OLIVE: VIEW-AUGMENTED LATENT PREDICTION WITH WAVEFORM RECONSTRUCTION FOR SPEECH SSL

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose **O**nline **L**atent prediction with **I**nvariant **V**iews and r**E**construction (OLIVE), a self-supervised speech representation learning framework that jointly optimizes analysis and synthesis objectives. OLIVE couples masked latent prediction and waveform reconstruction through a functional separation within a single encoder: reconstruction constrains early encoder features to retain signal-level information, while masked latent prediction shapes later contextual representations for robust downstream performance, with waveform view augmentations defining the invariances they learn. To the best of our knowledge, OLIVE is the first speech SSL model that directly reconstructs waveforms, retains its vocoder after pre-training, and remains competitive on discriminative tasks. Adding reconstruction largely preserves the speech recognition performance of masked latent prediction alone, while improving waveform reconstruction and speaker tasks. Adding waveform mixup and gain augmentations provides a controllable way to shift the representations further toward speaker, generation, and reconstruction tasks, while remaining competitive on recognition and semantic tasks.

1 INTRODUCTION

Recent advances in speech representation learning have been largely driven by self-supervised learning (SSL) on large-scale unlabeled audio (Mohamed et al., 2022; Liu et al., 2022). Models such as wav2vec 2.0 (Baevski et al., 2020b), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), and data2vec (Baevski et al., 2022; 2023) learn contextual representations that serve as general-purpose feature extractors, supporting both linguistic tasks such as automatic speech recognition and paralinguistic tasks including speaker and other non-linguistic acoustic attributes such as diarization and speech emotion recognition (Yang et al., 2021; Baroudi et al., 2024; Atmaja & Sasou, 2022; Purohit & Magimai-Doss, 2025). These methods inherit ideas from contrastive prediction (van den Oord et al., 2018), masked modeling (Devlin et al., 2019; Liu et al., 2020), quantization and discrete target discovery (Ling & Liu, 2020; Liu et al., 2021), and teacher–student distillation (Grill et al., 2020; Caron et al., 2021). Existing approaches are nevertheless predominantly focused on discriminative tasks: they aim to extract representations that are maximally informative for downstream predictive tasks, without explicitly preserving the information required for speech signal generation.

In contrast, speech communication fundamentally arises from two tightly coupled phenomena: *perception* (hearing) and *production* (speaking). Classical speech processing has long embraced this duality through the *analysis–synthesis* paradigm, where signals are decomposed into structured components and subsequently reconstructed (Atal & Hanauer, 1971; McAulay & Quatieri, 1986; Ghitza, 1986). This framework leverages the fact that speech production is governed by a well-defined physical and physiological process, making it amenable to explicit modeling (Fant, 1960; Titze, 1994; Zhang, 2016). Recent neural work has revisited this perspective by training decoders on frozen self-supervised representations, effectively forming a neural analysis–synthesis pipeline (Choi et al., 2021; Polyak et al., 2021). This has enabled a range of synthesis approaches for tasks such as voice conversion (Baas et al., 2023), text-to-speech (El Hajal et al., 2025), and speech enhancement (Irvin et al., 2023) that leverage structured information in intermediate self-supervised representations.

Despite this motivation, synthesis remains peripheral in most speech SSL systems, and coupling analysis and synthesis during pre-training is non-trivial. Reconstruction requires signal-level detail such as speaker and fine spectral structure, whereas discriminative SSL benefits from representations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

that abstract away such variability, so a reconstruction objective risks degrading the contextual representations that make SSL effective for recognition. UniWav, which trains an encoder jointly with a generative decoder, reports lower recognition performance than SSL models of similar scale as its main limitation (Liu et al., 2025). data2vec-SG keeps recognition close to data2vec, but treats reconstruction as a lightly weighted auxiliary loss on its top contextual layers, and its decoder is discarded after pre-training (Wang et al., 2023). It thus remains open how to reconstruct waveforms directly, retain a functional decoder, and preserve the discriminative strength of speech SSL.

In this work, we propose **OLIVE** (**O**nline **L**atent prediction with **I**nvariant **V**iews and r**E**construction), a self-supervised speech representation learning framework that jointly optimizes analysis and synthesis objectives. OLIVE resolves this tension through a *functional separation* within a single encoder: the synthesis branch constrains early local encoder features to preserve information for waveform reconstruction, while the analysis branch, based on masked latent prediction, shapes later contextual representations toward invariance for robust downstream analysis. This stands in contrast to purely discriminative models, which may discard information critical for generation, and purely generative approaches, which may underperform on discriminative tasks. Within the analysis branch, independently augmented waveform views for the student and teacher act as an explicit prior over the learned invariances, which can be adapted to the requirements of different applications. We validate OLIVE against established SSL baselines on a broad suite of downstream tasks and on direct waveform reconstruction, and study both the effect of joint training and the effect of waveform view augmentations. Code and model weights will be released upon publication.

Contributions.

- We introduce **OLIVE**, a speech SSL framework that jointly optimizes analysis and synthesis in a single pre-training stage; to our knowledge, it is the first speech SSL model that directly reconstructs waveforms, retains its vocoder after pre-training, and remains competitive on discriminative tasks.
- We propose a functional separation in which waveform reconstruction acts on early local encoder features, while later contextual representations are shaped primarily by masked latent prediction.
- We show that adding reconstruction largely preserves the recognition performance of masked latent prediction alone (4.9 vs. 4.8 WER), while improving reconstructability, yielding a perceptually effective integrated decoder, and restoring speaker information.
- We find that waveform view augmentations provide a controllable way to shift representations toward speaker, generation, and reconstruction tasks while remaining competitive on recognition and semantic tasks.

2 Related Work

Self-supervised speech representation learning. Early speech SSL methods learned from raw waveforms using contrastive objectives that predict latent representations of future or masked inputs from past context (van den Oord et al., 2018; Schneider et al., 2019; Baevski et al., 2020a). Wav2vec 2.0 introduced masked latent prediction with quantized targets and a contrastive loss at masked time steps (Baevski et al., 2020b). HuBERT replaced online quantization with masked prediction of offline clustered units (Hsu et al., 2021), while WavLM extended this line with denoising and large-scale pre-training for a broader set of speech tasks (Chen et al., 2022). Data2vec instead predicts continuous contextualized teacher representations, avoiding discrete targets (Baevski et al., 2022). Data2vec 2.0 improves the efficiency of this masked self-distillation setup by reusing teacher targets across masked views, encoding only unmasked steps, and using a lightweight convolutional decoder (Baevski et al., 2023). OLIVE uses masked latent prediction within a broader analysis–synthesis framework: the student and teacher receive independently augmented waveform views, and a synthesis branch performs waveform reconstruction from local encoder features.

View-based self-distillation. View-based self-supervised learning trains representations to agree across stochastic views of the same input (Schmidhuber & Prelinger, 1993). BYOL showed that an online network can predict the representation of a slowly updated target network without explicit negative pairs, learning invariances through data augmentations (Grill et al., 2020). In audio, BYOL-A adapts this principle using mixup and time–frequency transformations on log-mel spectrograms (Ni-

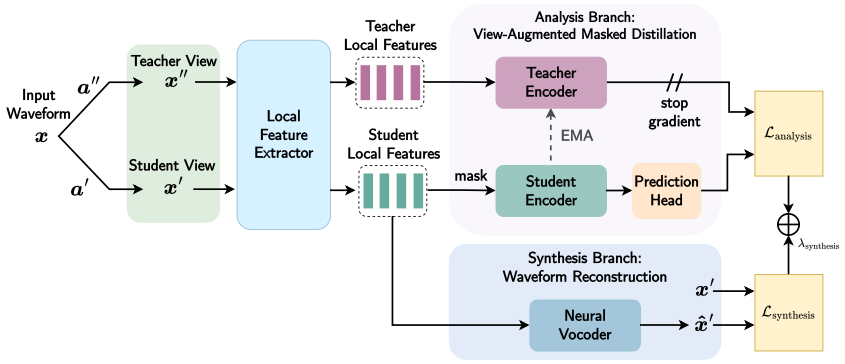


Figure 1: **OLIVE pre-training framework.** Two independently augmented waveform views are passed through a shared local feature extractor. The analysis branch performs view-augmented masked distillation: the student predicts contextual teacher targets from masked input features. The synthesis branch performs waveform reconstruction by conditioning a neural vocoder on the local student features. The final objective combines the analysis loss with a weighted synthesis loss.

izumi et al., 2021), and BYOL-S further studies bootstrapped speech representations across encoder architectures and pre-training data choices (Elbanna et al., 2022). WaveBYOL extends this approach to raw waveforms using time-domain augmentations, primarily targeting general audio tasks (Kim & Choi, 2023). Similar to data2vec, WavJEPA predicts latent targets from raw waveforms (Yuksel et al., 2025), following the broader JEPA framework (Assran et al., 2023). OLIVE builds on this line by combining masked latent prediction with time-domain waveform augmentations to explicitly induce invariances, while jointly optimizing a synthesis objective.

Speech reconstruction and analysis–synthesis. Recent work trains neural vocoders such as HiFi-GAN (Kong et al., 2020) on frozen self-supervised features, effectively forming neural analysis–synthesis pipelines (Choi et al., 2021; Polyak et al., 2021; Siuzdak et al., 2022; Guo et al., 2024), enabling synthesis methods that leverage self-supervised representations. Such methods typically treat synthesis as a downstream process, rather than integrating it into representation learning during pre-training. Neural audio codecs further explore this analysis–synthesis interface by learning compact latent representations that can be decoded back to waveforms (Zeghidour et al., 2021; Défossez et al., 2023). These works primarily target efficient storage, transmission, or speech language modeling, whereas OLIVE uses reconstruction as a training objective for representation learning.

Closer to speech SSL pre-training, data2vec-SG adds a decoder that reconstructs the clean waveform from the top contextual layers, and discards it after pre-training (Wang et al., 2023). UniWav conditions a flow-matching decoder on a weighted sum of all encoder layers to generate intermediate codec latents rather than waveforms, but reports lower recognition performance than SSL models of similar scale as its main limitation (Liu et al., 2025). Related two-stage approaches first learn latent representations and then train a separate reconstruction pathway (Ioannides et al., 2025), while masked generative pre-training methods such as Metis operate on SSL features for downstream speech generation (Wang et al., 2025). OLIVE differs along three axes: its decoder is conditioned on early local encoder features, leaving later contextual representations to be shaped primarily by the masked distillation objective; it reconstructs the input waveform directly; and it is retained as a usable vocoder after pre-training. Both objectives are optimized in a single pre-training stage, and OLIVE remains competitive with SSL models of similar scale on speech recognition (Section 4.2).

3 Method

OLIVE learns speech representations by jointly solving two self-supervised tasks, which we term *analysis* and *synthesis*. The framework is illustrated in Figure 1. The analysis task encourages the encoder to produce contextual representations that are invariant to augmentation-induced variability, while the synthesis task encourages early local encoder features to preserve acoustic information relevant to waveform reconstruction. We refer to this division of labor as a *functional separation*: the

two objectives share one encoder but act at different depths, so reconstruction can add signal-level information without directly constraining the contextual representations used for recognition.

3.1 Analysis: View-Augmented Masked Distillation

The analysis branch uses the teacher–student masked prediction objective of data2vec 2.0 (Baevski et al., 2023) with independently sampled waveform views for the student and teacher. The augmentation family acts as an explicit prior over the learned invariances (e.g., to environment, speaker, pitch, time, or gain) and can be adapted to applications with different requirements. We use $'$ to denote student views and $''$ to denote teacher views. Given an input waveform x, we sample two stochastic augmentations and form the student view $x' = a'(x)$ and the teacher view $x'' = a''(x)$. The student processes a masked view of x', while the teacher processes the unmasked view x''. The student is trained to predict contextual latent targets provided by the teacher at masked positions. We decompose the student and teacher encoders into a local feature extractor followed by a contextual encoder, and masking is applied to local latent features before contextualization.

Let e_θ denote the student local feature extractor with parameters θ, f_θ the student contextual encoder, and $e_{\bar{\theta}}, f_{\bar{\theta}}$ their teacher counterparts. Let $m(\cdot)$ denote the masking operation applied in latent space to the student local features. The local feature sequences over T frames and the corresponding contextual representations are

$$r'_{1:T} = e_\theta(x'), \quad \bar{r}''_{1:T} = e_{\bar{\theta}}(x''), \quad h'_{1:T} = f_\theta(m(r'_{1:T})), \quad \bar{h}''_{1:T} = f_{\bar{\theta}}(\bar{r}''_{1:T}). \tag{1}$$

The teacher parameters are updated as an exponential moving average (EMA) of the student parameters, $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$, with target decay rate $\tau \in [0, 1]$.

Teacher targets average the instance-normalized (IN) outputs of the top K teacher encoder layers. With $\bar{h}_t^{\prime\prime,\ell}$ the teacher representation at layer $\ell \in \{1, \ldots, L\}$ and time step t, the target is

$$z''_t = \frac{1}{K} \sum_{\ell=L-K+1}^{L} \mathrm{IN}\left(\bar{h}_t^{\prime\prime,\ell}\right). \tag{2}$$

The student predicts these targets at masked time steps, with stop-gradient $\mathrm{sg}(\cdot)$:

$$\mathcal{L}_{\text{analysis}} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \|p_\theta(h'_t) - \mathrm{sg}(z''_t)\|_2^2 \cdot d^{-1/2}, \tag{3}$$

where $\mathcal{M}$ is the set of masked frames, p_θ a prediction head, and d the representation dimension.

3.2 Synthesis: Waveform Reconstruction

The synthesis branch reconstructs the waveform from early encoder features with a HiFi-GAN vocoder (Kong et al., 2020). HiFi-GAN is an adversarially trained generator paired with waveform discriminators that distinguish real from generated audio. We replace the usual mel-spectrogram conditioning with the OLIVE encoder representation, so that reconstruction directly constrains the representation itself. Using the unmasked student local feature sequence $r'_{1:T}$ extracted from the waveform view x', the generator G_ψ with parameters ψ produces a waveform reconstruction $\hat{x}' = G_\psi(r'_{1:T})$.

The adversarial component uses two discriminator families. The multi-period discriminator (MPD) contains sub-discriminators with different periods, capturing harmonic and pitch-period structure. The multi-scale discriminator (MSD) operates at several temporal resolutions, encouraging realistic structure across scales (Kumar et al., 2019; Kong et al., 2020). The synthesis objective combines mel-spectrogram reconstruction, adversarial training, and feature matching.

The mel reconstruction loss is $\mathcal{L}_{\text{mel}} = \| \mathrm{mel}(x') - \mathrm{mel}(\hat{x}')\|_1$, where $\mathrm{mel}(\cdot)$ is the log-mel spectrogram transform. Let $\mathcal{D}_\phi = \{D_\phi^j\}_{j=1}^J$ be the discriminators. The adversarial losses are:

$$\mathcal{L}_{\text{gen}} = \sum_{j=1}^{J} \left(D_\phi^j(\hat{x}') - 1\right)^2, \qquad \mathcal{L}_{\text{disc}} = \sum_{j=1}^{J} \left[\left(D_\phi^j(x') - 1\right)^2 + \left(D_\phi^j(\mathrm{sg}(\hat{x}'))\right)^2\right]. \tag{4}$$

The feature-matching loss $\mathcal{L}_{\text{fm}} = \sum_{j=1}^{J} \sum_{i=1}^{T_j} N_{j,i}^{-1} \|D_{\phi,i}^{j}(x') - D_{\phi,i}^{j}(\hat{x}')\|_1$ compares the activations $D_{\phi,i}^{j}(\cdot)$ of layer i in sub-discriminator j, where T_j is the number of feature layers and $N_{j,i}$ the number of elements in that activation. The mel-spectrogram term anchors the reconstruction to the reference signal in a stable time-frequency space, while the adversarial and feature-matching terms encourage perceptually realistic details, with λ_{fm} and λ_{mel} weighting the corresponding terms in the generator objective $\mathcal{L}_{\text{synthesis}}^{G} = \mathcal{L}_{\text{gen}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}}$; the discriminator objective is $\mathcal{L}_{\text{synthesis}}^{D} = \mathcal{L}_{\text{disc}}$.

3.3 Joint Optimization

The final objective balances invariant representation learning with waveform reconstruction, $\mathcal{L}_{\text{OLIVE}} = \mathcal{L}_{\text{analysis}} + \lambda_{\text{synthesis}}\mathcal{L}_{\text{synthesis}}^{G}$, where $\lambda_{\text{synthesis}}$ controls the analysis–synthesis trade-off. This objective updates the student encoder, prediction head, and generator, while the teacher is updated by EMA and the discriminator parameters are updated separately.

4 Experiments

We describe OLIVE implementation details in Section 4.1, then evaluate OLIVE on a broad suite of downstream speech tasks in Section 4.2, and on waveform reconstruction in Section 4.3.

4.1 Implementation and Pre-training Details

We implement OLIVE in fairseq[1]. The analysis branch implements independently augmented student and teacher views, and uses masked-prediction configurations from the data2vec 2.0 speech implementation. This is coupled with waveform reconstruction in the joint variants. We pre-train on the 960-hour LibriSpeech dataset (Panayotov et al., 2015) and focus on *Base*-sized models due to computational constraints. We train and evaluate four variants: **OLIVE-A (Mix)** is an analysis-only variant using waveform mixup augmentations, and **OLIVE-A (Mix+Gain)** additionally applies gain perturbation. The joint variants, denoted **OLIVE-J**, optimize both masked prediction and waveform reconstruction: **OLIVE-J (Mask-only)** does not use waveform augmentations, while **OLIVE-J (Mix)** uses waveform mixup. All four variants share the same 93.8M-parameter analysis architecture; the OLIVE-J variants also share the same synthesis architecture and loss. To select which analysis-only variants to train, we first trained *Small*-sized models with hidden size $d = 384$ and evaluated their performance on a representative subset of downstream tasks (Appendix C.2). We describe the OLIVE pre-training setup in three parts: (i) the shared encoder and masked distillation objective, (ii) the waveform view augmentations used, and (iii) the synthesis branch in the joint models.

Encoder and masked distillation. OLIVE uses a raw-waveform speech encoder architecture following prior SSL models (Baevski et al., 2020b; 2023), with a 7-layer convolutional waveform encoder, a convolutional positional encoder, and a 12-layer Transformer (Vaswani et al., 2017) with hidden size $d = 768$. The waveform encoder has an overall stride of 320 samples, corresponding to 20 ms at 16 kHz. It is shared between the student and teacher networks. Teacher targets are formed by instance-normalizing and averaging the top $K = 8$ teacher layers, and the student predicts these targets with a convolutional prediction head. Similarly to data2vec 2.0, we use inverse-block masking and eight masked views per sample. Final models are trained for 400k updates with Adam and cosine learning-rate decay. The teacher EMA decay τ is annealed from 0.999 to 0.99999 over the first 75k updates. We use mixed precision for the analysis branch and FP32 for the synthesis branch for numerical stability. Further details and compute are in Appendices A and J.

Waveform view augmentations. We use waveform-level augmentations to define invariances for the analysis objective. We focus on perturbations related to acoustic environment and recording conditions, while avoiding transformations such as pitch shifting or time-scale modification that may remove information relevant to speaker identity, prosody, and paralinguistic tasks. **Mixup** mixes the input waveform with an utterance sampled from a memory bank using linear interpolation with a mixing ratio uniformly sampled from $[0, 0.4]$, encouraging robustness to background acoustic interference and following the general regularization principle of mixup-style training (Zhang et al.,

[1] https://github.com/facebookresearch/fairseq

2018; Niizumi et al., 2021). **Gain** applies a random global gain followed by clamping to $[-1, 1]$, encouraging invariance to recording gain and mild clipping artifacts. We omit gain perturbation in joint training, since clipping artifacts may adversely affect waveform reconstruction.

Synthesis branch. For OLIVE-J, the synthesis branch reconstructs the student view x' from its local encoder features with a HiFi-GAN V2 vocoder (Kong et al., 2020). Synthesis training uses 7040-sample waveform segments (0.44 s at 16 kHz) randomly sampled from the full extracted waveform. In the final OLIVE-J models, we set $\lambda_{\text{synthesis}} = 25$, chosen after ablations (Appendix C.1). A key design choice is which encoder representation conditions the vocoder. Earlier layers typically preserve the fine-grained acoustic detail needed for waveform reconstruction; prior work on synthesis-based speech enhancement found that vocoders conditioned on SSL local features can benefit from the acoustic detail available before deeper contextual abstraction (Irvin et al., 2023). Our ablations confirm this under matched separate-vocoder training (Appendix C.3), so we condition the vocoder on local encoder features. OLIVE-J optimizes 165.6M parameters during training, of which 70.7M belong to discriminators used only for adversarial training and discarded afterwards; waveform reconstruction then uses only the shared CNN feature extractor (4.2M) and the HiFi-GAN V2 generator (1.1M) (Appendix A).

4.2 Downstream Evaluation

We use the Speech Processing Universal PERformance Benchmark (SUPERB) (Yang et al., 2021; Tsai et al., 2022) using the S3PRL toolkit.[2] SUPERB evaluates frozen SSL models as general-purpose feature extractors. For each task, a learned weighted sum of all encoder layers is fed to a lightweight task head, and only the layer weights and the head are trained on labeled data.

Tasks. We evaluate OLIVE across the five SUPERB task categories: content, speaker, paralinguistic, semantic, and generation. The downstream tasks are phoneme recognition (PR), automatic speech recognition (ASR), keyword spotting (KS), query-by-example spoken term detection (QbE), speaker identification (SID), automatic speaker verification (ASV), speaker diarization (SD), emotion recognition (ER), intent classification (IC), slot filling (SF), speech translation (ST), speech enhancement (SE), source separation (SS), and voice conversion (VC). Unlike some prior work (Chen et al., 2022), we do not perform task-specific hyperparameter search due to compute constraints; all models are evaluated with the same default task configurations in S3PRL.

Baselines. All SSL baselines in the main comparison are Base-sized models trained on LibriSpeech 960h and have similar encoder architectures. Therefore, the comparison primarily isolates differences in pre-training objectives rather than data, architecture, or model capacity. We compare OLIVE with handcrafted log mel-filterbank acoustic features (FBANK) and widely used speech SSL baselines: wav2vec 2.0 (Baevski et al., 2020b), HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022), data2vec (Baevski et al., 2022), and data2vec 2.0 (Baevski et al., 2023). We use official public checkpoints and implementations from S3PRL. For data2vec-SG (Wang et al., 2023) and UniWav (Liu et al., 2025), no public checkpoints are provided. For completeness, we include paper-reported data2vec-SG results (†) on the overlapping generation tasks (see Appendix B for differences relative to each evaluation's data2vec baseline). We do not include UniWav in the main comparison because it differs substantially in model scale and pre-training data.

Scores. We report two aggregate scores. First, we use the SUPERB_s metric (Feng et al., 2023), which measures the fraction of the gap between the FBANK baseline and a task-specific reference score denoted as SOTA. For model u, task set T, and metric set I_t for task t, we compute

$$\text{SUPERB}_s(u) = \frac{1000}{|T|} \sum_{t \in T} \frac{1}{|I_t|} \sum_{i \in I_t} \frac{s_{t,i}(u) - s_{t,i}(\text{FBANK})}{s_{t,i}(\text{SOTA}) - s_{t,i}(\text{FBANK})}.$$

For tasks with multiple evaluation metrics, normalized metric scores are averaged within the task before averaging across tasks. We also report a WavLM-style aggregate score (Chen et al., 2022), $\text{WavLM}_s(u) = \frac{1}{|T|} \sum_{t \in T} \tilde{s}_t(u)$, which averages per-task utility scores $\tilde{s}_t$ after simple metric normalization: error rates such as WER and EER are converted to utilities using $100 - \text{error}$, QbE is scaled by 100, and higher-is-better metrics are used directly.

[2] https://github.com/s3prl/s3prl

Table 1: SUPERB downstream results on content, speaker, paralinguistic, and semantic tasks. Params denotes the encoder parameter count. **Bold**: best; underlined: second-best.

Model	Params	Content				Speaker			Paraling.	Semantics			
		PR	ASR	KS	QbE	SID	ASV	SD	ER	IC	SF		ST
		PER↓	WER↓	Acc.↑	MTWV↑	Acc.↑	EER↓	DER↓	Acc.↑	Acc.↑	F1↑	CER↓	BLEU↑
FBANK	–	82.1	23.1	8.3	0.0043	0.1	10.7	11.3	28.4	4.6	63.0	59.5	2.6
wav2vec 2.0 Base	95.04M	6.4	6.5	96.0	0.0634	66.5	5.8	6.6	62.5	92.6	87.0	26.5	14.9
HuBERT Base	94.70M	5.8	6.4	96.4	0.0748	67.3	5.6	6.2	65.1	98.2	88.0	25.1	15.8
WavLM Base	94.38M	5.6	6.2	97.0	**0.0813**	62.3	5.8	5.4	64.9	98.5	89.0	24.2	16.3
data2vec Base	93.16M	**3.8**	5.0	95.9	0.0609	54.9	7.0	7.0	65.4	98.6	89.0	23.4	17.5
data2vec 2.0 Base	93.16M	4.6	**4.8**	96.8	0.0556	45.5	5.5	6.3	63.9	99.0	89.0	23.4	**17.8**
OLIVE-A (Mix)	93.16M	6.4	6.4	97.0	0.0390	66.9	6.0	4.6	67.1	**99.1**	88.0	25.2	16.5
OLIVE-A (Mix+Gain)	93.16M	5.2	6.1	**97.3**	0.0459	63.0	**5.5**	**4.3**	**67.8**	98.9	88.0	24.4	16.8
OLIVE-J (Mask-only)	93.16M	4.7	4.9	97.0	0.0480	64.6	5.7	6.8	66.9	98.6	**89.4**	**22.3**	17.3
OLIVE-J (Mix)	93.16M	6.3	6.4	**97.3**	0.0445	**83.1**	5.8	4.6	64.6	98.1	88.0	25.8	15.4

Table 2: SUPERB downstream results on generation tasks, with overall SUPERB$_s$ and WavLM$_s$ scores. **Bold**: best; underlined: second-best.

Model	SE		SS	VC			Overall	
	PESQ↑	STOI↑	SI-SDRi↑	MCD↓	WER↓	ASV↑	SUPERB$_s$↑	WavLM$_s$↑
FBANK	2.83	94.3	9.2	8.21	41.40	87.0	0	34.2
data2vec-SG Base†	2.59	94.0	10.8	–	–	–	–	–
wav2vec 2.0 Base	2.92	94.8	10.3	7.50	11.30	97.8	795	65.5
HuBERT Base	3.00	94.9	9.9	7.48	10.90	98.0	836	66.5
WavLM Base	2.99	94.9	10.8	7.44	**8.95**	98.2	875	66.6
data2vec Base	2.94	94.8	9.9	**7.08**	9.65	**99.5**	805	65.9
data2vec 2.0 Base	2.96	94.9	10.4	7.28	10.40	**99.5**	835	65.3
OLIVE-A (Mix)	3.05	**95.2**	12.2	7.71	12.20	95.5	884	66.7
OLIVE-A (Mix+Gain)	3.06	**95.2**	12.2	7.59	12.10	96.2	**909**	66.8
OLIVE-J (Mask-only)	2.94	94.8	10.5	7.28	11.10	99.0	835	66.8
OLIVE-J (Mix)	**3.10**	95.1	**12.3**	7.61	11.60	98.2	908	**67.6**

Main results. Task-level results and overall SUPERB$_s$ and WavLM$_s$ scores are presented in Tables 1 and 2. The OLIVE variants achieve strong aggregate results across downstream tasks, with their different setups shifting the balance between recognition and tasks tied to fine-grained acoustic detail. Specifically, **OLIVE-A (Mix+Gain)** achieves the best SUPERB$_s$ score, whereas **OLIVE-J (Mix)** achieves the best WavLM$_s$ score and the strongest speaker and generation category scores. **OLIVE-J (Mask-only)** ties for the second-best WavLM$_s$ score, achieves strong ASR results (4.9 WER versus 4.8 for data2vec 2.0; see Appendix F for results across LibriSpeech splits), and leads on SF. Nonetheless, the augmented OLIVE variants remain competitive on ASR with wav2vec 2.0, HuBERT, and WavLM. Appendix E illustrates these differences through normalized per-task and category-level profiles. Appendix D visualizes layer-combination weights for a representative set of tasks, illustrating the emphasis on earlier layers for speaker and generation tasks and on later layers for content and semantic tasks. Overall, the functional separation retains strong recognition performance, while waveform view augmentations further steer the representations toward tasks that rely on acoustic detail and speaker information.

4.3 Waveform Reconstruction Evaluation

Setup. We evaluate waveform reconstruction directly by training a separate feature-conditioned HiFi-GAN V2 vocoder from scratch on the frozen features of each evaluated model, and comparing them with the integrated vocoder retained from each OLIVE-J variant. All separate vocoders use the same HiFi-GAN V2 architecture and training setup as the jointly trained vocoders. They are trained on the LibriSpeech training set with mixup augmentation (Appendix C.4) for 400k updates (Appendix A) and evaluated on LibriSpeech `test-clean`; results are thus matched comparisons under a common configuration, not fully converged upper bounds.

Models. The reconstruction baselines include mel-spectrogram, wav2vec 2.0, HuBERT, WavLM, data2vec, and data2vec 2.0 features. For SSL representations, we condition the vocoder on local encoder features, which preserve the most acoustic detail (Appendix C.3; Irvin et al., 2023) and match

Table 3: Waveform reconstruction on LibriSpeech test-clean with matched and integrated HiFi-GAN V2 vocoders (means; 95% CIs in Appendix H). **Bold**: best; underlined: second-best.

Model	Mel-L1 ↓	MCD ↓	LSD ↓	F_0 MAE ↓	STOI ↑	PESQ ↑	ViSQOL ↑	SI-SDR ↑	SNR ↑	UTMOS ↑
Reference audio										4.09
Mel spectrogram	0.677	6.24	0.829	14.4	0.833	1.42	3.59	−18.02	−2.60	2.98
wav2vec 2.0 Base	0.474	5.18	0.604	11.6	0.912	2.19	4.04	−7.53	−0.42	3.58
HuBERT Base	0.475	5.05	0.604	13.3	0.914	2.11	4.05	−11.78	−1.57	3.63
WavLM Base	0.470	5.01	0.599	13.1	0.917	2.21	4.05	−10.93	−1.31	3.68
data2vec Base	0.467	4.90	0.594	12.8	0.902	2.10	4.06	−10.64	−1.15	3.50
data2vec 2.0 Base	0.434	4.78	0.558	13.0	0.917	2.21	4.10	−9.67	−1.11	3.63
Frozen feature-conditioned vocoder										
OLIVE-A (Mix+Gain)	0.446	4.69	0.566	11.6	0.929	2.62	4.18	−3.36	1.06	3.70
OLIVE-J (Mask-only)	**0.410**	4.43	**0.525**	10.8	0.935	2.81	4.26	0.97	3.31	**3.86**
OLIVE-J (Mix)	0.421	**4.35**	0.528	**9.6**	**0.942**	**3.06**	**4.34**	**2.94**	**4.69**	3.83
Integrated vocoder										
OLIVE-J (Mask-only)	0.476	4.68	0.593	11.1	0.911	2.56	4.14	0.72	3.08	3.74
OLIVE-J (Mix)	0.579	4.56	0.688	10.0	0.921	2.88	4.21	2.50	4.25	3.78

the feature level used by the OLIVE-J synthesis branch. Among the analysis-only OLIVE variants, we report **OLIVE-A (Mix+Gain)** based on the layer ablation in Appendix C.3. For each of **OLIVE-J (Mask-only)** and **OLIVE-J (Mix)**, we report two reconstruction conditions. The *integrated vocoder* is the generator retained from joint pre-training. The *frozen feature-conditioned* vocoder is trained separately with the same HiFi-GAN V2 setup as the baselines on frozen features from the jointly pre-trained encoder; this isolates the reconstructability of the learned features.

Metrics. We report spectral, intrusive, and perceptual metrics. Spectral fidelity is assessed using log-mel spectrogram L1 distance (Mel-L1), log-spectral distance (LSD), and mel-cepstral distortion (MCD). Pitch fidelity is measured with frame-level fundamental-frequency mean absolute error (F_0 MAE). We additionally report intrusive intelligibility and quality metrics, namely STOI (Taal et al., 2011) and PESQ (Rix et al., 2001), as well as scale-invariant signal-to-distortion ratio (SI-SDR) and signal-to-noise ratio (SNR). We also include UTMOS (Saeki et al., 2022) and ViSQOL (Hines et al., 2015) as learned perceptual metrics. Generated and reference signals are loudness-matched.

Main results. Results are shown in Table 3. OLIVE substantially improves waveform reconstructability. We first observe that the analysis-only **OLIVE-A (Mix+Gain)** model, trained with waveform augmentations, is competitive with or better than strong analysis-only baselines on most reconstruction metrics. For the **OLIVE-J** models, vocoders conditioned on frozen features outperform the analysis-only baselines across all metrics under matched separate-vocoder training. **OLIVE-J (Mask-only)** with a frozen feature-conditioned vocoder achieves the best Mel-L1, LSD, and UTMOS scores. **OLIVE-J (Mix)** with a frozen feature-conditioned vocoder is strongest on all other metrics. Meanwhile, the integrated **OLIVE-J** vocoders perform strongly on perceptual and signal-level metrics while being less precise on Mel-L1 and LSD, which likely reflects the demands of joint training on the vocoder. Overall, joint analysis–synthesis pre-training makes the learned representation more informative for reconstruction, while yielding a perceptually effective decoder. Appendix I provides spectrogram visualizations of reconstructed waveforms, and Appendix G reports intelligibility (WER) and speaker similarity (SECS) of the resynthesized speech.

5 Analysis of Augmentation and Reconstruction

We analyze the effects of augmentation and reconstruction on the learned representations through a 2×2 comparison: data2vec 2.0 uses neither, OLIVE-A (Mix) adds augmentation, OLIVE-J (Mask-only) adds reconstruction, and OLIVE-J (Mix) combines both. Figure 2 shows the effect of each component and Figure 3 the resulting SUPERB category profiles.

Augmentation trades frame-level precision for invariance. Adding mixup has the same effect with or without reconstruction (Figure 3): speaker and generation scores increase, as does reconstruction quality on perceptual and signal-level metrics (e.g., PESQ 2.81 to 3.06 with reconstruction),

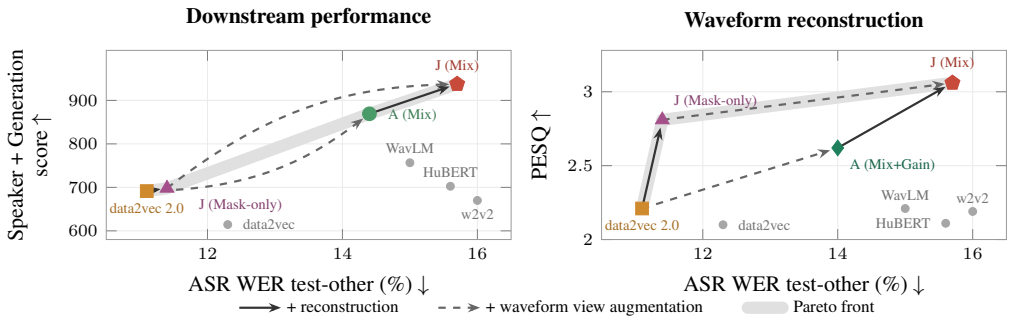


Figure 2: **Effects of waveform view augmentation and reconstruction.** x-axis: LibriSpeech `test-other` WER (no LM; Appendix F). Left: mean Speaker and Generation SUPERB score (normalized per task: FBANK = 0, best = 1000). Right: PESQ of frozen feature-conditioned vocoders, with OLIVE-A (Mix+Gain) replacing OLIVE-A (Mix). Upper left is better.

while content and semantic scores decrease slightly. Further, predicting across independently augmented views favors tasks that pool over segments or utterances (SID, SD, KS) at the expense of the frame-level detail required by PR, ASR, and QbE. Among generation tasks, SE and SS gain most, as mixup views resemble the interference these tasks remove. The SUPERB VC task behaves differently: being any-to-one, it rewards features that discard source-speaker identity, so the baselines with the lowest SID accuracy (data2vec, data2vec 2.0) obtain among the best VC scores.

Reconstruction restores the acoustic detail discarded by masked latent prediction. Masked latent prediction regresses averaged, normalized contextual targets and thus has no incentive to retain signal-level detail: the data2vec models have the lowest baseline SID accuracy (Table 1) and reconstruct waveforms no better than other SSL features (Table 3). Adding reconstruction improves every reconstruction metric in both settings (without augmentation, PESQ from 2.21 to 2.81 and SI-SDR from -9.7 to $+1.0$ dB) and SID accuracy (45.5 to 64.6 without augmentation, 66.9 to 83.1 with it), while PR and ASR are nearly unchanged. This matches OLIVE's functional separation: reconstruction acts on the local encoder features while masked prediction governs the contextual layers, so acoustic detail is recovered without the recognition cost of augmentation.

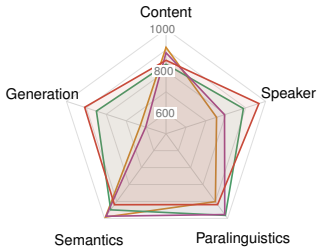


Figure 3: **SUPERB category scores:** **data2vec 2.0**, **OLIVE-A (Mix)** (+aug.), **OLIVE-J (Mask-only)** (+recon.), **OLIVE-J (Mix)** (both), normalized as in Figure 2.

The two components are complementary. Combined with augmentation, reconstruction further improves generation, and partly offsets the content cost of augmentation. Both OLIVE-J models therefore lie on the Pareto fronts of Figure 2: OLIVE-J (Mask-only) nearly matches data2vec 2.0 on recognition with much better reconstruction, and OLIVE-J (Mix) obtains the best speaker and generation scores and PESQ. In practice, reconstruction suffices for recognition-oriented applications, while speaker- and generation-oriented applications benefit from combining it with mixup augmentations.

6 Conclusion

We introduced **OLIVE**, a speech SSL framework that jointly optimizes analysis and synthesis during pre-training through a functional separation: reconstruction acts on early local encoder features, while later contextual representations are shaped by view-augmented masked latent prediction. Joint training improves speaker, generation, and waveform reconstruction performance over analysis-focused SSL representations, retains strong content and semantic performance, and yields a perceptually effective integrated decoder. In addition, waveform view augmentations offer a controllable way to adapt the learned invariances to different applications. Future work should examine larger models, more diverse data, and task-specific augmentations.

AI USE STATEMENT

We used LLMs for language proofreading and editing for readability, and for assistance with implementing figures, including plotting and figure-formatting code. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work.

ETHICS STATEMENT

OLIVE is a speech representation learning method designed for shared pre-training across speech analysis and generation tasks. Potential beneficial applications include speech enhancement, low-resource speech modeling, and other settings that benefit from a representation shared across discriminative and generative speech tasks. At the same time, methods that improve speech reconstruction may be misused for deepfakes, spoofing, or speaker impersonation. This work involves no human-subject experiments and uses only publicly available datasets under their respective licenses. Mitigating these risks will require continued progress on watermarking and other provenance mechanisms for synthetic speech, as well as reliable methods for distinguishing generated audio from authentic recordings.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix A specify the model variants, training data, optimization setup, masking, augmentations, and synthesis configuration, including the parameter breakdown and the separate vocoder training setup. The downstream evaluation uses the public S3PRL toolkit with default task configurations (Section 4.2), and the waveform reconstruction protocol and metrics are described in Section 4.3. The ablations behind our configuration choices are reported in Appendix C, and compute details in Appendix J. All experiments use publicly available data. Anonymized code and audio samples will be provided for review, and code and model weights will be released upon publication.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023.

Bishnu S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2B):637–655, 1971.

Bagus Tris Atmaja and Akira Sasou. Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, 10:124396–124407, 2022.

Matthew Baas, Benjamin van Niekerk, and Herman Kamper. Voice conversion with just nearest neighbors. In *Interspeech*, 2023.

Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *ICLR*, 2020a.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020b.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *ICML*, 2023.

Séverin Baroudi, Thomas Pellegrini, and Hervé Bredin. Specializing self-supervised speech representations for speaker segmentation. In *Interspeech*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Hwan Lee, Hoon Heo, and Kyogu Lee. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. In *NeurIPS*, 2021.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Karl El Hajal, Ajinkya Kulkarni, Enno Hermann, and Mathew Magimai-Doss. kNN retrieval for simple and effective zero-shot multi-speaker text-to-speech. In *NAACL*, 2025.

Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak. BYOL-S: Learning self-supervised speech representations by bootstrapping. In *HEAR: Holistic Evaluation of Audio Representations*, volume 166 of *Proceedings of Machine Learning Research*, 2022.

Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton, The Hague, 1960.

Tzu-hsun Feng, Annie Dong, Ching-Feng Yeh, Shu-wen Yang, Tzu-Quan Lin, Jiatong Shi, Kai-Wei Chang, Zili Huang, Haibin Wu, Xuankai Chang, et al. SUPERB@SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning. In *IEEE SLT*, 2023.

Oded Ghitza. Speech analysis/synthesis based on matching the synthesized and the original representations in the auditory nerve level. In *ICASSP*, 1986.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Yiwei Guo, Zhihan Li, Junjie Li, Chenpeng Du, Hankun Wang, Shuai Wang, Xie Chen, and Kai Yu. vec2wav 2.0: Advancing voice conversion via discrete token vocoders. *arXiv preprint arXiv:2409.01995*, 2024.

Andrew Hines, Jan Skoglund, Anil C. Kokaram, and Naomi Harte. ViSQOL: An objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Georgios Ioannides, Christos Constantinou, Aman Chadha, Aaron Elkins, Linsey Pang, Ravid Shwartz-Ziv, and Yann LeCun. JEPA as a neural tokenizer: Learning robust speech representations with density adaptive attention. In *NeurIPS 2025 Workshop UniReps: Unifying Representations in Neural Models*, 2025.

Bryce Irvin, Marko Stamenovic, Mikolaj Kegler, and Li-Chia Yang. Self-supervised learning for speech enhancement through synthesis. In *ICASSP*, 2023.

Sunghyun Kim and Yong-Hoon Choi. WaveBYOL: Self-supervised learning for audio representation from raw waveforms. *IEEE Access*, 11:8968–8977, 2023.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. MelGAN: Generative adversarial networks for conditional waveform synthesis. In *NeurIPS*, 2019.

Shaoshi Ling and Yuzong Liu. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659*, 2020.

Alexander H. Liu, Yu-An Chung, and James Glass. Non-autoregressive predictive coding for learning speech representations from local dependencies. In *Interspeech*, 2021.

Alexander H. Liu, Sang-gil Lee, Chao-Han Huck Yang, Yuan Gong, Yu-Chiang Frank Wang, James R. Glass, Rafael Valle, and Bryan Catanzaro. UniWav: Towards unified pre-training for speech representation learning and generation. In *ICLR*, 2025.

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP*, 2020.

Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W. Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.

Robert J. McAulay and Thomas F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4):744–754, 1986.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for audio: Self-supervised learning for general-purpose audio representation. In *IJCNN*, 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *ICASSP*, 2015.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. In *Interspeech*, 2021.

Tilak Purohit and Mathew Magimai-Doss. Emotion information recovery potential of a wav2vec 2.0 network fine-tuned for speech recognition. In *ICASSP*, 2025.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, 2022.

Jürgen Schmidhuber and Daniel Prelinger. Discovering predictable classifications. *Neural Computation*, 5(4): 625–635, 1993.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, 2019.

Hubert Siuzdak, Piotr Dura, Pol van Rijn, and Nori Jacoby. WavThruVec: Latent speech representation as intermediate features for neural speech synthesis. In *Interspeech*, 2022.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Ingo R. Titze. *Principles of Voice Production*. Prentice Hall, Englewood Cliffs, NJ, 1994.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, et al. SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL*, 2022.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Heming Wang, Yao Qian, Hemin Yang, Nauyuki Kanda, Peidong Wang, Takuya Yoshioka, Xiaofei Wang, Yiming Wang, Shujie Liu, Zhuo Chen, et al. data2vec-SG: Improving self-supervised learning representations for speech generation tasks. In *ICASSP*, 2023.

Yuancheng Wang, Jiachen Zheng, Junan Zhang, Xueyao Zhang, Huan Liao, and Zhizheng Wu. Metis: A foundation speech generation model with masked generative pre-training. In *NeurIPS*, 2025.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB: Speech processing universal performance benchmark. In *Interspeech*, 2021.

Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms. *arXiv preprint arXiv:2509.23238*, 2025.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2021.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.

Zhaoyan Zhang. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4):2614–2635, 2016.

A Model and Pre-training Details

Table 4 consolidates the main optimization, encoder, and synthesis-branch configurations used for OLIVE pre-training.

Table 4: OLIVE model architecture and pre-training configuration.

Component	Setting
Optimization and Pre-training Setup	
Implementation	fairseq
Pre-training data	LibriSpeech 960h unlabeled training set
Sampling rate	16 kHz
Model size	Base; 12 Transformer layers, hidden size $d = 768$
Small ablations	Hidden size $d = 384$; 100k updates
Optimizer	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$
Weight decay	0.01
Peak learning rate	7.5×10^{-4}
Learning-rate schedule	cosine decay after 8k warmup updates
Training updates	400k
Batching	1M raw samples/GPU; gradient accumulation 8
Hardware	2 NVIDIA H100 GPUs
Average realized batch size	72.7 utterances/update
Gradient clipping	disabled
Precision	FP16 analysis branch; FP32 synthesis branch
Encoder Architecture	
Waveform encoder	$[(512, 10, 5)] + 4 \times [(512, 3, 2)] + 2 \times [(512, 2, 2)]$
Waveform stride	320 samples / 20 ms
Effective receptive field	400 samples / 25 ms
Projection	$512 \rightarrow 768$
Positional encoder	5-layer convolutional positional encoder (kernel size 19 per layer; effective width 95; 16 groups)
Transformer encoder	12 layers, 12 heads, hidden size $d = 768$, MLP ratio 4
Dropout	encoder 0.1; attention 0.1; post-MLP 0.1; activation 0.0; input 0.0
Layer drop	0.05
Teacher target layers	top $K = 8$ layers after instance normalization
Prediction head	4-layer 1D convolutional decoder; dim 384; groups 16; kernel size 7; dropout 0.1
Masking	inverse-block masking; mask ratio 0.5; block width 5; adjustment 0.05; noise std. 0.01
Masked views per sample	8
EMA schedule	$0.999 \rightarrow 0.99999$ over 75k updates
OLIVE-J Synthesis Branch	
Vocoder	HiFi-GAN V2 generator
Conditioning representation	local encoder output
Conditioning dimension	768
Internal bottleneck	128
Upsampling rates / kernels	$[10, 8, 2, 2]$ / $[20, 16, 4, 4]$
Initial upsampling width	128
Residual block kernels / dilations	$[3, 7, 11]$ / $(1, 3, 5)$
Discriminators	HiFi-GAN MPD + MSD
MPD periods	$\{2, 3, 5, 7, 11\}$
MSD branches	3 sub-discriminators; spectral normalization in first branch
Synthesis segment length	7040 samples
Mel analysis	80 bins; FFT 1024; window 1024; hop 320
Loss weights	$\lambda_{\text{synthesis}} = 25$, $\lambda_{\text{fm}} = 1$, $\lambda_{\text{mel}} = 45$

Parameter Breakdown

Table 5 reports the parameter counts for the shared encoder and synthesis components.

Separate Vocoder Training

Separate feature-conditioned HiFi-GAN V2 vocoders (Section 4.3) are trained for 400k updates with batch size 72, Adam ($\beta_1 = 0.8$, $\beta_2 = 0.99$), learning rate 7.5×10^{-4}, and exponential learning-rate decay of 0.999. Training uses 7040-sample waveform segments (0.44 s at 16 kHz), 80-bin log-mel spectrograms, a 1024-point FFT, 1024-sample window, and 320-sample hop size. Training data are augmented with mixup, following the augmentation ablation in Appendix C.4.

Table 5: Parameter breakdown for OLIVE-J. The HiFi-GAN discriminators are used only during pre-training and are discarded afterwards.

Component	Parameters
Shared local encoder	
Convolutional waveform encoder	4.2M
Feature projection ($512 \rightarrow 768$)	0.4M
Analysis branch	
Relative positional convolutional encoder	3.5M
Transformer encoder	85.1M
Prediction decoder	0.6M
Synthesis branch	
HiFi-GAN generator	1.1M
HiFi-GAN multi-period discriminator	41.1M
HiFi-GAN multi-scale discriminator	29.6M
Totals (training-time)	
Shared local encoder	4.6M
Analysis branch	89.2M
Synthesis branch	71.8M
Full joint model	165.6M

B Relative Comparison with data2vec-SG

The data2vec-SG results in Table 2 are taken from Wang et al. (2023), whose baselines differ from ours (e.g., FBANK PESQ 2.55 versus 2.83), so absolute scores are not directly comparable. We therefore provide a relative comparison: Table 6 reports each model's difference from data2vec Base evaluated in the same setup.

Table 6: Relative comparison on SUPERB generation tasks. Δ is the difference from data2vec Base within the same evaluation. Upper block: reported by Wang et al. (2023); lower block: our evaluation (Table 2).

	SE		SS
Model	PESQ↑	STOI↑	SI-SDRi↑
Reported by Wang et al. (2023)			
FBANK	2.55	93.6	9.2
data2vec Base	2.57	93.9	9.8
data2vec-SG Base	2.59	94.0	10.8
Δ vs. data2vec	+0.02	+0.1	+1.0
Our evaluation			
FBANK	2.83	94.3	9.2
data2vec Base	2.94	94.8	9.9
OLIVE-J (Mask-only)	2.94	94.8	10.5
Δ vs. data2vec	0.00	0.0	+0.6
OLIVE-J (Mix)	3.10	95.1	12.3
Δ vs. data2vec	+0.16	+0.3	+2.4

C Ablations

This section presents ablations used to determine pre-training configuration choices.

C.1 Joint Loss Weight

To choose the synthesis loss weight for OLIVE-J, we trained 100k-step joint models with $\lambda_{\text{synthesis}} \in \{1, 25, 50\}$ and evaluated them on a representative subset of downstream discriminative tasks and waveform reconstruction

quality on the LibriSpeech `dev-clean` set. A non-unit synthesis weight is needed in practice because the analysis and synthesis objectives operate on different numerical scales. The analysis branch implements the masked regression term from Section 3.1 as a framewise feature loss over the masked set $\mathcal{M}$:

$$\mathcal{L}_{\text{analysis}} = \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \sum_{k=1}^{d} \frac{(p_\theta(h'_t)_k - \text{sg}(z''_t)_k)^2}{\sqrt{d}}. \tag{5}$$

Although the factor $1/\sqrt{d}$ stabilizes the magnitude of the regression term, the loss still accumulates errors over the d latent dimensions before averaging over masked frames, so moderate per-dimension errors can produce loss values on the order of several units. By contrast, the HiFi-GAN terms in $\mathcal{L}^G_{\text{synthesis}}$ are scalar objectives combined into the same joint loss. Since the analysis loss accumulates over d dimensions and masked frames, the two objectives operate at different numerical scales, which motivates a non-unit synthesis weight. Tables 7 and 8 (visualized in Figure 4) show that $\lambda_{\text{synthesis}} = 25$ provides the best balance: increasing the weight from 1 to 25 substantially improves reconstruction, while increasing it further to 50 weakens several downstream metrics and yields less consistent reconstruction gains.

Table 7: Synthesis loss weight ablation on a representative subset of SUPERB downstream tasks. All models are 100k-step OLIVE-J (Mix) variants trained on LibriSpeech 960h. **Bold**: best; underlined: second-best.

$\lambda_{\text{synthesis}}$	ASV EER↓	ER Acc↑	PR PER↓	QbE maxTWV ↑	VC MCD↓	VC WER↓	VC F_0 Corr↑
1	**0.0536**	0.666	**0.0647**	0.0515	7.55	10.6	0.282
25	0.0581	**0.673**	0.0652	0.0535	**7.52**	**10.2**	0.284
50	0.0643	0.667	0.0667	**0.0551**	7.55	12.1	**0.289**

Table 8: Synthesis loss weight ablation on waveform reconstruction quality using the integrated OLIVE-J (Mix) vocoder, evaluated on the LibriSpeech dev-clean set. All models are 100k-step OLIVE-J (Mix) variants. **Bold**: best; underlined: second-best.

$\lambda_{\text{synthesis}}$	Mel-L1↓	MCD↓	STOI↑	PESQ↑	F_0 MAE↓	UTMOS↑
1	1.149	5.06	0.873	2.27	6.74	3.51
25	1.086	**4.72**	0.891	2.66	5.01	**3.88**
50	**1.007**	4.93	**0.894**	**2.79**	**4.67**	3.78

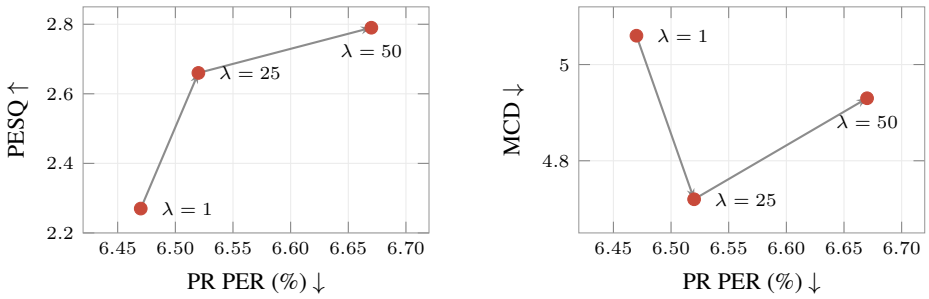


Figure 4: **Synthesis loss weight trade-off.** PR PER versus PESQ (left) and MCD (right) of the integrated vocoder for $\lambda_{\text{synthesis}} \in \{1, 25, 50\}$.

C.2 Variant Selection

To select which analysis-only variants to train, we first trained *Small*-sized models with hidden size $d = 384$ for 100k updates and evaluated their performance on a representative subset of content, speaker, paralinguistic, generation, and reconstruction tasks. Table 9 reports the resulting ablation used to select which variants to train at full *Base* scale in the main paper. The final column reports a normalized aggregate score across all metrics, under which Mix ranks first and Mix+Gain second among the OLIVE analysis-only variants.

Table 9: Small-model variant ablation used to select which analysis-only OLIVE augmentations to train at full scale. All models use hidden size $d = 384$ and are trained for 100k updates. Avg is the mean of all metrics after normalizing each to $[0, 100]$ across OLIVE variants, with lower-is-better metrics inverted. **Bold**: best; underlined: second-best.

Model	PR	ASV	SID	ER	VC			SE			Avg
	PER↓	EER↓	Acc↑	Acc↑	WER↓	MCD↓	F_0 Corr↑	PESQ↑	SI-SDR↑	STOI↑	↑
OLIVE-A (Gain) [small 100k]	0.0898	0.0726	0.4840	0.6270	0.1280	**7.3100**	0.2780	2.8900	**9.8100**	0.9460	25.77
OLIVE-A (Mix) [small 100k]	0.0863	**0.0661**	**0.4950**	**0.6520**	0.1210	7.5400	0.2780	**3.0100**	9.2600	**0.9490**	**57.75**
OLIVE-A (Mix+Gain) [small 100k]	**0.0812**	0.0678	0.4730	0.6380	**0.1090**	7.5000	**0.2820**	2.8800	9.3600	0.9480	52.01

C.3 HiFi-GAN Layer Ablation

This section reports how waveform reconstruction changes with the choice of conditioning layer. Table 10 compares separate HiFi-GAN V2 vocoders trained on frozen features from **OLIVE-A (Mix)**, **OLIVE-A (Mix+Gain)**, and **WavLM Base** at different encoder depths. In all cases, the best reconstruction quality is obtained from local encoder features, with quality degrading as conditioning moves deeper into the contextual encoder. Among the OLIVE analysis-only variants at the local encoder level, **OLIVE-A (Mix+Gain)** is slightly stronger on perceptual metrics and pitch accuracy. It is therefore selected for the final waveform reconstruction comparison.

Table 10: Layer-wise HiFi-GAN conditioning ablation. Separate HiFi-GAN V2 vocoders were trained for 100k updates on frozen features from OLIVE-A (Mix), OLIVE-A (Mix+Gain), and WavLM Base. "Local encoder features" denotes the output of the shared local encoder before the Transformer, and "Transformer k" denotes the output of the kth Transformer layer. **Bold**: best per model.

Model	Conditioning	Mel-L1↓	MCD↓	STOI↑	PESQ↑	F_0 MAE↓	UTMOS↑
OLIVE-A (Mix)	Local encoder features	**0.5556**	**5.1394**	**0.9044**	**2.2269**	**13.2972**	**3.4484**
OLIVE-A (Mix)	Transformer 1	0.5951	5.5717	0.8759	1.8747	14.4624	3.1674
OLIVE-A (Mix)	Transformer 2	0.6323	5.9658	0.8583	1.5139	14.9328	3.0849
OLIVE-A (Mix)	Transformer 3	0.6461	5.9678	0.8415	1.4350	16.6105	3.0969
OLIVE-A (Mix)	Transformer 4	0.6732	6.0818	0.8262	1.3332	19.4579	3.0786
OLIVE-A (Mix)	Transformer 5	0.7036	6.3123	0.8070	1.2558	21.6921	2.9327
OLIVE-A (Mix+Gain)	Local encoder features	**0.5611**	**5.1712**	**0.9070**	**2.2426**	**12.6522**	**3.5567**
OLIVE-A (Mix+Gain)	Transformer 1	0.5947	5.5902	0.8802	1.8566	13.8737	3.3040
OLIVE-A (Mix+Gain)	Transformer 2	0.6357	5.8769	0.8549	1.4910	16.3225	3.0375
OLIVE-A (Mix+Gain)	Transformer 3	0.6316	5.8134	0.8618	1.5785	15.2443	3.2385
OLIVE-A (Mix+Gain)	Transformer 4	0.6638	6.1259	0.8465	1.4499	16.6433	3.1341
OLIVE-A (Mix+Gain)	Transformer 5	0.7110	6.4318	0.8121	1.3221	19.7228	2.9569
WavLM Base	Local encoder features	**0.5422**	**5.3733**	**0.8928**	**1.9573**	**13.9462**	**3.3709**
WavLM Base	Transformer 1	0.5596	5.5364	0.8811	1.7441	14.9699	3.3503
WavLM Base	Transformer 2	0.6080	5.8671	0.8411	1.4173	17.3569	2.8881
WavLM Base	Transformer 3	0.6127	5.8017	0.8386	1.3218	20.3122	2.8329
WavLM Base	Transformer 4	0.6498	6.1132	0.8133	1.2944	24.1076	2.8239
WavLM Base	Transformer 5	0.6876	6.4009	0.7867	1.1658	26.2397	2.9338

C.4 HiFi-GAN Augmentation Ablation

Table 11 compares matched separate HiFi-GAN V2 vocoders trained on the same frozen **OLIVE-A (Mix+Gain)** local encoder features, with and without waveform mixup augmentation applied during training. Mixup improves all reported reconstruction metrics, so we use it for all separate vocoder training in the main evaluation.

Table 11: HiFi-GAN augmentation ablation for separate vocoder training. We compare matched HiFi-GAN V2 vocoders trained on the same frozen **OLIVE-A (Mix+Gain)** local encoder features, with and without waveform mixup augmentation. **Bold**: better.

Training data	Mel-L1↓	MCD↓	STOI↑	PESQ↑	F_0 MAE↓	UTMOS↑
No augmentation	0.8251	6.4177	0.7541	1.2384	16.9843	1.9400
Waveform mixup	**0.5297**	**4.9983**	**0.9045**	**2.2566**	**13.0708**	**3.3836**

D SUPERB LAYERWISE PROFILES

Figure 5 shows representative learned layer combination weights for four SUPERB tasks. ASR and slot filling place most of their weight on later contextual layers, whereas speaker verification and speech enhancement rely more heavily on earlier layers.

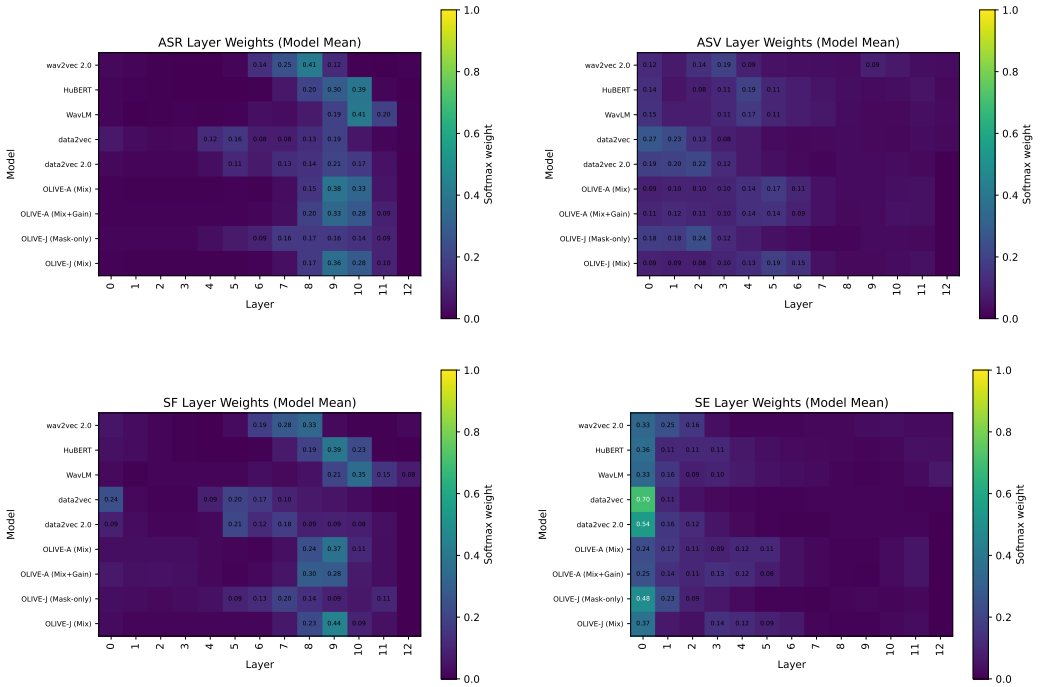


Figure 5: **Representative SUPERB layer-combination weights.** Automatic speech recognition (ASR) and slot filling (SF) emphasize later contextual layers, while automatic speaker verification (ASV) and speech enhancement (SE) place more weight on earlier layers.

E SUPERB RADAR PROFILES

Figure 6 shows normalized per-task SUPERB profiles for the OLIVE variants and the SSL baselines, together with the category-level comparison.

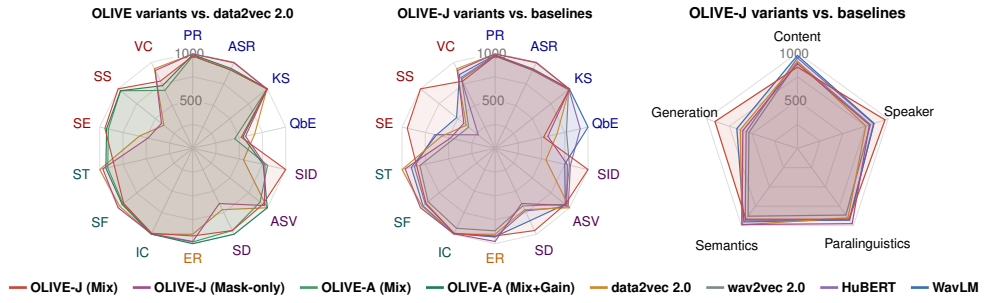


Figure 6: **SUPERB radar comparisons.** Left and center: normalized per-task profiles. Right: aggregate category-level comparison. Per-task metrics are normalized so that FBANK is 0 and the best value in Tables 1 and 2 is 1000; multi-metric tasks are averaged within task, and category scores average the tasks within each category.

F ASR FINE-TUNING EVALUATION

Table 12: ASR fine-tuning results on LibriSpeech using `train-clean-100` as the labeled fine-tuning set. Word error rates (%) are reported with and without a 4-gram language model; lower is better. **Bold**: best; underlined: second-best.

	No LM				4-gram LM			
Model	dev-clean	dev-other	test-clean	test-other	dev-clean	dev-other	test-clean	test-other
FBANK	22.7	52.1	23.1	53.9	14.4	41.9	15.0	43.7
wav2vec 2.0 Base	6.0	16.3	6.5	16.0	4.1	12.6	4.9	12.6
HuBERT Base	5.9	15.5	6.4	15.6	4.1	12.1	4.7	12.0
WavLM Base	5.6	15.0	6.2	15.0	3.8	11.6	4.6	11.6
data2vec Base	4.7	12.0	5.0	12.3	3.3	9.5	3.9	9.7
data2vec 2.0 Base	**4.5**	**11.0**	**4.8**	**11.1**	**3.2**	**8.6**	**3.8**	**8.8**
OLIVE-A (Mix)	5.9	14.7	6.4	14.4	4.2	11.4	4.7	11.2
OLIVE-A (Mix+Gain)	5.6	14.0	6.1	14.0	3.8	11.0	4.6	11.0
OLIVE-J (Mask-only)	**4.5**	11.4	4.9	11.4	3.3	9.0	3.9	9.0
OLIVE-J (Mix)	6.2	15.6	6.4	15.7	4.3	12.1	5.0	12.2

We complement the SUPERB ASR evaluation by reporting word error rate on the standard LibriSpeech development and test splits (`dev-clean`, `dev-other`, `test-clean`, and `test-other`), using `train-clean-100` for fine-tuning under the same default task settings, with and without a 4-gram language model; the results are shown in Table 12. The strongest recognition numbers are obtained by data2vec 2.0, with **OLIVE-J (Mask-only)** close behind on every split. This suggests that the reconstruction objective, which acts mainly on earlier encoder features, can enhance the information retained for reconstruction without degrading the discriminative quality of later contextual representations, while maintaining strong downstream performance. The OLIVE models trained with augmented waveform views show a recognition trade-off compared to data2vec 2.0 and OLIVE-J (Mask-only), but remain competitive with other Base-sized SSL systems. **OLIVE-A (Mix+Gain)**, for example, stays close to WavLM Base on most ASR splits while performing better on several speaker and generation metrics. It achieves the highest overall SUPERB_s score, while **OLIVE-J (Mix)** achieves the highest overall WavLM_s score.

G INTELLIGIBILITY AND SPEAKER PRESERVATION OF RESYNTHESIZED SPEECH

Table 13 complements the signal-level metrics of Table 3 by measuring whether resynthesized speech preserves linguistic content and speaker identity. All SSL-conditioned vocoders yield similar intelligibility, with the frozen feature-conditioned OLIVE-J vocoders obtaining the lowest WER, while speaker similarity separates the models more clearly: every OLIVE variant exceeds all SSL baselines in SECS, and OLIVE-J (Mix) obtains the highest SECS with both the frozen feature-conditioned and the integrated vocoder. Waveform view augmentations further increase SECS over OLIVE-J (Mask-only), consistent with the speaker-task gains observed on SUPERB.

H WAVEFORM RECONSTRUCTION CONFIDENCE INTERVALS

Tables 14 and 15 report the waveform reconstruction results of Table 3 together with the margins of error from 95% bootstrap confidence intervals.

I WAVEFORM RECONSTRUCTION SPECTROGRAMS

Figures 7 and 8 show spectrograms that complement Table 3. Figure 7 illustrates three reference utterances reconstructed by the frozen feature-conditioned and integrated OLIVE-J (Mix) vocoders. Figure 8 illustrates one reference utterance reconstructed by all separately trained, matched HiFi-GAN V2 vocoders conditioned on local encoder features from the baseline and OLIVE models.

Table 13: Intelligibility and speaker preservation of resynthesized speech on all 2,620 LibriSpeech test-clean utterances, using the same vocoders as Table 3. WER is computed with Whisper large-v3; SECS is the mean cosine similarity between ECAPA-TDNN speaker embeddings of the resynthesized and reference utterances. **Bold**: best; underlined: second-best.

Model	WER (%) ↓	SECS ↑
Mel spectrogram	4.23	0.565
wav2vec 2.0 Base	3.10	0.645
HuBERT Base	2.96	0.668
WavLM Base	3.06	0.670
data2vec Base	3.09	0.664
data2vec 2.0 Base	3.01	0.676
Frozen feature-conditioned vocoder		
OLIVE-A (Mix+Gain)	3.05	0.696
OLIVE-J (Mask-only)	2.95	0.709
OLIVE-J (Mix)	**2.92**	**0.736**
Integrated vocoder		
OLIVE-J (Mask-only)	3.07	0.693
OLIVE-J (Mix)	3.04	0.722

Table 14: Waveform reconstruction on LibriSpeech test-clean with matched and integrated HiFi-GAN V2 vocoders: spectral and pitch metrics. Values are mean ± margin of error from 95% bootstrap confidence intervals. **Bold**: best; underlined: second-best.

Model	Mel-L1 ↓	MCD ↓	LSD ↓	F_0 MAE (Hz) ↓
Mel spectrogram	0.677 ± 0.004	6.24 ± 0.03	0.829 ± 0.005	14.4 ± 0.3
wav2vec 2.0 Base	0.474 ± 0.003	5.18 ± 0.03	0.604 ± 0.004	11.6 ± 0.3
HuBERT Base	0.475 ± 0.002	5.05 ± 0.02	0.604 ± 0.003	13.3 ± 0.3
WavLM Base	0.470 ± 0.002	5.01 ± 0.02	0.599 ± 0.003	13.1 ± 0.3
data2vec Base	0.467 ± 0.002	4.90 ± 0.02	0.594 ± 0.003	12.8 ± 0.3
data2vec 2.0 Base	0.434 ± 0.002	4.78 ± 0.02	0.558 ± 0.003	13.0 ± 0.3
Frozen feature-conditioned vocoder				
OLIVE-A (Mix+Gain)	0.446 ± 0.002	4.69 ± 0.02	0.566 ± 0.003	11.6 ± 0.2
OLIVE-J (Mask-only)	**0.410 ± 0.002**	4.43 ± 0.01	**0.525 ± 0.002**	10.8 ± 0.2
OLIVE-J (Mix)	0.421 ± 0.002	**4.35 ± 0.02**	0.528 ± 0.002	**9.6 ± 0.2**
Integrated vocoder				
OLIVE-J (Mask-only)	0.476 ± 0.002	4.68 ± 0.01	0.593 ± 0.002	11.1 ± 0.3
OLIVE-J (Mix)	0.579 ± 0.003	4.56 ± 0.02	0.688 ± 0.003	10.0 ± 0.2

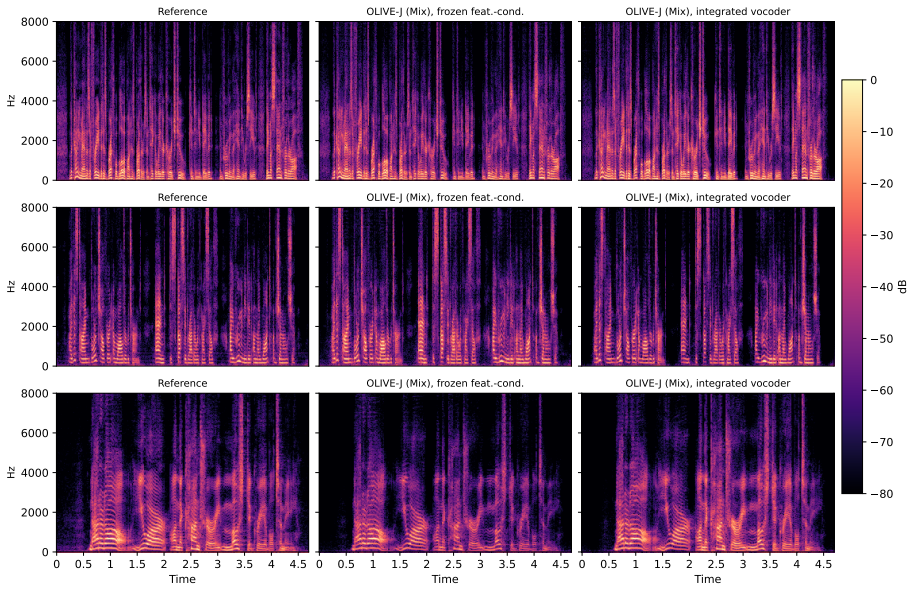


Figure 7: Spectrograms for three reference utterances and reconstructions from the frozen feature-conditioned and integrated OLIVE-J (Mix) vocoders.

Table 15: Waveform reconstruction on LibriSpeech test-clean with matched and integrated HiFi-GAN V2 vocoders: intelligibility, perceptual, and signal-level metrics. Values are mean ± margin of error from 95% bootstrap confidence intervals. **Bold**: best; underlined: second-best.

Model	STOI ↑	PESQ ↑	ViSQOL ↑	SI-SDR ↑	SNR ↑	UTMOS ↑
Reference audio						4.09 ± 0.01
Mel spectrogram	0.833 ± 0.001	1.42 ± 0.01	3.59 ± 0.01	-18.02 ± 0.15	-2.60 ± 0.02	2.98 ± 0.02
wav2vec 2.0 Base	0.912 ± 0.001	2.19 ± 0.01	4.04 ± 0.01	-7.53 ± 0.21	-0.42 ± 0.06	3.58 ± 0.02
HuBERT Base	0.914 ± 0.001	2.11 ± 0.01	4.05 ± 0.01	-11.78 ± 0.17	-1.57 ± 0.04	3.63 ± 0.02
WavLM Base	0.917 ± 0.001	2.21 ± 0.01	4.05 ± 0.01	-10.93 ± 0.20	-1.31 ± 0.04	3.68 ± 0.02
data2vec Base	0.902 ± 0.001	2.10 ± 0.01	4.06 ± 0.01	-10.64 ± 0.21	-1.15 ± 0.04	3.50 ± 0.02
data2vec 2.0 Base	0.917 ± 0.001	2.21 ± 0.01	4.10 ± 0.01	-9.67 ± 0.18	-1.11 ± 0.04	3.63 ± 0.02
Frozen feature-conditioned vocoder						
OLIVE-A (Mix+Gain)	0.929 ± 0.001	2.62 ± 0.01	4.18 ± 0.01	-3.36 ± 0.18	1.06 ± 0.08	3.70 ± 0.02
OLIVE-J (Mask-only)	0.935 ± 0.001	2.81 ± 0.01	4.26 ± 0.01	0.97 ± 0.15	3.31 ± 0.08	**3.86 ± 0.01**
OLIVE-J (Mix)	**0.942 ± 0.001**	**3.06 ± 0.01**	**4.34 ± 0.01**	**2.94 ± 0.16**	**4.69 ± 0.10**	3.83 ± 0.02
Integrated vocoder						
OLIVE-J (Mask-only)	0.911 ± 0.001	2.56 ± 0.01	4.14 ± 0.01	0.72 ± 0.14	3.08 ± 0.08	3.74 ± 0.02
OLIVE-J (Mix)	0.921 ± 0.001	2.88 ± 0.02	4.21 ± 0.01	2.50 ± 0.14	4.25 ± 0.09	3.78 ± 0.02

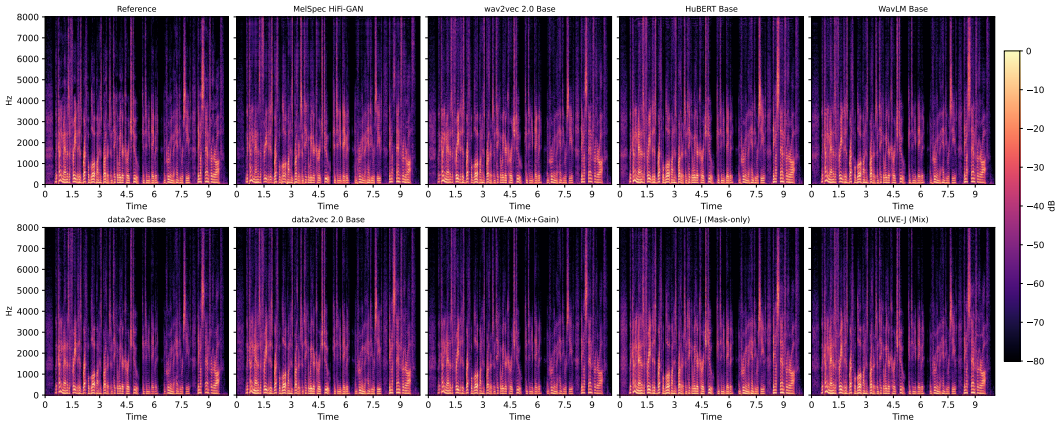


Figure 8: Spectrograms for one reference utterance and reconstructions from matched HiFi-GAN V2 vocoders conditioned on local features from the baseline and OLIVE models.

J Compute Details

All main OLIVE pre-training runs used 2 NVIDIA H100 80GB GPUs. The final analysis-only models required 123 training hours per run, and the final OLIVE-J models required 206 training hours per run. Vocoder training and downstream task evaluation used Multi-Instance GPU (MIG) 20GB and 40GB partitions of H100 GPUs.