



# Exploration on HuBERT with Multiple Resolutions

Jiatong Shi<sup>1</sup>, Yun Tang<sup>2</sup>, Hirofumi Inaguma<sup>2</sup>, Hongyu Gong<sup>2</sup>, Juan Pino<sup>2</sup>, Shinji Watanabe<sup>1</sup>

<sup>1</sup>Language Technologies Institute, Carnegie Mellon University    <sup>2</sup>Meta AI

{jiatongs, swatanab}@cs.cmu.edu

## Abstract

Hidden-unit BERT (HuBERT) is a widely-used self-supervised learning (SSL) model in speech processing. However, we argue that its fixed 20ms resolution for hidden representations would not be optimal for various speech-processing tasks since their attributes (e.g., speaker characteristics and semantics) are based on different time scales. To address this limitation, we propose utilizing HuBERT representations at multiple resolutions for downstream tasks. We explore two approaches, namely the parallel and hierarchical approaches, for integrating HuBERT features with different resolutions. Through experiments, we demonstrate that HuBERT with multiple resolutions outperforms the original model. This highlights the potential of utilizing multiple resolutions in SSL models like HuBERT to capture diverse information from speech signals.

**Index Terms:** speech self-supervised learning, multi-resolution HuBERT, Hidden-unit BERT.

## 1. Introduction

In recent years, self-supervised learning (SSL) models for speech processing have demonstrated impressive performance on a range of tasks [1]. These models can leverage unlabeled speech data to learn general-purpose knowledge, rather than relying solely on supervised training with paired labels. As a result, speech SSLs have emerged as a powerful tool for speech processing, offering a promising alternative to traditional supervised learning approaches.

HuBERT [2] is one of the most prominent speech self-supervised learning (SSL) models, according to the SUPERB benchmark [3–5]. During training, HuBERT employs an offline clustering step to generate pseudo labels and uses a masked language model (MLM) objective. Like many speech processing systems, HuBERT begins by converting the raw waveform into hidden states using convolutional (conv.) layers, resulting in a fixed 20ms resolution for its representation.

HuBERT can be used as a feature extractor or directly fine-tuned as an encoder. In the feature extraction approach, the pre-trained model is used to extract high-level features from speech signals, which are then fed into a downstream task-specific model such as a classifier or a regression model. This approach reduces the computational cost during training [6, 7]. On the other hand, fine-tuning the pre-trained HuBERT model as an encoder is a popular approach, which further improves the performance at the cost of training massive encoder parameters. In this approach, the pre-trained model is further trained on the downstream task data, either by updating all the model parameters or just the last few layers.

Although the HuBERT representation has demonstrated strong performance, the empirical selection of a 20ms resolu-

tion raises concerns regarding its optimality for diverse speech-related tasks.<sup>1</sup> In contrast, the literature also suggests that modeling speech at multiple resolutions is preferable for speech recognition [8–16], speaker verification [17–19], speech enhancement [20, 21], and voice conversion [22]. Two mainstream approaches have emerged: one that focuses on parallel processing [8–14, 17], and the other that utilizes hierarchical frameworks such as U-net [18, 21–26].

The parallel paradigm is based on observations of multiple parallel processing streams in the human speech cognitive system [8, 9]. To formalize multi-stream signals, a common method is to use parallel encoders that consider multi-resolution signals. For example, [13] employs two encoders based on recurrent neural network (RNN) and convolution neural network (CNN)-RNN, respectively. Both encoders use the same input features, while the second applies CNN to transform features into a different temporal resolution.

The second hierarchical approach, in contrast, serializes the aggregation of multi-resolution information. An example of this approach is the U-net-like architecture, which is based on an encoder-decoder structure [15, 16, 18, 21, 22]. The encoder processes high-resolution features initially and downsamples them to prioritize low-resolution features. Conversely, the decoder starts from low-resolution features and upsamples them to refine information in high resolution. To ensure stability, corresponding blocks with the same resolution in the encoder and decoder are connected with residual connections.

In this work, we propose using HuBERT representations at different resolutions (HuBERT-MR) for downstream speech tasks. In our experiments, we evaluate both the parallel and the hierarchical approaches to efficiently utilize HuBERT of different resolutions. Experiments show that our proposed method could get significantly better performances over the original HuBERT at 20ms resolution. In some tasks, the HuBERT with multi-resolution can even achieve reasonable performances compared to large models, even with less training data and fewer parameters.

## 2. HuBERT with Multiple Resolutions

Let  $S \in \mathbb{R}^{1 \times L}$  be a speech signal with length  $L$ . The HuBERT model  $H$  consists of two modules: a conv. feature extractor and an  $N$ -layer transformer encoder. The conv. block first converts  $S$  into a sequence of vectors  $X^0 = [x_1^0, \dots, x_T^0] \in \mathbb{R}^{T \times D}$ , where  $T$  is the number of frames and  $D$  is the dimension of each frame. The resulting feature sequence  $X^0$  is then passed to the transformer encoder, which produces a sequence of feature rep-

<sup>1</sup>It's worth noting that this choice of resolution is derived from an ASR conversion involving downsampling, which is specific to that particular task.

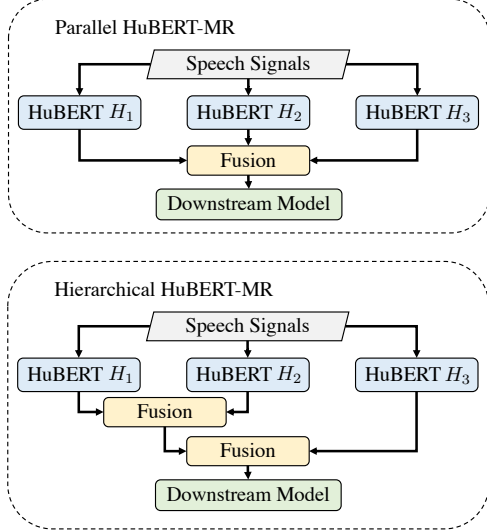


Figure 1: *HuBERT-MR-P* and *HuBERT-MR-H*. In *HuBERT-MR-P* (shown in the upper figure), three *HuBERT* models are fused in parallel. In contrast, *HuBERT-MR-H* (shown in the lower figure) fuses *HuBERT* models hierarchically, with features of low resolutions being fused earlier. Details about the framework design and *Fusion* modules can be found in Section 2.1 and Section 2.2, respectively.

representations  $X^i = [x_1^i, \dots, x_T^i] \in \mathbb{R}^{T \times D}$  at the  $i$ -th transformer layer. Each frame  $x_t^i$  corresponds to a fixed time interval  $R$ , where  $R \cdot T = L$ .<sup>2</sup> We refer to  $R$  as the resolution of features.

By controlling the stride of the conv. feature extractor, we can obtain a range of resolutions ( $R_1, \dots, R_K$ ) and correspondingly,  $K$  distinct *HuBERT* models ( $H_1, \dots, H_K$ ). In the next subsections, we discuss the application of the parallel and hierarchical approaches discussed in Sec. 1 to merge  $K$  *HuBERT* models for downstream tasks. For the easiness of discussion, we consider  $K = 3$  as an example and all *HuBERT* models ( $H_1, H_2, H_3$ ) use the same model configuration for all encoder modules so that the feature dimension for both models is  $D$ .

### 2.1. Parallel *HuBERT-MR* (*HuBERT-MR-P*)

As explained in Sec. 1, the parallel approach employs parallel encoders to process input signals at different resolutions. Therefore, we use three *HuBERT* models ( $H_1, H_2$ , and  $H_3$ ) with resolutions  $R_1, R_2$ , and  $R_3$ , respectively, to obtain layerwise features from the layer before the top encoder layer to the last encoder layer:  $(X_1^0, \dots, X_1^N)$ ,  $(X_2^0, \dots, X_2^N)$ , and  $(X_3^0, \dots, X_3^N)$ . These feature tensors have shapes  $X_1^i \in \mathbb{R}^{T_1 \times D}$ ,  $X_2^i \in \mathbb{R}^{T_2 \times D}$ , and  $X_3^i \in \mathbb{R}^{T_3 \times D}$ , respectively. Noted that  $T_1, T_2, T_3$  are different. The illustration of *HuBERT-MR-P* is shown in Figure 1. We define multi-resolution features  $X_{MR-P}$  as follows:

$$X_{MR-P} = \sum_{i=0}^N (w_{1,i} \cdot \text{UP}_1(X_1^i) + w_{2,i} \cdot \text{UP}_2(X_2^i) + w_{3,i} \cdot \text{UP}_3(X_3^i)), \quad (1)$$

where  $w_{1,i} \in [0, 1]$ ,  $w_{2,i} \in [0, 1]$ , and  $w_{3,i} \in [0, 1]$  are learnable weights that sum up to one (i.e.,  $\sum_{i=0}^N (w_{1,i} + w_{2,i} + w_{3,i}) = 1$ ). The functions  $\text{UP}_1, \text{UP}_2$ , and  $\text{UP}_3$  upsample the representations to the greatest common divisor of  $R_1, R_2$ , and

<sup>2</sup>In practical situations, it is necessary to incorporate rounding procedures that take into account edge cases.

Table 1: *Configurations of the pre-trained HuBERT with multi-resolutions. The convolution (Conv.) module is represented in [(kernel-size, stride)\* layer-number].*

ID	Res.(ms)	Param.	Conv. Module
A	20	94.7	(10,5)*1 + (3,2)*4 + (2,2)*2
B	40	95.2	(10,5)*1 + (3,2)*4 + (2,2)*3
C	100	97.3	(10,5)*2 + (3,2)*4 + (2,2)*2

$R_3$ , denoted as  $R_{1,2,3}$ , to ensure matching feature lengths. Finally, we can use  $X_{MR-P} \in \mathbb{R}^{T_{MR} \times D}$  for various downstream tasks, where  $T_{MR} = L // R_{1,2,3}$ . The UP functions can be any upsampling functions, including simple methods such as repeating the features along the time axis, as used in [27], or more complex methods such as transposed conv. networks.

### 2.2. Hierarchical *HuBERT-MR* (*HuBERT-MR-H*)

As described in Sec. 1, the hierarchical approach models multiple resolutions in a sequential manner. Unlike U-net-based methods [18, 21, 22], we do not perform additional feature encoding as the *HuBERT* models with different resolutions are already pre-trained. Instead, as shown in Figure 1, we adopt the decoder architecture inspired by U-net and fuse the *HuBERT* representations from low to high resolution. Assuming  $R_1 > R_2 > R_3$ , we first fuse the outputs from  $H_1$  and  $H_2$ , and then we further fuse  $H_3$  for additional downstream models.

The fusion module combines the representations from two different resolutions into a single stream. Specifically, we use a conv. module and a de-conv. module for each feature, respectively. Note that additional conv. modules improve the stability of the fusion as observed in our experiments. Given input features  $X_1^N \in \mathbb{R}^{T_1 \times D}$  and  $X_2^N \in \mathbb{R}^{T_2 \times D}$ , we first apply conv. modules with residual connections and then employ transposed conv. modules to align their resolutions. The resulting feature  $X_{MR-H}^{1:2}$  is defined as:

$$X_{MR-H}^{1:2} = \text{DeConv}_1(\text{Conv}_1(X_1^N)) + \text{DeConv}_2(\text{Conv}_2(X_2^N)). \quad (2)$$

Then, we further apply a conv. module to  $X_3^N$  and use transposed conv. modules to compute  $X_{MR-H}^{1:3} \in \mathbb{R}^{T_{MR} \times D}$  as:

$$X_{MR-H}^{1:3} = \text{DeConv}_{1,2}(X_{MR-H}^{1:2}) + \text{DeConv}_3(\text{Conv}_3(X_3^N)). \quad (3)$$

We can then use the feature  $X_{MR-H}^{1:3}$  for downstream tasks.

## 3. Experiments

### 3.1. Pre-trained *HuBERT*

To evaluate the effectiveness of *HuBERT* models with different resolutions, we trained three *HuBERT* models by modifying their conv. feature extractor. The configurations of these models are presented in Table 1. We trained all *HuBERT* models following the same procedure as *HuBERT*-base in [2], except for changes in label rates and corresponding conv. modules. We conducted two iterations of training for each *HuBERT*, where the first iteration was trained on Mel frequency cepstral coefficients (MFCC) clusters, and the second iteration was trained using the intermediate features' clusters. The final dimension of each *HuBERT* was set to  $D = 768$ . We pre-trained all *HuBERT* models using the Librispeech dataset [28].

### 3.2. Experimental Setups

**SUPERB benchmark:** In our experiments, we evaluate *HuBERT-MR* on the SUPERB benchmark [3–5]. According to

Table 2: *HuBERT-MR-P on SUPERB benchmark. Detailed tasks and evaluation metrics are discussed in Sec. 3.2. Proposed HuBERT-MR-P is introduced in Sec. 2.1.*

Model	Res.(ms)	PR(↓)	ASR(↓)	ER(↑)	IC(↑)	SID(↑)	SD(↓)	SV(↓)	SE(↑)	ST(↑)
HuBERT	20	5.41	6.42	<b>64.92</b>	98.34	81.42	5.88	5.11	<b>2.58</b>	15.53
wav2vec2	20	5.74	6.43	63.43	92.35	75.18	6.08	6.02	2.55	14.81
HuBERT-MR-P	(100,40,20)	<b>4.83</b>	<b>5.48</b>	63.76	<b>98.51</b>	<b>83.23</b>	<b>5.75</b>	<b>5.10</b>	2.55	<b>16.18</b>

Table 3: *The summation of layer weights of HuBERT with different resolutions in the HuBERT-MR-P model, which was evaluated in the SUPERB benchmark as shown in Table 2.*

HuBERT	ASR	SV	ST	Avg. Tasks
100ms	0.21	0.16	0.21	0.18
40ms	0.33	0.28	0.37	0.32
20ms	0.46	0.56	0.42	0.50

the SUPERB benchmark policy, we do not to include additional trainable parameters except for the layerwise weights. Therefore, we mainly focused on HuBERT-MR-P for the SUPERB tasks. We use the repeating method as the upsampling function UP, as described in Sec. 2.1.

We evaluate most of the SUPERB tasks, including understanding tasks (phone recognition (PR), automatic speech recognition (ASR), intent classification (IC), and speech translation (ST)), speaker-related tasks (speaker identification (SID), speaker verification (SV), and speaker diarization (SD)), and frontend processing (speech enhancement (SE)). To better understand the results, we also show the performances on wav2vec2-base for comparison [2, 29].

**ASR Fine-tuning:** We evaluate the ASR fine-tuning task on the Librispeech 100-hour train set with dev-clean and test-clean for development and testing, respectively. The training utilizes fairseq toolkit [30]. For all models, we train 100k steps with a maximum token number of 1M to form mini-batches. The training uses an AdamW optimizer with 8k warmup steps and a learning rate of  $2e-5$ . We compared HuBERT-MR-P and HuBERT-MR-H with base HuBERT models, as well as wav2vec2 models and HuBERT-large. Instead of repeating, we use transposed convolution for the UP function of HuBERT-MR-P. For simplification, we use a linear projection for CTC loss computation as the downstream module needed for HuBERT-MR.

We present not only the Viterbi decoding results but also the results after language model rescoring. For decoding, we used Flashlight for wav2letter decoding [31] and applied a beam size of 500 with a beam threshold of 100 and a language model weight of 2 for language model rescoring. The language model used was trained on the 4-gram language model training corpus of Librispeech [28].

**Evaluation metrics** We generally follow the evaluation metrics for SUPERB tasks. We use phone error rate (PER) for PR; word error rate (WER) for ASR; accuracy for ER, IC, and SID; diarization error rate (DER) for SD; equal error rate (EER) for SV; PESQ for SE; BLEU for ST. While for ASR fine-tuning, we use WER. Meanwhile, for efficiency concerns, we also report Floating Point Operations Per Second (FLOPs) and Multiply-Accumulate Operations (MACs) to models. The calculation procedure follows the SUPERB challenge [5].

Table 4: *Fine-tuning results on Librispeech-100h (comparison with baselines). Results with language model rescoring are in brackets. HuBERT-MR-P is explained in Sec 2.1 and HuBERT-MR-H is discussed in Sec 2.2.*

Model	Res.(ms)	WER(↓)
HuBERT	20	7.73( 3.81)
HuBERT	40	12.38( 4.90)
HuBERT	100	98.37(97.87)
HuBERT-MR-P	(100,20)	6.99( 3.70)
HuBERT-MR-P	(40,20)	7.13( 3.75)
HuBERT-MR-P	(100,40,20)	6.53( 3.61)
HuBERT-MR-H	(100,20)	6.59( 3.59)
HuBERT-MR-H	(40,20)	7.01( 3.71)
HuBERT-MR-H	(100,40,20)	<b>6.11( 3.31)</b>

### 3.3. Experimental Results

Table 2 presents the experimental results of the SUPERB benchmark. In most tasks, HuBERT-MR-P showed significant improvements over the original HuBERT, with the exception of ER and SE. We also analyzed the weight contribution of HuBERT with different representations of HuBERT-MR-P on ASR, SV, ST, and the average overall tasks, which are presented in Table 3. Among the tasks, SE has the lowest weight for 100ms HuBERT (0.15), while ER has the highest weight for 100ms HuBERT (0.26). The results indicate that HuBERT features from multiple resolutions provide additional benefits and can significantly contribute to various types of tasks.

The experimental results of ASR fine-tuning are presented in Tables 4 and 5. Table 4 compares the performance of HuBERT-MR with the original HuBERT models. The following observations can be found:

- Both HuBERT-MR-P and HuBERT-MR-H outperform the base HuBERT model trained with resolutions of 100ms, 40ms, and 20ms.
- Although the HuBERT model trained with resolutions of 100ms and 40ms does not achieve similar performance to the one trained with 20ms, their features appear to be complementary to each other, resulting in improved performance for all HuBERT-MR models. We observe that the 100ms-based HuBERT model does not perform well in the task, likely due to the feature sequence being too short for effective CTC-based training.

In Table 5, we compare the performance of HuBERT-MR-H to that of the HuBERT-large and wav2vec2 models. Our findings are as follows:

- HuBERT-MR-H outperforms both the base versions of HuBERT and wav2vec2, highlighting the superior performance of this method.
- Although HuBERT-MR-H is a significant improvement over the base HuBERT model, there is still some performance gap when compared to HuBERT-large. This difference could be

Table 5: Fine-tuning results on Librispeech-100h (comparison with other models). Results with language model rescoring are in brackets. \* indicates the large model setting. The unlabeled column shows the number of hours used for SSL pre-training. HuBERT-MR-H is discussed in Sec 2.2. Noted that the HuBERT base model with 20ms is our baseline.

Model	Res.(ms)	Unlabeled(h)	Param.(M)	MACs(G)	FLOPs(T)	WER(↓)
HuBERT	20	960	94.7	1669	3.34	7.73(3.81)
wav2vec2	20	960	95.0	1669	3.34	6.54(4.33)
wav2vec2*	20	60K	317.4	4326	8.66	5.90(3.45)
HuBERT*	20	60K	316.6	4324	8.66	<b>5.40(2.82)</b>
HuBERT-MR-H	(100,40,20)	960	298.4	3454	6.91	6.11(3.31)

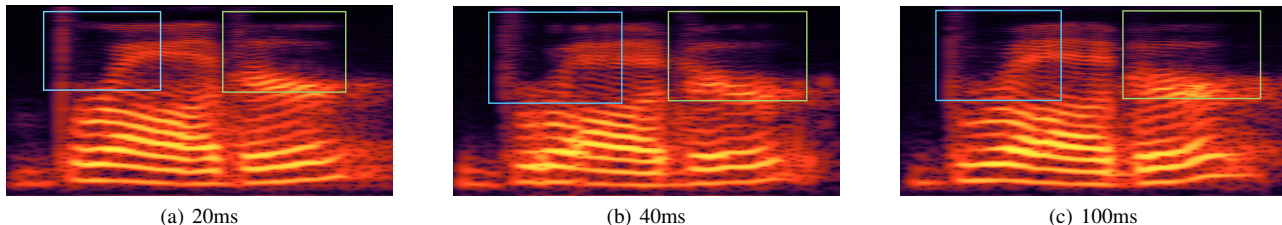


Figure 2: Speech re-synthesis using features from HuBERT base models at different resolutions. The high-resolution features capture better envelope information in the time domain (shown in the blue box), while the low-resolution features provide more detailed information in the frequency domain (shown in the green box). See Sec. 3.4 for experimental details and discussion.

due to the limited training data (960 Librispeech training sets versus 60K Librilight [32]) and fewer pre-training iterations (all three base HuBERT models use two iterations, while HuBERT-large uses three iterations).

- HuBERT-MR-H has a similar parameter size to both HuBERT-large and wav2vec2-large after combining three HuBERT models. However, it requires less computational overhead compared to other large models. This reduction is mainly due to the  $O(T^2)$  complexity of the transformer layers in computing intermediate hidden representations [33]. While HuBERT-MR-H has lower-resolution networks in its sub-module, it can save computational effort, as shown in MACs and FLOPs in Table 5.

### 3.4. Further Discussion

Our experiments show that HuBERT models with different resolutions can extract features from the same speech source in distinct ways that are useful for downstream tasks. To investigate these differences, we analyzed the features extracted by three pre-trained HuBERT models with 100ms, 40ms, and 20ms resolutions. Specifically, we extracted the 6<sup>th</sup> layer representations and used them as input features to train a HiFi-GAN vocoder [34] with the ParallelWaveGAN toolkit [35, 36].<sup>3</sup> We trained the vocoder on the LJSpeech dataset and adapted the upsampling modules to match the resolution of each HuBERT model. The vocoder was trained for 50k steps using the same configuration as the ParallelWaveGAN toolkit. Finally, we generated and compared the spectrograms of synthesized test-set speech produced from different representations, as shown in Figure 2.<sup>4</sup> The followings are some interesting findings:

- HuBERT features at different resolutions are capable of producing high-quality re-synthesized speech. Despite not performing well on ASR fine-tuning tasks (as shown in Table 4), HuBERT with 100ms resolution exhibits excellent speech re-synthesis quality. This suggests that the feature still contains

the necessary information in the speech.

- As shown in Figure 2, high-resolution HuBERT features capture better envelope information in each frame of the speech, while low-resolution features have a more detailed formant presentation. This leads us to hypothesize that high-resolution HuBERT may have a better understanding in the time domain, while low-resolution features have more detailed information in the frequency domain. This property is similar to Short-time Fourier transformation with different window sizes and shifts, to some extent.

## 4. Conclusion

In this study, we revisit the use of HuBERT with multiple resolutions, recognizing that the original 20ms resolution may not be optimal for various downstream tasks. To address this, we propose HuBERT-MR, which integrates information from three HuBERT base models pre-trained with different resolutions. We examine two approaches for integration: a parallel approach (HuBERT-MR-P) and a hierarchical approach (HuBERT-MR-H). We evaluate HuBERT-MR-P on the SUPERB benchmark and both HuBERT-MR models on ASR fine-tuning. Our experiments demonstrate that the HuBERT-MR models significantly improve model performance on various downstream tasks, indicating that pre-trained features from multiple resolutions are complementary. Furthermore, we find that HuBERT-MR can outperform larger models in some scenarios, even with less pre-training data and fewer parameters. To further highlight the differences among HuBERT features at different resolutions, we conduct speech re-synthesis with the HiFi-GAN vocoder. Our results demonstrate that the features do differ across resolutions, while all retain the essential information for intelligibility. We believe this work offers valuable insights into the potential benefits of considering multi-resolution SSL in the field.

## 5. Acknowledgement

This work was supported by a Meta AI SRA grant. J. Shi and S. Watanabe are funded in part of the Delta project, supported by the NSF (award OCI 2005572), and the State of Illinois.

<sup>3</sup><https://github.com/kan-bayashi/ParallelWaveGAN>.

<sup>4</sup>Synthesized audio examples can be found in the <https://www.dropbox.com/s/61ap65iegi93i1/audio-samples-resynthesis.zip>.

## 6. References

- [1] A. Mohamed *et al.*, “Self-supervised speech representation learning: A review,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.
- [2] W.-N. Hsu *et al.*, “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [3] S.-w. Yang *et al.*, “SUPERB: Speech Processing Universal Performance Benchmark,” in *Proc. Interspeech*, 2021, pp. 1194–1198.
- [4] H.-S. Tsai *et al.*, “SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities,” in *Proc. ACL*, 2022, pp. 8479–8492.
- [5] T.-h. Feng *et al.*, “SUPERB@ SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning,” in *Proc. SLT*, 2023, pp. 1096–1103.
- [6] X. Chang *et al.*, “An exploration of self-supervised pretrained representations for end-to-end speech recognition,” in *Proc. ASRU*, 2021, pp. 228–235.
- [7] D. Berrebbi *et al.*, “Combining Spectral and Self-Supervised Features for Low Resource Speech Recognition and Translation,” in *Proc. Interspeech*, 2022, pp. 3533–3537.
- [8] S. H. Mallidi and H. Hermansky, “Novel neural network based fusion for multistream asr,” in *Proc. ICASSP*, 2016, pp. 5680–5684.
- [9] S. H. R. Mallidi *et al.*, “A practical and efficient multistream framework for noise robust speech recognition,” Ph.D. dissertation, Johns Hopkins University, 2018.
- [10] H. Hermansky, “Multistream recognition of speech: Dealing with unknown unknowns,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1076–1088, 2013.
- [11] K. J. Han *et al.*, “Multistream cnn for robust acoustic modeling,” in *Proc. ICASSP*, 2021, pp. 6873–6877.
- [12] J. Luo *et al.*, “Multi-quartznet: Multi-resolution convolution for speech recognition with multi-layer feature fusion,” in *Proc. SLT*, 2021, pp. 82–88.
- [13] R. Li *et al.*, “Multi-stream end-to-end speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 646–655, 2019.
- [14] Y. Kong *et al.*, “Multi-channel automatic speech recognition using deep complex unet,” in *Proc. SLT*, 2021, pp. 104–110.
- [15] A. Andrusenko, R. Nasretdinov, and A. Romanenko, “Uconv-conformer: High reduction of input sequence length for end-to-end speech recognition,” *arXiv preprint arXiv:2208.07657*, 2022.
- [16] S. Kim *et al.*, “Squeezeformer: An efficient transformer for automatic speech recognition,” in *Proc. NeurIPS*.
- [17] W. Yao *et al.*, “Multi-stream convolutional neural network with frequency selection for robust speaker verification,” *arXiv preprint arXiv:2012.11159*, 2020.
- [18] Z. Gao, M.-W. Mak, and W. Lin, “Unet-densenet for robust far-field speaker verification,” *Proc. Interspeech 2022*, pp. 3714–3718, 2022.
- [19] M. Burchi and V. Vielzeuf, “Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition,” in *Proc. ASRU*, 2021, pp. 8–15.
- [20] Y. Zhang *et al.*, “Research on speech enhancement algorithm based on sa-unet,” in *2019 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, 2019, pp. 818–8183.
- [21] T. Zhao *et al.*, “Unet++-based multi-channel speech dereverberation and distant speech recognition,” in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021, pp. 1–5.
- [22] R. Li *et al.*, “Unet-tts: Improving unseen speaker and style transfer in one-shot voice cloning,” in *Proc. ICASSP*, 2022, pp. 8327–8331.
- [23] X. Xiang, X. Zhang, and H. Chen, “A nested u-net with self-attention and dense connectivity for monaural speech enhancement,” *IEEE Signal Processing Letters*, vol. 29, pp. 105–109, 2021.
- [24] Y. Xian *et al.*, “A multi-scale feature recalibration network for end-to-end single channel speech enhancement,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 143–155, 2020.
- [25] G. Liu *et al.*, “Cp-GAN: Context pyramid generative adversarial network for speech enhancement,” in *Proc. ICASSP*, 2020, pp. 6624–6628.
- [26] X. Xiang, X. Zhang, and H. Chen, “A convolutional network with multi-scale and attention mechanisms for end-to-end single-channel speech enhancement,” *IEEE Signal Processing Letters*, vol. 28, pp. 1455–1459, 2021.
- [27] J. Shi *et al.*, “Bridging speech and textual pre-trained models with unsupervised ASR,” *arXiv preprint arXiv:2211.03025*, 2022.
- [28] V. Panayotov *et al.*, “Librispeech: An asr corpus based on public domain audio books,” in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [29] A. Baevski *et al.*, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [30] M. Ott *et al.*, “Fairseq: A fast, extensible toolkit for sequence modeling,” *NAACL HLT 2019*, p. 48, 2019.
- [31] J. D. Kahn *et al.*, “Flashlight: Enabling innovation in tools for machine learning,” in *Proc. ICML*, 2022, pp. 10 557–10 574.
- [32] J. Kahn *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *Proc. ICASSP*, 2020, pp. 7669–7673.
- [33] Y. Meng *et al.*, “On compressing sequences for self-supervised speech models,” in *Proc. SLT*, 2023.
- [34] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [35] T. Hayashi *et al.*, “ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in *Proc. ICASSP*, 2020, pp. 7654–7658.
- [36] T. Hayashi *et al.*, “ESPnet2-TTS: Extending the edge of TTS research,” *arXiv preprint arXiv:2110.07840*, 2021.