Improving Spoken Language Modeling with Phoneme Classification: A Simple Fine-tuning Approach

Anonymous ACL submission

Abstract

Recent progress in Spoken Language Modeling has demonstrated the feasibility of learning language directly from speech. Generating speech through a pipeline that operates at the text level typically loses nuances, intonations, and non-verbal vocalizations. Modeling directly from speech opens up the path to more natural and expressive systems. On the other hand, speech-only systems tend to trail behind text-based language models in terms of their semantic abilities. We show that fine-tuning speech representation models on phoneme classification leads to more context-invariant representations, which in turn improve downstream language modeling performance.

1 Introduction and related work

Recent advances in Self-supervised Speech Representation Learning (SSL) (Mohamed et al., 2022; Chen et al., 2022; Hsu et al., 2021; Baevski et al., 2020) have enabled the development of label-free representations that are valuable for various downstream tasks (wen Yang et al., 2021). These representations can be discretized and treated as pseudo-text, allowing for the training of language models directly from raw audio (Lakhotia et al., 2021), which capture both prosody and linguistic content (Kharitonov et al., 2022). Applications of these audio-based language models include dialogue modeling (Nguyen et al., 2023b), emotion conversion (Polyak et al., 2021), and direct speech-to-speech translation (Lee et al., 2022). They can be trained not only on discretized SSL representations but also on continuous word-size tokens (Algayres et al., 2023) or a combination of acoustic and semantic tokens (Borsos et al., 2023). However, these models still lag behind their text-based counterparts in terms of capturing semantics when trained with similar data quantity (Nguyen et al., 2020). Recent approaches tackled this issue by jointly training speech and text Language Models (LMs) (Nguyen et al., 2024; Maiti et al., 2024; Chou et al., 2023) or by using existing LMs as a warm initialization (Hassid et al., 2023).

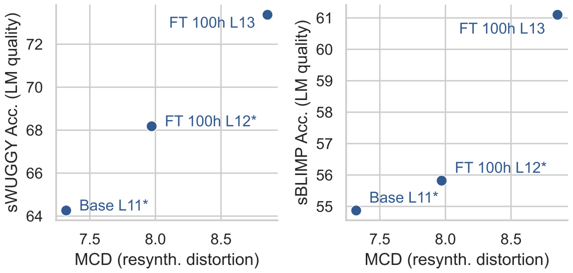

Figure 1: Trade-off between language modeling and expressive resynthesis. *: embeddings initialized from unit centroids.

Unfortunately, SSL units are not invariant to irrelevant acoustic variations, which hinders downstream language modeling. Recent works have addressed this issue for background noise (Chen et al., 2022), speech rate change (Gat et al., 2023), and speaker change (Qian et al., 2022; Chang et al., 2023; Chang and Glass, 2023). However, contextual variations due to coarticulation remain a challenge (Hallap et al., 2023): SSL units align more closely with contextual phone states (Young et al., 1994) rather than linguistic units (Dunbar et al., 2022), which may affect the LM's capacity to learn higher-order representations of language.

Here, we test a simple idea: using supervised fine-tuning from a phoneme classification task to help the model remove its contextual dependency. We first show that fine-tuned models learn representations that are much more context-invariant than the original SSL representations, even with as little as a few hours of labels. Next, we show that these representations can be used to train an LM that outperforms the standard approach. We then evaluate whether the fine-tuned representations have retained their expressive power by measuring the distortion when resynthesizing expressive speech.

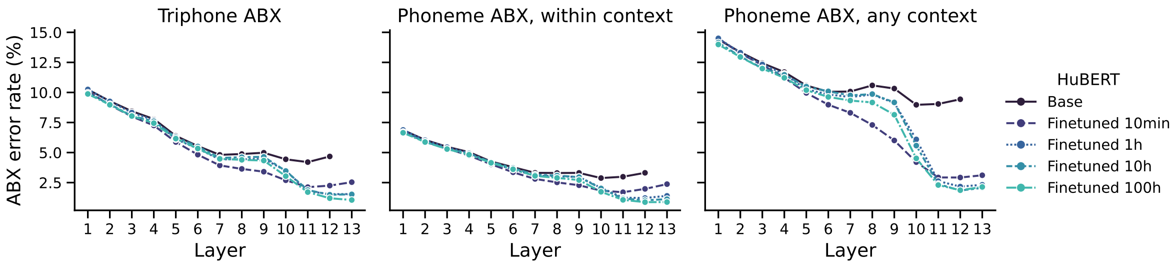

Figure 2: ABX error rate averaged across subset (dev-clean, dev-other) and speaker (within, across) conditions.

2 Method

2.1 Phoneme classification

We started from the pretrained HuBERT (Hsu et al., 2021) Base model, with 95M parameters, and fine-tuned it on a phoneme classification task. Instead of using a CTC objective (Graves et al., 2006), we resorted to frame-level phoneme classification. Our main focus is not the quality of the phonemic transcription of the audio. Our goal is that individual representations are well aligned with the phonemic content of the audio source. By operating at the frame level, we impose a strong condition on each frame and do not lose the temporal component. Such an approach comes at the cost of needing forced alignments from the training dataset, when using CTC only demands sentence-level transcripts. We added one fully connected layer on top of the HuBERT backbone that maps the 768-dimensional representation to our phoneme space of dimension 40. We fine-tuned this model on LibriSpeech `train-clean-100` (Panayotov et al., 2015). We also reported results for models fine-tuned on LibriLight Limited 10 h, 1 h, and 10 min (Kahn et al., 2020). The forced alignments are those used in Nguyen et al. (2020), obtained with the Abkhazia library[1]. The fine-tuning hyperparameters are derived from those used in Hsu et al. (2021) for ASR. We trained for 20 000 steps with a batch size of 32 on a single NVIDIA V100 GPU.

2.2 Quantization

We selected the best layer in terms of Triphone ABX score for the standard HuBERT base and the model fine-tuned on `train-clean-100`. We trained k-means models on the features of a 10 h subset of `train-clean-100` extracted from those layers, with $k = 500$. We also quantized the logits of the fine-tuned model by simply setting the labels as the predicted phonemes for each frame.

2.3 Language modeling

Finally, we trained LMs on the discretized units. The language model is a 3-layer LSTM, following the low-budget baseline of Nguyen et al. (2020), only changing the embedding dimension from 200 to 768. It was trained on the discrete units of LibriSpeech 960 h, for 30 000 steps on a single NVIDIA V100 GPU. This 26M parameters language model is two orders of magnitude smaller both in terms of number of parameters and hours of training data than Spoken LMs like TWIST (Hassid et al., 2023) or SpiRit-LM (Nguyen et al., 2024). Our fine-tuned units can in principle benefit any other LM, including these larger ones.

2.4 Speech resynthesis

For speech resynthesis, we trained a HiFi-GAN (Kong et al., 2020; Polyak et al., 2021) on the EXPRESSO dataset (Nguyen et al., 2023a), conditioned on the HuBERT discrete speech units and one-hot speaker embeddings from one of EXPRESSO's voices. We trained for 250 000 steps on two NVIDIA V100 GPUs and followed the other hyperparameters used in EXPRESSO. In this setup the HiFi-GAN has a different training domain than the HuBERT, the k-means, and the LM, which were trained on the audiobooks of LibriSpeech.

2.5 Evaluation metrics

We evaluate continuous and discrete units using ABX discriminability (Schatz et al., 2013; Schatz, 2016). This task quantifies the discriminability between two sound categories, A and B, as the probability that a token x of category A will be closer to another $a \in A$ than to a $b \in B$. The dissimilarity function is the dynamic time-warping aligned angular distance between the model's representations of two sounds. The ABX error rate is calculated by averaging the discriminabilities for all pairs of categories and subtracting it from 1. In the standard evaluation, each token is a triphone and triphones differ only by the central phoneme

[1] `https://github.com/bootphon/abkhazia`

	Triphone ABX ↓	Phoneme ABX ↓	
		W/in ctx	Any ctx
Continuous			
wav2vec 2.0 Base L6	5.41	3.78	11.55
WavLM Base L11	3.57	2.54	8.26
ContentVec$_{100}$ L12	3.84	2.54	6.89
HuBERT + Spin$_{2048}$ L12	3.05	2.31	7.63
Continuous			
Base L11	4.20	2.98	9.04
FT 100h L12	1.20	**0.87**	**1.87**
FT 100h L13	**1.05**	0.88	2.14
Centroid			
Base L11	4.54	3.84	7.34
FT 100h L12	1.65	1.92	2.76
One-hot			
Base L11	7.81	12.23	30.00
FT 100h L12	4.02	6.51	26.88
FT 100h L13	4.08	4.78	5.40

Table 1: ABX error rate on selected layers averaged across subset and speaker conditions. Without quantization, when considering the k-means centroid and with one-hot encoding. For each condition, the best score is in **bold** and the second best is underlined.

in a triplet. In the "within speaker" condition, a, b, and x come from the same speaker, while in the "across speaker" condition, a and b come from the same speaker, and x from another one.

Following Hallap et al. (2023), we also evaluate our models on the Phoneme ABX task, where each token is a phoneme. We examine two conditions: "within context" (constant preceding and following phonemes) and "any context" (no constraints on context). This task assesses context-invariance in speech representations, revealing that current self-supervised systems struggle with context independence. Notably, in Hallap et al. (2023) the performance drop when removing the constant context condition is larger than the gaps observed in speaker independence or clean versus less-clean speech conditions. By fine-tuning at a frame level without taking into account the context, our approach is a way to directly tackle this issue.

We evaluate spoken language modeling at the lexical and syntactic levels using the sWUGGY and sBLIMP metrics from the ZeroSpeech 2021 challenge (Nguyen et al., 2020). sWUGGY is a "spot-the-word" task, where the network is presented with a word and a matching non-word, and evaluated on its ability to assign a higher probability to the true word. We also report results for "in-vocab" pairs, which only contains words from LibriSpeech. sBLIMP assesses the network's ability to prefer grammatically correct sentences over incorrect ones, given a pair of matching sentences.

We evaluate content preservation in resynthesized speech by following (Nguyen et al., 2023a) and running wav2vec 2.0 Large ASR (Baevski et al., 2020) on the resynthesized speech, reporting the Word Error Rate (WER). We assess this on EXPRESSO-READ the reading subset of EXPRESSO – in-domain for the vocoder but out-of-domain for the HuBERT backbone and the k-means module – and on LibriSpeech, which is out-of-domain for the vocoder. On EXPRESSO the target voice is the same as the input voice, while on LibriSpeech the target voice is sampled from the four voices. We also compute the mel cepstral distortion (MCD) (Kubichek, 1993) between the original and resynthesized samples of EXPRESSO-READ using Sternkopf and Taubert (2024).

3 Results

3.1 Results at the phonemic level

As shown in Figure 2, we computed the ABX error rate for each Transformer layer of the base model and the fine-tuned models, including the added fully connected layer (layer 13). We calculated both triphone- and phoneme-level ABX error rates. Fine-tuning mainly improves the last layers' ABX error rates, with near-perfect scores for the 10h and 100h fine-tuned models in the "within context" condition. SSL representations generally struggle more in the "any context" condition: there the gain in error rate is the most significant, dropping from 9.4% to 2.4% after fine-tuning on as little as 10 minutes. Fine-tuning pushes representations to become more context-independent.

We selected the best layers for the base model (layer 11) and fine-tuned 100h model (layer 12) based on the Triphone ABX score, as well as the last layer of the fine-tuned 100h model (layer 13). We trained k-means on these representations and report the results in Table 1. We compare these to the ABX error rates of the best layers of wav2vec 2.0 (Baevski et al., 2020), WavLM (Chen et al., 2022), ContentVec$_{100}$ (Qian et al., 2022) and HuBERT + Spin$_{2048}$ (Chang et al., 2023). For the centroid scores, each representation is replaced by the continuous representation of the closest centroid in k-means. For the one-hot scores, each representation is replaced by a one-hot vector with a 1 at its label position. We use the same distance to compute the ABX as for continuous representations. In the case of the base model's layer 11 (Base L11) and the fine-tuned 100h model's layer 12 (FT 100h

	WER ↓					MCD ↓
	dev-clean	dev-other	test-clean	test-other	EXPRESSO-READ	EXPRESSO
Original audio	1.69	3.55	1.86	3.89	11.90	-
Base L11	3.82	11.37	4.12	11.26	20.93	7.32
FT 100h L12	4.36	10.75	4.62	10.90	23.03	7.97
FT 100h L13	5.78	11.90	5.97	12.12	23.80	8.85

Table 2: Resynthesis evaluation. WER is computed using a wav2vec 2.0 ASR system on the resynthetized output. MCD compares the cepstral representation of the inputs and outputs.

	sWUGGY ↑		sBLIMP ↑
	all	in-vocab	
GSLM (6k h)	-	68.7	57.1
AudioLM (60k h)	71.5	83.7	64.7
TWIST-7B (150k h)	74.6	84.4	62.1
Base L11	64.26	70.87	54.87
FT 100h L12	68.18	77.55	55.82
FT 100h L13	73.37	85.20	61.10
Gold phonemes	81.58	94.75	62.77
Init from centroids			
Base L11	64.78	71.56	54.83
FT 100h L12	68.85	78.66	56.17

Table 3: Zero-shot language comprehension scores (in %), for LMs with an embedding table either initialized randomly or from the unit centroids.

L12), the representations are of dimension 768, while for the fine-tuned 100h model's layer 13 (FT 100h L13) they have a dimension of only 40. Fine-tuning improves both triphone and phoneme ABX scores, particularly in reducing the context effect in the "any context" condition, as observed earlier. In the case of the ABX of one-hot representations, the error rates increase across all conditions, but the highest increase is when the context is not shared between the phones in the triplet. This is a sign that the k-means clusters not only are organized according to the phonemes but also to the surrounding context. Clusters are grouped according to their most probable phoneme, and within each group, clusters encode different contexts. By going from centroid representations to one-hot representations, all 500 clusters are now equidistant, which leads to the dramatic loss in "any context" compared to the more modest ones in the other two conditions.

3.2 Results above the phonemic level

We report in Table 3 the zero-shot sWUGGY (lexical level) and sBLIMP (syntactic level) scores for the base and fine-tuned models, as well as for an LSTM trained on the gold phonemes. Following the observation regarding the ABX error rates of the centroids, which remained within 1 percentage point of the standard continuous units, we train LSTMs by initializing their embedding table directly with the associated centroid representation of dimension 768. Apart from this change, the training process is the same between the two conditions. As can be seen, fine-tuning for phoneme classification improves spoken language modeling in terms of zero-shot comprehension evaluations. Overall, the gap between training from speech and training with golden phonemes is now halved. Fine-tuning for phoneme classification results in models that are on par in terms of lexical comprehension with much larger baselines, which were trained on orders of magnitude more of data.

However, Table 2 shows that this comes at the cost of the quality of resynthesis. Notably, there is a cost in content preservation, illustrated by the WER. It exists both for the LibriSpeech dataset and for the EXPRESSO-READ, while these two datasets correspond to the training domain of different components of our pipeline. Figure 1 makes directly visible the trade-off between language modeling and speech generation quality.

4 Conclusion

We showed that fine-tuning SSL representations with a phoneme classification task is an effective and simple procedure to improve context independence. This leads to improvements in the performance of LMs trained on these units. And we also found that initializing the embeddings of the discrete tokens of the LMs with the centroids of the units further helps with LM scores. This shows that the units found are meaningfully placed relative to one another in this representation space. Our work also highlights the trade-off between language modeling (which works best with abstract units), and speech generation (which works best with specific units). Fine-tuning on phoneme classification can adjust this trade-off.

5 Limitations

Further work is needed to improve on the tradeoff, perhaps by combining SSL, resynthesis, and fine-tuning objectives concurrently. More comprehensive studies could explore the role of the encoder in the spoken language modeling pipeline by examining the impact of fine-tuning methods on downstream language modeling, comparing various SSL and supervised speech models. Another important direction to consider is the application of this method in a multilingual setting. The benefits of fine-tuning are visible after training on as little as a few hours of aligned data, making it applicable to low resource languages.

References

Robin Algayres, Yossi Adi, Tu Nguyen, Jade Copet, Gabriel Synnaeve, Benoît Sagot, and Emmanuel Dupoux. 2023. Generative spoken language model based on continuous word-sized audio tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3008–3028, Singapore. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Heng-Jui Chang and James Glass. 2023. R-Spin: Efficient Speaker and Noise-invariant Representation Learning with Acoustic Pieces. *Preprint*, arxiv:2311.09117.

Heng-Jui Chang, Alexander H. Liu, and James Glass. 2023. Self-supervised Fine-tuning for Improved Content Representations by Speaker-invariant Clustering. In *Proc. INTERSPEECH 2023*, pages 2983–2987.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore. Association for Computational Linguistics.

Ewan Dunbar, Nicolas Hamilakis, and Emmanuel Dupoux. 2022. Self-supervised language learning from raw audio: Lessons from the zero resource speech challenge. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1211–1226.

Itai Gat, Felix Kreuk, Tu Anh Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. 2023. Augmentation invariant discrete representation for generative spoken language modeling. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 465–477, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar. 2023. Evaluating context-invariance in unsupervised speech representations. In *Proc. INTERSPEECH 2023*, pages 2973–2977.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis CONNEAU, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 63483–63501. Curran Associates, Inc.

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

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

R. Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022. Textless speech-to-speech translation on real data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 860–872, Seattle, United States. Association for Computational Linguistics.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-Weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330.

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

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *Preprint*, arxiv:2011.11588.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023a. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Proc. INTERSPEECH 2023*, pages 4823–4827.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023b. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. SpiRit-LM: Interleaved Spoken and Written Language Model. *Preprint*, arxiv:2402.05755.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*, pages 3615–3619.

Kaizhi Qian, Yang Zhang, Heting Gao, Junrui Ni, Cheng-I Lai, David Cox, Mark Hasegawa-Johnson, and Shiyu Chang. 2022. ContentVec: An improved self-supervised speech representation by disentangling speakers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18003–18017. PMLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Thomas Schatz. 2016. *ABX-Discriminability Measures and Applications*. Theses, Université Paris 6 (UPMC).

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair ABX task: analysis of the classical MFC/PLP pipeline. In *Proc. Interspeech 2013*, pages 1781–1785.

Jasmin Sternkopf and Stefan Taubert. 2024. mel-cepstral-distance.

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. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

S.J. Young, J.J. Odell, and P.C. Woodland. 1994. Tree-based state tying for high accuracy modelling. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

A Appendix

A.1 Fine Tuning Results

Table 5 presents the frame-level accuracy and Phone Error Rate (PER) for models fine-tuned on increasing labeled data quantity. The PER was computed by deduplicating consecutive predictions, without using a Language Model. For reference, the HuBERT base in SUPERB (wen Yang et al., 2021), trained with the CTC objective and with a frozen backbone, has a PER of 5.41% on `test-clean`.

	dev-clean	dev-other	test-clean	test-other
	Frame Classification Accuracy ↑			
FT 10min	88.80	83.78	88.80	84.29
FT 1h	91.36	87.35	91.24	87.66
FT 10h	93.01	89.03	92.96	89.31
FT 100h	94.36	90.36	94.28	90.75
	Phone Error Rate ↓			
FT 10min	8.45	15.82	8.87	15.30
FT 1h	4.68	9.59	5.15	9.25
FT 10h	3.64	8.70	4.02	8.38
FT 100h	2.83	7.53	3.15	7.07

Table 5: Fine Tuning Results (in %)

A.2 Discrete units quality

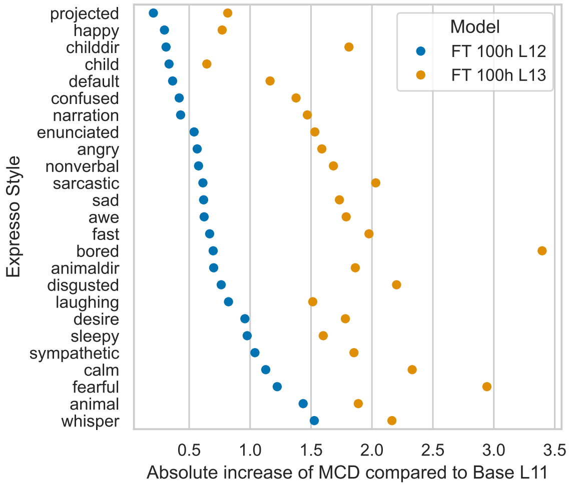

Figure 3: Difference between the MCD of the fine-tuned models and Base L11 on EXPRESSO for each style.

In addition to the ABX scores reported in Section 3.1, the quality of the discrete units and their relationship to phonemes can also be assessed with the three metrics proposed in Hsu et al. (2021): Cluster Purity, Phone Purity, and PNMI. Cluster purity is the conditional probability of a k-means label given a phone label, phone purity is the conditional probability of a phone label given a k-means label, and PNMI is the phone-normalized mutual information between units and phone labels. The units are obtained from the cluster assignments given by the k-means with 500 clusters trained on the output of the considered model. The evaluation is done on the combination of LibriSpeech `dev-clean` and `dev-other`. We have for the Base L11 and FT 100h L12 models: a PNMI of 0.669 and 0.846, Cluster Purity of 0.093 and 0.131, and Phone Purity of 0.685 and 0.858, respectively.

A.3 Resynthesis evaluation with another ASR system

We report Table 4 the Word Error Rate for resynthesis on the evaluation datasets using Whisper large-v3 (Radford et al., 2023) instead of wav2vec 2.0 as the ASR system. The differences between models are consistent with those in Table 2.

A.4 Resynthesis quality by expressive style

The drop in resynthesis quality by going from the standard model to the fine-tuned ones is further detailed is Figure 3. For each expressive style in EXPRESSO, the fine-tuned models exhibit a higher MCD compared to Base L11. The difference is the most prominent for styles capturing more non-verbal vocalizations such as “whisper” or “bored”.

	dev-clean	dev-other	test-clean	test-other	EXPRESSO-READ
Original audio	2.07	3.76	2.03	3.91	3.33
Base L11	3.84	11.61	4.03	11.38	6.58
FT 100h L12	4.24	10.97	4.34	10.67	7.95
FT 100h L13	5.72	11.76	5.68	11.84	9.68

Table 4: WER using Whisper large-v3 (in %)