A Large-Scale Evaluation of Speech Foundation Models

Shu-wen Yang, Heng-Jui Chang*, Zili Huang*, Andy T. Liu*, Cheng-I Lai*, Haibin Wu*, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Abdelrahman Mohamed, Shang-Wen Li, Shinji Watanabe, Hung-yi Lee

***Abstract*—The foundation model paradigm leverages a shared foundation model to achieve state-of-the-art (SOTA) performance for various tasks, requiring minimal downstream-specific data collection and modeling. This approach has been proven crucial in the field of Natural Language Processing (NLP). However, the speech processing community lacks a similar setup to explore the paradigm systematically. To bridge this gap, we introduce Speech processing Universal PERformance Benchmark (SUPERB). SUPERB represents an ecosystem designed to evaluate foundation models across a wide range of speech processing tasks. It facilitates the sharing of results on an online leaderboard, fostering collaboration from a community-driven benchmark database which helps new development cycles. We present a unified framework for solving all the speech processing tasks in SUPERB with the frozen foundation model followed by task-specialized lightweight prediction heads. Combining our results with community submissions, we verify that the framework is simple yet effective as the SOTA foundation model shows competitive generalizability across most of the SUPERB tasks. Lastly, we conduct a series of analyses to offer an in-depth understanding of SUPERB and speech foundation models, from the information flows across tasks inside the models to the statistical significance and robustness of the benchmark. Our benchmark results suggest foundation models should be powerful, robust, and disentangled, in order to be universally applicable for most kinds of speech processing.**

***Index Terms*—speech, foundation model, self-supervised learning, representation learning, task generalization, benchmark, evaluation, SUPERB, S3PRL**

I. Introduction

DEVELOPING well-performing deep learning networks has become costly, involving data collection, modeling, computing power, and training time. This repetition for each specific use case is time and cost-prohibitive for both academic and industrial researchers. To address this, the foundation model paradigm [1] proposes a framework that transfers knowledge from a centralized foundation model for downstream use cases. Scaling up the foundation model with more data and computing resources improves performance on numerous downstream tasks simultaneously. Self-Supervised Learning (SSL) has emerged as a promising technique for developing foundation models [2], [3]. This technique involves pre-training a model with a substantial amount of parameters and unlabeled data to learn powerful, general, and transferable representations. Since after pre-training once, the model achieves state-of-the-art (SOTA) downstream performances after fine-tuning on various tasks, SSL appears as a remedy for democratizing SOTA deep learning research and deployment for people from different backgrounds. The success of this approach has been witnessed in both Natural Language Processing (NLP) [2] and Computer Vision (CV) [3].

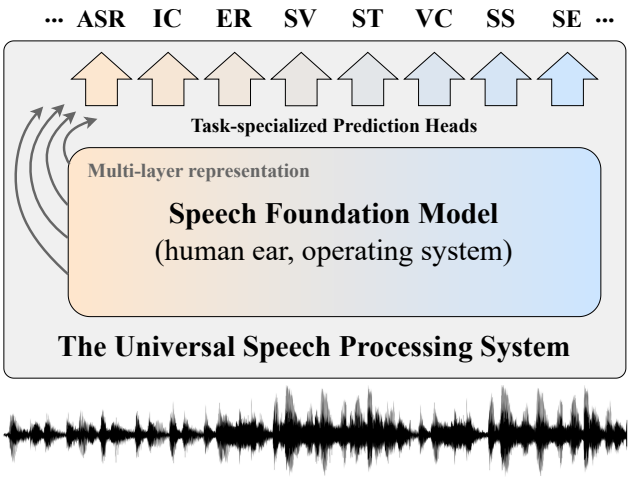


Fig. 1: The figure presents the concept of a speech foundation model. As the core component of a universal speech processing system, the model handles waveforms from different tasks, processes them into high-level representations, and feeds the representations into different prediction heads to produce predictions for various tasks.

SSL has been explored in speech [4]–[14], with studies applying SSL models to various applications [8], [11]. However, these studies used different datasets and setups, making it challenging to gain insights from *meaningful comparisons*. Furthermore, unlike NLP, where foundation models are assessed across multiple tasks and benchmarks like GLUE [15], speech SSL evaluation often narrows down to specific tasks. Despite this approach pushing task-specific performance, it overlooks SSL's potential of developing universal foundation models which handle diverse data conditions and generalize to new tasks without costly downstream-specific modeling, pre-training and data collection. LEAF [16] and HEAR [17] advocated the development of universal models that emulate human perception to replace FBANK, but the models were primarily evaluated on audio classification tasks. Little effort has been devoted to benchmarking the *universality* of speech

*Equal contribution; sorted alphabetically

SSL models.

To address the limitations mentioned above, we introduce Speech processing Universal PERformance Benchmark (SUPERB), highlighting the following three features:

1) **Comprehensive coverage**: SUPERB brings standardization to comparing SSL models across a diverse range of 15 speech processing tasks.
2) **Task generalizability**: SUPERB encourages the development of speech foundation models to solve numerous tasks simultaneously compared to single-task experts.
3) **Community standard**: SUPERB adopts popular tasks from speech communities, and follows the conventional evaluation protocol to align with common research interests.

Existing works have proposed several benchmarks for speech SSL [18], [19]. Compared to these efforts, SUPERB explicitly emphasizes the *task generalizability* and remains to be the benchmark with the most comprehensive task coverage across most kinds of speech processing tasks, from discriminative to generative tasks, which we believe is essential to foster the progress on universal speech foundation models.

The 15 SUPERB tasks include the popular Automatic Speech Recognition (ASR), Speaker Verification (SV), Emotion Recognition (ER), Voice Conversion (VC), Speech Enhancement (SE), etc., as shown by Fig [1]. We study the *frozen foundation models followed by task-specialized prediction heads* for solving SUPERB tasks in this work with the following considerations:

1) **Inclusive**: The computing barrier is significantly reduced, making it more affordable for researchers from diverse backgrounds.
2) **Scalable**: The final system is significantly smaller as we only need to save the lightweight prediction heads, leading to a substantial reduction of the parameter size when scaling up the task number.
3) **Versatile**: By mounting all the prediction heads to the foundation model, we automatically acquire a multi-task model capable of dealing with all kinds of speech tasks.

We defined the standardized task design, provided the baseline model results, and released the offline evaluation software in [20], [21]. In this work, we extend our previous work with the following contributions with some takeaways for future researchers to develop SSL:

- Combined with the released software, we further provide a complete platform consisting of an online leaderboard for public submissions and comparative analysis supports including statistical and visualization tools (Section IV). After launching the submission system, we *received 14 submissions*, suggesting that our platform is becoming an active community.
- We scale the evaluation from the original *14 models to 36 models*, together with the efforts from our community, to provide a comprehensive coverage of the existing speech SSL literature and track the latest research (Section V-B).
- We point out common pitfalls of leveraging or evaluating speech foundation models and provide corresponding suggestions accordingly, including (1) extracting the last-layer feature without a throughout examination of the model layers (Section V-A), (2) interpreting the information flow within model layers with the misleading trained weights from the weighted-sum technique (Section VI-A), and (3) ranking models solely basing on the score of each task regardless of the statistical significance (Section VII).
- We analyze the updated leaderboard results along with the layer-wise benchmarking results and obtain several findings, including: (1) SOTA foundation models possess various high-quality information but often struggle to handle out-of-domain ASR datasets, noisy waveforms, or feature disentanglement (Section V-B3). (2) Incorporating vector quantization into network architectures is beneficial for content-centric tasks but diminishes the model's universality on SUPERB tasks (Section V-B4). (3) The foundation models consistently learn speaker characteristics at prior layers, followed by emotional information, and then linguistic signals. (Section VI-B1) (4) In the VC task, we suggest conducting the single-layer benchmarking for the consistently better performance (Section VI-B2). (5) WavLM Large learns good disentanglement concurrently with better universal features compared to other models, while Data2vec Large achieves SOTA disentanglement capability and content accessibility by sacrificing speaker characteristics severely (Section VI-B3 and Section VI-B4). (6) The relative performance of SUPERB is robust against various changes in data conditions VIII.

II. Related Work

Several works have proposed different methods for evaluating speech SSL models. The ZeroSpeech series [18] focuses on the intrinsic evaluation for different levels of content information from phonetics, lexicon, to semantics, with the linguistically-motivated ABX and ABX-based metrics. The SLUE series focuses on evaluating SSL models' capability for solving spoken language understanding (SLU) tasks like named entity recognition, sentiment analysis, and spoken question answering [22], [23]. [24] proposed a benchmark for evaluating the paralinguistic information inside the speech SSL models, including masked speech detection and dysarthria classification.

Besides these benchmarks focusing on a specific aspect of speech, some works proposed to standardize the evaluation of different aspects of speech in a single benchmark. [25] proposed to benchmark SSL models with SV, ER, and SLU tasks in English through fine-tuning the entire SSL model. LeBenchmark [19] established a multi-task SSL benchmark for French with ASR, SLU, ER, and speech translation (ST). FLEURS [26] and XTREME-S [27] extend the multi-task evaluation frameworks to the multi-lingual setting.

Compared to these efforts, the SUPERB series cover broader aspects of speech processing, including content (ASR), semantics (ST), speaker (SV), paralinguistics (ER), denoising (SE), and disentanglement (VC). The original SUPERB [20] addresses 10 discriminative tasks, with the follow-up SUPERB-SG [21] introducing 5 additional tasks for semantic and

generative capabilities. These 15 tasks define the *public set* of SUPERB. The SUPERB Challenge [28] introduces the *hidden set* for partial tasks to prevent overfitting SSL development on the public set, where the corpora for the hidden set are privately recorded and the participants submit the models to the hidden set committee for evaluation. ML-SUPERB [29] extends the framework from English to the multilingual setting covering 143 languages with the consideration of ASR and language identification (LID) as the first step.

Despite the increasing interest and adoption of the 15-task public set of SUPERB [14], [30], [31], the original work [20], [21] presents the standardized task design and the evaluation results on limited models without detailed analyses and suggestions for the benchmark adoption. In this work, we extend [20], [21] with an online interactive platform, up-to-date foundation model evaluation, and in-depth analyses for the benchmark itself along with the current status of speech foundation model development.

III. Speech Processing Universal PERformance Benchmark

We establish and release Speech processing Universal PERformance Benchmark (SUPERB), a benchmark dedicated to evaluate the task generalizability of speech foundation models, encompassing most aspects of speech processing. We collect 15 well-known tasks: Phoneme Recognition (**PR**), Keyword Spotting (**KS**), Speaker Identification (**SID**), Emotion Recognition (**ER**), Intent Classification (**IC**), Automatic Speech Recognition (**ASR**), Speaker Verification (**SV**), Query-by-example (**QbE**), Slot Filling (**SF**), Speaker Diarization (**SD**), Out-of-domain ASR (**OOD-ASR**), Speech Translation (**ST**), Voice Conversion (**VC**), Source Separation (**SS**), and Speech Enhancement (**SE**). The tasks are selected to investigate several aspects of speech, ranging from content (ASR), speaker (SV), semantics (ST), paralinguistics (ER), generation (SS), denoising (SE), to disentanglement (VC). We described each *task design* including the task target, assessed properties, adopted corpus, corpus statistics, and evaluation metrics in [20], [21], and organize them together in the supplementary material. The material also describes the *downstream adaptation* for the foundation models to solve these tasks, including the prediction head we use, and the optimization loss. Note that in principle SUPERB only defines the *task design* and we welcome researchers to explore various effective ways for *downstream adaptation*.

IV. Platform Design

As shown by Fig 2a, we design our platform to let the users easily reproduce our results, evaluate customized foundation models, extend the benchmark database themselves, analyze the model characteristics by comparing to others for the model limitation, and foster future development. To achieve this, our platform consists of the following three components: a software, a website, and numerous helpful artifacts.

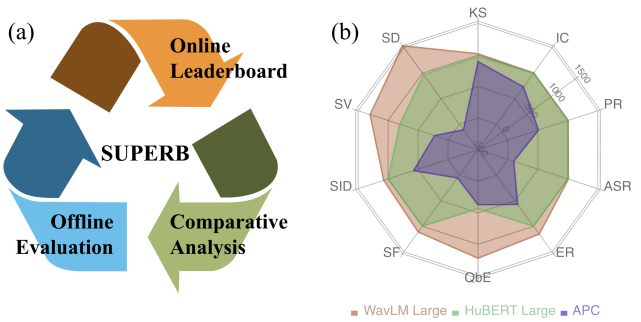


Fig. 2: (a) demonstrates the SUPERB platform's development cycle for speech foundation models. Starting from the offline evaluation, the user develops a new speech foundation model and evaluates it with S3PRL across a wide range of standardized SUPERB tasks. The downstream prediction files are auto-generated at this stage. Then, the user submits the prediction files to the online leaderboard to extend the benchmark database and compare it with others. Finally, the user analyzes the performance difference across SUPERB tasks with the website's visualization tools and S3PRL's statistical tools to gain insight into the future improvement direction. (b) gives an example of the website's radar visualization tool for comparing WavLM Large, HuBERT Large, and APC on several tasks. All scores are normalized to be higher for better performances.

A. Software

We implement the SUPERB evaluation tasks in the S3PRL toolkit[1] [5], [6], [20], as it supports numerous speech foundation models as the easily reusable modules. The foundation models in S3PRL are termed as *upstream* models and are designed to be independent of the downstream task implementation, so that one can easily switch between any upstream and downstream combination or add one's own foundation model as a new upstream and evaluate it with all the tasks in S3PRL. This upstream addition implementation can be done once and all the tasks are supported automatically. The software is tightly integrated with the SUPERB official website so that once the evaluation of each task is done, the corresponding submission files are automatically generated, and the users can upload the submission files to our website by themselves and easily help extend the benchmark database. Compared with the initial release of SUPERB [20], [21], we upgraded the codebase to support Kaldi-style [32] data preparation for most of the tasks. This makes the software much more versatile for different experimental settings including switching corpus and conducting few-shot learning. We also upgraded to support a unified interface for adding distortion on recordings for all the tasks by simply writing configurations, including reverberation and additive noises. The distortions are applied on-the-fly and deterministic (reproducible) to save disk space. Finally, we add the statistical tools for analyzing the difference significance between models for all the tasks.

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

TABLE I: Weighted-sum benchmark results of wav2vec 2.0 and HuBERT on SID and IC using different fine-tuning learning rates. The learning rates with $*$ denote the default learning rate in S3PRL. Bold fonts mark the best learning rates for different models. We searched from 1e-1 to 1e-7 in log-scale and show the partial results due to the space limit.

Models	SID (acc)				IC (acc)	
	1e-1	1e-2	1e-3	1e-4*	1e-3	1e-4*
wav2vec 2.0 Base	NaN	74.28	**75.18**	66.72	92.12	**92.35**
wav2vec 2.0 Large	NaN	84.38	**86.15**	82.71	**95.28**	93.22
HuBERT Base	**81.42**	81.01	70.09	67.37	**98.34**	97.81
HuBERT Large	NaN	86.94	**90.33**	86.94	98.63	**98.76**

B. Website

Our website[2] serves an important functionality of the SUPERB benchmark: an actively extending benchmark database, so that SUPERB is not only a static leaderboard with our own results. Upon the time of submission, our online leaderboard receives 14 submissions, suggesting that our platform is becoming an active community. To reduce the participation barrier, the website also accepts submissions with partial results when evaluating all the tasks is too expensive. The website further provides helpful visualization tools to compare different models' detailed characteristics by scatter and radar charts, as shown by Fig 2b.

C. Artifacts

Modern deep learning systems are complicated and hard to reproduce even when the code is released, since a slight misconfiguration of the hyper-parameters or a wrong package version might lead to worse performances. Hence, we release our fine-tuning logs (Tensorboard files), hyper-parameters and well-trained downstream prediction heads for several well-known foundation models[3], so that the user can debug their training procedure with our official ones. Furthermore, we release the downstream prediction files for several SOTA models, so the user can analyze the improvement significance with the statistical tools in S3PRL.

V. Main Result

We present the benchmark results of 36 speech foundation models [4]–[7], [9]–[14], [30], [31], [33]–[44] on 15 speech processing tasks. This is the most comprehensive standardized evaluation database for speech SSL to our best knowledge. In this section, we first compare two downstream adaptation protocols we tried for foundation models. Then, we present the full benchmark results.

Note that it is important to explicitly search for the suitable learning rates for different foundation models instead of directly using the default one in S3PRL, since different models favor different learning rates as shown by Table I. We search from 1e-1 to 1e-7 in log-scale in the following experiments.

[2]https://superbbenchmark.org/

[3]github.com/s3prl/s3prl/blob/main/s3prl/downstream/docs/superb_artifacts.md

A. Downstream Adaptation Protocols

Before diving into full benchmarking, we first examine the best way to utilize a frozen foundation model. We compare (1) the last layer representation and (2) the weighted sum over all layers of representation. In this work, we do not consider fine-tuning the whole foundation model due to the significant computational cost and parameter inefficiency associated with handling a large number of tasks. Table II shows that in most cases weighted-sum is better than the last-layer representation, either equally good or significantly better. Conversely, most of the highlighted failing cases have only slight differences.

It is worth noting that, wav2vec 2.0 shows serious differences when switching from the last-layer to the weighted-sum protocol for all the tasks: 80%, 67% and 58% relative improvement on PR, SID and IC respectively. [45] reported this behavior on the content-centric tasks, and our results demonstrate that the poor performances are consistent across most of the SUPERB tasks including the speaker and emotion-related tasks. This phenomenon suggests that the last layer of wav2vec 2.0 possesses less useful information and one should always try the weighted-sum protocol when evaluating wav2vec 2.0. We believe this can be a pitfall worth explicitly pointing out since there are works [17], [46] on benchmarking or leveraging frozen wav2vec 2.0 following the conventional last-layer protocol.

The weighted-sum protocol also shows improvement on HuBERT, especially on the SID task with a 25% relative improvement. Overall, we observe that the weighted-sum protocol is a much more reliable protocol, yielding stable and competitive results for most of the foundation models. Intuitively, the trainable weights automatically determine the informative layers for each task and ensemble them usually results in a better performance. Hence, we set this protocol as default for SUPERB benchmark.

B. Full Benchmark Result

We present the main results following the weighted-sum protocol in Fig 3, and discuss several important aspects in this section. We also mark the best and the worst model in Table III.

1) Hand-crafted acoustic feature vs. SSL on the task generalizability: Firstly, we compare all the SSL models with a classic acoustic feature, FBANK, to assess their task generalizability. We choose FBANK as the baseline because it is widely used in various speech processing tasks and has usually been adopted as the baseline for evaluating general-purpose frontends, e.g. LEAF [16].

Table III shows that FBANK is not the worst feature in many cases. By further confirming with Fig 3, FBANK performs competitive on several tasks, including SV, SE and SS. Several models, including PASE+ [12], TERA [6], Modified CPC [36], and Discrete BERT [44], perform worse than FBANK on SV, despite showing competitive performance on other tasks like Discrete BERT on ASR. On SE, NPC [34], DeCoAR 2 [7], and Discrete BERT perform poorly and more models fail on SS, including APC [4], VQ-APC [33], vq-wav2vec [10], and Discrete BERT. Even though Data2vec [42] Base demonstrates

TABLE II: The last layer representation v.s. weighted-sum over all layers. In each cell, the upper number represents the last layer; the lower number represents the weighted-sum. Bold fonts highlight the cases when weighted-sum is worse.

Models	PR	KS	QbE	ASR	SID	ASV	SD	IC	SF		ER
	per ↓	acc ↑	mtwv ↑	wer ↓	acc ↑	eer ↓	der ↓	acc ↑	f1 ↑	cer ↓	acc ↑
PASE+ [12]	58.88	82.37	0.07	24.92	35.84	**10.91**	**8.52**	**30.29**	60.41	62.77	57.64
	58.87	82.54	0.72	25.11	37.99	**11.61**	**8.68**	**29.82**	62.14	60.17	57.86
APC [4]	41.85	**91.04**	2.68	21.61	59.79	8.81	10.72	74.64	**71.26**	**50.76**	58.84
	41.98	**91.01**	3.10	21.28	60.42	8.56	10.53	74.69	**70.46**	**50.89**	59.33
TERA [6]	**47.53**	88.09	8.7e-3	18.45	**58.67**	16.49	**9.54**	48.8	63.28	57.91	54.76
	49.17	89.48	0.13	18.17	**57.57**	15.89	**9.96**	58.42	67.50	54.17	56.27
wav2vec [9]	32.39	94.09	3.07	16.40	44.88	9.83	10.79	78.91	**77.52**	**41.75**	58.17
	31.58	95.59	4.85	15.86	56.56	7.99	9.9	84.92	**76.37**	**43.71**	59.79
wav2vec 2.0 Base [11]	28.37	92.31	8.8e-2	9.57	45.62	9.69	7.48	58.34	79.94	37.81	56.93
	5.74	96.23	2.33	6.43	75.18	6.02	6.08	92.35	88.30	24.77	63.43
HuBERT Base [13]	6.85	95.98	7.36	6.74	64.84	7.22	6.76	95.94	86.24	28.52	62.94
	5.41	96.30	7.36	6.42	81.42	5.11	5.88	98.34	88.53	25.20	64.92

TABLE III: The best and the last model on each metric.

Task	Metric	Best	Last
PR	PER	Data2vec Large	FBANK
KS	ACC	Unispeech SAT Large	FBANK
IC	ACC	Unispeech SAT Large	FBANK
SID	ACC	WavLM Large	FBANK
ER	ACC	Unispeech SAT Large	FBANK
ASR	WER	Data2vec Large	PASE+
OODASR	WER	WavLM Large	Mockingjay
SF	F1	WavLM Large	Mockingjay
	CER	WavLM Large	PASE+
QBE	MTWV	Unispeech SAT Base+	Mockingjay
ASV	EER	Unispeech SAT Large	TERA
SD	DER	WavLM Large	Mockingjay
ST	BLEU	WavLM Large	FBANK
VC	MCD	CoBERT Base	PASE+
	WER	CoBERT Base	FBANK
	ASV	vq-wav2vec	PASE+
SE	PESQ	WavLM Large	Discrete BERT
	STOI	WavLM Large	Discrete BERT
SS	SISDRi	Unispeech SAT Large	Discrete BERT

leading performance on most tasks, it slightly underperforms FBANK on SS. Hence, we conclude that despite SSL showing promising results on the reported datasets and tasks, it is still challenging to outperform FBANK in terms of task generalizability.

2) The SOTA models on different tasks vary: By searching through the darkest (best) cell of each column in Fig 3, we find that no single model can top all the tasks simultaneously.

Firstly, we compare PR and OOD-ASR. Data2vec Large is the best model for the LibriSpeech-based PR[4], while WavLM Large ranks first on all the OOD-ASR datasets, demonstrating the effectiveness of the extra 34k hours of pre-training data for the out-of-domain scenarios.

When considering the speaker-related tasks of SID and ASV and the mixture-related tasks of SD and SS, WavLM Large and Unispeech SAT [43] Large achieve the highest rankings on SID and ASV, thanks to their pre-training task design that focuses on distinguishing between speakers in mixed signals. Consequently, these two models also outperform the others on SD and SS, as most of the SSL models lack exposure to mixture data during their pre-training phase.

On VC, the leading models are different from other tasks. We adopt Mean Cepstral Distortion (MCD) as the primary metric as it was shown well correlated to the human perception [47]. The top three models are CoBERT Base, Discrete BERT and Data2vec (Base and Large). The VC prediction head converts speech from a source speaker to a target speaker while preserving the underlying spoken contents, given the foundation model features and the target speaker. The task design thus requires the features with pure linguistic signals which disentangle the source speaker characteristics [47]. By considering VC, PR, and SID jointly, Discrete BERT achieves average performance on PR (14.32% PER) but performs poorly on SID (33.63% ACC). When comparing Data2vec Base/Large to HuBERT Base/Large and WavLM Base/Large, the performances on PR and ASR are similar, but the accuracy on SID is much lower (70.21/79.24% ACC). A similar phenomenon is observed when comparing CoBERT Base to HuBERT Base. These three models contain valuable content information but offer limited accessibility to speaker information, and become the SOTA on VC. Consequently, the representation with excellent content and speaker information simultaneously cannot easily rule the VC task, like wav2vec 2.0. This observation does not imply that it is impossible to pre-train a universal model to excel in content, speaker, and disentanglement tasks concurrently. Rather, our benchmark setting necessitates the foundation model to explicitly disentangle the information within its layers while preserving different types of information. We breakdown this analysis in more details in Section VI-B3.

Finally, in SE task, WavLM Large rank first due to its pre-training task also involving mixtures of speech and noises. In conclusion, WavLM stands out as the closest model to the concept of a foundation model, but it performs suboptimally

[4]The ASR performances between Data2vec Large and WavLM [14] Large is not significant according to Table V

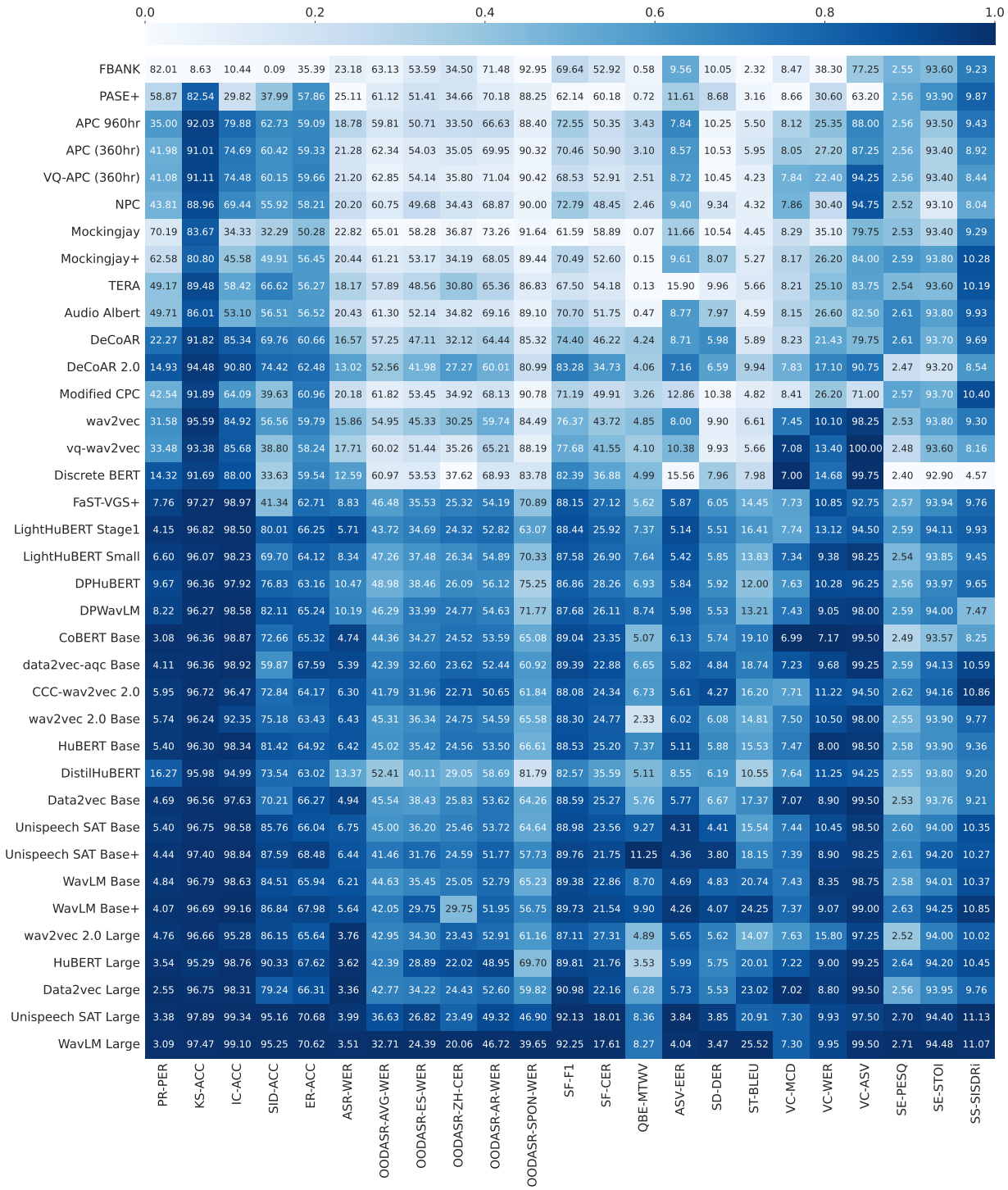


Fig. 3: The full benchmark results of 36 foundation models on 15 speech processing tasks. Each column represents a metric of a task. The heatmap reflects the performance linearly, and the darker cells of the same task always indicate better performance.

on VC for the disentanglement capability.

3) The failing patterns of SOTA models: We discuss the failing patterns of several SOTA models. Discrete BERT, one of the VC SOTA, improves upon wav2vec on LibriSpeech-based PR and ASR, while performing significantly worse on OOD-ASR. When examining each OOD-ASR dataset individually, we observe a slight improvement (0.8%) upon wav2vec in spontaneous English ASR, but a severe degradation when transferring to other languages. This result suggests that in spite of showing effectiveness for disentanglement, the codebook ID input may be more susceptible to domain shifts compared to continuous representation.

wav2vec 2.0 [11] Base, as the end-to-end upgraded version of Discrete BERT, surpasses it in most tasks except VC. When comparing wav2vec 2.0 and HuBERT [13], the latter shows slight improvements over the former across most tasks, and

this behavior is consistent for both the Base and Large variants. WavLM further outperforms HuBERT consistently in both model variants. However, despite possessing various high-quality information including much better linguistic signals on PR and ASR, these three SOTA models still fall behind the SOTA established by Discrete BERT in VC, suggesting that further efforts could be made for disentanglement.

On SE, we observe that CCC-wav2vec 2.0 [40], with only the standard Base model size, significantly outperforms wav2vec 2.0 Base and Large. The result again suggests the importance of explicitly considering noisy input during the pre-training stage.

Data2vec is of good performance on the content and semantic tasks, while performing worse for speaker recognition [5], denoising and separation. The behavior is consistent for the Base and the Large variants. This purity of content information again makes it ideal for the VC task, but not ideal for the speaker tasks as a foundation model. Interestingly, Data2vec is the only model achieving SOTA disentanglement capability without adopting the VQ technique. Additionally, Data2vec Large exhibits poor transferability to OOD-ASR datasets when compared to WavLM Large, which highlights the limitation of pre-training speech foundation models solely with the standard LibriSpeech dataset.

In conclusion, we point several weaknesses of the SOTA models in the SUPERB evaluation framework, which might be worth considering when designing new models:

- **Out-of-domain transferability**: Unable to transfer to out-of-domain ASR datasets besides LibriSpeech.
- **Robustness**: Unable to handle noisy waveforms.
- **Disentanglement**: Unable to disentangle the content and speaker information while preserving both pieces of information.

4) The impact of vector quantization: We notice a consistent pattern regarding the impact of incorporating vector quantization (VQ) into network architectures on the task generalizability of models. This observation can be illustrated by comparing pairs such as APC versus VQ-APC and wav2vec versus vq-wav2vec, Discrete BERT. While DeCoAR [35] and DeCoAR 2.0 cannot be directly compared due to the significant difference in parameter size, it is worth noting that DeCoAR 2.0 underperforms compared to DeCoAR on certain tasks, possibly due to the inclusion of an additional VQ module.

In comparison to APC, the quantized variant VQ-APC demonstrates improvements in content-related tasks such as PR, ASR, KS, and QbE. However, it also exhibits degradation in SID and SV tasks. Consistent with the discussion about disentanglement made in Section V-B2, this improvement in content-related tasks and the corresponding decline in speaker-related tasks contribute to an enhancement in the VC task, resulting in a reduction of the MCD from 8.05 to 7.84. However, VQ-APC experiences a degradation in the SS task.

[5]Note that despite Data2vec Large shows slight improvement on SD upon wav2vec 2.0 Large, our significance analysis in Table V shows the improvement is not significant. Conversely, the results of SID and SV both show that Data2vec Large is worse, getting 79% ACC and 5.73% EER.

A similar phenomenon occurs among wav2vec, vq-wav2vec, and Discrete BERT regarding VQ. We observe that vq-wav2vec exhibits degradation compared to wav2vec in most tasks. Specifically, on SID, vq-wav2vec shows a significant degradation from 56.56% to 38.8% ACC, while experiencing a slight degradation on PR from 31.58% to 33.48% PER. However, in the VC task, vq-wav2vec improves upon wav2vec, reducing the score from 7.45 to 7.08 MCD. This suggests that the uneven degradation in content and speaker still leads to an improvement in the disentanglement VC task. For Discrete BERT, the impact of quantization is more pronounced. As only the quantized codebook ID is used as input, there is a substantial information loss from the raw waveform. Discrete BERT outperforms wav2vec and vq-wav2vec in PR and ASR, but performs significantly worse in SID. The purity of the linguistic signal in Discrete BERT allows it to achieve the second ranking in VC. However, Discrete BERT ranks last in other generative tasks such as SE and SS.

Finally, while DeCoAR 2.0 outperforms DeCoAR across most tasks, it performs worse on the speaker-related SD and generative tasks SE and SS. Combining these findings, we infer that adding the VQ module after scaling up the network from the original DeCoAR might be the root cause. Based on the experiences in the VC literature [47], [48], it can be inferred that the incorporation of VQ within network architectures generally yields improved content information while sacrificing speaker information. This phenomenon results in the formation of a disentangled content-centric representation, which brings notable advantages to the VC task. However, it should be noted that VQ often leads to suboptimal performance in generative tasks such as SE and SS.

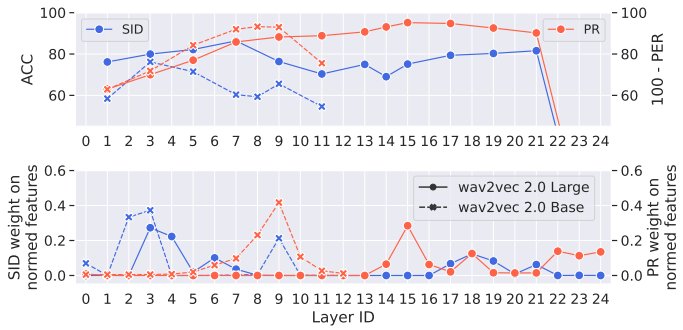
In conclusion, when constructing a speech foundation model, VQ may not be the most preferable design choice. It is not an indispensable component for achieving competitive content information and disentanglement capability, as evidenced by the success of Data2vec, and its inclusion can significantly impair speaker information, and hinder the effectiveness in generative tasks. Nevertheless, if the primary focus is exclusively on content and semantic tasks, VQ can still be considered as a desirable component.

VI. Layer-wise Analysis of Speech Foundation Models

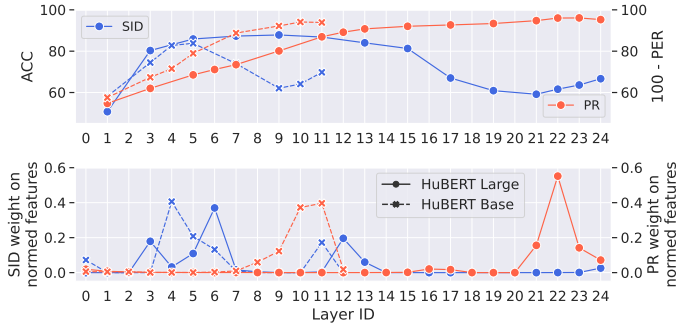
We further dive into several well-known models' layer-wise properties on several representative tasks. We first verify the correctness of a common analysis protocol for inferring each layer's importance to a task, followed by an in-depth analysis of the information flow inside speech foundation models.

A. Analysis on the trained layer-weights

After fine-tuning a trainable weighted-sum over all the layers along with the downstream prediction head, we acquire a weight for each layer contributing to the best performance on the development set. We term these trained weights as *layer-weights*. It is widely adopted to analyze each layer's importance to a task by these layer-weights [14], [29], [41],



(a) wav2vec 2.0 Base and Large. The ACC and $100 - PER$ of wav2vec 2.0 Large's final two layers are all near 0.



(b) HuBERT Base and Large.

Fig. 4: Comparing each layer's performance to layer-weights after the weighted-sum fine-tuning. The blue lines are for SID; the red lines are for PR. The solid lines are for the Large models; the dashed lines are for the Base models.

under the hypothesis that the weights are proportional to each layer's true performance. Before adopting this analysis protocol, we first examine the correlation between the layer-weights and the true performance of each layer on a few tasks to explore whether the analysis tool can lead to a reliable conclusion.

Since some models possess different numerical scales across layers, which essentially affect the layer-weights and the interpretation in our preliminary experiments[6]. We consider another benchmarking setting to factor out the effect of the feature numerical scale, *normalized benchmarking*. In a normalized benchmarking, we first normalize each layer of features by a non-trainable layer-norm across the hidden size dimension, and then take the *normalized features* for the trainable weighted-sum. Following the original benchmarking setting, the trainable floating points for all the layers are normalized into a valid probability distribution with a *softmax* function and initialized as a uniform distribution.

[6]This phenomenon is especially observable for the Large variants of wav2vec 2.0, HuBERT and WavLM. These models' last layer features are in very small numerical values compared to the other layers, and the layer-weight of the last layer is extremely large. However, the last layer of wav2vec 2.0 does not contain any useful information according to Fig 4a and Fig 5. Hence, we infer that the layer-weights might serve two functionalities jointly: (1) normalizing the numerical scale across layers, (2) identifying the informative layers. As an importance analysis tool, we only care about the functionality (2), hence the raw layer-weight from the default benchmarking is not an appropriate choice.

TABLE IV: Spearman's ρ between layer-weights from a *normalized benchmarking* and the true layer performances. The *Score* is designed to be higher for better performance.

Task	PR	SID	ER	VC	SE
Score	100 - per	acc	acc	-mcd	pesq
ρ	0.393	0.494	0.371	-0.693	0.711
p-value	0.031	0.007	0.041	0	0

The results for wav2vec 2.0 Base/Large and HuBERT Base/Large are presented in Fig 4. The performances of these four models on PR and SID with the normalized benchmarking do not show significant differences compared to the default benchmarking setting in SectionV-B. We show layer-weights for all the layers, while due to the huge computation cost of layer-wise benchmarking, we only benchmark the odd layers.

On the Base models, the layer-weights roughly reflect the true performance on PR and SID, despite some inconsistency exists. E.g. The 5-th layer of wav2vec 2.0 Base on SID is better than the 9-th layer, but the layer-weights suggest the opposite. When considering the Large models, the inconsistency between the layer performances and the layer-weights becomes more severe. On SID, the layer-weights fail to locate the best layer for both wav2vec 2.0 Large and HuBERT Large.

Furthermore, on both PR and SID the layer-weights fail to reflect the smooth information change inside the speech foundation models. For HuBERT Large, the speaker information rises smoothly and reaches the peak at layer 9, and then it decreases smoothly. Conversely, the layer-weights suggest that HuBERT Large's layer 3, layer 6 and layer 12 are the best three layers for SID, and layer 7$\sim$11 are the layers with poor speaker information.

Finally, we compute the Spearman's rank correlation coefficient (Spearman's ρ) between the layer performances and the layer-weights. We use $1 - PER$ and ACC as the layer performance of PR and SID to compute the correlation. The ρ values of the Base models, Large models, and both variants are 0.7, 0.32, and 0.4 respectively, suggesting that layer-weights are not well-correlated with the layer performances especially on the Large models. By further examining wav2vec 2.0 Large and HuBERT Large on ER, VC and SE, Table IV shows that the Large models' layer-weights are not proportional to layer performances for all the tasks except SE. On VC, they are even negatively correlated, suggesting that the weighted-sum optimization process was misguided by the unwanted information, which we will elaborate on further in Section VI-B2.

In conclusion, we verify the inconsistency between the trained layer-weights and the true performance of each layer across several tasks. As a result, we opted to benchmark each layer solely for a reliable understanding on the information flow inside speech foundation models in Section VI-B.

B. Layer-wise single-layer benchmarking

In Fig 5, we present the single-layer benchmark results for some representative models: APC, TERA, DeCoAR 2.0, wav2vec 2.0 Base, wav2vec 2.0 Large, HuBERT Base, HuBERT Large, Data2vec Large and WavLM Large. Due to

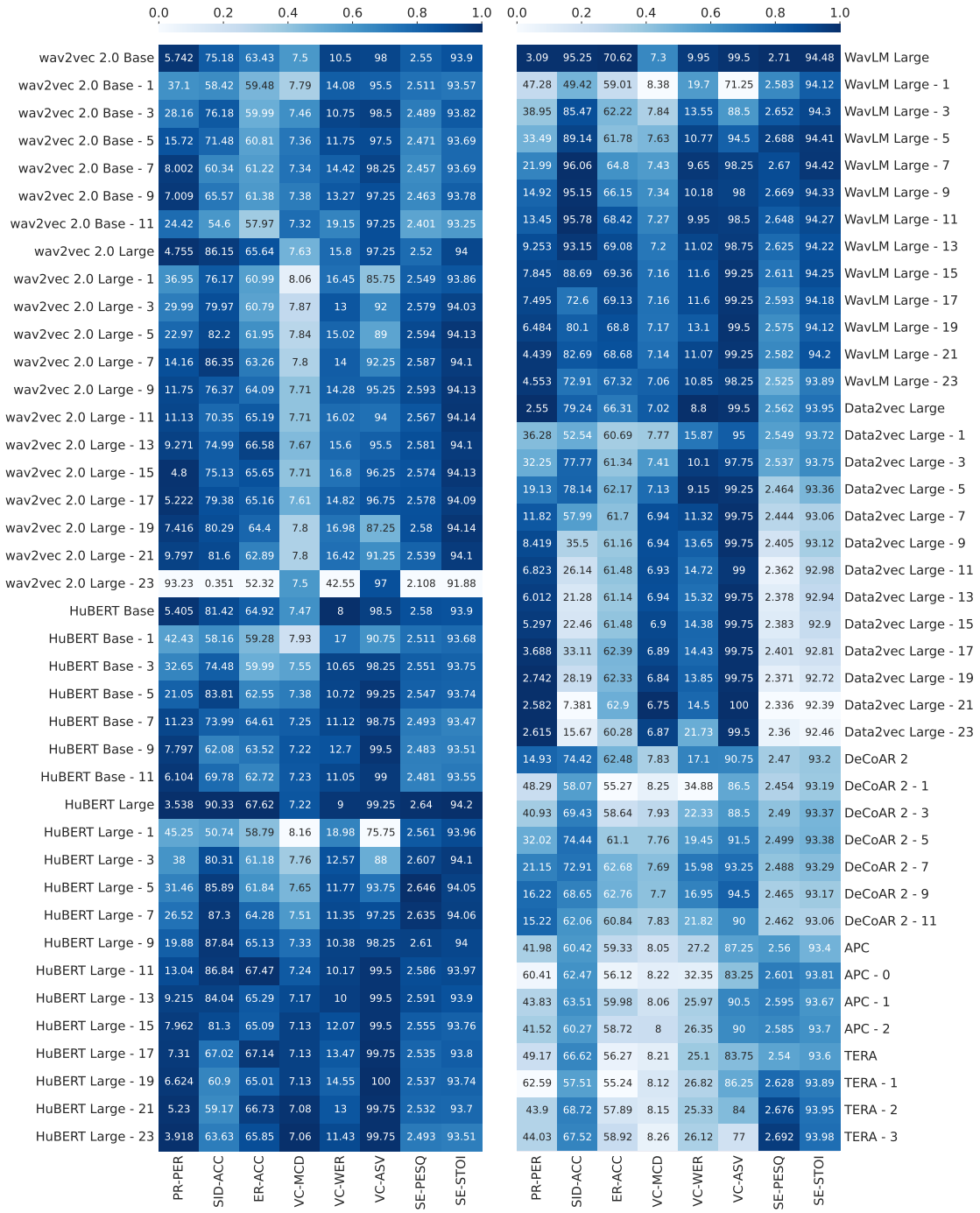


Fig. 5: Layer-wise benchmarking of 9 speech foundation models on 5 tasks. The model name's suffix number represents the layer ID, and the corresponding row is the result of single-layer benchmarking. We benchmark on the odd layers for all the models except APC and TERA since they have fewer layers. For the row without the suffix number, it means the model is trained with the weighted-sum protocol.

the huge computational cost, we benchmark the odd layers on a subset of representative tasks: PR, SID, ER, VC and SE. Intuitively, these tasks represent models' capability on content, speaker, emotion, disentanglement (content/speaker), and denoising (noise/speech discrimination in our setting).

1) Information flow inside foundation models: We observe that different information locates at different layers. The information changes inside speech foundation models present a similar trend across different models. The lower layers benefit the SE task which requires the manipulation of STFT masks when considering HuBERT Base/Large, WavLM Large, Data2vec Large, APC, and DeCoAR. For wav2vec 2.0 Base/Large and TERA, the SE differences between layers are less obvious. In the middle layers, the information changes to benefit SID, followed by ER. Finally, the higher layers benefit the PR task.

2) Pushing limits with single-layer benchmarking: We observe that sometimes the single-layer benchmarking outperforms the weighted-sum benchmarking. The examples include APC and TERA on PR; wav2vec 2.0 Base and HuBERT Base on SID; DeCoAR 2.0 and TERA on ER; wav2vec 2.0 Large and HuBERT Large on PESQ for SE. This phenomenon has also been reported in [49], which focuses on analyzing the content information across layers. Here we observe that the phenomenon is consistent across several tasks.

It is worth mentioning that, on VC, all the models in Fig 5 improve significantly with single-layer benchmarking, and the well-performing layers are with low PER on PR and low ACC on SID. Combined with our observation in Section VI-A, we infer that these models all possess the layer with rich speaker characteristics and the weighted-sum protocol inevitably exposes the unwanted source speaker information to the downstream prediction head, leading to a suboptimal transferability to the target speaker.

3) Feature disentanglement across layers: When considering the relation between PR, SID and VC across layers, we find that HuBERT Base/Large, WavLM Large, and Data2vec Large all show a clear disentanglement that the layers with excellent content information get poor speaker information, and vice versa. wav2vec 2.0 Large shows only a weak disentanglement compared to HuBERT Large especially in Fig 4. HuBERT Large's better disentanglement leads to a better single layer performance and a better weighted-sum performance on VC with 7.06 and 7.22 MCD respectively, compared to wav2vec 2.0 Large's 7.61 and 7.63 MCD. HuBERT Large achieves better disentanglement without sacrificing the performances on content and speaker related tasks, as verified by our main result in Fig 3 that HuBERT Large is better than wav2vec 2.0 Large in most tasks. The same phenomenon can be observed when comparing WavLM Large against wav2vec 2.0 Large.

Combined with the reported autoencoder-style behavior of wav2vec 2.0 [45] and the rising SID accuracy at the final layers in Fig 5, we infer that the objective of wav2vec 2.0, InfoNCE, might lead to poorer disentanglement capability. Since the network's final layers must be similar to the positive samples from the prior layers hence both ends contain low-level entangled information, and make the network hard to cleanly separate the content and speaker information across model layers.

These results confirm that our benchmark settings necessitate speech foundation models possessing various types of rich information explicitly disentangled across layers, and it is promising to deliver such a model with different learning objectives as demonstrated by HuBERT Large.

It is worth noting that despite wav2vec 2.0 Base being worse than wav2vec 2.0 Large on most tasks including PR and SID in Fig 3, the former uncommonly outperforms the latter on VC with 7.5 and 7.63 MCD respectively. This outlier could be explained by the better disentanglement capability of wav2vec 2.0 Base as verified in Fig 4a when comparing the 8-th layer of wav2vec 2.0 Base and the 14-th layer of wav2vec 2.0 Large. The reason behind the Base model's better disentanglement compared to the Large model is yet to be explored, but this phenomenon reveals an important fact that the disentanglement capability does not improve trivially as the model size or the pre-training data amount scales up, and more efforts could be put into developing powerful, disentangled, and scalable speech foundation models.

4) SOTA disentanglement capability of Data2vec Large: By examining the layer-wise performances of Data2vec Large, we find that Data2vec achieves high VC performance by sacrificing speaker information throughout the model layers. Furthermore, the 21-th layer reaches the highest disentangled content representation with 2.58% PER for PR and only 7.38% ACC for SID. Compared to recent efforts focusing on speaker disentanglement, where both ContentVec [50] and Spin [51] report SUPERB PR performance above 4% PER and SID performance above 10% ACC, Data2vec Large show especially strong disentanglement. In spite of not being directly comparable due to the Large model size, the results still suggest that Data2vec Large is ideal for the tasks requiring pure linguistic information including voice conversion, acoustic unit discovery tasks in ZeroSpeech [18] and the discrete unit-based systems [52]. We verify Data2vec Large's disentanglement capability with our VC task and achieve the new SOTA of 6.75 MCD with single-layer benchmarking on the 21-th layer. The result is surprising as Data2vec was not designed for disentanglement and did not adopt the VQ technique. We leave explaining and exploring Data2vec Large's disentanglement characteristic as future work.

VII. Significance of SUPERB

We analyze the statistical significance for the current SUPERB leaderboard. Since our submission system requires the downstream prediction of each task, we can analyze the significance between models including community submissions. Since the downstream predictions are from the same testing set and a testing utterance is evaluated twice with two models, we adopt the paired tests:

- **Paired t-test**: For the tasks with recording-level or query-level continuous metrics. We compute the query-level MTWV for QBE and adopt the recording-level metrics for the other tasks, including PER, WER, slot-type F1, DER, BLEU, PESQ, STOI, SISDRi, and MCD.
- **McNemar test**: For the tasks with categorical recording prediction, including ACC for SID and EER for SV.

For SE, we report both PESQ and STOI as there is no apparent better choice. The scores of them are frequently very closed and merit a discussion on their significance. For VC, we report the primary metric MCD [47]. For OOD-ASR, we compute the p-value separately for SBCSAE, Common Voice Spanish, Mandarin, and Arabic. Then, we average 4 p-values as the final p-value.

Note that two overall scores are not required to be hugely different to be significant. When the improvement is small but consistent across all testing utterances, the difference is significant; while when the overall improvement is large but inconsistent across utterances (i.e. with serious degradation on some utterances), the overall difference is unreliable and insignificant.

We present the results in Table V for four well-known SOTA models: wav2vec 2.0 Large, HuBERT Large, Data2vec Large

TABLE V: The p-values of the paired t-test or McNemar test for the Large models. The bold cells mark the cases when the difference is insignificant (p-value > 0.05).

		HuBERT	W2V2	Data2vec	WavLM	
	per	PR		ASR		wer
HuBERT	3.29	×	**.8045**	.0195	**.2852**	3.76
W2V2	4.75	0	×	.0404	**.4269**	3.62
Data2vec	2.55	0	0	×	**.1917**	3.44
WavLM	3.22	**.1855**	0	0	×	3.36
	acc	KS		QbE		mtwv
HuBERT	95.29	×	**.1192**	.0018	0	3.53
W2V2	96.27	.009	×	.0174	0	5.06
Data2vec	96.75	0	**.1289**	×	0	6.28
WavLM	97.47	0	0	0	×	8.86
	acc	IC		ER		acc
HuBERT	98.76	×	.0028	.0005	.0354	67.58
W2V2	95.68	0	×	**.5558**	0	65.64
Data2vec	98.31	**.0827**	0	×	0	65.29
WavLM	99.31	.0035	0	0	×	68.87
	slot-f1	SF		SD		der
HuBERT	89.81	×	**.1569**	**.0513**	0	5.75
W2V2	86.94	0	×	**.5412**	0	5.62
Data2vec	90.98	0	0	×	0	5.53
WavLM	92.21	0	0	.0001	×	3.24
	acc	SID		SV		eer
HuBERT	90.33	×	.0029	**.0550**	0	5.99
W2V2	86.15	0	×	**.5267**	0	5.65
Data2vec	76.77	0	0	×	0	5.73
WavLM	95.49	0	0	.0002	×	3.77
	wer	OOD-ASR (avg)		ST		bleu
HuBERT	42.28	×	0	0	0	20.23
W2V2	42.90	0	×	0	0	12.78
Data2vec	42.71	0	**.4368**	×	.0113	23.02
WavLM	32.66	0	0	0	×	26.56
	pesq	SE (pesq)		SE (stoi)		stoi
HuBERT	94.18	×	0	0	0	2.64
W2V2	94.04	.0036	×	0	0	2.52
Data2vec	93.95	0	.0444	×	0	2.56
WavLM	94.51	0	0	.0002	×	2.70
	sisdri	SS		VC		mcd
HuBERT	10.45	×	0	0	0	7.22
W2V2	10.02	0	×	0	0	7.63
Data2vec	9.76	0	0	×	0	7.02
WavLM	11.07	0	0	0	×	7.3
	wer	OOD-ASR (es)		OOD-ASR (ar)		wer
HuBERT	28.89	×	0	0	0	48.95
W2V2	34.3	0	×	**.3826**	0	52.91
Data2vec	34.22	0	**.4102**	×	0	52.6
WavLM	24.39	0	0	0	×	46.72
	cer	OOD-ASR (zh)		OOD-ASR (spon)		wer
HuBERT	22.02	×	0	0	0	69.7
W2V2	23.43	0	×	**.9544**	0	61.16
Data2vec	24.43	0	0	×	0	59.82
WavLM	20.06	0	0	0	×	39.65

and WavLM Large. On SV, when taking HuBERT Large as the baseline, wav2vec 2.0 Large improves 0.34 EER which is significant, while Data2vec Large improves 0.26 EER which is insignificant. The decision boundary is hard to infer without statistical tools. On QBE, wav2vec 2.0 Large and Data2vec Large show a 1.22 MTWV difference which is significant, while wav2vec 2.0 Large and HuBERT Large show a 1.53 MTWV difference which is instead insignificant, suggesting that a larger difference on the overall scores do not always lead to more significant results. Hence, it is important to analyze the significance explicitly with the downstream predictions.

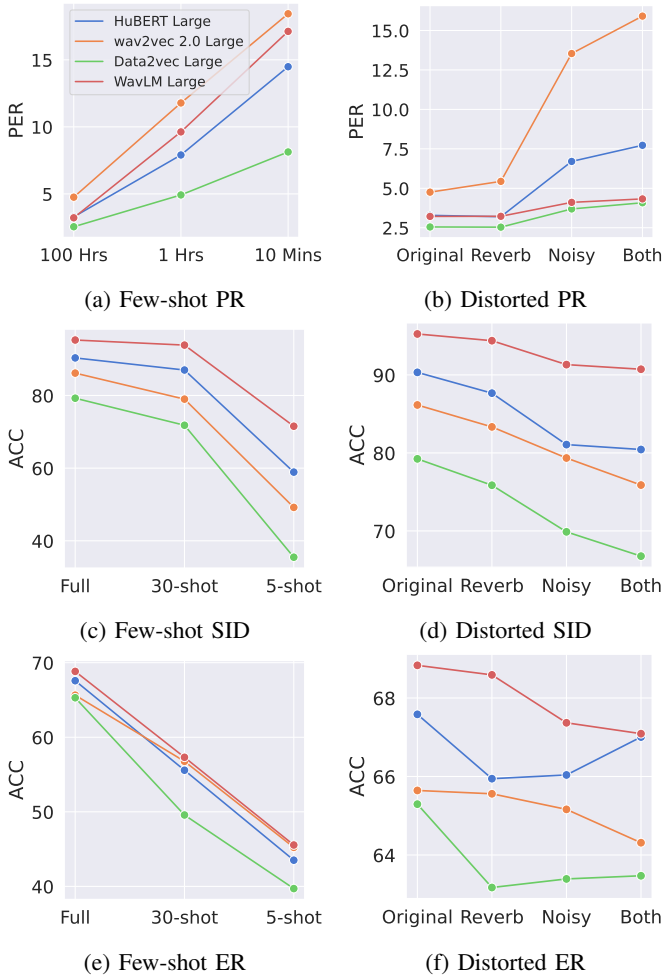
The majority of results obtained on LibriSpeech ASR are deemed insignificant. Although WavLM appears to rank first in terms of WER for ASR, the significance test reveals that this is not the case within our ASR setting. The results further suggest that the evaluation of speech foundation models on the standard LibriSpeech dataset is saturating. As a comparison, the differences in OOD-ASR are most significant, even for each separate dataset, and WavLM improves upon others significantly on all the OOD-ASR datasets.

The differences in SV and SD are also frequently insignificant except WavLM. In terms of DER scores, Data2vec Large ranks ahead of wav2vec 2.0 Large, followed by HuBERT Large. However, the p-values indicate that their performances are statistically equal. These results once again highlight the misleading nature of ranking models based solely on task scores, as even a minor random disturbance can result in a noticeable alteration in the ranking. Interestingly, despite that the PESQ, STOI and SISDRi scores on SE and SS are highly similar for all the models, they all pass the significance test, suggesting that the improvement is small but consistent. On VC, all the MCD results are significant.

We conclude that insignificant results exist and it is unreliable to rank models solely according to the score on each task. We release the downstream predictions of these SOTA models, along with the code to calculate the significance of all tasks in S3PRL. We recommend the participants to always explicitly consider the statistical significance when evaluating with SUPERB benchmark and compare to these released baselines.

VIII. Robustness of SUPERB

We discuss the robustness of the proposed benchmark to understand the transferability of the conclusion derived from the benchmark results. We examined the robustness of SUPERB-SG in [21], and extend the examination to the tasks defined in SUPERB [20] in this work. Due to the space limit, we select representative tasks, PR, SID and ER for content, speaker and paralinguistic information respectively. We discuss two types of condition variations: low-resource and distorted recordings. For the low-resource condition, we consider two levels. For PR, we randomly sample 1 hour and 10 minutes of recordings from the LibriSpeech train-clean-100 subset for the few-shot and extreme few-shot conditions respectively; we randomly sample 30 and 5 utterances from each speaker for SID; we randomly sample 30 and 5 utterances from each emotion category for ER. For the distorted condition, we consider adding additive noises, reverberation and both. For additive noise, the WHAM! [53] dataset's training, validation, and testing sets are applied to the training, validation, and testing sets of PR, SID and ER respectively. The SNR for each noise addition is randomly sampled from 3, 6, and 9 dB. For



(a) Few-shot PR (b) Distorted PR

(c) Few-shot SID (d) Distorted SID

(e) Few-shot ER (f) Distorted ER

Fig. 6: PR, SID and ER under few-shot and distorted conditions with HuBERT Large, wav2vec 2.0 Large, Data2vec Large and WavLM Large.

reverberation, we leverage *SoX*[7] to add the reverberation with the *freeverb* algorithm. We randomly sample from 0∼80 for reverberance. When both additive noise and reverberation are applied, we follow the same settings above. We present the results for HuBERT Large, wav2vec 2.0 Large, WavLM Large and Data2vec Large in Fig 6.

Firstly, Fig 6 shows that for PR, SID and ER, different condition changes do not lead to significantly different rankings, except that wav2vec 2.0 Large shows slightly better robustness than HuBERT Large in the low-resource ER. In the few-shot PR, we find that Data2vec Large shows better robustness in the low-resource conditions according to its smoother slope compared to all the others. Despite WavLM Large and HuBERT Large show similar performance in the default PR setting with 100 hours of data, HuBERT Large is more robust against the few-shot 1-hour and 10-minute settings. On the other hand, WavLM is much more robust against the additive noises in the distorted PR. This suggests that while the models might exhibit similar performance in the default SUPERB setting, they could possess very different robustness characteristics. In few-shot SID and distorted SID, the default SUPERB can perfectly reflect the performance.

In conclusion, similar to the results in [21], the default experimental settings of SUPERB provide a valuable reflection of the models' relative performance in various scenarios, albeit with a few exceptions. We further observe that each model exhibits different levels of robustness against different conditions, and the default SUPERB evaluation may not always capture these characteristics comprehensively due to the saturating performance. Nonetheless, the findings from SUPERB offer promising insights into the models' capabilities across different scenarios.

IX. Conclusion

We present SUPERB benchmark, an interactive platform for evaluating speech foundation models. The standardized 15 tasks cover a wide range of speech processing, including both discriminative and generative tasks. The 36 evaluated models provide comprehensive baselines. We point out common pitfalls of benchmarking and analyzing speech foundation models, including inappropriate ranking between models regardless of statistical significance, and the inaccurate implication of the learned layer-weight. Our results suggest that in order to be universal on the SUPERB benchmark, the foundation model should possess various high-quality information and be robust to domain shifts and noisy conditions. Furthermore, different types of information should be explicitly disentangled inside the model. Finally, analyses of SUPERB's significance and robustness reveal the need for a harder version of SUPERB, which we leave for future work. We open-source all the materials to lower the barrier for reproduction, benchmarking, submission, and analysis. We welcome researchers to join our active community and drive the research frontier together.

[7]https://sox.sourceforge.net/

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th ICML*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607.

[4] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Interspeech*, 2019, pp. 146–150.

[5] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," *ICASSP*, 2020.

[6] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[7] S. Ling and Y. Liu, "Decoar 2.0: Deep contextualized acoustic representations with vector quantization," *arXiv preprint arXiv:2012.06659*, 2020.

[8] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[9] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in *Interspeech*, 2019.

[10] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *ICLR*, 2020.

[11] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[12] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP*, 2020, pp. 6989–6993.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "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.

[14] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[15] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in *EMNLP*, 2018, pp. 353–355.

[16] N. Zeghidour, O. Teboul, F. de Chaumont Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *ICLR*, 2021.

[17] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally *et al.*, "HEAR: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[18] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[19] S. Evain, H. Nguyen, H. Le, M. Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet *et al.*, "Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech," in *INTERSPEECH 2021: Conference of the International Speech Communication Association*, 2021.

[20] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[21] H.-S. Tsai, H.-J. Chang, W.-C. Huang, Z. Huang, K. Lakhotia, S.-w. Yang, S. Dong, A. Liu, C.-I. Lai, J. Shi *et al.*, "SUPERB-SG: Enhanced speech processing universal performance benchmark for semantic and generative capabilities," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8479–8492.

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

[23] S. Shon, S. Arora, C.-J. Lin, A. Pasad, F. Wu, R. Sharma, W.-L. Wu, H.-Y. Lee, K. Livescu, and S. Watanabe, "SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks," *arXiv preprint arXiv:2212.10525*, 2022.

[24] J. Shor, A. Jansen, W. Han, D. Park, and Y. Zhang, "Universal paralinguistic speech representations using self-supervised conformers," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3169–3173.

[25] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[26] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[27] A. Conneau, A. Bapna, Y. Zhang, M. Ma, P. von Platen, A. Lozhkov, C. Cherry, Y. Jia, C. Rivera, M. Kale *et al.*, "Xtreme-s: Evaluating cross-lingual speech representations," *arXiv preprint arXiv:2203.10752*, 2022.

[28] T.-h. Feng, A. Dong, C.-F. Yeh, S.-w. Yang, T.-Q. Lin, J. Shi, K.-W. Chang, Z. Huang, H. Wu, X. Chang *et al.*, "SUPERB@SLT 2022: Challenge on generalization and efficiency of self-supervised speech representation learning," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1096–1103.

[29] J. Shi, D. Berrebbi, W. Chen, H.-L. Chung, E.-P. Hu, W. P. Huang, X. Chang, S.-W. Li, A. Mohamed, H.-y. Lee *et al.*, "ML-SUPERB: Multilingual speech universal performance benchmark," *arXiv preprint arXiv:2305.10615*, 2023.

[30] Y. Peng, Y. Sudo, S. Muhammad, and S. Watanabe, "Dphubert: Joint distillation and pruning of self-supervised speech models," *arXiv preprint arXiv:2305.17651*, 2023.

[31] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "Lighthubert: Lightweight and configurable speech representation learning with once-for-all hidden-unit bert," *arXiv preprint arXiv:2203.15610*, 2022.

[32] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[33] Y.-A. Chung, H. Tang, and J. Glass, "Vector-quantized autoregressive predictive coding," in *Interspeech*, 2020, pp. 3760–3764.

[34] A. H. Liu, Y.-A. Chung, and J. Glass, "Non-autoregressive predictive coding for learning speech representations from local dependencies," *arXiv preprint arXiv:2011.00406*, 2020.

[35] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6429–6433.

[36] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP*, 2020, pp. 7414–7418.

[37] C. Meng, J. Ao, T. Ko, M. Wang, and H. Li, "Cobert: Self-supervised speech representation learning through code representation learning," *arXiv preprint arXiv:2210.04062*, 2022.

[38] P. Peng and D. Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," *arXiv preprint arXiv:2202.03543*, 2022.

[39] V. S. Lodagala, S. Ghosh, and S. Umesh, "data2vec-aqc: Search for the right teaching assistant in the teacher-student training setup," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[40] ——, "Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1–8.

[41] H.-J. Chang, S.-w. Yang, and H.-y. Lee, "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7087–7091.

[42] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1298–1312.

[43] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li *et al.*, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6152–6156.

[44] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7694–7698.

[45] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[46] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira *et al.*, "Towards learning universal audio representations," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4593–4597.

[47] W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A comparative study of self-supervised speech representation based voice conversion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1308–1318, 2022.

[48] D.-Y. Wu and H.-y. Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7734–7738.

[49] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[50] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised

speech representation by disentangling speakers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 003–18 017.

[51] H.-J. Chang, A. H. Liu, and J. Glass, "Self-supervised fine-tuning for improved content representations by speaker-invariant clustering," *arXiv preprint arXiv:2305.11072*, 2023.

[52] A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang *et al.*, "Direct speech-to-speech translation with discrete units," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3327–3339.

[53] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," *Proc. Interspeech 2019*, pp. 1368–1372, 2019.