On the Nature of Discrete Speech Representations in Multilingual Self-supervised Models

Badr M. Abdullah **Mohammed Maqsood Shaik** **Dietrich Klakow**
Language Science and Technology (LST), Saarland University, Germany
{ babdullah, mmshaik, dietrich }@lsv.uni-saarland.de

1 Overview and Research Question

Self-supervision has emerged as an effective paradigm for learning representations of spoken language from raw audio without explicit labels or transcriptions. Self-supervised speech models, such as wav2vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), have shown significant promise in improving the performance across different speech processing tasks. One of the main advantages of self-supervised speech models is that they can be pre-trained on a large sample of languages (Conneau et al., 2020; Babu et al., 2022), which facilitates cross-lingual transfer for low-resource languages (San et al., 2021).

State-of-the-art self-supervised speech models include a quantization module that transforms the continuous acoustic input into a sequence of discrete units. One of the key questions in this area is whether the discrete representations learned via self-supervision are language-specific or language-universal. In other words, we ask: *do the discrete units learned by a multilingual speech model represent the same speech sounds across languages or do they differ based on the specific language being spoken?* From the practical perspective, this question has important implications for the development of speech models that can generalize across languages, particularly for low-resource languages. Furthermore, examining the level of linguistic abstraction in speech models that lack symbolic supervision is also relevant to the field of human language acquisition (Dupoux, 2018).

2 Approach

To answer our research question, we conduct a series of experiments with spoken language identification (SLID) as a probing task. Our intuition is that if we can accurately predict the language of a short speech sample (∼10 sec) from its discretized representation, this would suggest that the model has learned language-specific discrete units that are unique to each language. On the other hand, a difficulty in predicting the language would suggest that the model has learned a common set of discrete units that are shared across multiple languages.

Experimental Data. We use a balanced subset of the Common Voice speech corpus (Ardila et al., 2020) consisting of 16 languages that span diverse sub-groups within the Indo-European language family, namely: Romance (Catalan, Portuguese, French, Spanish, Italian), Germanic (German, Dutch, Swedish, Frisian), Slavic (Ukrainian, Russian, Polish), Celtic (Welsh, Breton), Hellenic (Greek), and Indo-Iranian (Persian). Our language sample exhibits a considerable degree of typological diversity with respect to various phonological features, including the Consonant-Vowel Ratio, which is high in Russian but low in German, French, and Swedish (Maddieson, 2013). In addition, stress location patterns are highly variable in Russian and Spanish, but fixed in languages such as Greek, Persian, and Welsh (Goedemans and van der Hulst, 2013). We use ∼6.75, ∼3.75, ∼4.25 hours per each language for training, validation, and evaluation sets, respectively. A speech sample in our study is an utterance of a few seconds of read speech.

SLID Classifiers. For the set of languages in our study, we obtain discrete presentations from two pre-trained speech models: (1) monolingual English wav2vec 2.0 (W2V2), and multilingual model XLSR-53 (XLSR) (Conneau et al., 2020). We use the English W2V2 model to establish a comparison with a model that did not observe the languages in our study during pre-training.

Baseline. We use the majority class as a baseline, which corresponds to chance performance since our training and evaluation dataset are balanced.

Discrete Classifiers. Next, we train three different SLID classifiers on the discretized representations

Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023), pages 159–161
May 6, 2023 ©2023 Association for Computational Linguistics

of utterances in our study from both W2V2 and XLSR: (1) a Naive Bayes (NB) classifier, and (2) a linear classifier based on multi-class logistic regression (LC-D), and (3) a unidirectional LSTM (LSTM-D). NB and LC-D discard the sequential nature of representations and view each speech sample as a bag of discrete units. With the LSTM-D classifier, we can examine how much we gain by incorporating sequential information when decoding the language ID from the discrete sequence.

Continuous Classifiers. To investigate the effect of the discretization step on the extractability of language ID information from the model representations, we need to compare to SLID classifiers trained on continuous representations. To this end, we train linear classifiers on the representations from all transformer layers (after applying mean pooling). In this abstract we focus on classifiers trained on the output of the local convolutional encoder (LC-C0) and the contextualized transformer layer that yielded highest accuracy in both model (LC-CX). We also train a unidirectional LSTM on the sequence of contextualized vectors (LSTM-CX), identical to those used to train LC-CX.

Skyline. Finally, we fine-tune the pre-trained models to predict language ID to establish a reasonable upper-bound of the performance on the SLID task.

3 Preliminary Results

Activated Discrete Units. First, we find that the set of activated units are nearly identical across the languages in our study, which implies that the models do not learn units that are predictable features of the identity of the spoken language.

SLID Experiments. Table 1 shows the results of our SLID experiments. We observe that the non-sequential classifiers trained on discrete units (NB and LC-D) yield only modest improvements over the majority class baseline. This indicates that the languages in our study exhibit similar distributions over the discrete units. We do not observe considerable differences between the monolingual W2V2 and multilingual XLSR models in this case. However, W2V2 surprisingly outperforms XLSR for the sequential discrete classifier (LSTM-D), which indicates either that the monolingual model is more successful at approximating the languages' phonotactics or that the multilingual model projects the audio frames onto a shared discrete space where language identity is more difficult to extract compared to the monolingual model.

	Classifier	Accuracy (%)	
		W2V2	XLSR
Baseline	Majority class	6.25	6.25
Discrete	Naive Bayes	11.84	13.28
	LC-D	13.89	12.78
	LSTM-D	**39.78**	32.10
Continuous	LC-C0	22.00	22.57
	LC-CX	47.04	59.54
	LSTM-CX	58.70	**59.80**
Skyline	Fine-tuned	*54.96*	*59.72*

Table 1: The performance of spoken language identification using different classifiers.

Discrete vs. Continuous Classifiers. If we consider the performance of the continuous classifiers, we observe a higher accuracy compared to their discrete counterparts. This demonstrates the ease of extraction for the language ID information from the continuous representations. Moreover, we find that sequential models (e.g., LSTMs) trained on the representations from a middle layers to be successful in predicting the language ID compared to lower and higher layers in the transformer, which indicates the language ID information emerges as a product of the contextualization in the transformer block. This is evident in our results since the linear classifier on middle layer representations (LSTM-CX) in XLSR performs as good as the skyline fine-tuning setting. It is worth pointing out that XLSR has observed the languages in our study during pre-training, which can explain the high accuracy in predicting the language via a linear classifier from continuous representations in the middle layers.

4 Conclusion

We summarized the findings of our experiments whereby we investigate the nature of the discrete units in multilingual, self-supervised speech models. We employed language identification as a probing task and demonstrated the difficulty of predicting the language of an utterance from its discretized representation. Our findings support the hypothesis that latent, discretized speech representations in self-supervised models correspond to sub-phonetic events that are shared across the world's languages, rather than language-specific, abstract phonemic categories.

Acknowledgement

This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project ID 232722074 – SFB 1102.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Rob Goedemans and Harry van der Hulst. 2013. Fixed stress locations (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

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.

Ian Maddieson. 2013. Consonant-vowel ratio (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth, and Dan Jurafsky. 2021. Leveraging pretrained representations to improve access to untranscribed speech from endangered languages. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*, pages 1094–1101. IEEE.