can achieve the best SLU performance and shows that pre-training approaches rooted in self-supervised learning are more potent than those grounded in supervised learning. [17] develop compositional end-to-end SLU systems that initially transform spoken utterances into a series of token representations, followed by token-level classification using a NLU system. [6] suggest a comprehensive approach that combines a multilingual pretrained speech model, a text model, and an adapter to enhance speech understanding. [18] incorporate semantics into speech encoders and present a task-agnostic unsupervised technique for integrating semantic information from large language models (LLMs) [19], [20] into self-supervised speech encoders without the need for labeled audio transcriptions.

III. Method

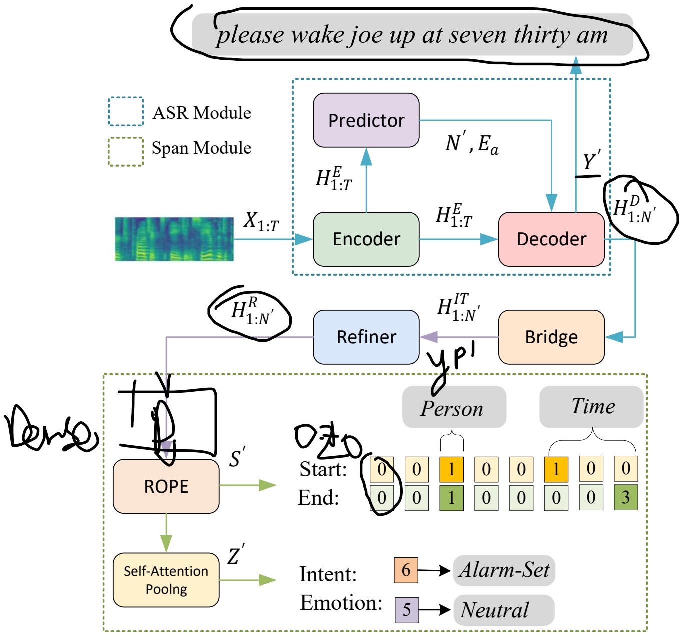

Fig. 1. The structure of our proposed framework, JSRSL.

Figure 1 shows the full overview of our proposed framework, JSRSL. Follow [21], we adopt this parallel Transformer for non-autoregressive end-to-end speech recognition as our ASR module framework. In addition, we use a *Refiner* module to connect the ASR module and span [22], and a *Bridge* module to connect the ASR module and *Refiner* module. The *Bridge* module can convert acoustic representation into text representation. The *Refiner* module can strengthen the text representation of upstream inputs and extract richer features for structured learning. Before doing structure learning, we apply rotary embedded encoding (ROPE) [23] into span. ROPE can leverage the boundary information of span by injecting relative position information.

A. SLU Representation Learning

Let X be a speech sequence with T frames, $X = \{x_1, x_2, x_3, \ldots, x_T\}$. Y is a sequence of tokens, and its length is N. Each token is in the vocabulary V, $Y = \{y_1, y_2, y_3, \ldots, y_N \mid y_i \in V\}$. The ASR module gets X as input and outputs the decoding result Y' and the *Decoder*'s hidden representation $H^D_{1:N'}$.

The *Refiner* module is essentially a NLU model, and it requires a text vector as input. It will not function correctly if the acoustic hidden representation $H^D_{1:N'}$ is fed directly. If we use the decoded token sequence as input, the computation graph will be truncated, preventing the joint training of the ASR module and *Refiner* module. Therefore, a *Bridge* module is needed to implement the conversion from acoustic hidden representation $H^D_{1:N'}$ to initial text hidden representation $H^{IT}_{1:N'}$. Our *Bridge* module first calculates the probability distribution of the token sequence Y^p based on $H^D_{1:N'}$, then it maps the distribution to the embedding layer of the *Refiner* module to convert Y^p into the initial text representation $H^{IT}_{1:N'}$. The specific operation of the map is to multiply Y^p and the weights of embedding layer $W^e_{1:V}$ in a matrix.

$$Y^p = Softmax(W^p_V H^D_{1:N'} + b^p_V)$$

$$H^{IT}_{1:N'} = Matmul(Y^p, W^e_{1:V})$$

V×1

Scale

Subsequently, we feed $H^{IT}_{1:N'}$ into the *Refiner* module for strengthening text representation, resulting in a refined text features $H^R_{1:N'}$. It is used as the input features for SLU downstream tasks. The *Refiner* is implemented of multiple stacked Transformer layers.

B. Span-based Structure Learning

We utilize span to extract the token sequence of each specific information in a given speech sequence X along with their ~~corresponding~~ types. Let S be the pointer sequence, the set of pointer labels be Ω, ~~and~~ $S = \{s_1, s_2, s_3, \ldots, s_N \mid s_i \in \Omega\}$. The pointer sequence corresponds one-to-one to the token sequence obtained from the speech transcription. We define two pointer sequences, namely the starting pointer sequence S_{start} and the ending pointer sequence S_{end}. The former is used to locate the head position of the specific information, $\Omega_{start} = \{0, 1\}$, and the latter is used to locate the tail position of the specific information and identify its type, $\Omega_{end} = \{0, 1, 2 \ldots N_{type}\}$.

For ordinary structure extraction task, we firstly use a linear layer called *Dense* to double the dimension of $H^R_{1:N'}$, which is aimed at make two pointers use different feature vectors. The output of *Dense* is evenly divided into two parts, d_1 and d_2. Then, ROPE is employed to embed relative position information for the pointer sequence.

$$d_1, d_2 = Dense(H^R_{1:N'})$$

t1m m×2n

d1 m×n

d2 m×n

$$S_{start}^{'}, S_{end}^{'} = ROPE(d_1), ROPE(d_2)$$

For classification tasks, we treat them as sentence-level span. Let Z be the classification result, and the set of classification result be Ω_{cls}, $\Omega_{cls} = \{1, 2, \ldots, N_{type}\}$. Firstly, we use ROPE to embed relative position information and get the token-level classification features: $Z_{1:N'}$.

$$Z_{1:N'} = ROPE(H_{1:N'}^{R})$$

Drawing inspiration from [24], a modified self-attention pooling mechanism is utilized to comprehensively extract sentence-level classification result $Z^{'}$ from token-level features. The process of the self-attention pooling mechanism is as follows, $A_{1:N'}^{Z}$ is attention value of $Z_{1:N'}$ calculated by a linear layer.

Q K

$$A_{1:N'}^{Z} = Softmax(W_1 Z_{1:N'} + b_1)$$

$$Z^{'} = \arg\max_{N_{type}}(W_{N_{type}}(\sum_{N'} A_{1:N'}^{Z} \times Z_{1:N'}) + b_{N_{type}})$$

V

C. Joint Optimization

The adopted ASR module calculates three loss functions when training: the cross-entropy (CE), the mean absolute error (MAE), and the minimum word error rate (MWER) loss. CE and MWER are used to optimize the model's transcription ability, while MAE guides the predictor to convergence. According to [21], the loss function of the ASR part is:

$$\mathcal{L}_{asr} = \gamma\mathcal{L}_{CE} + \mathcal{L}_{MAE} + \mathcal{L}_{werr}^{N}(x, y^*)$$

$$\mathcal{L}_{werr}^{N}(x, y^*) = \sum_{y_i \in Sample} \widehat{P}(y_i \mid x)[\mathcal{W}(y_i, y^*) - \widehat{W}]$$

We also use CE to optimize the model's ability for structure prediction. The total loss function can be formulated as follows, where α is used to control the proportion of ASR loss and structured loss, $\alpha \in (0, 1)$.

$$\mathcal{L}_{total} = (1 - \alpha)\mathcal{L}_{asr} + \alpha\mathcal{L}_{sp}$$

IV. Experiment

A. Configuration

We employ AISHELL-NER [5] and SLURP's NER (slot filling) subset for NER, and SLURP's IC subset for IC [25]. Following the previous work [26] that used SLURP, the provided synthetic data is utilized for training. For assessment metrics, SLURP-NER utilizes WER and SLURP-F1, SLURP-IC employs WER and micro-F1, while AISHELL-NER uses CER and micro-F1. Our baseline, Seq2Seq, applies the sequence-to-sequence method to the JSRSL's ASR module and annotates like [5]. Another baseline, Pipe, is the concatenation of the JSRSL's ASR module and a NLU module, predicting structure by span. The models chosen as benchmarks all adopt sequence generation method by adding special symbols in transcribed text or generating structured sequence directly.

We build the experiment environment based on Funasr [27] and ModelScope, utilizing 220M Chinese and English Paraformer for experiments. The pre-training parameters of BERT-base [28] is adopted to initialize the *Refiner* of JSRSL and the NLU module of Pipe. The α of the loss is set to 0.5. The models are trained until converges. We consistently use the Adam optimizer with a learning rate of 5e-5.

B. Main Result

TABLE I
COMPARISON RESULTS WITH BENCHMARKS IN AISHELL-NER AND SLURP. IN SLURP, ALL MODELS CONDUCT JOINT TRAINING AND EVALUATION FOR NER AND IC TASKS. THE BERT IN [5]'S PIPELINE MODEL ARE PRE-TRAINED. THE POST AED USES 1K EXTERNAL DATA.

Model	Paradigm	AISHELL-NER	
		CER (↓)	F1 (↑)
Transformer [5]	E2E	9.25	64.34
Conformer [5]	E2E	4.79	73.37
Transformer+BERT* [5]	Pipeline	9.25	65.95
Conformer+BERT* [5]	Pipeline	4.79	74.90
Seq2Seq (Ours)	E2E	3.48	76.17
Pipe (Ours)	Pipeline	1.76	77.98
JSRSL (Ours)	E2E	**1.71**	**80.85**

Model	Paradigm	SLURP		
		WER (↓)	SLU F1 (↑)	IC F1 (↑)
RNNT→BertNLU [29]	Pipeline	15.20	72.35	82.45
Espnet SLU [26]	E2E	-	71.90	86.30
PostDec AED* [6]	E2E	-	80.08	89.33
Conformer-RNNT [7]	E2E	14.20	77.22	90.10
Conformer-CTC [7]	E2E	14.50	69.30	85.50
TDT 0-6 [15]	E2E	-	**80.61**	89.28
TDT 0-8 [15]	E2E	-	79.90	90.07
Seq2Seq (Ours)	E2E	18.83	68.84	86.63
Pipe (Ours)	Pipeline	13.61	76.62	86.67
JSRSL (Ours)	E2E	**13.14**	**80.17**	**91.03**

Table I presents the comparison results with baselines and benchmarks. Before the joint training of Paraformer and BERT, the Paraformer in JSRSL is pre-trained by audio-text pairs from AISHELL-NER or SLURP and the BERT is pre-trained by the vocabulary of Paraformer and the text-span pairs of corresponding dataset. The results show that the proposed model, based on span, outperforms the E2E and pipeline baselines and benchmarks using the sequence-to-sequence method in both structure extraction and ASR. By using span, we can distinguish between ASR and SLU tasks, avoid the impact of extra structured annotations on transcription accuracy and allow for more precise extraction of structured information. In addition, Pipe is worse than JSRSL due to issues with error propagation of pipeline model, indicating that the E2E model is more promising. In SLURP, TDT models and PostDEC AED have the best SLU performance in the benchmarks. TDT models adopt the token and duration transformer architecture, which enhances the ability of sequence generation and accelerates the inference speed of the original RNN-Transducer [30] through joint learning of token and duration prediction. "0-X" refers to the maximum duration of duration configurations as "X". The performance of TDT is comparable to that of our JSRSL, but it is influenced by specific configurations. TDT 0-6 performs better on SLURP-

TABLE II
ABLATION RESULTS IN AISHELL-NER AND SLURP. SAMPLER IS A COMPONENT IN PARAFORMER USED TO ENHANCE TRAINING EFFECTIVENESS.

Model		AISHELL-NER				SLURP-NER				SLURP-IC	
	Refiner	CER (↓)	Precision (↑)	Recall (↑)	F1 (↑)	WER (↓)	Precision (↑)	Recall (↑)	SLU F1 (↑)	WER (↓)	F1 (↑)
JSRSL	×	1.76	**83.21**	**78.15**	**80.46**	10.79	**80.25**	73.37	**76.66**	**13.84**	**88.19**
w/o ROPE	×	1.78	82.50	78.00	80.06	10.81	79.33	**73.44**	76.27	13.87	87.93
w/o Sampler	×	**1.75**	82.98	78.00	80.32	**10.72**	80.14	73.20	76.51	14.05	**88.19**
JSRSL	✓	1.77	81.89	79.27	80.45	**11.17**	**83.26**	**77.46**	**80.26**	**13.30**	**89.90**
w/o pretrain	✓	**1.71**	**82.36**	**79.54**	**80.85**	12.17	82.41	76.36	79.27	13.53	89.67

NER, while TDT 0-8 performs better on SLURP-IC. PostDec AED uses an adaptor to concatenate the speech encoder and NLU model. The adaptor tries to minimize the distance between the speech representation and the text representation. Before finetune the SLU system, it must undergo adapter pretraining, which optimizes the loss between predicted text and real text. Although PostDec AED's performance is just slightly inferior to our JSRSL, its training process is more complex and requires a large amount of external data.

C. Ablation Study

TABLE III
ASR RESULTS (CER ON AISHELL-NER AND WER ON SLURP) AMONG PARAFORMER, WHISPER, XLSR-53 AND OUR JSRSL WITH SIMILAR SIZE.

Model	Params	AISHELL-NER	SLURP
XLSR-53 [31]	315M	4.12	22.10
Whisper [32]	244M	5.54	16.89
Paraformer	220M	**1.76**	**13.61**
JSRSL (Ours)	314M	**1.71**	**13.14**

Table II shows the ablation results. The proposed model shows a certain improvement after introducing ROPE. In addition, in SLURP-NER and AISHELL-NER, we find that *Sampler* does not significantly improve model's ASR performance, but rather increases its spoken understanding ability. In AISHELL-NER and SLURP-IC, the introduction of refinement module leads to a further decrease in the model's CER or WER and an increase in F1 score. However, in SLURP-NER, we find that with refinement module, while JSRSL enhances SLU F1, WER increases. But overall, the introduction of refinement modules is beneficial for improving model performance. In AISHELL-NER, the JSRSL with pre-trained Paraformer does not achieve better performance, possibly due to overfitting. Nevertheless, in SLURP, pre-trained ASR module improves the overall ability of JSRSL. Additionally, to assess the effectiveness of the adopted ASR module, we carry out an ASR experiment detailed in Table III. The results indicate that Paraformer's ASR performance on AISHELL-NER and SLURP surpasses that of Whisper and XLSR-53. This demonstrates the robustness of our model's ASR module. Furthermore, JSRSL achieves better ASR capability after joint training.

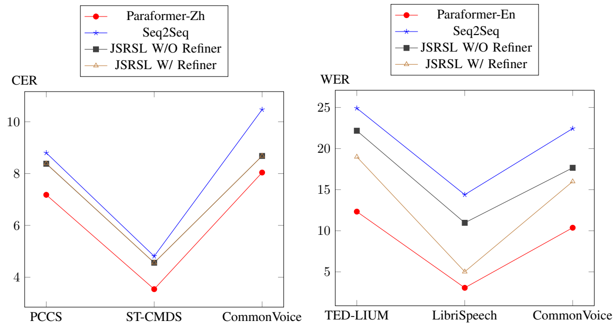

Fig. 2. Out-of-Distribution experiment result. The results on the left are in the Chinese datasets and the results on the right are in the English datasets.

D. OOD Analysis

This section conducts Out-of-Distribution (OOD) experiment with unseen ASR datasets. The Chinese ASR datasets are Primewords Chinese Corpus Set 1 (PCCS) [33], Free ST Chinese Mandarin Corpus (ST-CMDS), and Common Voice (Zh) [34]. The English ASR datasets include Librispeech [35], Ted-LIUM [36], and Common Voice (En). Each dataset uniformly uses the first 8000 training samples for the experiment. When evaluating the Seq2Seq, the special symbols that indicate name entities are ignored. Figure 2 presents the OOD results. The original Chinese or English Paraformer demonstrates better generalization ability and achieves the lowest CER or WER among all datasets. Although all trained models' transcription performance has decreased, compared to Seq2Seq, which uses a sequence generation method, the span-based models exhibit lower recognition error rates. What's more, JSRSL with refinement module performs better than that without refinement module. This indicates that the proposed method results in less performance loss in speech recognition, making it more suitable for joint ASR and structure prediction.

V. CONCLUSION

This paper proposes a method called JSRSL for jointly ASR and structure prediction based on span. This approach aims to ensure accurate transcribing and understanding of speech simultaneously. Experiments have shown that the proposed scheme is superior to traditional sequence-to-sequence method in both transcription and extraction capabilities, and achieves SOTA performance on the datasets used for the experiments.

References

[1] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[2] S. Arora, H. Futami, J.-w. Jung, Y. Peng, R. Sharma, Y. Kashiwagi, E. Tsunoo, K. Livescu, and S. Watanabe, "Universlu: Universal spoken language understanding for diverse tasks with natural language instructions," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 2754–2774.

[3] A. Pasad, F. Wu, S. Shon, K. Livescu, and K. Han, "On the use of external data for spoken named entity recognition," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, Conference Proceedings, pp. 724–737.

[4] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, "End-to-end named entity recognition from english speech," *Organization*, vol. 2299, no. 1473, p. 3772, 2020.

[5] B. Chen, G. Xu, X. Wang, P. Xie, M. Zhang, and F. Huang, "Aishell-ner: Named entity recognition from chinese speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, Conference Proceedings, pp. 8352–8356.

[6] P. Denisov and N. T. Vu, "Leveraging multilingual self-supervised pretrained models for sequence-to-sequence end-to-end spoken language understanding," *arXiv preprint arXiv:2310.06103*, 2023.

[7] K. Singla, S. Jalalvand, Y.-J. Kim, A. Moreno Daniel, S. Bangalore, A. Ljolje, and B. Stern, "1spu: 1-step speech processing unit," *arXiv e-prints*, p. arXiv: 2311.04753, 2023.

[8] S. Shon, F. Wu, K. Kim, P. Sridhar, K. Livescu, and S. Watanabe, "Context-aware fine-tuning of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2023, Conference Proceedings, pp. 1–5.

[9] S. Shon, K. Kim, P. Sridhar, Y.-T. Hsu, S. Watanabe, and K. Livescu, "Generative context-aware fine-tuning of self-supervised speech models," *arXiv preprint arXiv:2312.09895*, 2023.

[10] Z. Zhang, Y. Wu, H. Zhao, Z. Li, S. Zhang, X. Zhou, and X. Zhou, "Semantics-aware bert for language understanding," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 9628–9635.

[11] S. He, Z. Li, H. Zhao, and H. Bai, "Syntax for semantic role labeling, to be, or not to be," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2061–2071.

[12] Z. Li, S. He, H. Zhao, Y. Zhang, Z. Zhang, X. Zhou, and X. Zhou, "Dependency or span, end-to-end uniform semantic role labeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6730–6737.

[13] H. Jing, Z. Li, H. Zhao, and S. Jiang, "Seeking common but distinguishing difference, a joint aspect-based sentiment analysis model," *arXiv preprint arXiv:2111.09634*, 2021.

[14] L. Peng, Z. Li, and H. Zhao, "Sparse fuzzy attention for structured sentiment analysis," *arXiv preprint arXiv:2109.06719*, 2021.

[15] H. Xu, F. Jia, S. Majumdar, H. Huang, S. Watanabe, and B. Ginsburg, "Efficient sequence transduction by jointly predicting tokens and durations," *ArXiv*, vol. abs/2304.06795, 2023.

[16] Y. Peng, S. Arora, Y. Higuchi, Y. Ueda, S. Kumar, K. Ganesan, S. Dalmia, X. Chang, and S. Watanabe, "A study on the integration of pre-trained ssl, asr, lm and slu models for spoken language understanding," in *2022 IEEE Spoken Language Technology Workshop (SLT).* IEEE, 2023, Conference Proceedings, pp. 406–413.

[17] S. Arora, S. Dalmia, B. Yan, F. Metze, A. W. Black, and S. Watanabe, "Token-level sequence labeling for spoken language understanding using compositional end-to-end models," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, Conference Proceedings, pp. 5419–5429.

[18] D. Xu, S. Dong, C. Wang, S. Kim, Z. Lin, A. Shrivastava, S.-W. Li, L.-H. Tseng, A. Baevski, G.-T. Lin, H.-y. Lee, Y. Sun, and W. Wang, "Introducing semantics into speech encoders," in *Annual Meeting of the Association for Computational Linguistics*, 2022, Conference Proceedings.

[19] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] Z. Li, S. Zhang, H. Zhao, Y. Yang, and D. Yang, "Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer," *arXiv preprint arXiv:2307.00360*, 2023.

[21] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," *arXiv preprint arXiv:2206.08317*, 2022.

[22] J. Fu, X.-J. Huang, and P. Liu, "Spanner: Named entity re-/recognition as span prediction," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, Conference Proceedings, pp. 7183–7195.

[23] J. Su, A. Murtadha, S. Pan, J. Hou, J. Sun, W. Huang, B. Wen, and Y. Liu, "Global pointer: Novel efficient span-based approach for named entity recognition," *arXiv preprint arXiv:2208.03054*, 2022.

[24] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, Conference Proceedings, pp. 7927–7931.

[25] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "Slurp: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, Conference Proceedings, pp. 7252–7262.

[26] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, and B. Yan, "Espnet-slu: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2022, Conference Proceedings, pp. 7167–7171.

[27] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, and Z. Xiao, "Funasr: A fundamental end-to-end speech recognition toolkit," *arXiv preprint arXiv:2305.11013*, 2023.

[28] 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.

[29] A. Raju, M. Rao, G. Tiwari, P. Dheram, B. Anderson, Z. Zhang, C. Lee, B. Bui, and A. Rastrow, "On joint training with interfaces for spoken language understanding," *arXiv preprint arXiv:2106.15919*, 2021.

[30] M. Ghodsi, X. Liu, J. Apfel, R. Cabrera, and E. Weinstein, "Rnn-transducer with stateless prediction network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 7049–7053.

[31] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning.* PMLR, 2023, Conference Proceedings, pp. 28 492–28 518.

[33] L. Primewords Information Technology Co., "Primewords chinese corpus set 1," 2018, https://www.primewords.cn.

[34] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[35] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP).* IEEE, 2015, pp. 5206–5210.

[36] A. Rousseau, P. Deléglise, and Y. Esteve, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.