

# CHINESE ASR AND NER IMPROVEMENT BASED ON WHISPER FINE-TUNING

Hao Yang, Min Zhang, Shimin Tao, Miaomiao Ma, Ying Qin

yanghao30@huawei.com

2012 Labs, Huawei Technologies CO., LTD  
Beijing, China

**Index Terms**—Chinese ASR, Chinese NER, Whisper fine-tuning

## ABSTRACT

Based on 680k hours of weakly supervised multilingual and multi-task speech transcription/translation data, Whisper[1] has developed a robust system for both Automated Speech Recognition (ASR) and Speech Translation (ST). Whisper provides a simple model architecture based on Mel spectrum + two-layer convolution + Seq2seq Transformer architecture, which is easy to fine-tune on conditional generation tasks. This paper analyzes how to fine-tune Chinese ASR[2] and NER tasks based on Whisper, including (1) how to design different prompts for different generative tasks; (2) how to train ASR and NER tasks at the same time; (3) whether the performance can be further improved by using weak supervision for data enhancement. Experiments based on AISHELL[3] and AISHELL-NER[4] data, and multi-task fine-tuning based on Whisper can effectively improve the performance of ASR and NER.

## I. INTRODUCTION

With the rapid development of Transformer models and pre-trained models, the ASR models with SOTA performance include Conformer[5] [6], U2[7] [8], Wav2Vec 2.0[9], HuBERT[10], Whisper and so on. Conformer is comprised of two macaron-like feed-forward layers (FFN) with half-step residual connections sandwiching the multi-headed self attention and convolution modules, which is followed by a post layernorm. Conformer achieves WER of 2.1 without using a language model and 1.9 with an external language model on LibriSpeech test datasets. Wav2Vec 2.0 is the first model that proves learning powerful representations from speech audio alone followed by fine-tuning on transcribed speech can outperform the best semi-supervised methods while being conceptually simpler. Wav2Vec 2.0 masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned. And experiments using all labeled data of LibriSpeech achieve 1.8 WER on the clean test sets.

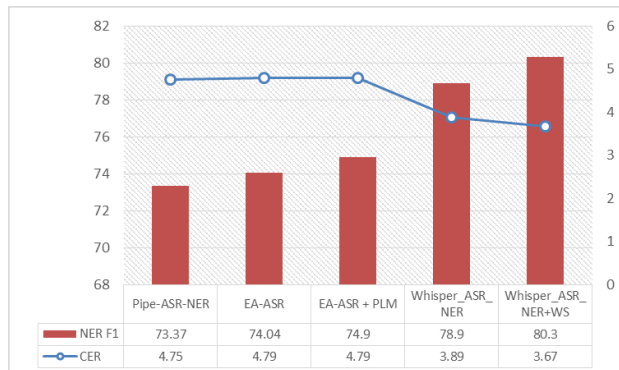


**Fig. 1.** ASR and NER performance comparison, PIPE-ASR-NER is a pipeline system with ASR and NER[2]. EA-ASR is E2E entity-aware ASR, [4]. EA-ASR + PLM is E2E entity-aware ASR with PLM to enhance NER performance[4]. Whisper-ASR-NER is multitask prompt finetuning based on Whisper. And Whisper-ASR-NER-WS is weak supervision augmented with CommonVoice. Note that a lower CER and higher NER F1 score indicate better performance.

When lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100-hour subset while using 100 times less labeled data. Using just ten minutes of labeled data and pre-training on 53k hours of unlabeled data still achieves 4.8 WER.

With the development of deep learning speech technology, Speech NER is another very attractive technology direction. Speech NER has two typical architectures, pipeline architecture and end-to-end (E2E) architecture. The pipeline architecture with two steps is implemented by cascading the ASR and NER models [3] [2]. Firstly, the ASR system converts the speech into a text sequence, and then the NER system marks NEs based on the sequence. The typical problem of this cascaded system is error propagation and amplification [11][12]. If the text output by the ASR is too noisy, the entity recognition accuracy will be greatly reduced. [4] show that the NER F1 score of this cascaded

model is 73%, while the NER F1 score of the ground truth text is 93%. In the E2E speech NER architecture, the NER is implemented by tagging spans with braces, and speech NER training is just a sequence-to-sequence process [13]. Data scarcity is the major problem of the E2E speech NER architecture, whose NER F1 score is 69.47%. Applying the pre-trained language model (PLM) can improve E2E speech NER by using a large amount of pre-training data, increasing the NER F1 score from 69.47% to 74.9% [14][15]. However, the E2E ASR-NER model has a negative influence on the performance of ASR, increasing the CER from 4.75% to 4.79%. As pointed out by [4], in terms of E2E speech NER, both ASR and NER scores are not good enough due to data scarcity.

Different from Wav2vec2 with unsupervised data, Whisper[1] achieves human level performance in ASR with 680k hours of weakly supervised data. It is built on the transformer architecture, stacking encoder blocks and decoder blocks with the attention mechanism propagating information between both. It will take the audio recording, split it into 30-second chunks and process them one by one. For each 30-second recording, it will encode the audio using the encoder section and save the position of each word said, and leverage this encoded information to find what was said using the decoder. Whisper was able to achieve generalization using a zero-shot performance and exhibited about 50% more robustness and fewer errors than supervised ASR models such as SOTA on CoVoST2. Even more impressive is the fact that Whisper was able to achieve human-level performance across diverse ASR tasks on different languages.

The SOTA effect of ASR is obtained by using Whisper multi-task learning. We try to fine-tune multi-tasking on the basis of Whisper. We achieved optimal ASR and NER performance through normal speech recognition tasks, speech entity tasks with Prompt, and weakly supervised speech recognition and speech entity tasks. In general, the innovations of this article include:

- 1) In addition to the normal ASR fine-tuning of Wav-Text, the Speech NER task is fine-tuned in Prompt mode.
- 2) Through the multi-task learning of ASR and Speech NER, the experimental performance is better than that of single tasks.
- 3) By strengthening the data of weak supervision, the performance is greatly improved, and the multi-task fine-tuning and weak supervision can effectively improve the robustness of the model.

## II. RELATED WORK

### II-A. ASR Model

For ASR models, a sequence of speech is input and a sequence of text is predicted. With deep learning and Transformer development, the ASR can be efficiently modeled

using Transformer [16] or Conformer [17]. The loss function can be enhanced by the CTC/Attention joint approach and Language Model (LM) [18]. The ASR function can be described with speech sequence  $X$  and text sequence  $Y$ .

$$\begin{aligned} X &= [x_1, x_2, \dots, x_m] \\ Y &= [y_1, y_2, \dots, y_n] \end{aligned} \quad (1)$$

$$P(Y|X) = \prod_h P(y_t|X, y_1, \dots, y_{t-1}) \quad (2)$$

$$\begin{aligned} \mathcal{L} &= -\lambda \log P_{CTC}(Y|X) - (1 - \lambda) \log P_{Attn}(Y|X) \\ &\quad - \theta \log P_{LM}(Y) \end{aligned} \quad (3)$$

### II-B. NER Model

NER is usually modeled as sequence tagging problem[19][13]. In the BIO tagging format, the B- prefix tags tokens that are at the beginning of an entity, I- tags tokens inside an entity, while o- tags tokens that are not part of any entity. Given tokenized sequence  $Y$  and NER Tagging  $T$ , the NER model can be depicted as:

$$\begin{aligned} Y &= [y_1, y_2, \dots, y_n] \\ T &= [t_1, t_2, \dots, t_n] \end{aligned} \quad (4)$$

$$P(T|Y) = \prod_t P_{BIO}(T_t|Y, T_1, \dots, T_{t-1}, T_{t+1}, \dots, T_n) \quad (5)$$

$$\mathcal{L} = CrossEntropy(T_{pred}, T_{groud\_truth}) \quad (6)$$

### II-C. Multi-task Learning

With unified input and output formats[20], ASR and NER tasks are fine-tuned together like multi-task learning. For simplicity, two losses are added with a threshold  $\lambda$

$$\mathcal{L} = \lambda * \mathcal{L}_{ASR} + (1 - \lambda) * \mathcal{L}_{NER} \quad (7)$$

## III. PROPOSED APPROACH

Because Whisper is a prompt-based multi-task trained model, the multi-task mode needs to be fine-tuned based on Whisper. On the one hand, a task is converted into a sequence-to-sequence task, and on the other hand, a prompt condition needs to be designed for different tasks.

### III-A. Prompt ASR Fine-tuning in Chinese

Whisper improves language prompt and transcript/translation prompt. For Chinese Prompt ASR fine-tuning, only the prompt header needs to be added to the tokenizer.

For example,  $\langle |zh| \rangle \langle |transcript| \rangle$  英特尔首席执行官布莱恩·克扎尼奇表示，北京携手张家口申办冬奥会。 , which means Brian Kerzanich, Intel's chief executive, said, Beijing joins hands Zhangjiakou bid to host the Winter Olympic Games.

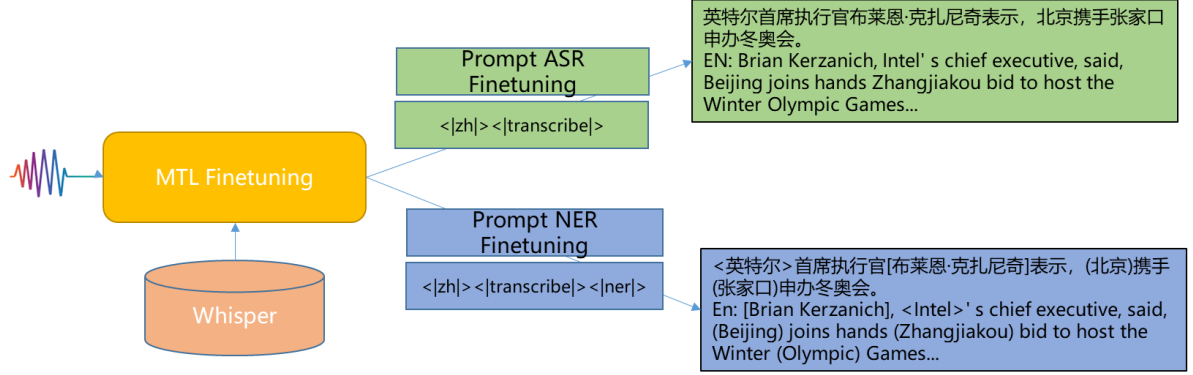


Fig. 2. Multi-task prompt fine-tuning for Chinese ASR and NER based on Whisper

### III-B. Prompt NER Fine-tuning in Chinese

Whisper is used to fine-tune the Speech NER data. First, we change the Sequence tagging mode of the NER data to the sequence-to-sequence mode. Based on research methods such as AISHELL-NER, we can add `[]`, `<>`, and `()` before and after the entity to identify *PER*/*ORG*/*LOC* entities. To fine-tune the prompt, we add the prompt header `<|zh|> <|transcript|> <|ner|>` to the tokenizer.

For example, `<|zh|> <|transcript|> <|ner|>` 英特尔首席执行官布莱恩·克扎尼奇表示，北京携手张家口申办冬奥会。 , which means `[BrianKerzanich], <Intel>'s chief executive, said, (Beijing) joins hands (Zhangjiakou) bid to host the Winter (Olympic) Games.`

Brian Kerzanich is *PER*, Intel is *ORG*, and Beijing, Zhangjiakou and Olympic are *LOCs*.

### III-C. Weak Supervision Augmentation

We have used different ASR data and Speech NER data to perform mixed fine-tuning to test whether different tasks can improve each other's performance.

## IV. EXPERIMENTS

For Chinese ASR and Speech NER experiments, the comparison methods include (1) Conformer ASR, (2) Pipeline-ASR-NER [4], which is a pipeline speech NER system, (3) EA-ASR [4], which is an E2E Speech NER system, (4) EA-ASR+PLM, augmented with PLM (BERT) to improve NER accuracy.

### IV-A. ASR NER and Weak Supervision Datasets

**ASR and Speech NER Dataset** ASR dataset is AISHELL-1, which contains 150-hour speech data <sup>1</sup>. And the related speech NER dataset is provided by [4] in 2022, which contains 15k *PER*, 20k *LOC* and 21k *ORG* entities <sup>2</sup>. NER tags are initially added by NER models, and then manually checked and confirmed by professional language experts. For detailed data, see Table I.

<sup>1</sup><http://www.openslr.org/33/>

<sup>2</sup><https://github.com/Alibaba-NLP/AISHELL-NER/tree/main/data>

|       | Sentence | PER    | LOC    | ORG    |
|-------|----------|--------|--------|--------|
| Train | 120,098  | 15,842 | 20,693 | 21,455 |
| Dev   | 14,326   | 1,902  | 2,588  | 2,731  |
| Test  | 7,176    | 898    | 1,330  | 1,165  |

Table I. AIShellNER dataset distribution

**Weak Supervision Dataset** The Common Voice dataset consists of a unique MP3 and corresponding text file. Many of the 24,210 recorded hours in the dataset also include demographic metadata like age, gender, and accent that can help improve the accuracy of speech recognition engines. We choose 'ZH-CN' for our weak supervision dataset in Table II. As it is only used as a weak supervision dataset, no NEs are tagged.

|       | Sentence | PER | LOC | ORG |
|-------|----------|-----|-----|-----|
| Train | 29,056   | -   | -   | -   |
| Dev   | 10,581   | -   | -   | -   |
| Test  | 10,581   | -   | -   | -   |

Table II. CommonVoice-11.0, ZH-CN dataset

**Speech NER Metrics** Similar to the works [4] [21] [22], CER/WER is used as the ASR accuracy metric and F1 as the NER accuracy metric. For CER/WER metrics, D, I, S indicate the number of characters to be deleted, inserted and substituted respectively. N indicates the number of all characters. For F1 metrics, precision and recall are used to indicate correct percentage upon ground truth and predicted truth respectively. Our experiments select Micro average because of class imbalance as shown in Table I.

$$CER = (D + I + S)/N \quad (8)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

### IV-B. Speech NER Performance comparison

According to the performance comparison results in Fig 1, Whisper-ASR-NER (CER 3.89, NER F1 78.9%) outperforms Pipeline-ASR-NER (CER 4.75, NER F1 73.37%) and

EA-ASR (CER 4.79, NER F1 74.04). Weak supervision can also help improve CER and NER accuracy, with CER decreased to 3.67 while NER F1 promoted to 80.3%.

#### IV-C. Fine-tuning ASR Results

##### IV-C1. Categorized NER Results

Like [4] has done, 3 different NE categories, including PER, LOC and ORG, are examined in Table III. Our PER and ORG F1 is SOTA, especially PER F1, which is significantly improved by 27%.

| Model           | NER | Precision | Recall | F1           |
|-----------------|-----|-----------|--------|--------------|
| GT              | PER | 95.75     | 95.43  | 95.59        |
| +               | LOC | 94.54     | 95.18  | 94.86        |
| BERT            | ORG | 90.90     | 88.32  | 89.59        |
| EA-ASR          | PER | 58.18     | 55.44  | 56.78        |
| +               | LOC | 86.05     | 86.11  | <b>86.08</b> |
| PLM(BERT)       | ORG | 75.59     | 74.50  | 75.04        |
| Whisper-ASR-NER | PER | 78.31     | 79.73  | <b>78.83</b> |
| +               | LOC | 82.39     | 83.53  | 82.93        |
| WS              | ORG | 76.88     | 75.41  | <b>76.27</b> |

Table III. Categorized NER Performance

#### IV-D. Best Epoch Selection

As can be seen in Table IV, Whisper-ASR-NER + WS can achieve best performance with just 7 epoches, with CER 3.67 and NER F1 80.30%.

| Epoch | F1           | WER         |
|-------|--------------|-------------|
| 0     | 0.733        | 6.63        |
| 1     | 0.741        | 6.42        |
| 2     | 0.767        | 6.19        |
| 3     | 0.773        | 5.78        |
| 4     | 0.782        | 5.32        |
| 5     | 0.791        | 4.96        |
| 6     | 0.795        | 4.54        |
| 7     | <b>0.803</b> | <b>3.67</b> |
| 8     | 0.798        | 3.89        |
| 9     | 0.785        | 3.92        |
| 10    | 0.76         | 4.02        |

Table IV. Best epoch selection

## V. CONCLUSION

This paper proposes multi-task prompt based fine-tuning with Whisper, to achieve good performance both on ASR and NER. Experiments show multi-task fine-tuning can help both tasks promote the performance. Like Whisper, this policy is compatible with the data-enhanced pattern. Our future works will focus on fine-tuning with more languages and more speech tasks[23].

## VI. REFERENCES

- [1] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.
- [2] Alan Akbik, Tanja Bergmann, and Roland Vollgraf, “Pooled contextualized embeddings for named entity recognition,” in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, p. 724–728.
- [3] Alan Akbik, Duncan Blythe, and Roland Vollgraf, “Contextual string embeddings for sequence labeling,” in *COLING 2018, 27th International Conference on Computational Linguistics*, 2018, pp. 1638–1649.
- [4] Boli Chen, Guangwei Xu, Xiaobin Wang, Pengjun Xie, Meishan Zhang, and Fei Huang, “AISHELL-NER: Named Entity Recognition from Chinese Speech,” Feb. 2022, arXiv:2202.08533 [cs].
- [5] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, “Conformer: Convolution-augmented transformer for speech recognition,” 2020.
- [6] Yinglu Li, Minghan Wang, Jiaxin Guo, Xiaosong Qiao, Yuxia Wang, Daimeng Wei, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying Qin, “The HW-TSC’s Offline 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), 2022, pp. 239–246, Association for Computational Linguistics.
- [7] Minghan Wang, Jiaxin Guo, Yuxia Wang, Daimeng Wei, Hengchao Shang, Chang Su, Yimeng Chen, Yinglu Li, Min Zhang, Shimin Tao, and Hao Yang, “Di-former: Directional Transformer for Neural Machine Translation,” Dec. 2021, arXiv:2112.11632 [cs].
- [8] Minghan Wang, Jiaxin Guo, Yinglu Li, Xiaosong Qiao, Yuxia Wang, Zongyao Li, Chang Su, Yimeng Chen, Min Zhang, Shimin Tao, Hao Yang, and Ying 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), 2022, pp. 247–254, Association for Computational Linguistics.
- [9] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations,” *arXiv:2006.11477 [cs, eess]*, Oct. 2020, arXiv: 2006.11477.
- [10] Benjamin van Niekirk, Marc-Andre Carbonneau, Julian Zaidi, Matthew Baas, Hugo Seute, and Herman 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, May 2022, pp. 6562–6566, IEEE.
- [11] Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn

- Shah, “End-to-End Named Entity Recognition from English Speech,” in *Proc. Interspeech 2020*, 2020, pp. 4268–4272.
- [12] Hao Yang, Shimin Tao, Minghan Wang, Min Zhang, Daimeng Wei, Shuai Zhao, Miaomiao Ma, and Ying Qin, “CCDC: A Chinese-Centric Cross Domain Contrastive Learning Framework,” in *Knowledge Science, Engineering and Management*, Gerard Memmi, Baijian Yang, Linghe Kong, Tianwei Zhang, and Meikang Qiu, Eds., Cham, 2022, pp. 225–236, Springer International Publishing.
- [13] Hao Yang, Gengui Xie, Ying Qin, and Song Peng, “Domain Specific NMT based on Knowledge Graph Embedding and Attention,” in *2019 21st International Conference on Advanced Communication Technology (ICACT)*. 2019, pp. 516–521, IEEE.
- [14] Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang, “FLAT: Chinese NER using flat-lattice transformer,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 6836–6842.
- [15] Hao Yang, Yao Deng, Minghan Wang, Ying Qin, and Shiliang Sun, “Humor Detection based on Paragraph Decomposition and BERT Fine-Tuning,” in *AAAI Workshop 2020*, 2020.
- [16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, “Conformer: Convolution-augmented Transformer for Speech Recognition,” in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.
- [18] Takaaki Hori, Shinji Watanabe, Yu Zhang, and William Chan, “Advances in joint ctc-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM,” in *Proc. Interspeech 2017*, 2017, pp. 949–953.
- [19] Yang Hao, Chen Junliang, Meng Xiangwu, and Qiu Bingyu, “Dynamically traveling web service clustering based on spatial and temporal aspects,” 2007, pp. 348–357, Springer.
- [20] Yue Zhang and Jie Yang, “Chinese NER using lattice LSTM,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1554–1564.
- [21] Sahar Ghannay, Antoine Caubriere, Yannick Esteve, Antoine Laurent, and Emmanuel Morin, “End-to-end named entity extraction from speech,” *arXiv preprint arXiv:1805.12045*, 2018.
- [22] Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah, “End-to-end Named Entity Recognition from English Speech,” May 2020, arXiv:2005.11184 [cs, eess].
- [23] Guan-Ting Lin, Chan-Jan Hsu, Da-Rong Liu, Hung-Yi Lee, and Yu 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.