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

Maxime Poli[1] Emmanuel Chemla[1]
Emmanuel Dupoux[1,2] [1]ENS - PSL, EHESS, CNRS [2]Meta FAIR [maxime.poli@ens.psl.eu

Abstract

Spoken language models trained on discrete speech units have recently emerged as an alternative to traditional automatic speech recognition systems. However, the units learned through self-supervised learning are not guaranteed to align with linguistically meaningful categories such as phonemes. In this work, we propose a simple fine-tuning approach to improve the linguistic quality of discrete units by adding a phoneme classification objective to a pretrained speech encoder. We show that this additional supervision improves phoneme discriminability while preserving the overall structure of the learned representation. When used for spoken language modeling, the improved units yield better performance both at the phonemic level and above, including in speech resynthesis. Our results suggest that incorporating minimal phonemic supervision is an effective way to enhance discrete speech representations for downstream language modeling tasks.

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

IMAGE NOT PROVIDED

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

IMAGE NOT PROVIDED

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

trained not only on discretized SSL representations but also on continuous word-size tokens (Algayres et al., 2023) or on 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), with scaling laws up to three orders of magnitude slower (Cuervo and Marxer, 2024). 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).

One hypothesis for the data inefficiency of spoken language models is that they must at the same time perform language modeling and process irrelevant acoustic variations. 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, 2024). 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) than with 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 on 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 a LM that outperform the standard approach.

We then evaluate whether the fine-tuned representations have retained their expressive power by measuring the distortion when resynthesizing expressive speech. We release the code and models at https://github.com/bootphon/spokenlm-phoneme.

2 Method

2.1 Phoneme classification

We start from a pretrained self-supervised speech model and fine-tune it on a phoneme classification task. Concretely, we add a linear projection layer on top of the encoder representations and train the model to predict frame-level phoneme labels using a cross-entropy loss. During fine-tuning, we update the parameters of the encoder jointly with the classifier.

Phoneme labels are obtained from a forced alignment using a standard ASR system. We experiment with different amounts of labeled data, ranging from a few hours to the full training set, in order to evaluate the impact of supervision quantity on the learned representations.

2.2 Quantization

After fine-tuning, we discretize the continuous representations into a sequence of discrete units. We use a k-means clustering algorithm applied to the encoder outputs to obtain a fixed inventory of units. The cluster centroids are learned on a subset of the training data and then used to quantize the full dataset.

2.3 Language modeling

The resulting discrete sequences are used to train a spoken language model. We train a Transformer-based autoregressive model to predict the next unit given the previous ones. The language model is trained using a standard cross-entropy objective over the discrete vocabulary.

2.4 Speech resynthesis

To evaluate the expressive quality of the learned units, we resynthesize speech from the discrete sequences. We train a neural vocoder conditioned on the discrete units to reconstruct the waveform. The quality of the resynthesized speech is assessed using automatic metrics.

2.5 Evaluation metrics

We evaluate the phonemic discriminability of the representations using the ABX error rate. For language modeling, we report perplexity on held-out data. For speech resynthesis, we measure the distortion between the original and reconstructed waveforms using standard objective metrics.

3 Results

3.1 Results at the phonemic level

We first evaluate the impact of phoneme fine-tuning on the quality of the learned representations at the phonemic level. We report ABX error rates averaged across subsets (dev-clean, dev-other) and speaker conditions (within, across). Fine-tuned models consistently outperform the original self-supervised representations, with substantial reductions in ABX error, even when trained with limited labeled data.

The improvements are observed across all experimental settings, indicating that the representations become more invariant to contextual variation while preserving phonemic distinctions. Increasing the amount of labeled data further enhances performance, although gains tend to saturate beyond a certain point.

3.2 Results above the phonemic level

We then assess the impact of improved units on spoken language modeling. Language models trained on units derived from fine-tuned representations achieve lower perplexity compared to those trained on standard self-supervised units. This improvement holds across different vocabulary sizes and training conditions.

Finally, we evaluate speech resynthesis quality. While fine-tuning improves phonemic discriminability and language modeling performance, we observe a trade-off between linguistic abstraction and expressive fidelity. Resynthesized speech from fine-tuned units may exhibit slightly higher distortion compared to the baseline, especially for highly expressive styles. Nevertheless, the degradation remains moderate, suggesting that phoneme supervision enhances linguistic content without severely compromising expressive information.

4 Conclusion

In this work, we proposed a simple fine-tuning approach to improve discrete speech representations for spoken language modeling. By adding a phoneme classification objective to a pretrained self-supervised encoder, we encouraged the representations to become more invariant to contextual variation while preserving phonemic distinctions. Our experiments showed consistent improvements in phoneme discriminability and downstream language modeling performance, even with limited labeled data. Although a slight trade-off was observed in expressive speech resynthesis, the overall gains suggest that minimal phonemic supervision is an effective way to enhance the linguistic quality of discrete speech units for spoken language modeling tasks.

5 Limitations

Our approach relies on supervised phoneme labels obtained from forced alignment, which may introduce biases from the ASR system used for annotation. The quality of the fine-tuned representations therefore depends on the accuracy of these alignments. In addition, while we demonstrate improvements in phonemic discriminability and language modeling performance, the evaluation is conducted on a limited set of datasets and languages. Further experiments would be necessary to assess the generalization of the method across diverse linguistic settings.

Another limitation concerns the trade-off between linguistic abstraction and expressive fidelity. Although phoneme supervision improves context invariance, it may reduce the amount of paralinguistic information retained in the representations, which can affect speech resynthesis quality in highly expressive conditions. Exploring alternative objectives that better balance these aspects remains an avenue for future work.

Acknowledgments

We thank the members of the Laboratoire de Sciences Cognitives et Psycholinguistique for helpful discussions and feedback. This work was supported by [ILLEGIBLE].

References

References

[1] Algayres, R., et al. (2023). [ILLEGIBLE].

[2] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems.*

[3] Borsos, Z., et al. (2023). [ILLEGIBLE].

[4] Chang, Y.-N., et al. (2023). [ILLEGIBLE].

[5] Chang, Y.-N., and Glass, J. (2024). [ILLEGIBLE].

[6] Chen, S., et al. (2022). [ILLEGIBLE].

[7] Chou, C.-F., et al. (2023). [ILLEGIBLE].

[8] Cuervo, S., and Marxer, R. (2024). [ILLEGIBLE].

[9] Dunbar, E., et al. (2022). [ILLEGIBLE].

[10] Gat, I., et al. (2023). [ILLEGIBLE].

[11] Hallap, A., et al. (2023). [ILLEGIBLE].

[12] Hassid, M., et al. (2023). [ILLEGIBLE].

[13] Hsu, W.-N., et al. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing.*

[14] Kharitonov, E., et al. (2022). [ILLEGIBLE].

[15] Lakhotia, K., et al. (2021). Generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics.*

[16] Lee, A., et al. (2022). [ILLEGIBLE].

[17] Maiti, S., et al. (2024). [ILLEGIBLE].

[18] Mohamed, A., et al. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing.*

[19] Nguyen, T., et al. (2020). [ILLEGIBLE].

[20] Nguyen, T., et al. (2023b). [ILLEGIBLE].

[21] Nguyen, T., et al. (2024). [ILLEGIBLE].

[22] Polyak, A., et al. (2021). [ILLEGIBLE].

[23] Qian, Y., et al. (2022). [ILLEGIBLE].

[24] wen Yang, S., et al. (2021). [ILLEGIBLE].

[25] Young, S., et al. (1994). The HTK book. *Cambridge University Engineering Department.*

A Appendix

A.1 Fine-tuning results

We report additional results for phoneme classification fine-tuning under different data regimes. Across all configurations, fine-tuned models show consistent improvements in ABX error rates compared to the original self-supervised representations. Gains are observed even with a limited number of labeled hours, and performance increases gradually with more supervision.

A.2 Discrete units quality

We further analyze the quality of the discrete units obtained after quantization. Units derived from fine-tuned representations exhibit better alignment with phonemic categories and reduced contextual variability. Clustering quality is evaluated using intrinsic metrics and phoneme classification accuracy on held-out data.

A.3 Resynthesis evaluation with another ASR system

To assess robustness, we evaluate resynthesized speech using a different automatic speech recognition system. Results confirm the main findings: fine-tuned representations improve linguistic accuracy while maintaining acceptable reconstruction quality.

A.4 Resynthesis quality by expressive style

We also analyze resynthesis quality across different expressive styles. While slight increases in distortion are observed for highly expressive speech, the overall degradation remains moderate, suggesting that phoneme supervision does not severely compromise expressive information.