MauBERT: Universal Phonetic Inductive Biases for Few-Shot Acoustic Units Discovery

Angelo Ortiz Tandazo[◇†] **Manel Khentout**[◇] **Youssef Benchekroun**[§]
Thomas Hueber[†*] **Emmanuel Dupoux**[◇§*]

[◇]ENS, PSL Research University, EHESS, CNRS, Paris, France
[†]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, Grenoble, France
[§]Meta AI Research, France
angelo.ortiz.tandazo@ens.psl.eu

Abstract

This paper introduces MauBERT, a multilingual extension of HuBERT that leverages articulatory features for robust cross-lingual phonetic representation learning. We continue HuBERT pre-training with supervision based on a phonetic-to-articulatory feature mapping in 55 languages. Our models learn from multilingual data to predict articulatory features or phones, resulting in language-independent representations that capture multilingual phonetic properties. Through comprehensive ABX discriminability testing, we show MauBERT models produce more context-invariant representations than state-of-the-art multilingual self-supervised learning models. Additionally, the models effectively adapt to unseen languages and casual speech with minimal self-supervised fine-tuning (10 hours of speech). This establishes an effective approach for instilling linguistic inductive biases in self-supervised speech models.

1 Introduction

Is it possible to automatically discover the linguistic units of an unknown language from raw audio only? Doing so would be of great help to linguists or speech technologists working on low-resource or unwritten languages (Chen et al., 2024a; Mohamed et al., 2022; Żelasko et al., 2022; Chen et al., 2023; Zhang et al., 2021), or to cognitive modellers trying to understand how children learn their native language before learning to read and write (Kuhl, 1993; Werker et al., 2007). This question has been addressed using a variety of approaches under the Zero Resource Speech Challenge series (Versteegh et al., 2015; Dunbar et al., 2017, 2022), yielding impressive progress alongside unresolved questions.

Much of this progress stems from advances in self-supervised learning (SSL) techniques (Oord et al., 2019; Baevski et al., 2020; Hsu et al., 2021), which have produced speech representations that capture phonetic structure better than traditional features like MFCCs or mel filterbanks. This is evidenced by improved discriminability in the learnt representation spaces: two instances of the syllable 'bit' lie closer together than one instance of 'bit' and one instance of 'bet', even across different speakers (Schatz, 2016; Schatz et al., 2013). Further evidence comes from the success of quantisation of these representations, yielding low-bitrate discrete codes suitable for training generative language models that produce novel utterances in the target language (Lakhotia et al., 2021; Borsos et al., 2023; Défossez et al., 2024; Rouard et al., 2025).

However, current approaches face two limitations. First, units discovered through speech SSL do not correspond one-to-one with linguistic units like phones, syllables or words. After clustering, these units are typically shorter and more numerous than standard linguistic units: 20–40 ms long vs. 70 ms for phonemes, and $N = 100$–1000 vs. 30–80 for phonemes (Lavechin et al., 2025; Schatz et al., 2021). Moreover, they lack full invariance to speaker identity (de Seyssel et al., 2022; Mohamed et al., 2024) and phonetic context (Hallap et al., 2023), suggesting they capture acoustic events rather than abstract linguistic units. As a result, they produce codes with higher bitrates than phonemic transcriptions: about 100–150 bit/s versus 50–70 bit/s (Lakhotia et al., 2021; Dunbar et al., 2022). Second, current SSL algorithms require massive amounts of clean speech: Hsu et al. (2021) uses 960 hours of clean English audio, Zanon Boito et al. (2024) uses 90 k hours, and Chen et al. (2024b) uses 1 M hours. Such quantities are unavailable for low-resource languages, and notably, children acquire their language's phonetics with far less than 1000 h of much noisier input.

One avenue for improving SSL models involves pre-training universal models (Conneau et al., 2021), with recent work expanding both language

[*] Equally contributed as senior authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 568–585

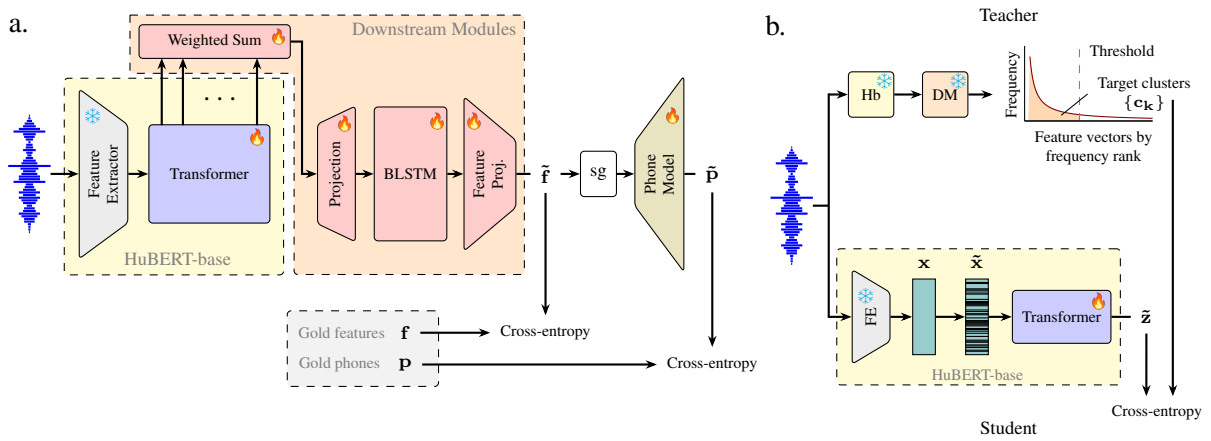

Figure 1: **a. Multilingual training.** MAUBERT-feat is trained to recognise ternary-valued articulatory features and phones using an encoder (HuBERT-base), downstream modules (weighted sum, up-projection, two-layer BLSTM, feature projection), and a phone model (two-layer perceptron); the feature states receive no gradients from the phone recognition loss due to the stop-gradient operator (sg). **b. Self-supervised fine-tuning.** *Top:* Offline clustering is applied to one of the layers of MAUBERT, the teacher network, on an unseen language; *bottom:* the MAUBERT Transformer, the student network, is then trained to predict the corresponding clusters of masked input.

coverage and training data (Babu et al., 2022; Zanon Boito et al., 2024; Pratap et al., 2024; Chen et al., 2024b). Inspired by the International Phonetic Alphabet (IPA), another research direction explores how phonetically-informed target signals influence learnt representations and their cross-lingual transferability (Wang et al., 2022; Ma et al., 2023; Feng et al., 2023), suggesting that explicit phonological supervision enhances speech models' cross-lingual capabilities.

In this paper, we explore the hypothesis that standard SSL algorithms lack **strong inductive biases** necessary for learning invariant speech representations from limited audio data in new languages. Following the universal pre-training and phonetically-informed research lines, we propose transforming a monolingual pre-trained SSL model (specifically, HuBERT-base trained on English) into a universal SSL model with strong inductive biases by fine-tuning it on **universal IPA phonemes and features** across 55 diverse languages. By directly addressing the limitations of prior work, such as the need for large datasets and the absence of explicit phonological supervision, our method provides a practical and linguistically informed solution for both speech technology and language documentation in low-resource settings. We evaluate this model, coined MAUBERT, on the ZRC2017 challenge, which presents 5 languages with less than 10 h of training data (English, French, German, Mandarin, Wolof). To increase the evaluation's diversity and validity, we extend the ZRC2017 benchmark with 5 typologically diverse languages (Swahili, Tamil, Thai, Turkish, Ukrainian). Evaluation employs the within- and across-speaker ABX metrics from ZRC2017, supplemented with metrics measuring invariance to contextual allophony (Hallap et al., 2023).

Our main contributions are twofold: (i) We demonstrate that multilingual supervised fine-tuning of HuBERT for articulatory feature or phone prediction creates robust multilingual phonetic representations with strong zero-shot transfer capabilities. (ii) Our resulting models enable effective adaptation to unseen languages and casual speech with minimal self-supervised fine-tuning, achieving strong speaker and contextual invariance in new languages with only 10 h of unlabelled data. As a by-product, our method also yields candidate phoneme and feature sets for unseen languages, with potential applications for linguistic analyses of low-resource languages. The code with the data processing, methods and baselines can be found at https://github.com/bootphon/s3prl.

2 Related Work

Multilingual Speech Representation Learning. The field of multilingual speech processing has grown rapidly with large-scale semi- or self-supervised learning models that showed the potential for cross-lingual representation learning with little to no supervision (Wang et al., 2021; Conneau et al., 2021). Recent studies have expanded language coverage (Babu et al., 2022), diversified

data sources (Pratap et al., 2024), and improved efficiency (Zanon Boito et al., 2024) and robustness to noise (Chen et al., 2024b). These multilingual SSL models build upon foundational work in self-supervised speech representation learning (Baevski et al., 2020; Hsu et al., 2021; Chen et al., 2022) and have been evaluated with multi-task frameworks like SUPERB (Yang et al., 2021; Shi et al., 2023). Meanwhile, other studies have explored the impact of phonetically-informed targets on learnt representations and their cross-lingual transferability (Wang et al., 2022; Ma et al., 2023; Feng et al., 2023). Inspired by the downstream framework of SUPERB, this work extends HuBERT-base for articulatory feature prediction.

Articulatory Features in Speech Processing. Early work established the use of articulatory features (AFs) in speech processing (Deng and Erler, 1991; Elenius and Takacs, 1991; Eide et al., 1993), demonstrated that supervised learning can be used to automatically extract phonological features from raw (and continuous) speech (Papcun et al., 1992; King and Taylor, 2000) and produced robust articulatory/phonological feature-based speech technologies (Kirchhoff, 1999; Livescu et al., 2007; Frankel et al., 2007). The development of systematic feature inventories, particularly PanPhon (Mortensen et al., 2016), has provided practical computational tools for cross-linguistic analysis. This has enabled recent efforts to explore AFs in multilingual contexts, demonstrating their effectiveness for zero-shot multilingual speech synthesis (Staib et al., 2020) or showing their utility for cross-lingual speech recognition in low-resource languages (Feng et al., 2023).

Evaluation of Speech Representations. More specific subtasks have been developed as alternatives to downstream-based evaluation, offering clearer insights into unsupervised language learning. A prominent example is the ABX discriminability evaluation (Schatz, 2016), which assesses whether learnt representations can distinguish between different phonetic units in a way that reflects human perceptual boundaries. The Zero Resource Speech Challenge series (Versteegh et al., 2015; Dunbar et al., 2017) has systematically applied ABX evaluation to assess unsupervised speech representations, establishing benchmarks for phonetic discrimination across diverse languages and speakers. While ABX testing shows sufficient correlation with downstream performance to serve as a model comparison proxy, traditional ABX evaluation has not assessed other types of invariance, like speaking rate or speech style variations (Dunbar et al., 2022). A recent extension has begun addressing this limitation by measuring context invariance (Hallap et al., 2023). The present work builds on this extension and adds the comparison between read and casual speech.

3 MAUBERT

In this section, we introduce our **M**ultilingual **a**rticulatory hidden-**u**nit **BERT** (MAUBERT) models (Figure 1). We describe the base architecture for multilingual training (§3.1), and the self-supervised fine-tuning approach (§3.2).

3.1 Multilingual Pre-Training

MAUBERT models are based on multilingual, continual learning of a pre-trained self-supervised speech model for articulatory feature (AF) or phone recognition (Figure 1a). We re-train HuBERT (Hsu et al., 2021) using the VoxCommunis Corpus (Ahn and Chodroff, 2022), and the associated featural annotations extracted with PanPhon (Mortensen et al., 2016).

We propose two versions of MAUBERT: FEAT and PHONE. The former incorporates an AF bottleneck (Figure 1a), while the latter directly predicts phones without intermediate AFs[1].

Encoder. We use the pre-trained HuBERT-base model as our *encoder*. The convolutional feature extractor is kept frozen, but the Transformer encoder is trainable. We extract the feature extractor's output after layer normalisation and dropout, as well as the outputs from each of the 12 Transformer encoder layers. The input masking is disabled during continual pre-training following SUPERB's downstream framework (Yang et al., 2021).

Downstream modules. Preliminary experiments revealed that a simple linear layer on top of HuBERT was insufficient for the feature recognition task. Following the SUPERB framework for ASR (Yang et al., 2021), we instead leverage a weighted sum of HuBERT's intermediate representations fed into a BLSTM. Ablation experiments confirmed the importance of this contextual architecture: replacing the BLSTM with non-contextual networks

[1]The feature projection in Figure 1a is replaced with a phone projection, and the phone model is dropped.

Eval. lang.	MAUBERT variant	Feat. acc. ↑	Phone acc. ↑	PER ↓
Train	FEAT	**95.60**	72.28	30.64
	PHONE	92.72	**82.72**	**28.69**
Dev	FEAT	**92.35**	51.20	50.46
	PHONE	88.57	**67.15**	**48.38**

Table 1: Feature and phone evaluation of MAUBERT on the held-out test set of the 55 training languages and zero-shot performance on the 5 development languages. All scores are in %.

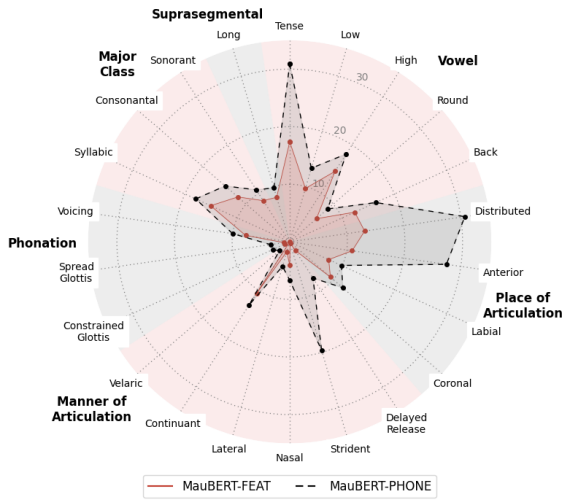

Figure 2: Classification errors across articulatory features (↓) for both MAUBERT variants on the 5 development languages. All values are in %.

consistently degraded performance on both the multilingual recognition tasks and the phonetic probing, motivating our adoption of the SUPERB design.

Concretely, we first compute a weighted sum of the intermediate representations from our encoder and up-project them to a 1024-dimensional space. These representations are then processed through a bidirectional two-layer LSTM. Finally, we down-project the concatenated forward and backward output states into task-specific spaces: a 22-dimensional AF space for MAUBERT-FEAT and a 3293-dimensional phone space for MAUBERT-PHONE.

Phone model. Given the non-injective nature of the feature-to-phone mapping, for MAUBERT-FEAT, we jointly learn a phone model consisting of a two-layer perceptron. Since we want the pre-training to be led by the feature recognition task only, a stop gradient operator prevents the feature hidden states from receiving any gradients from the phone recognition loss.

Model	# Params	# Langs	# Hours	Seen dev	Seen test
MMS	965 M	1406	491 k	5	5
XEUS	577 M	4057	1 M	5	5
mHuBERT-147	95 M	147	90 k	5	4
HuBERT-base	95 M	1	960	0	1
MAUBERT (ours)	141–144 M	55	788	0	1[2]

Table 2: Comparison of speech models by number of parameters, number of languages, training data size, and development and test languages seen during training (continual learning for MAUBERT).

3.2 Self-Supervised Fine-Tuning

We employ self-supervised fine-tuning to adapt MAUBERT models to unseen languages with limited or no labelled data. This approach generates pseudo-labels through clustering of learnt representations and applies masked language modelling (Figure 1b), enabling MAUBERT to adapt to the acoustic patterns of new languages.

We use four methods to generate pseudo-labels: K-means, frequent features, frequent phones and all phones. As Hsu et al. (2021), we apply K-means clustering with $K = 100$ to representations from encoder Transformer layers (HuBERT-base, MAUBERT) or downstream module layers (MAUBERT variants). For MAUBERT-FEAT, we extract the top K most frequent feature vectors (*feat. freq.*) from the articulatory feature space (top of Figure 1b). For both MAUBERT variants, we extract the top K most frequent phones (*phone freq.*) or all phones from pre-training data (*all phones*). See Appendix C for details.

4 MAUBERT Multilingual Pre-Training

This section describes the multilingual training and evaluation of MAUBERT variants for articulatory feature and phone recognition.

4.1 Data Processing

We use the VoxCommunis Corpus, which provides phone-level annotations for a subset of Common Voice (Ardila et al., 2020) obtained with Montreal Forced Aligner (McAuliffe et al., 2017). Of the 63 covered languages, 55 are used for supervised articulatory feature prediction (totalling 788.4 h hours), 5 serve as development languages, and 3 are discarded as they were test languages. (Refer

[2]The backbone of our models being HuBERT-base, some English influence might remain in our models' weights.

to §5.3 for the development and test languages and to Appendix A for more data processing details.)

Using PanPhon's feature table[3], ternary-feature[4] annotations are derived from the phone-level annotations. Annotated segments incompatible with PanPhon are manually fixed (*e.g.* [tʃ] → [t͡ʃ], [bʱ] → [b̤], [g] → [ɡ]). Finally, we collapse the IPA table by keeping only distinct feature *vectors* (*e.g.* [æ̈], [e˕], [ɐ˕], [a̟] and [æ] are all represented by the same feature vector), which reduces the table size from 6367 to 3293 segment representatives. These representatives are then used for both phone recognition and feature recognition (underlying feature values).

4.2 Training Details

We train MAUBERT variants for feature or phone recognition across the 55 languages drawn from the VoxCommunis Corpus. Due to PanPhon's ternary feature representation, we exclude MAUBERT-FEAT predictions that correspond to zero-valued target features. Furthermore, to handle *multiphthongs* (hypernym of diphthong), we use a uniform heuristic so that the duration of the resulting *monophthongs* is roughly the same.

The FEAT and PHONE variants are trained to minimise binary and multiclass cross-entropy losses, respectively, at the frame level with the Adam optimiser (Kingma and Ba, 2015). We use one V100 GPU for 40 k steps with a tri-stage learning rate schedule (4 k for warmup and 16 k for decay) that peaks at 5×10^{-5}. Following Conneau et al. (2021), we employ a language up-sampling strategy to balance the amount of data between low-resource and high-resource languages. (See Appendix A for more details.)

4.3 Evaluation and Results

We evaluate our MAUBERT models using three speech recognition metrics: frame-wise feature accuracy, frame-wise phone accuracy and phone error rate (PER) without deduplication heuristics. For feature accuracy, we compute scores over non-zero features only, excluding zero-valued target features as in training. Since MAUBERT-PHONE lacks an explicit feature space, we extract feature vectors from predicted phones using PanPhon's feature table.

Table 1 shows results for both MAUBERT variants on held-out test sets from the 55 training languages and the 5 development languages. Both variants exhibit superior performance on training languages, particularly for phone-level metrics. When transitioning from training to development languages, phone accuracy drops by 15 % to 21 % and PER increases by approximately 20 %, while feature accuracy shows more resilience with only 3–4 % degradation. The FEAT variant consistently outperforms the PHONE variant in articulatory feature prediction across all languages (see Figure 2). However, this advantage does not translate to improved phone recognition performance, and the PHONE variant exhibits an even greater phone prediction advantage on development languages compared to training languages. Note that the PER gap in favour of the PHONE variant is stable across languages.

5 Few-shot Language Adaptation

In this section, we assess the linguistic relevance of MAUBERT's learnt representations by evaluating their phonetic invariance across languages and speaking styles, in a zero-shot or few-shot setting.

5.1 Language Adaptation Setting

Modes. We compare how SSL models encode speech in a new language in three modes: zero-shot, supervised fine-tuning and self-supervised fine-tuning. All the baselines and our two MAUBERT models are evaluated in zero-shot mode, while only the monolingual baseline and our two models are evaluated in the fine-tuning modes (on the 10 h training split) for fairness[5]. In the supervised mode, the models are trained to predict the ground-truth phones of masked inputs (MPR) or without masking at all (PR). In the self-supervised mode, a clustering step first produces discrete pseudo-labels, which are later used as targets for masked prediction.

Baselines. We compare MAUBERT against several baselines, including traditional acoustic features (MFCCs), the monolingual HuBERT-base backbone, and three self-supervised models trained on massively multilingual data: MMS-1B (Pratap

[3]We exclude PanPhon's two tonal features from the 24 AFs since VoxCommunis alignments lack tone segments.

[4]Features take '+', '-' or '0' values, with zero indicating context-dependent values (*e.g.* `high` for [ɾ]) or irrelevance to the phone (*e.g.* `strident` for vowels).

[5]HuBERT-base and our MAUBERT models are trained on two to three orders of magnitude less data than the multilingual baselines.

Systems			Development languages							Test languages (ZRC2017)						
Model	Layer	# units	triphone ABX ↓		phoneme ABX ↓ within ctx		phoneme ABX ↓ any ctx		*avg.*	triphone ABX ↓ 1 s		triphone ABX ↓ 10 s		triphone ABX ↓ 120 s		*avg.*
			WS	AS	WS	AS	WS	AS		WS	AS	WS	AS	WS	AS	
Zero-shot																
MFCC	-	39	20.00	29.00	13.23	22.36	18.05	26.33	21.49	14.78	25.58	14.70	25.33	14.70	25.32	20.07
MMS-1B	34	1280	9.37	10.74	4.76	6.02	10.53	11.37	8.80	7.58	9.02	6.91	7.91	6.91	7.83	7.69
XEUS	18	1024	6.14	7.15	3.58	4.52	9.28	9.45	6.69	**4.67**	**5.68**	**4.19**	**4.91**	**4.29**	**4.99**	**4.79**
mHuBERT-147	7	768	7.37	8.64	3.70	4.80	9.00	9.51	7.17	6.93	8.13	5.75	6.49	6.67	7.78	6.96
HuBERT-base	11	768	6.77	8.18	3.77	4.92	8.55	9.19	6.90	6.21	7.42	5.31	6.21	5.62	6.62	6.23
MAUBERT																
FEAT	9	768	5.49	6.52	**2.95**	3.81	5.97	6.47	5.20	5.86	6.84	4.78	5.57	4.86	5.68	5.60
PHONE	proj	1024	**5.42**	**6.46**	2.96	**3.79**	**5.49**	**6.12**	**5.04**	5.36	6.44	4.68	5.58	4.68	5.60	5.39
supervised FT (10 h)																
HuBERT-base																
+ PR	ws	768	4.87	6.13	2.30	3.09	3.65	4.17	4.04	5.52	6.67	4.10	4.99	4.51	5.49	5.21
+ MPR	11	768	4.26	4.98	2.05	2.62	3.94	4.30	3.69	4.26	4.84	3.25	3.73	3.89	4.36	4.05
MAUBERT																
FEAT + MPR	11	768	3.65	**4.38**	1.83	**2.28**	3.17	3.44	3.13	**3.81**	**4.26**	2.86	3.25	3.28	3.71	3.53
PHONE + MPR	12	768	**3.58**	4.49	**1.79**	2.30	**2.88**	**3.35**	**3.07**	3.92	4.61	**2.57**	**3.08**	**2.86**	**3.32**	**3.39**
self-supervised FT (10 h)																
HuBERT-base																
+ K-means (L11)	10	768	5.71	6.64	3.15	4.09	7.13	7.58	5.72	5.65	6.38	4.79	5.40	5.09	5.77	5.51
MAUBERT-FEAT																
+ K-means (L9)	10	768	**4.72**	**5.50**	2.58	3.31	5.08	5.59	4.46	5.01	**5.56**	4.19	4.71	4.38	5.00	4.81
+ K-means (feat)	9	768	5.00	5.81	2.69	3.41	5.29	5.69	4.65	5.16	5.92	4.30	4.98	4.51	5.20	5.01
+ feat. freq.	9	768	4.88	5.65	2.63	3.28	5.24	5.66	4.56	4.99	5.80	4.19	4.86	4.39	5.09	4.89
+ phone freq.	9	768	5.01	5.90	2.62	3.35	5.21	5.62	4.62	5.09	5.87	4.31	5.01	4.53	5.24	5.01
MAUBERT-PHONE																
+ K-means (proj)	10	768	4.91	5.71	2.66	3.32	4.93	5.55	4.51	**4.84**	5.62	4.17	4.81	4.38	5.15	4.83
+ K-means (phone)	10	768	4.88	5.83	2.70	3.40	5.29	5.79	4.65	5.52	6.16	4.14	4.76	4.28	4.86	4.95
+ phone freq.	10	768	4.77	5.78	**2.49**	3.17	**4.82**	**5.26**	**4.38**	5.11	5.79	4.09	4.72	4.24	4.86	**4.80**
+ all phones	10	768	4.88	5.84	**2.49**	**3.16**	4.85	5.28	4.42	5.15	5.89	**4.05**	**4.70**	**4.20**	**4.83**	**4.80**

Table 3: Acoustic discriminability scores (lower is better) over 5 development languages (sw, ta, th, tr, uk) and, as test languages, the 5 languages from the Zero Resource Speech Challenge 2017 (en, fr, zh, de, wo). The best layer for each model is selected based on the average ABX score on the development languages. The best scores are in **bold** and the second best are underlined.

et al., 2024), mHuBERT-147 (Zanon Boito et al., 2024), and XEUS (Chen et al., 2024b). Table 2 shows a brief comparison of the training data between the baselines and our models.

Implementation. For the supervised fine-tuning, we train the models for 20 k steps on one V100 GPU with a tri-stage learning rate schedule (2 k for warmup and 8 k for decay). We use the Adam optimiser with a peak learning rate at 1×10^{-4}. For the self-supervised fine-tuning, we train the Transformer encoder for 50 k steps on one H100 GPU. We use the Adam optimiser with a linear decay schedule (8 % for warmup, then linear decay back to zero) that peaks at 5×10^{-6}.

5.2 Metric

We employ the ABX discriminability test to measure phonetic invariance (Schatz, 2016). It evaluates speech representations by comparing distances between three triphones: *A*, *X* (same linguistic unit as *A*), and *B* (different unit). The test is considered successful when the distance between *A* and *X* is smaller than that between *A* and *B*. The test comprises two variants: a triphone-based version that examines complete triphone representations, and a phoneme-based version that focuses exclusively on central phone representations.

The speaker condition varies between two scenarios: *within*-speaker (all triphones share the speaker) and *across*-speaker (only *A* and *B* share the speaker). In addition, contextual conditions across all three items (*A*, *B*, and *X*) can be manipulated: *within*-context (where all items share identical surrounding phonetic context) versus *any*-context (where surrounding contexts may differ).

We compute all the ABX scores with the CPU backend of `fastabx` (Poli et al., 2025a).

5.3 Language Data

Following the Zero Resource Speech Challenge 2017 (Dunbar et al., 2017), we curate ABX-ready datasets for five *development languages* from the VoxCommunis Corpus: Swahili, Tamil, Thai, Turkish and Ukrainian. The ABX datasets consist of three splits for each language: a 10 h training set, a validation set and a test set. We select the best parameters, hyperparameters and layers of the various

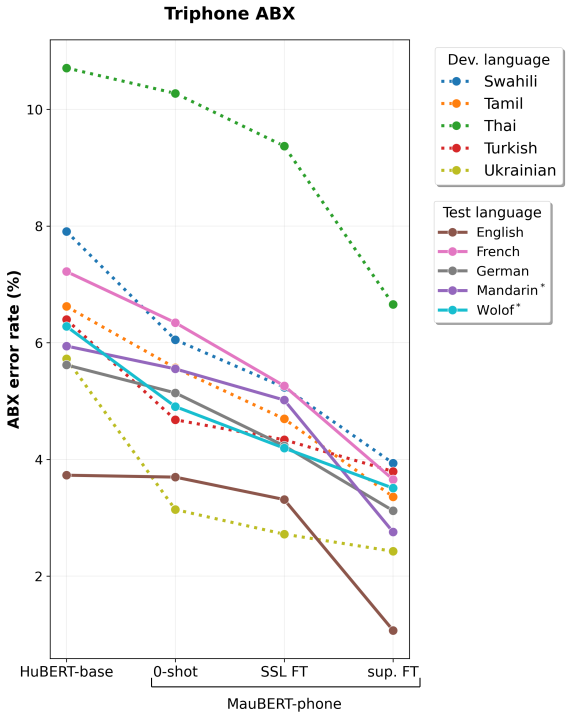

Figure 3: Reduction of the triphone ABX error rates across the 5 development languages and 5 test languages between the base HuBERT model, and MAUBERT, tested in zero-shot and after masked fine-tuning (10 h) with or without labels on a new language. The two speaker conditions are averaged, and the 10 s subset is chosen for the test languages. *Mandarin and Wolof only have 1.5 and 1.8 h of training data, resp.

models according to their impact on the average ABX score (triphone-based ABX, within-context phoneme ABX and any-context phoneme ABX) on the ABX test sets.

We use both the development and surprise languages from the aforementioned Zero Resource Speech Challenge 2017, namely English, French, Mandarin, German and Wolof, as *test languages* (hereafter referred to as ZRC2017). The amount of speech in the original training set ranges from 2.3 h for Mandarin to 35.3 h for English. We thus extract training and validation splits of up to 10 h. In line with the desiderata of the challenge, we keep the original test subsets of differing length (1 s, 10 s and 120 s) to evaluate the effect of context length (triphone-based ABX only). We evaluate only the best configuration for each model on these languages.

Additionally, we curate an ABX dataset of casual speech in English and French, sourcing high-quality recorded conversations of native speakers. The dataset possesses the same three-split structure as the development languages.

Systems	Read		Casual	
	WS	AS	WS	AS
MMS	6.75	8.99	13.47	17.47
XEUS	4.46	5.78	8.69	11.29
mHuBERT-147	5.29	6.83	10.62	13.72
HuBERT-base	4.71	6.24	9.41	12.48
ours				
MAUBERT-FEAT	4.45	5.75	9.43	12.11
+ K-means (L9)	3.95	5.00	**8.25**	**10.66**
MAUBERT-PHONE	4.29	5.75	9.23	11.92
+ phone freq.	**3.69**	**4.89**	8.43	10.85

Table 4: Triphone-based ABX error rates across registers (read vs. spontaneous) for English and French in zero-shot mode. Our two MAUBERT variants are also tested after self-supervised fine-tuning on 10 h.

5.4 Few-shot Language Adaptation Results

Our experimental results demonstrate the competitive performance of our MAUBERT models across multiple evaluation scenarios.

Multilingual Training Benefits. The top of Table 3 illustrates the phonetic invariance performance in zero-shot mode achieved by MAUBERT models through multilingual pre-training. Our models attain particularly strong results in the any-context phoneme-based ABX tasks, and the MAUBERT-PHONE model delivers the best overall zero-shot performance (5.22 % against 5.74 % for XEUS). Further, Figure 3 confirms the performance improvements of the multilingual pre-training as well as the proposed self-supervised fine-tuning: 6.62 % for the HuBERT-base baseline vs. 5.54 % for MAUBERT-PHONE in zero-shot and 4.84 % after *phone freq.* MPR in triphone ABX with similar audio lengths.

Cross-linguistic Performance Patterns. Table 3 also shows that development languages present greater challenges than test languages when evaluated under comparable conditions (triphone ABX with similar audio lengths) across models and modes. This pattern indicates varying degrees of phonetic complexity across language families and suggests that our model selection strategy (detailed in §5.3) based on development languages' performance provides a robust foundation for cross-lingual generalisation. Figure 3 reinforces this observation, with development languages (shown with dotted lines) generally exhibiting higher error rates and more variable performance across the

training progression compared to test languages (solid lines), suggesting the former present more challenging phonetic discrimination tasks and may represent more diverse or complex phonological systems.

Supervised Fine-tuning Efficacy. Supervised fine-tuning yields substantial improvements in ABX error rates. Particularly striking is the effectiveness of predicting the ground-truth phones of masked inputs (MPR), which reduces ABX error rates compared to standard phone prediction (PR), especially for triphone-based ABX. The MAUBERT-PHONE + MPR configuration achieves the best supervised performance (3.07 % on development languages, 3.39 % on test languages), representing a significant 38 % relative improvement over the zero-shot baseline. Figure 3 illustrates this systematic improvement pattern across all languages, with supervised fine-tuning showing the most notable gains (3.43 % average ABX score). Remarkably, fine-tuning effectiveness appears largely independent of training data quantity: low-resource language Wolof achieves comparable error rates to high-resource languages, indicating robust few-shot adaptation capabilities.

Self-supervised Fine-tuning Analysis. While self-supervised fine-tuning approaches show consistent improvements over zero-shot performance, a performance gap remains compared to the fully-supervised standard. Among the clustering strategies, our phone frequency-based approach demonstrates some gains over standard K-means clustering, particularly excelling in phoneme-level discrimination tasks and longer temporal contexts (10 s and 120 s triphone ABX). MAUBERT-PHONE with phone frequency clustering achieves the best self-supervised performance (4.59 % average ABX score), highlighting the value of linguistically-informed clustering strategies.

Speech Register Adaptation Results. Table 4 reveals nuanced domain-specific patterns across read versus casual speech. In zero-shot mode, our models perform slightly better than multilingual baselines on read speech (MAUBERT-PHONE: 5.02 % vs. XEUS: 5.12 %) but show reversed performance on casual speech (10.58 % vs. 9.99 % for XEUS), reflecting the inherent difficulty of spontaneous speech processing with its increased phonetic variability and reduced articulatory precision. However, self-supervised fine-tuning not only amplifies our advantage on read speech (4.29 %) but also recovers competitive performance on casual speech (9.64 %), demonstrating the robustness of our adaptation approach across speech domains.

5.5 Phonetic Inventory Discovery Results

Two of the MAUBERT SSL methods consist in assigning a feature or a phoneme set to a new language as a target for SSL fine-tuning. These methods amount to discovering the *phonetic* inventories of previously unseen languages[6]. Following Żelasko et al. (2022), we leverage the frequency distribution of (discrete) articulatory feature vectors produced by MAUBERT-FEAT, where high-frequency combinations likely correspond to actual phones in the language inventory[7].

Table 7 reveals a clear trade-off between precision and recall across different threshold strategies. The top-100 approach achieves consistently high recall (at least 0.825 for four out of five languages), successfully capturing most phonemes in the target inventories. However, this comes at the cost of precision (0.270–0.390), indicating substantial inclusion of spurious feature vectors. Conversely, the optimised frequency threshold approach significantly improves precision (0.778–0.872) while maintaining reasonable recall (0.532–0.810), suggesting more accurate phonetic identification with fewer false positives.

The superior F_1 performance of optimised thresholds over fixed thresholds underscores the importance of adaptive, data-driven approaches to inventory discovery. (See Table 8 for some inventory examples with F_1-optimal thresholds.)

6 Discussion

Broader Impact. Our demonstration that effective phonetic models can be developed for low-resource languages with minimal training data (as evidenced by Wolof performance with less than 2 h of data) is an encouraging signal towards more linguistic inclusion in computational models. In addition, our frequency-based methodology offers particular value for endangered language documentation, where traditional phonological analysis may

[6] MAUBERT-FEAT can only predict monophthongs due to the splitting of *multiphthongs* during training.

[7] The inventory consists of all the phones observed in VoxCommunis. Most phones appear in the 'CV dictionaries' on https://mfa-models.readthedocs.io/en/latest/dictionary/index.html.

be impractical, providing linguists with not only a multilingual articulatory feature recogniser but also an automated tool for initial phonetic hypothesis generation that can guide subsequent detailed analysis. However, the superior performance of high-resource languages like English also highlights the importance of linguistic diversity in training data, since the imbalance thereof could persist through evaluation.

Future Work. Several promising research directions emerge from our findings. The counter-intuitive relationship between training data quantity and fine-tuning effectiveness suggests that investigation into optimal data selection strategies could yield significant improvements, potentially focusing on phonetically diverse rather than simply large datasets. Given that speaker and content information are encoded at different layers of HuBERT's encoder, selectively targeting a subset of hidden layers during multilingual pre-training is another promising avenue worth exploring. The domain adaptation capabilities demonstrated in our casual speech experiments indicate potential for developing more robust models through multi-domain training paradigms. Furthermore, extending the self-supervised fine-tuning beyond the encoder to encompass the entire MAUBERT architecture could address current limitations by enabling end-to-end adaptation of both the pre-trained representations and the downstream articulatory feature prediction modules, potentially leading to improved performance on target languages and domains, and better phonetic inventory discovery.

7 Conclusion

This work presents MAUBERT, a multilingual extension of HuBERT that demonstrates competitive phonetic discrimination capabilities across diverse languages while revealing important insights about cross-lingual representation learning. Our results establish that multilingual supervised pre-training creates robust phonetic foundations that enable effective few-shot adaptation to new languages (10 hours of speech) with or without supervision. The demonstrated effectiveness on both read and spontaneous speech, coupled with strong performance on low-resource languages, positions this work as a significant step towards more inclusive multilingual speech technologies.

Limitations

The evaluation is constrained to the ABX discrimination task, which, while established as a standard phonetic benchmark, may not fully capture the nuanced linguistic representations as required for other linguistic levels (*e.g.* syntax and semantics). The performance gap between self-supervised and supervised fine-tuning methods suggests that clustering- or frequency-based approaches, despite their linguistic motivation, remain suboptimal compared to gold-standard supervision.

Acknowledgments

We thank the reviewers for their valuable feedback, which helped improve the clarity and quality of the paper. This work was granted access to the HPC resources of IDRIS under the allocation 2024-AD011014739R1 made by GENCI, and was supported in part by Agence Nationale de Recherche (ANR-17-EURE-0017 FrontCog, ANR-10-IDEX-0001-02 PSL and ANR-23-IACL-0006 France 2030). ED in his EHESS role and MK were funded by an ERC grant (InfantSimulator). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 5286–5294, Marseille, France. European Language Resources Association.

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 Twelfth Language Resources and Evaluation Conference, pages 4218–4222, Marseille, France. European Language Resources Association.

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, pages 2278–2282.

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.

Can Balioglu, Alexander Erben, Martin Gleize, Artyom Kozhevnikov, Ilia Kulikov, and Julien Yao. 2023. fairseq2.

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.

Chih-Chen Chen, William Chen, Rodolfo Joel Zevallos, and John E Ortega. 2024a. Evaluating self-supervised speech representations for indigenous American languages. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6444–6450.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In Findings of the Association for Computational Linguistics: ACL 2023, pages 4969–4983, Toronto, Canada. Association for Computational Linguistics.

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.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024b. Towards robust speech representation learning for thousands of languages. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In Interspeech 2021, pages 2426–2430.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In Interspeech 2022, pages 1402–1406.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: A speech-text foundation model for real-time dialogue. Preprint, arXiv:2410.00037. Version 2.

Li Deng and Kevin Erler. 1991. Microstructural speech units and their hmm representation for discrete utterance speech recognition. In [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, pages 193–196 vol.1.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 323–330.

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.

Ellen Eide, Jan Robin Rohlicek, Herbert Gish, and Sanjoy Mitter. 1993. A linguistic feature representation of the speech waveform. In 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, pages 483–486 vol.2.

Kjell Elenius and Gy Takacs. 1991. Phoneme recognition with an artificial neural network. In 2nd European Conference on Speech Communication and Technology (Eurospeech 1991), pages 121–124.

Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, and Yuxuan Wang. 2023. Language-universal phonetic representation in multilingual speech pretraining for low-resource speech recognition. In Interspeech 2023, pages 1384–1388.

Joe Frankel, Mathew Magimai-Doss, Simon King, Karen Livescu, and Özgür Çetin. 2007. Articulatory feature classifiers trained on 2000 hours of telephone speech. In 8th Annual Conference of the International Speech Communication Association, INTERSPEECH 2007, Antwerp, Belgium, August 27-31, 2007, pages 2485–2488. ISCA.

Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar. 2023. Evaluating context-invariance in unsupervised speech representations. In 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, and 1 others. 2023. Textually pretrained speech language models. Advances in Neural Information Processing Systems, 36:63483–63501.

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.

Simon King and Paul Taylor. 2000. Detection of phonological features in continuous speech using neural networks. Computer Speech & Language, 14(4):333–353.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Katrin Kirchhoff. 1999. Robust Speech Recognition Using Articulatory Information. PhD dissertation, University of Bielefeld.

Patricia K. Kuhl. 1993. Innate Predispositions and the Effects of Experience in Speech Perception: The Native Language Magnet Theory, pages 259–274. Springer Netherlands, Dordrecht.

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.

Marvin Lavechin, Maureen de Seyssel, Hadrien Titeux, Guillaume Wisniewski, Hervé Bredin, Alejandrina Cristia, and Emmanuel Dupoux. 2025. Simulating early phonetic and word learning without linguistic categories. Developmental Science, 28(2):e13606.

Alexander H. Liu, Heng-Jui Chang, Michael Auli, Wei-Ning Hsu, and Jim Glass. 2023. DinoSR: Self-Distillation and Online Clustering for Self-supervised Speech Representation Learning. Advances in Neural Information Processing Systems, 36:58346–58362.

Karen Livescu, Ozgur Cetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Haggerty, Bronwyn Woods, Joe Frankel, Mathew Magimai-Doss, and Kate Saenko. 2007. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, volume 4, pages IV–621–IV–624.

Ziyang Ma, Zhisheng Zheng, Guanrou Yang, Yu Wang, Chao Zhang, and Xie Chen. 2023. Pushing the limits of unsupervised unit discovery for SSL speech representation. In Interspeech 2023, pages 1269–1273.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In Proc. Interspeech 2017, pages 498–502.

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.

Mukhtar Mohamed, Oli Danyi Liu, Hao Tang, and Sharon Goldwater. 2024. Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations. In Interspeech 2024, pages 3625–3629.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. Preprint, arXiv:1807.03748. Version 2.

George Papcun, Judith Hochberg, Timothy R. Thomas, François Laroche, Jeff Zacks, and Simon Levy. 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. The Journal of the Acoustical Society of America, 92(2):688–700.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(85):2825–2830.

Maxime Poli, Emmanuel Chemla, and Emmanuel Dupoux. 2025a. fastabx: A library for efficient computation of abx discriminability. Preprint, arXiv:2505.02692.

Maxime Poli, Manel Khentout, Angelo Ortiz Tandazo, Ewan Dunbar, Emmanuel Chemla, and Emmanuel Dupoux. 2026. Discophon: Benchmarking the unsupervised discovery of phoneme inventories with discrete speech units. Preprint, arXiv:2603.18612.

Maxime Poli, Mahi Luthra, Youssef Benchekroun, Yosuke Higuchi, Martin Gleize, Jiayi Shen, Robin Algayres, Yu-An Chung, Mido Assran, Juan Pino, and Emmanuel Dupoux. 2025b. Spidr: Learning fast and stable linguistic units for spoken language models without supervision. Transactions on Machine Learning Research.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25(97):1–52.

Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. 2025. Continuous audio language models. Preprint, arXiv:2509.06926. Version 2.

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

Thomas Schatz, Naomi H. Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. Proceedings of the National Academy of Sciences, 118(7):e2001844118.

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 Interspeech 2013, pages 1781–1785.

Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Shinji Watanabe. 2023. ML-SUPERB: Multilingual speech universal performance benchmark. In Interspeech 2023, pages 884–888.

Marlene Staib, Tian Huey Teh, Alexandra Torresquintero, Devang S. Ram Mohan, Lorenzo Foglianti, Raphael Lenain, and Jiameng Gao. 2020. Phonological features for 0-shot multilingual speech synthesis. In Interspeech 2020, pages 2942–2946.

Maarten Versteegh, Roland Thiollière, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In Interspeech 2015, pages 3169–3173.

Chengyi Wang, Yiming Wang, Yu Wu, Sanyuan Chen, Jinyu Li, Shujie Liu, and Furu Wei. 2022. Supervision-guided codebooks for masked prediction in speech pre-training. In Interspeech 2022, pages 2643–2647.

Chengyi Wang, Yu Wu, Yao Qian, Kenichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. 2021. Unispeech: Unified speech representation learning with labeled and unlabeled data. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 10937–10947. PMLR.

Janet F. Werker, Ferran Pons, Christiane Dietrich, Sachiyo Kajikawa, Laurel Fais, and Shigeaki Amano. 2007. Infant-directed speech supports phonetic category learning in English and Japanese. Cognition, 103(1):147–162.

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 Interspeech 2021, pages 1194–1198.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A compact multilingual hubert model. In Interspeech 2024, pages 3939–3943.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2021. Uwspeech: Speech to speech translation for unwritten languages. Proceedings of the AAAI Conference on Artificial Intelligence, 35(16):14319–14327.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu

Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. Preprint, arXiv:2205.01068.

Piotr Żelasko, Siyuan Feng, Laureano Moro Velázquez, Ali Abavisani, Saurabhchand Bhati, Odette Scharenborg, Mark Hasegawa-Johnson, and Najim Dehak. 2022. Discovering phonetic inventories with crosslingual automatic speech recognition. Computer Speech & Language, 74:101358.

A Data

We use the annotations of Common Voice (Ardila et al., 2020) made by Ahn and Chodroff (2022). At the time of download (21 August 2024), the dataset consisted of 63 languages. Five languages (Swahili, Tamil, Thai, Turkish and Ukrainian) were held out for hyperparameter tuning, and three languages (French, Mandarin and Hong Kong Mandarin) were discarded because of their presence in the test set. Appendix A contains the list of the 55 languages kept for training.

Training languages. We filter out utterances containing spn segments, which indicate alignment errors from Montreal Forced Aligner (McAuliffe et al., 2017), or those that are excessively long (non-silent phones exceeding 0.5 s). We retain only utterances lasting between 2 and 20 s. We fix misplaced diacritics that were incorrectly attached to adjacent phones in four languages. We also handle IPA characters that PanPhon (Mortensen et al., 2016) does not recognise by mapping them to their proper equivalents (*e.g.* [g] becomes [g]). To prevent high-resource languages from dominating the training data for MAUBERT, we limit each language to a maximum of 50 h, yielding a total of 788.4 h of multilingual training speech.

Development languages. We apply the same preprocessing pipeline used for the training languages to the development languages. Next, we combine the three original Common Voice splits (train, dev, and test) and create three new splits with the following specifications for each language: (i) the test set contains 7.3–14.0 h of audio uniformly distributed across 20 speakers, (ii) the training set contains 8.3–9.5 h uniformly distributed across 10 speakers, and (iii) the validation set contains 8.5–9.7 h hours of audio. We ensure that speakers are completely disjoint across all newly created splits.

Test languages. We use the five languages from the Zero Resource Challenge 2017 (Dunbar et al., 2017): English, French, Mandarin, German, and Wolof. From the original long-form recordings (each corresponding to a different speaker), we extract[8] training and validation splits of up to 10 h per language through the following process: (i) apply voice activity detection using official challenge alignments, (ii) segment recordings into 2–20 s

[8]The challenge training set contains at least 21 h of speech, except for Mandarin and Wolof, which have 2.3 and 3.0 h of speech, respectively.

IETF code	Language	# Hours
ab	Abkhaz	22.4
am	Amharic	0.1
ba	Bashkir	49.7
be	Belarusian	50.0
bg	Bulgarian	6.2
bn	Bengali	30.5
ca	Catalan	50.0
ckb	Central kurdish	6.6
cs	Czech	24.9
cv	Chuvash	0.5
dv	Maldivian	2.5
el	Greek	2.1
eu	Basque	50.0
gn	Guarani	1.5
ha	Hausa	2.2
hi	Hindi	4.6
hsb	Upper sorbian	1.5
hu	Hungarian	49.4
hy-AM	Armenian	0.4
id	Indonesian	7.4
it	Italian	50.0
ja	Japanese	12.1
ka	Georgian	50.0
kk	Kazakh	0.0
kmr	Northern kurdish	4.8
ko	Korean	0.6
ky	Kyrgyz	2.2
lij	Ligurian	0.7
lt	Lithuanian	9.4
ml	Malayalam	1.4
mn	Mongolian	3.1
mr	Marathi	3.6
mt	Maltese	2.2
myv	Erzya	1.9
nan-tw	Taiwanese hokkien	2.0
pa-IN	Punjabi	1.1
nl	Dutch	40.4
pl	Polish	28.9
pt	Portuguese	23.8
ro	Romanian	3.9
ru	Russian	37.3
rw	Kinyarwanda	50.0
sk	Slovak	3.3
sl	Slovenian	1.3
sq	Albanian	0.1
sr	Serbian	1.4
sv-SE	Swedish	8.2
tk	Turkmen	1.1
tt	Tatar	9.3
ug	Uyghur	15.2
ur	Urdu	0.1
uz	Uzbek	50.0
vi	Vietnamese	1.4
yo	Yoruba	1.9
yue	Cantonese	3.3
	Total	788.4

Table 5: List of 55 languages with their amount of speech included in the training set.

clips including silences of up to 1 s, and (iii) assign each speaker's clips exclusively to either training or validation splits to ensure speaker disjointness. This yields training sets of 1.5–10.0 h and validation sets of 0.7–10.0 h, with Mandarin having the smallest splits and European languages having the largest.

B Training

Table 6 lists the (hyper-) parameters used for multilingual feature recognition. All the (hyper-) parameters for self-supervised and supervised fine-tuning can be found in the released code.

Language Up-sampling. During multilingual pre-training, we draw from the multinomial distribution $p_l \sim (\frac{n_l}{N})^\alpha$, where n_l is the number of audios of language l, N is the training set size, and α is the up-sampling factor controlling the importance between high- and low-resource languages.

Length grouping. To reduce unused representations in batches, we split the multilingual data into buckets of audio of roughly the same length.

C Clustering methods

Ground-truth phones. We use the collapsed list of segments from PanPhon for the development languages, and the list of unique phonemes from the official alignments for the test languages.

K-means. We run the `MiniBatchKMeans` algorithm from `scikit-learn` (Pedregosa et al., 2011) on the training set for each development and test language. We select three different representations: (i) the best-performing layer from zero-shot mode, (ii) the feature logits for MAUBERT-FEAT, (iii) and the phone logits (after reducing to only the phones seen during training) for MAUBERT-phone.

Predicted phones. First, we determine the most likely phone label for each frame in the training set. We then prune the phone prediction layer by removing the output heads corresponding to phones absent from the training set (*all phones* labels). For the *phone freq.* labels, we further restrict the prediction layer to the K most frequent phones in the training set, discarding the remaining heads and fine-tuning the layer to produce highly confident predictions for these K phones.

Predicted features. For the *feat. freq.* labels, we extract predicted articulatory feature logits for each frame in the training set, apply a sigmoid activation, and hard-threshold the resulting values at 0.5 to obtain binary feature vectors. We then compute the frequency of each unique binary vector across the training set and retain the K most common ones. For pseudo-labelling, each frame is assigned to its nearest neighbour among these K frequent vectors under the ℓ_1 distance, with ties broken in favour of vectors with fewer zero-valued features.

Paramenters	Value
Model	
Up-projection dimension	1024
BLSTM layers	2
BLSTM dimension	1024
BLSTM dropout	0.2
BLSTM layer normalisation	No
Phone MLP hidden dimension	1024
Phone MLP activation function	GELU
Features	
Diphthong feature strategy	split
Zero values loss	ignore

Hyper-Paramenters	Value
Data	
Up-sample factor (α)	0.7
Batch size	32
Optimizer	
Name	Adam
Peak learning rate	5×10^{-5}
Betas	(0.9, 0.98)
Weight decay	No
Epsilon	1×10^{-8}
Warmup steps	4000
Hold steps	16 000
Decay steps	20 000
Mixed precision	fp16

Table 6: Model parameters and training hyper-parameters used for MAUBERT-FEAT.

D Inventory discovery supplement

Language	Inventory size	Top 100			F_1-optimal		
		Prec.	Recall	F_1	Prec.	Recall	F_1
Swahili	40	0.330	0.825	0.471	0.824	0.700	0.757
Tamil	35	0.300	0.857	0.444	0.815	0.629	0.710
Thai	42	0.390	0.929	0.549	0.872	0.810	0.840
Turkish	47	0.270	0.574	0.367	0.781	0.532	0.633
Ukrainian	38	0.350	0.921	0.507	0.778	0.737	0.757

Table 7: Precision, recall and F_1 score for the inventory discovery on the development languages for the top-100 threshold and the best threshold for F_1 score.

Lang.	Correctly predicted phones	Missing phones
th	m, i, k, j, u, a, p, w, n, t, l, s, b, ŋ, e, o, h, d, f, ɛ, ɔ, iː, aː, uː, ɾ, eː, kʰ, pʰ, tʰ, ɛː, ɔː, t͡ɕ, ɤ, ʉː	ʔ, oː, ɯ, t͡ɕʰ, ɯː, ɤː, ʉ, a̯
tr	m, i, k, j, u, a, p, b, e, o, g, h, f, t͡ʃ, ʃ, d͡ʒ, ɾ, t, n, d, ɯ, y, s, l, z	mː, kː, jː, pː, bː, gː, hː, fː, t͡ʃː, ʃː, d͡ʒː, v, vː, ɾː, tː, nː, ʒ, dː, sː, œ, lː, zː
uk	m, i, k, j, u, p, b, ɾ, ɛ, ʃ, t, x, t͡ʃ, n, ʒ, d, a, s, tʲ, lʲ, v, z, t͡s, sʲ, ɾʲ, l, nʲ, t͡sʲ	g, f, ɔ, d͡ʒ, ɪ, fi, d͡z, dʲ, zʲ, d͡zʲ

Table 8: Phonetic inventory prediction using an F_1-optimal threshold for Thai, Turkish and Ukrainian. The language inventories comprise all the phones observed in the alignments from VoxCommunis.

E Spoken language modelling evaluation

To complement the ABX phonetic probing results, we assess whether the discrete units produced by MAUBERT lead to better spoken language models (sLMs) through using these units as input features to a spoken LM trained in a larger dataset (6 k hours of Libri-Light).

We use our zero-shot models and SSL models fine-tuned on the English adaptation set (10 h) as frame-level encoders. A K-means model (with V clusters) is trained on the resulting embeddings to produce discrete tokens, which are then used to train an OPT-125M language model (Zhang et al., 2022) with fairseq2 (Balioglu et al., 2023), following the architectural choices of Hassid et al. (2023) and Poli et al. (2025b). Training uses the 6 k hours subset of Libri-Light (Kahn et al., 2020) on 8 GPUs, with a context length of 2048 tokens, batches of at most 81 920 tokens, for 25 k steps. The learning rate is set to 1×10^{-2} with a linear warmup of 1000 steps and cosine annealing; remaining hyper-parameters follow OPT-125M defaults. We choose the checkpoint with the lowest validation loss.

We compare MAUBERT against the monolingual SSL models: wav2vec 2.0 (Baevski et al., 2020), HuBERT-base (Hsu et al., 2021), WavLM-base (Chen et al., 2022) and DinoSR (Liu et al., 2023), all trained on 1 k hours of English. We evaluate sLMs on three tasks: sWUGGY (lexical), sBLIMP (syntactic), and tSC (semantic consistency), reporting scores as percentages. We consider two vocabulary sizes: $V = 256$ (Table 9) and $V = 100$ (Table 10). For Table 9, the MAUBERT-PHONE zero-shot (all phones) variant uses a one-hot encoding over all 370 observed phones in all our data; for Table 10, the MAUBERT-PHONE SSL fine-tuned (logits) variant uses a one-hot encoding over all 87 observed phones in the English adaptation data.

Model	Units	sWUGGY		sBLIMP	tSC
		all	in-vocab		
Monolingual models					
wav2vec 2.0†	K-means (L6)	62.29	68.50	53.34	65.97
WavLM-base	K-means (L11)	**70.07**	**80.36**	56.48	**71.15**
HuBERT-base	K-means (L11)	65.01	72.96	55.34	68.22
DinoSR†	codebook (L5)	60.10	64.56	57.05	69.44
Ours					
MAUBERT-FEAT					
zero-shot	K-means (L9)	63.24	69.78	54.71	68.11
zero-shot	feat. freq.	55.55	58.74	51.52	61.16
+ self-supervised FT	K-means (L10)	67.34	76.05	55.26	66.72
MAUBERT-PHONE					
zero-shot	K-means (proj)	61.67	66.87	53.32	65.60
zero-shot◇	one-hot (all phones)	57.91	62.22	52.20	63.30
+ self-supervised FT	K-means (L10)	67.74	76.46	**57.34**	67.84

† Fetched from Poli et al. (2025b).

Table 9: **Spoken language modelling scores in English ($V = 256$).** Monolingual self-supervised learning baselines on top. ◇All phones seen in the data are kept ($V = 370$). Best scores in **bold**, second best underlined.

Model	Units	sWUGGY		sBLIMP	tSC
		all	in-vocab		
MAUBERT-FEAT					
zero-shot	K-means (L9)	64.63	71.85	56.15	68.00
zero-shot	feat. freq.	56.00	59.31	51.46	60.10
+ self-supervised FT	K-means (L10)	67.08	75.36	55.89	67.74
+ self-supervised FT	one-hot (logits)	66.83	75.34	56.24	70.41
MAUBERT-PHONE					
zero-shot	K-means (proj)	63.63	70.66	55.11	67.09
zero-shot	phone freq.	58.49	63.20	52.88	65.01
+ self-supervised FT	K-means (L10)	**67.65**	**76.48**	**58.83**	**70.46**
+ self-supervised FT◇	one-hot (logits)	60.79	66.69	54.65	69.18

Table 10: **Spoken language modelling scores in English ($V = 100$).** ◇Only the phones seen during fine-tuning are used as vocabulary ($V = 87$). Best scores in **bold**, second best underlined.

Results. Tables 9 and 10 reveal a nuanced picture that does not always mirror the ABX probing results. In zero-shot mode, MAUBERT variants underperform the monolingual HuBERT-base and WavLM-base baselines on sWUGGY and sBLIMP, despite their superior phonetic discriminability in ABX. This gap is consistent across both vocabulary sizes and suggests that the richer cross-lingual phonetic structure encoded by MAUBERT does not translate straightforwardly into better lexical or syntactic priors for English-specific language modelling. After self-supervised fine-tuning, however, both MAUBERT variants recover competitive performance, approaching WavLM-base on sWUGGY (67–68 % vs. 70 %) and matching or exceeding HuBERT-base across all three tasks. This mirrors the pattern observed in ABX scores, where self-supervised fine-tuning substantially closes the gap with toplines. The MAUBERT-PHONE variant with K-means fine-tuning achieves the strongest sLM performance among our models at $V = 100$ (Table 10: 67.65 % sWUGGY, 76.48 % sBLIMP), consistent with its leading ABX scores in Table 3.

Taken together, these results suggest that while MAUBERT's multilingual phonetic inductive biases do not directly boost English sLM performance in zero-shot mode, the representations become competitive after even minimal self-supervised adaptation, paralleling the conclusions drawn from ABX probing. This conclusion is only tentative, because the base speech model was trained in English itself and is therefore difficult to disentangle the effect of the pre-training from our fine-tuning. Further studies should start with a base universal model trained on languages other than English, and/or tested on a larger set of held-out languages.

F DiscoPhon benchmark

DiscoPhon (Poli et al., 2026) is a benchmark designed to evaluate the ability of speech representation models to encode phonemic information on unlabelled data. It consists of 6 development languages (German, Swahili, Tamil, Thai, Turkish and Ukrainian) and 6 test languages (Mandarin, English, Basque, French, Japanese and Wolof), selected to span a diverse range of phonological categories. These language splits differ from those used in §5. While MAUBERT has not seen any of the development languages, it retains some English influence from the HuBERT backbone and some Japanese bias from the 12.1 h of Japanese in the training data.

We compare MAUBERT against the same probing baselines as in §5: MMS-1B, XEUS, mHuBERT-147, and HuBERT-base, using the best layers identified in Table 3. We report results under three conditions: (i) many-to-one with 256 units, where each unit resulting from the clustering process is mapped to the most probable phoneme (Table 11); (ii) one-to-one with ground-truth vocabulary size, where each phoneme is assigned a unique unit (Table 12); and (iii) one-to-one with fine-tuned models predicting the ground-truth vocabulary size, yielding a phoneme–unit bijection (Table 13). We evaluate using Phone Error Rate (PER), R-value, F_1, Phoneme-Normalised Mutual Information (PNMI), and triphone-based ABX discriminability on continuous representations.

Results. The DiscoPhon results largely corroborate and extend the ABX findings from §5. In zero-shot mode, both MAUBERT variants substan-

	dev languages					test languages				
	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	ABX c. ↓	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	ABX c. ↓
Zero-shot										
MMS-1B	111.50	13.08	60.12	59.16	11.63	112.94	9.70	61.16	62.00	9.69
XEUS	71.28	44.10	62.48	61.04	7.54	70.23	42.20	63.68	63.42	6.02
mHuBERT-147	70.16	47.87	66.51	63.96	9.16	71.21	43.88	68.01	67.14	7.33
HuBERT	88.20	37.05	65.20	60.80	8.74	83.51	36.99	66.67	65.02	6.72
MAUBERT-FEAT	**56.52**	**59.96**	72.75	**68.98**	7.18	**56.28**	**57.34**	74.28	**72.10**	5.54
MAUBERT-PHONE	59.43	58.09	**73.39**	68.59	**7.00**	60.84	54.85	**74.92**	71.60	**5.20**
Self-supervised FT (10 h)										
HuBERT + K-means (L11)	73.19	46.66	67.95	64.95	6.92	71.66	44.41	68.73	68.40	5.43
MAUBERT-FEAT										
+ K-means (L9)	49.07	64.25	74.56	70.87	5.91	47.43	62.84	76.06	74.04	4.60
+ K-means (feat)	42.47	70.78	75.47	71.44	6.15	40.64	68.81	77.37	74.96	4.62
+ feat. freq.	40.71	71.48	75.71	71.68	5.98	37.08	71.48	77.90	75.02	4.53
+ phone freq.	38.31	74.20	77.61	72.44	6.04	36.83	72.33	79.66	75.58	4.66
MAUBERT-PHONE										
+ K-means (proj)	50.98	62.98	74.84	70.93	6.10	50.33	61.01	76.54	73.90	4.53
+ K-means (phone)	44.44	70.66	76.90	71.32	5.95	38.40	72.19	79.62	75.28	4.43
+ phone freq.	**34.09**	**78.29**	**80.06**	**73.38**	5.93	32.70	76.78	82.05	**76.60**	**4.41**
+ all phones	35.22	77.69	79.92	72.99	**5.89**	**32.51**	**77.09**	**82.18**	76.48	**4.41**

Table 11: **DiscoPhon benchmark's many-to-one scores (256 units).** Fine-tuned models predict 100 or 370 clusters (MAUBERT-PHONE + all phones MPR predicts all the seen phones). Units are mapped to the most probable phoneme for evaluation. Target layer as in Table 3. Results (in %) averaged across dev. and test languages, resp. Triphone-based ABX averaged between within- and across-speaker conditions. K-means clustering: layer for pseudo-labels to be predicted between parentheses. Best scores in **bold**, second best underlined.

tially outperform all baselines across PER, R-value, F_1, and PNMI in the many-to-one condition (Table 11), with MAUBERT-FEAT and MAUBERT-PHONE reducing PER by at least 15 % relative to the best multilingual baseline on development (mHuBERT-147) on test languages (XEUS). This advantage is consistent with the phonetic invariance improvements observed in Table 3, and reinforces the conclusion that multilingual articulatory supervision yields genuinely better phoneme-level representations, not merely better triphone discriminability.

After self-supervised fine-tuning, MAUBERT variants further strengthen their lead. In the many-to-one condition (Table 11), MAUBERT-PHONE with phone frequency clustering achieves the best PER and R-value across both development and test languages (34.1 % PER, R-value 78.3 % on dev; 32.7 % PER, R-value 76.8 % on test), paralleling its top ABX performance. The feature frequency and phone frequency methods consistently outperform standard K-means, echoing the pattern from §5.4, where linguistically motivated clustering strategies showed advantages in longer temporal contexts.

The one-to-one condition (Tables 12 and 13) is considerably more demanding, as each phoneme must be assigned a dedicated unit. Here, the ranking among methods is preserved, but absolute scores deteriorate sharply for all models, highlighting the difficulty of achieving a clean phoneme–unit bijection without gold supervision. Notably, the K-means (phone) variant incurs a significant ABX cost in Table 13, suggesting that forcing a bijective mapping via K-means can hurt representational quality despite improving phoneme assignment metrics. Nonetheless, MAUBERT variants with phone- or feature-frequency-based fine-tuning remain the strongest systems (Table 12: MAUBERT-PHONE + phone freq. reaches 73.4 % PER, 58.2 % R-value on dev), and the ground-truth vocabulary fine-tuning condition (Table 13) shows that phone-frequency MPR yields the best trade-off between PER and R-value.

	dev languages				test languages			
	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	PER ↓	R-value ↑	F_1 ↑	PNMI ↑
Zero-shot								
MMS-1B	272.84	−106.13	42.91	42.64	270.69	−107.90	43.86	45.74
XEUS	206.75	−53.59	44.46	43.40	215.05	−63.63	45.14	46.21
mHuBERT-147	220.10	−59.96	46.59	46.31	213.57	−58.99	48.09	49.17
HuBERT	195.82	−40.63	50.33	47.68	199.22	−47.29	50.79	50.78
MAUBERT-FEAT	134.26	6.13	58.50	56.26	146.20	−8.15	57.34	58.75
MAUBERT-PHONE	**123.24**	**16.73**	**62.58**	**58.65**	**128.29**	**5.77**	**61.85**	**61.16**
Self-supervised FT (10 h)								
HuBERT + K-means (L11)	175.23	−24.01	53.38	51.25	179.07	−33.25	53.11	53.59
MAUBERT-FEAT								
+ K-means (L9)	129.80	8.84	59.44	57.31	144.22	−7.10	57.85	58.94
+ K-means (feat)	88.39	44.93	66.51	61.25	102.86	29.27	65.28	63.67
+ feat. freq.	86.80	47.02	67.45	61.42	100.97	32.16	65.82	63.70
+ phone freq.	91.70	42.05	66.21	62.64	103.51	26.46	65.38	64.96
MAUBERT-PHONE								
+ K-means (proj)	134.66	4.27	59.75	58.20	142.09	−7.41	58.63	60.06
+ K-means (phone)	85.52	49.94	68.82	62.33	90.40	42.32	69.27	65.16
+ phone freq.	**73.35**	**58.24**	**72.49**	**64.84**	88.76	39.20	69.57	**67.31**
+ all phones	74.99	57.78	72.25	64.12	**85.96**	**44.09**	**70.63**	66.89

Table 12: **DiscoPhon benchmark's one-to-one scores ($(|\mathcal{P}|+1)$ units).** Fine-tuned models predict 100 or 370 clusters (MAUBERT-PHONE + all phones MPR predicts all the seen phones). Each phoneme is mapped to a single unit. Target layer as in Table 3, best scores in **bold**, second best underlined.

	dev languages					test languages				
	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	ABX c. ↓	PER ↓	R-value ↑	F_1 ↑	PNMI ↑	ABX c. ↓
HuBERT + K-means (L11)	176.81	−24.97	52.84	48.96	9.18	182.38	−33.29	52.73	52.02	6.73
MAUBERT-FEAT										
+ K-means (L9)	121.14	16.94	60.96	58.28	7.02	133.60	1.74	59.37	60.45	5.26
+ K-means (feat)	104.38	40.92	62.63	47.65	11.16	110.65	31.75	64.03	53.30	8.08
+ feat. freq.	86.26	53.96	65.72	46.44	8.71	89.50	48.24	67.36	51.43	6.66
+ phone freq.	39.30	81.98	79.27	59.88	8.29	40.22	82.71	82.69	62.29	6.47
MAUBERT-PHONE										
+ K-means (proj)	106.68	29.59	65.64	**60.90**	**6.66**	114.07	17.68	64.34	63.34	**4.76**
+ K-means (phone)	101.66	52.07	64.56	48.87	17.67	103.08	46.18	67.10	55.38	11.41
+ phone freq.	**37.59**	**82.93**	**80.24**	60.83	9.20	**38.30**	**83.82**	**83.75**	**63.73**	6.85

Table 13: **DiscoPhon benchmark's one-to-one scores ($|\mathcal{P}|+1$ units).** Fine-tuned models predict the ground-truth vocabulary size, resulting in a phoneme-unit bijection. The target layer corresponds to the logits from the MPR task (on top of HuBERT's last layer). K-means clustering: layer for pseudo-labels to be predicted between parentheses. Best scores in **bold**, second best underlined.