Interspeech 2024
1-5 September 2024, Kos, Greece

# OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer

*Yifan Peng*[1], *Jinchuan Tian*[1], *William Chen*[1], *Siddhant Arora*[1], *Brian Yan*[1], *Yui Sudo*[2], *Muhammad Shakeel*[2], *Kwanghee Choi*[1], *Jiatong Shi*[1], *Xuankai Chang*[1], *Jee-weon Jung*[1], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, USA [2]Honda Research Institute Japan, Japan

yifanpen@andrew.cmu.edu, swatanab@andrew.cmu.edu

## Abstract

Recent studies have highlighted the importance of fully open foundation models. The Open Whisper-style Speech Model (OWSM) is an initial step towards reproducing OpenAI Whisper using public data and open-source toolkits. However, previous versions of OWSM (v1 to v3) are still based on standard Transformer, which might lead to inferior performance compared to state-of-the-art speech encoder architectures. This work aims to improve the performance and efficiency of OWSM without additional data. We present a series of E-Branchformer-based models named OWSM v3.1, ranging from 100M to 1B parameters. OWSM v3.1 outperforms its predecessor, OWSM v3, in most evaluation benchmarks, while showing an improved inference speed of up to 25%. We further reveal the emergent ability of OWSM v3.1 in zero-shot contextual biasing speech recognition. We also provide a model trained on a subset of data with low license restrictions. We will publicly release the code, pre-trained models, and training logs.[1]

**Index Terms**: speech foundation models, speech recognition, speech translation, branchformer

## 1. Introduction

Large speech foundation models have gained popularity recently. Owing to the scaling of model and data sizes as well as the knowledge sharing across languages and tasks, these massively multilingual and multitasking models achieve state-of-the-art (SOTA) performance in various speech processing tasks [1–3]. OpenAI Whisper [1] is one of the most widely used speech foundation models, which releases pre-trained model weights at five scales from 39M to 1.5B parameters. However, the full development pipeline, including the training data details and model learning dynamics, is unavailable to the public, which could lead to data leakage and concerns about fairness and bias. Recent studies have advocated for open-source reproduction of foundation models, including large language models (LLMs) [4–6], self-supervised speech models [7, 8], and Whisper-style speech models [9].

The Open Whisper-style Speech Model (OWSM) [9] is an initial step towards reproducing Whisper-style training using public datasets and an open-source toolkit ESPnet [10]. It supports multilingual automatic speech recognition (ASR), any-to-any speech translation (ST), language identification (LID), and utterance-level alignment. It also publicly releases all scripts, pre-trained model weights, and training logs. To match the design of OpenAI Whisper, the three versions in [9], OWSM v1, v2, and v3, adopt the standard Transformer [11] architecture. However, it can lead to suboptimal performance compared to more advanced encoders such as Conformer [12], Branchformer [13], and E-Branchformer [14].

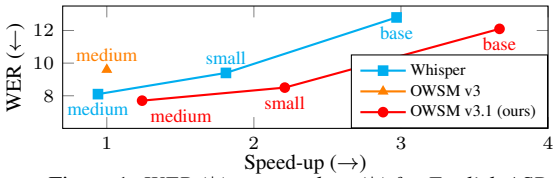


Figure 1: *WER (↓) vs. speed-up (↑) for English ASR.*

In this work, our goal is to improve the performance and efficiency of the previous OWSM v3 using the same amount of training data (see Figure 1 for English ASR results). We conduct preliminary experiments to compare Transformer, Conformer, and E-Branchformer encoders and select E-Branchformer due to its faster convergence. We then present new OWSM v3.1 models at three scales: base (101M), small (367M), and medium (1.02B). To stabilize the training of large E-Branchformer models, we propose a piecewise-linear learning rate schedule. Results on extensive benchmarks show that OWSM v3.1 outperforms the previous OWSM v3 in 8 of 9 English ASR, 10 of 11 multilingual ASR, 13 of 19 ST, and 3 of 4 SLUE-PERB [15] test sets. Additionally, OWSM v3.1 is 24% faster for English ASR and 16% to 25% faster for ST during inference, owing to the smaller decoder. Figure 1 shows that OWSM v3.1 even achieves a better trade-off between performance and efficiency than Whisper. Furthermore, we reveal that OWSM v3.1 has the emergent ability in zero-shot contextual biasing ASR. To extend the accessibility of our model, we provide a small-sized model trained on a subset of data with low restrictions. We will publicly release the code, pre-trained models, and training logs to promote transparency and open science.

## 2. OWSM v3.1

### 2.1. Model architecture

Whisper [1] and OWSM v3 [9] adopt the Transformer encoder-decoder architecture [11]. More advanced speech encoders such as Conformer [12] and Branchformer [13, 14] have achieved superior results in various speech processing tasks [16, 17]. It is thus natural and promising to explore them in large speech foundation models. In this work, we demonstrate the effectiveness and scalability of E-Branchformer [14] up to a scale of 1B parameters. E-Branchformer is an enhanced Branchformer [13], which utilizes parallel branches to capture local and global information and merges them with convolutions. In Whisper-style training, the input audio has a fixed length of 30s, so we simply use the sinusoidal absolute positional encoding. Table 1 summarizes the model configurations. The proposed OWSM v3.1 mostly follows the design of OWSM v3, except for the encoder.

[1]https://www.wavlab.org/activities/2024/owsm/



10.21437/Interspeech.2024-1194

Table 1: *Model architectures and training setups. LR (low restriction) is a small-sized model trained on a subset of data with low license restrictions.*

| | Whisper [1] | | | OWSM v3 [9] | OWSM v3.1 (ours) | | | |
|---|---|---|---|---|---|---|---|---|
| | base | small | medium | medium | base | small | medium | LR |
| **Model architectures** | | | | | | | | |
| Params | 74M | 244M | 769M | 889M | 101M | 367M | 1.02B | 367M |
| Encoder | Transformer | | | Transformer | E-Branchformer | | | |
| Decoder | Transformer | | | Transformer | Transformer | | | |
| Layers | 6 | 12 | 24 | 24 | 6 | 9 | 18 | 9 |
| Hidden | 512 | 768 | 1024 | 1024 | 384 | 768 | 1024 | 768 |
| Heads | 8 | 12 | 16 | 16 | 6 | 12 | 16 | 12 |
| **Training setups** | | | | | | | | |
| Data (h) | 680K | | | 180K | 180K | | | 70K |
| Languages | 99 | | | 151 | 151 | | | 143 |
| GPU hours | unknown | | | 30.7K | 2.3K | 3.2K | 24.6K | 3.2K |
| Max LR | 1e-3 | 5e-4 | 2.5e-4 | 2.5e-4 | 1e-3 | 5e-4 | 2e-4 | 5e-4 |

We modify the hidden size and the number of layers to adjust the size of the model. We provide three variants to investigate the scaling behavior, including base (101M), small (367M), and medium (1.02B). Although slightly larger than OWSM v3 and Whisper at the same scale, OWSM v3.1 models exhibit faster inference speeds (see Figure 1, Table 4, and Table 5), mainly due to the smaller decoder.

### 2.2. Data preparation

We prepare training data using scripts publicly released by [9]. Table 1 shows the amount of data and the number of languages. Please refer to [9] for more details. We perform the following preprocessing to make the text transcripts more consistent, which affects only a very small amount of data.

- We exclude WSJ from the training data due to its different speaking and annotation styles, in which the punctuation is explicitly uttered and annotated as a word.
- AMI [18] and VoxForge [19] provide uppercase transcripts. We convert them to lowercase. Other data remain unchanged.
- We merge two language codes "cmn" and "zho" into "zho".

Our base, small, and medium models are trained on all 180K hours of data. To extend the accessibility of our model, we also train a small-sized model using a subset of data with low restrictions (LR): AMI (CC-BY-4.0) [18], CommonVoice (CC0-1.0) [20], FLEURS (CC-BY-4.0) [21], KsponSpeech (MIT) [22], LibriSpeech (CC-BY-4.0) [23], Multilingual LibriSpeech (CC-BY-4.0) [24], and VCTK (CC-BY-4.0) [25]. This subset contains 70K hours of ASR data but no ST data.

### 2.3. Training setups

Our models are implemented in ESPnet [10] with PyTorch [26]. We use FlashAttention [27] to improve training efficiency. The batch size is 256. Our base, small, and medium models are trained for approximately 3 entire passes of the 180K hours of data using 16, 16, and 64 NVIDIA A100 GPUs (40GB), respectively. The low-restriction model follows the setup of OWSM v3.1 small, but uses only 70K hours of data. Table 1 shows the estimated GPU hours, assuming a stable GPU cluster.

We find it difficult to train models on massively multilingual, multitasking, and long-form speech data.[2] A typical strategy to improve convergence is to use a very small learning rate at the beginning of training. However, with the linear warmup schedule, we have to greatly reduce the peak learning rate or increase the warmup steps, both leading to inferior performance according to our preliminary explorations. To alleviate this issue, we propose a piecewise-linear warmup schedule that slowly increases the learning rate at the beginning and more quickly later. Specifically, the learning rate is linearly increased to a very small value (e.g., 5e-5) in the first 30K steps and then linearly increased to the peak learning rate in another 30K steps. After warmup, it is decreased exponentially in the same way as the vanilla version. The proposed piecewise-linear schedule enables successful training of OWSM v3.1.

[2]Based on our experience, this is mainly due to the 30s long-form data format. Even small models have a hard time converging.

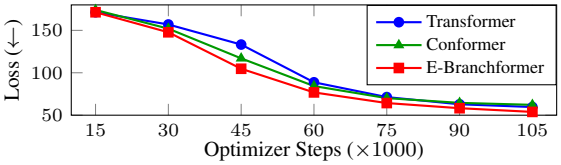

Figure 2: *Validation loss curves of three encoders.*

Table 2: *WER (↓) of English ASR. **Bold**: the best result. Underlined: OWSM v3.1 outperforms OWSM v3. CV: CommonVoice. LS: LibriSpeech. MLS: Multilingual LibriSpeech.*

| Test set | Whisper | | | OWSM v3 | OWSM v3.1 (ours) | | | |
|---|---|---|---|---|---|---|---|---|
| | base | small | medium | medium | base | small | medium | LR |
| CV [20] | 25.2 | 15.7 | **11.9** | 14.5 | 21.5 | 14.3 | 12.6 | 12.3 |
| FLEURS [21] | 12.4 | 9.6 | **6.4** | 10.9 | 14.8 | 10.3 | 9.0 | 10.8 |
| LS clean [23] | 5.1 | 3.3 | 2.8 | 2.7 | 3.6 | 2.5 | 2.4 | **2.1** |
| LS other [23] | 12.0 | 7.7 | 6.5 | 6.0 | 9.1 | 5.8 | **5.0** | 5.2 |
| MLS [24] | 13.4 | 9.1 | 10.2 | 7.4 | 12.0 | 8.1 | 7.1 | **7.0** |
| SWBD [28] | 25.7 | 22.2 | 19.4 | 17.2 | 22.9 | 17.4 | **16.3** | 31.5 |
| TEDLIUM [29] | 6.3 | **4.6** | 5.1 | 4.8 | 7.8 | 5.0 | 5.1 | 9.2 |
| VoxPopuli [30] | 10.2 | 8.5 | **7.6** | 9.2 | 12.0 | 9.1 | 8.4 | 13.8 |
| WSJ [31] | 5.0 | 4.3 | **2.9** | 13.4 | 5.3 | 3.8 | 3.5 | 4.9 |
| Ave. WER (↓) | 12.8 | 9.4 | 8.1 | 9.6 | 12.1 | 8.5 | **7.7** | 10.8 |
| Speed-up (↑) | 2.97x | 1.81x | 0.94x | 1.00x | **3.67x** | 2.21x | 1.24x | 2.50x |

## 3. Experiments

### 3.1. Comparison of encoder architectures

We first compare different encoders by training small-sized models on 10% of the training data. These models use the same decoder but different encoders: Transformer, Conformer, or E-Branchformer. Their overall model sizes are kept the same to ensure a fair comparison (366M, 367M, and 367M, respectively). Figure 2 shows the validation losses within the first 105K steps.[3] E-Branchformer converges faster than the others, which is consistent with prior work [17]. Hence, we adopt E-Branchformer in our main experiments.

### 3.2. English speech recognition

Table 2 shows English ASR results. Figure 1 visualizes the average word error rate (WER) versus speed-up measured on an NVIDIA A40 GPU. We follow [9] to perform greedy search and apply the Whisper text normalizer before scoring. We have the following observations: (1) Compared to the previous OWSM v3, the proposed OWSM v3.1 medium model performs better in 8 of 9 test sets. The improvement is especially large in CommonVoice, FLEURS, LibriSpeech, Switchboard, VoxPopuli, and WSJ.[4] This verifies the effectiveness of our E-

[3]It takes more than a week for the model to fully converge with 16 GPUs. Due to budget and time limits, we only compare their convergence speeds based on the first 105K steps.

[4]As discussed in [9], the WSJ training data is used by OWSM v3, but its transcripts are fully uppercased. The model might treat it as another

Table 3: *WER/CER (↓) of multilingual ASR. Training data sizes (in hours) are also shown. OWSM v3.1 uses the same amount of training data as OWSM v3.* ***Bold****: the best result.* *Underlined: OWSM v3.1 outperforms OWSM v3.*

| Test set | Language | Metric | Whisper | | | | OWSM v3 | | OWSM v3.1 (ours) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | data | base | small | medium | data | medium | base | small | medium |
| MLS [24] | Spanish | WER | 11.1K | 14.5 | 9.1 | **6.1** | 2.0K | 11.7 | 18.5 | 10.8 | 9.0 |
| | French | | 9.8K | 25.2 | 13.6 | **9.7** | 2.5K | 14.1 | 24.2 | 14.1 | 12.1 |
| | German | | 13.3K | 19.9 | 11.5 | **8.1** | 3.7K | 11.9 | 18.7 | 12.4 | 10.8 |
| | Dutch | | 2.1K | 30.9 | 18.2 | **12.2** | 1.7K | 17.7 | 28.6 | 19.7 | 18.1 |
| | Italian | | 2.6K | 32.9 | 21.3 | **15.6** | 0.7K | 24.5 | 33.7 | 21.8 | 20.2 |
| | Portuguese | | 8.6K | 23.5 | 13.8 | **8.9** | 0.3K | 28.2 | 44.9 | 26.7 | 21.6 |
| | Polish | | 4.3K | 25.2 | 12.5 | **6.8** | 0.3K | 37.0 | 49.7 | 28.5 | 25.2 |
| AISHELL-1 [32] | Chinese | CER | 23.4K | 39.1 | 25.1 | 15.7 | 16.0K | 7.1 | 12.2 | 7.5 | **6.4** |
| KsponSpeech clean [22] | Korean | | 8.0K | 27.0 | 24.0 | 17.6 | 1.0K | 20.5 | 23.8 | 17.2 | **16.7** |
| KsponSpeech other [22] | | | | 22.9 | 15.4 | **12.8** | | 22.6 | 26.1 | 18.9 | 18.9 |
| ReazonSpeech [33] | Japanese | | 7.1K | 54.1 | 32.5 | 25.3 | 18.9K | 11.3 | 11.2 | 8.5 | **7.9** |
| Average WER/CER (↓) | | | - | 28.7 | 17.9 | **12.6** | - | 18.8 | 26.5 | 16.9 | 15.2 |

Table 4: *BLEU (↑) of X-to-En ST on CoVoST-2 [34]. Training data sizes (in hours) are also shown. OWSM v3.1 uses the same amount of training data as OWSM v3.* ***Bold****: the best result.* *Underlined: OWSM v3.1 outperforms OWSM v3.*

| Source | Whisper | | | | OWSM v3 | | OWSM v3.1 (ours) | | |
|---|---|---|---|---|---|---|---|---|---|
| | data | base | small | medium | data | medium | base | small | medium |
| German | 4.3K | 11.4 | 25.0 | **33.6** | 0.2K | 16.2 | 7.3 | 15.1 | 17.1 |
| Spanish | 6.7K | 19.2 | 32.8 | **39.7** | 0.1K | 20.5 | 10.0 | 19.3 | 22.3 |
| French | 4.5K | 13.1 | 26.4 | **34.4** | 0.3K | 21.7 | 11.1 | 20.3 | 22.7 |
| Catalan | 0.2K | 9.7 | 21.7 | **29.2** | 0.1K | 16.8 | 9.0 | 16.2 | 18.4 |
| Ave. BLEU (↑) | 13.4 | 26.5 | **34.2** | - | 18.8 | 9.4 | 17.7 | 20.1 | |
| Speed-up (↑) | 2.14x | 1.80x | 0.98x | - | 1.00x | **3.23x** | 2.26x | 1.16x | |

Table 5: *BLEU (↑) of En-to-X ST on CoVoST-2 [34].* ***Bold****: the best result.* *Underlined: OWSM v3.1 outperforms OWSM v3.*

| Target | Training Data (h) | OWSM v3 | OWSM v3.1 (ours) | | |
|---|---|---|---|---|---|
| | | medium | base | small | medium |
| German | 14.0K | **25.4** | 14.6 | 22.8 | **25.4** |
| Catalan | 0.4K | **20.0** | 7.7 | 15.9 | 19.6 |
| Chinese | 13.7K | **33.4** | 14.5 | 26.7 | 32.1 |
| Persian | 0.8K | 9.5 | 3.0 | 7.7 | **10.1** |
| Estonian | 0.4K | **7.8** | 1.8 | 5.8 | 7.7 |
| Mongolian | 0.4K | 3.1 | 1.0 | 3.3 | **4.6** |
| Turkish | 0.9K | 6.1 | 1.2 | 4.8 | **6.5** |
| Arabic | 0.9K | 6.6 | 1.6 | 5.1 | **7.2** |
| Swedish | 0.4K | 19.9 | 8.1 | 16.6 | **20.3** |
| Latvian | 0.4K | 6.3 | 1.3 | 4.4 | **6.4** |
| Slovenian | 0.4K | 8.6 | 0.7 | 5.7 | **9.0** |
| Tamil | 0.4K | 0.0 | 0.0 | 0.0 | 0.0 |
| Japanese | 1.0K | 17.3 | 8.7 | 16.4 | **19.6** |
| Indonesian | 0.4K | 14.5 | 5.1 | 12.4 | **16.1** |
| Welsh | 0.4K | **15.9** | 4.5 | 11.6 | 15.3 |
| Ave. BLEU (↑) | | 13.0 | 4.9 | 10.6 | **13.3** |
| Speed-up (↑) | | 1.00x | **3.00x** | 2.43x | 1.25x |

Table 6: *WER (↓) of long-form ASR on TEDLIUM.* ***Bold****: the best result.* *Underlined: OWSM v3.1 outperforms OWSM v3.*

| Whisper | | | OWSM v3 | OWSM v3.1 (ours) | | |
|---|---|---|---|---|---|---|
| base | small | medium | medium | base | small | medium |
| 5.3 | 4.4 | **3.8** | 9.2 | 9.6 | 6.7 | 5.7 |

Branchformer encoder. (2) OWSM v3.1 even achieves lower average WERs than Whisper at each scale, demonstrating its competitive performance, although trained on much less English ASR data (73K vs. 438K hours). (3) OWSM v3.1 is faster during inference than the others at the same scale, primarily due to the smaller decoder. (4) Our small-sized low-restriction (LR) model achieves reasonable performance considering that it is trained on a subset of data (see Section 2.2).

### 3.3. Multilingual speech recognition

Table 3 presents multilingual ASR results. We perform greedy decoding and apply the Whisper text normalizer before calculating word or character error rates (WER/CER). We observe that OWSM v3.1 medium outperforms OWSM v3 in 10 of 11 test sets in various languages, usually by a large margin. Specifically, the average error rate is reduced from 18.8% to 15.2%. Compared to Whisper, OWSM v3.1 still falls behind in many European languages due to limited training data. In contrast, when the data are sufficient (e.g. Chinese and Japanese), OWSM v3.1 achieves strong performance and outperforms Whisper. This reveals the importance of the quantity of training data. In the future, we will include more data from public sources like YODAS [35] to further improve OWSM.

### 3.4. Speech translation

We evaluate ST on CoVoST-2 test sets [34]. For English-to-X, we utilize all 15 directions. For X-to-English, we report the results of directions where OWSM has more than 100 hours of training data. For other directions with very limited training data like Japanese- or Chinese-to-English, OWSM usually does not work [9]. We also record the average decoding time of each test set on an NVIDIA A40 GPU and calculate the relative decoding speed compared to OWSM v3.

For X-to-English (shown in Table 4), the proposed OWSM v3.1 medium achieves consistently higher BLEU scores than OWSM v3. The average BLEU is improved from 18.8 to 20.1. OWSM v3.1 is also 16% faster than OWSM v3 during inference. Compared to Whisper, OWSM v3.1 performs still worse due to limited training data. But OWSM v3.1 has a faster inference speed than Whisper at each scale, thanks to the larger time shift in the encoder (40 ms vs. 20 ms) and the smaller decoder.

For English-to-X (shown in Table 5), OWSM v3.1 outperforms OWSM v3 in 9 of 15 directions. The average BLEU is slightly improved from 13.0 to 13.3 and the inference speed is 25% faster. Note that Whisper cannot perform translation in these directions.

### 3.5. Long-form speech recognition

Table 6 presents long-form English ASR results on the TEDLIUM test set [29]. Similar to [1, 9], OWSM takes an entire audio recording as input and generates transcripts in chunks. Each chunk has a fixed length of 30s and is gradually shifted

low-resource language, which leads to poor results. In v3.1, we exclude WSJ during training and achieve a significantly lower WER.

Table 7: *Accuracy % (↑) of LID on FLEURS [21].*

| Whisper | | | OWSM v3 | OWSM v3.1 (ours) | | |
|---|---|---|---|---|---|---|
| base | small | medium | medium | base | small | medium |
| 47.6 | 53.1 | 54.8 | **81.4** | 41.9 | 67.1 | 75.6 |

Table 8: *F1 scores (↑) of SLU tasks on SLUE-PERB [15].*

| Task | Metric | OWSM v3 | OWSM v3.1 (ours) |
|---|---|---|---|
| Sentiment Analysis | F1 score | **60.1** | 56.2 |
| Named Entity Recognition | F1 score | 54.8 | **65.8** |
| Named Entity Localization | frame-F1 | 40.5 | **50.4** |
| Dialogue Act Classification | F1 score | 56.5 | **64.8** |

based on the predicted timestamps. The proposed OWSM v3.1 medium achieves a WER of 5.7%, compared to 9.2% of OWSM v3. This demonstrates the robustness of OWSM v3.1 against long-form audio; the predicted timestamps might also be more accurate. OWSM v3.1 still falls behind Whisper, likely because (1) our training data is only around a quarter of Whisper's training data, and (2) many public datasets used by OWSM do not provide unsegmented long-form data and we have to use the segmented short audio for training, which leads to a mismatch between training and inference. In the future, we will add more long-form data to mitigate this issue.

### 3.6. Language identification

Table 7 shows the accuracy of language identification on the FLEURS test set. We notice a degradation of OWSM v3.1 compared to the previous OWSM v3, but OWSM v3.1 medium is still much better than Whisper medium because our model uses the massively multilingual FLEURS and CommonVoice data for training. We also find that OWSM v3.1 benefits more from scaling up compared to Whisper. From base to medium, the accuracy of OWSM v3.1 is almost doubled (41.9% to 75.6%), while the accuracy of Whisper is only slightly increased (47.6% to 54.8%). A possible reason is that OWSM supports more languages for ASR and language pairs for ST, which is more challenging for smaller models to learn.

### 3.7. Spoken language understanding via fine-tuning

Pre-trained speech models can be applied to downstream tasks via fine-tuning, which generally improves performance [36]. We take spoken language understanding (SLU) as an example and evaluate OWSM on the recently proposed SLUE-PERB benchmark [15]. Specifically, the pre-trained speech encoder is frozen and a randomly initialized shallow decoder is trained on task-specific SLU data. The model is then evaluated on the corresponding SLU test data. This evaluation procedure is similar to the widely used SUPERB benchmark [37]. We consider four SLU tasks, i.e., sentiment analysis (SA), named entity recognition (NER), named entity localization (NEL), and dialog act classification (DAC). As shown in Table 8, the proposed OWSM v3.1 medium outperforms the previous v3 model by a large margin in NER, NEL, and DAC, confirming the strong capacity of our E-Branchformer encoder.

### 3.8. Emergent ability for zero-shot contextual biasing

OWSM generates ASR or ST hypotheses conditioned on an optional text prompt. During training, the previous sentence in the same recording is used as a prompt according to the probability of 0.5. During inference, the user can provide a prompt to potentially adjust the output. An application of this feature is

Table 9: *WER (↓) of zero-shot contextual biasing.*

| OWSM v3.1 | LibriSpeech test-clean | | | LibriSpeech test-other | | |
|---|---|---|---|---|---|---|
| | WER | U-WER | B-WER | WER | U-WER | B-WER |
| base | 3.88 | 2.45 | 15.47 | 9.48 | 6.89 | 32.17 |
| + biasing | 4.37 | 3.09 | 14.79 | 12.49 | 10.45 | 30.36 |
| small | 2.68 | 1.63 | 11.27 | 6.16 | 4.21 | 23.27 |
| + biasing | 2.58 | 1.75 | 9.32 | 5.89 | 4.48 | 18.34 |
| medium | 2.59 | 1.61 | 10.61 | 5.31 | 3.52 | 21.12 |
| + biasing | 2.24 | 1.62 | 7.31 | 5.03 | 3.86 | 15.35 |

zero-shot contextual biasing, which aims to improve the ASR performance of rare words by providing a list of biasing words containing true targets and many distractions [38]. We evaluate OWSM v3.1 models on the LibriSpeech biasing test sets created by [38]. Specifically, we use 100 biasing words separated by spaces as the prompt and perform greedy decoding. Unlike Section 3.2, we do not use any text normalizer to match the condition in [38]. Contextual biasing aims to reduce the biased WER (B-WER) while maintaining the unbiased WER (U-WER). Table 9 shows the WERs of our three models. Compared to ASR without biasing, the base model shows minor improvements on B-WER but much larger degradations on U-WER, indicating that it cannot distinguish between useful contextual information and distractions. In contrast, small and medium models greatly reduce B-WER and mostly maintain U-WER, demonstrating that these models can extract and utilize useful contextual information in a zero-shot manner. The phenomenon that the smaller OWSM performs very poorly in zero-shot biasing ASR while larger ones perform well reveals that **speech foundation models also have the emergent ability**, which has been widely observed in LLMs [39].

## 4. Conclusion and future work

We present OWSM v3.1, a family of Open Whisper-style Speech Models based on E-Branchformer, ranging from 100M to 1B parameters. Although trained on the same amount of data, OWSM v3.1 achieves better results than the previous OWSM v3 in the vast majority of evaluation sets, while showing up to 25% faster inference speeds. We further investigate the emergent ability of speech foundation models using zero-shot contextual biasing ASR, which verifies the benefit of scaling up. To extend the accessibility of our model, we provide a model trained on a subset of data with low license restrictions. We will publicly release the code, pre-trained model weights, and training logs to promote transparency and facilitate the development of foundation models in the speech field.

A limitation is that this work does not enhance the quantity or quality of training data, which might lead to suboptimal performance in low-resource languages. Future research directions include exploring the impact of data diversity on model performance, adding more public data like YODAS [35] for better performance, compressing the pre-trained model for better efficiency [40–45], and exploring various downstream applications such as SLU [36, 46] and speech language models [47, 48].

## 5. Acknowledgements

We use PSC Bridges2 and NCSA Delta via ACCESS CIS210014, by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

# 6. References


[1] A. Radford, J. W. Kim, T. Xu, *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023.

[2] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[3] L. Barrault, Y.-A. Chung, M. C. Meglioli, *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[4] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv:2302.13971*, 2023.

[5] Z. Liu, A. Qiao, W. Neiswanger, *et al.*, "Llm360: Towards fully transparent open-source llms," *arXiv preprint arXiv:2312.06550*, 2023.

[6] D. Groeneveld, I. Beltagy, P. Walsh, *et al.*, "Olmo: Accelerating the science of language models," *arXiv preprint arXiv:2402.00838*, 2024.

[7] W. Chen, X. Chang, Y. Peng, *et al.*, "Reducing Barriers to Self-Supervised Learning: HuBERT Pre-training with Academic Compute," in *Proc. Interspeech*, 2023.

[8] W. Chen, J. Shi, B. Yan, *et al.*, "Joint prediction and denoising for large-scale multilingual self-supervised learning," in *Proc. ASRU*, 2023.

[9] Y. Peng, J. Tian, B. Yan, *et al.*, "Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data," in *Proc. ASRU*, 2023.

[10] S. Watanabe, T. Hori, S. Karita, *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017.

[12] A. Gulati, J. Qin, C.-C. Chiu, *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech*, 2020.

[13] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. ICML*, 2022.

[14] K. Kim, F. Wu, Y. Peng, *et al.*, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. SLT*, 2023.

[15] S. Arora, R. Sharma, A. Pasad, *et al.*, "SLUE-PERB: A Spoken Language Understanding Performance Benchmark and Toolkit," in *ASRU SPARKS Workshop*, 2023.

[16] P. Guo, F. Boyer, X. Chang, *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. ICASSP*, 2021.

[17] Y. Peng, K. Kim, F. Wu, *et al.*, "A Comparative Study on E-Branchformer vs Conformer in Speech Recognition, Translation, and Understanding Tasks," in *Proc. Interspeech*, 2023.

[18] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Lang. Res. Eval.*, vol. 41, pp. 181–190, 2007.

[19] *VoxForge: http://www.voxforge.org/*.

[20] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv:1912.06670*, 2019.

[21] A. Conneau *et al.*, "FLEURS: Few-Shot Learning Evaluation of Universal Representations of Speech," in *Proc. SLT*, 2022.

[22] J.-U. Bang *et al.*, "Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition," *Applied Sciences*, vol. 10, no. 19, p. 6936, 2020.

[23] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[24] V. Pratap *et al.*, "MLS: A large-scale multilingual dataset for speech research," *arXiv:2012.03411*, 2020.

[25] J. Yamagishi *et al.*, *CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*, 2019.

[26] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019.

[27] T. Dao, D. Y. Fu, S. Ermon, *et al.*, "Flashattention: Fast and memory-efficient exact attention with io-awareness," in *Proc. NeurIPS*, 2022.

[28] J. Godfrey *et al.*, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. ICASSP*, 1992.

[29] F. Hernandez *et al.*, "Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *Speech & Computer*, 2018, pp. 198–208.

[30] C. Wang *et al.*, "VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation," in *Proc. ACL*, 2021.

[31] D. B. Paul and J. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*, 1992.

[32] H. Bu *et al.*, "AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline," in *Proc. O-COCOSDA*, 2017.

[33] Y. Yin, D. Mori, *et al.*, *ReazonSpeech: A Free and Massive Corpus for Japanese ASR*, 2023.

[34] C. Wang *et al.*, "CoVoST 2 and Massively Multilingual Speech Translation," in *Interspeech*, 2021.

[35] X. Li, S. Takamichi, T. Saeki, *et al.*, "Yodas: Youtube-oriented dataset for audio and speech," in *Proc. ASRU*, 2023.

[36] Y. Peng, S. Arora, Y. Higuchi, *et al.*, "A Study on the Integration of Pre-trained SSL, ASR, LM and SLU Models for Spoken Language Understanding," in *Proc. SLT*, 2022.

[37] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, 2021.

[38] D. Le, M. Jain, G. Keren, *et al.*, "Contextualized streaming end-to-end speech recognition with trie-based deep biasing and shallow fusion," in *Proc. Interspeech*, 2021.

[39] J. Wei, Y. Tay, R. Bommasani, *et al.*, "Emergent abilities of large language models," *Trans. Mach. Learn. Res.*, vol. 2022, 2022.

[40] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *Proc. ICASSP*, 2022.

[41] C.-I. J. Lai, Y. Zhang, A. H. Liu, *et al.*, "PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition," in *Proc. NeurIPS*, 2021.

[42] Y. Peng, K. Kim, F. Wu, *et al.*, "Structured pruning of self-supervised pre-trained models for speech recognition and understanding," in *Proc. ICASSP*, 2023.

[43] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "DPHuBERT: Joint Distillation and Pruning of Self-Supervised Speech Models," in *Proc. Interspeech*, 2023.

[44] Y. Peng, J. Lee, and S. Watanabe, "I3D: Transformer Architectures with Input-Dependent Dynamic Depth for Speech Recognition," in *Proc. ICASSP*, 2023.

[45] S. Gandhi, P. von Platen, and A. M. Rush, "Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling," *arXiv preprint arXiv:2311.00430*, 2023.

[46] S. Arora, H. Futami, J.-w. Jung, *et al.*, "UniverSLU: Universal spoken language understanding for diverse classification and sequence generation tasks with a single network," *arXiv preprint arXiv:2310.02973*, 2023.

[47] M. Wang, W. Han, I. Shafran, *et al.*, "SLM: Bridge the thin gap between speech and text foundation models," in *Proc. ASRU*, 2023.

[48] C. Tang, W. Yu, G. Sun, *et al.*, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.