From 'snippet-lects' to doculects and dialects: Leveraging neural representations of speech for placing audio signals in a language landscape

Séverine Guillaume, Guillaume Wisniewski, Alexis Michaud

From 'Snippet-lects' to Doculects and Dialects: Leveraging Neural Representations of Speech for Placing Audio Signals in a Language Landscape

Séverine Guillaume[1], *Guillaume Wisniewski*[2], *Alexis Michaud*[1]

[1] Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle – Institut National des Langues et Civilisations Orientales (INALCO)
[2] Université de Paris Cité, Laboratoire de Linguistique Formelle (LLF), CNRS, F-75013 Paris, France
severine.guillaume@cnrs.fr, guillaume.wisniewski@u-paris.fr, alexis.michaud@cnrs.fr

Abstract

XLSR-53, a multilingual model of speech, builds a vector representation from audio, which allows for a range of computational treatments. The experiments reported here use this neural representation to estimate the degree of closeness between audio files, ultimately aiming to extract relevant linguistic properties. We use max-pooling to aggregate the neural representations from a 'snippet-lect' (the speech in a 5-second audio snippet) to a 'doculect' (the speech in a given resource), then to dialects and languages. We use data from corpora of 11 dialects belonging to 5 less-studied languages. Similarity measurements between the 11 corpora bring out greatest closeness between those that are known to be dialects of the same language. The findings suggest that (i) dialect/language can emerge among the various parameters characterizing audio files and (ii) estimates of overall phonetic/phonological closeness can be obtained for a little-resourced or fully unknown language. The findings help shed light on the type of information captured by neural representations of speech and how it can be extracted from these representations.
Index Terms: pre-trained acoustic models, language documentation, under-resourced languages, similarity estimation

1. Introduction

The present research aims to contribute to a recent strand of research: exploring how pre-trained multilingual speech representation models like XLSR-53 [1] or HuBERT [2] can be used to assist in the linguistic analysis of a language [3]. XLSR-53, a multilingual model of speech, builds a vector representation from an audio signal. The neural representation is different in structure from that of the audio recording. Whereas wav (PCM) audio consists in a vector of values in the range [-1:+1], at a bit-depth from 8 to 32 and a sampling rate on the order of 16,000 Hz, the XLSR-53 neural representation contains 1,024 components, at a rate of 47 frames per second. The size of the vector representation is on the same order of magnitude as that of the audio snippet, and the amount of information can be hypothesized to be roughly comparable. But the neural representation, unlike the audio format, comes in a vector form that is tractable to a range of automatic treatments building on the vast body of work in data mining and machine learning. The neural representation of speech holds potential for an epistemological turning-point comparable to the introduction of the spectrogram 8 decades ago [4, 5, 6].

The experiments reported here use the neural representation yielded by XLSR-53 (used off-the-shelf, without fine-tuning, unlike [7, 8]) as a means to characterize audio: estimating the degree of closeness between audio signals, and (ultimately) extracting relevant linguistic properties, teasing them apart from other types of information, e.g. technical characteristics of the recordings. We start out from 5-second audio snippets, and we pool neural representations (carrying out *mean pooling*, i.e. averaging across frames) to progress towards the level of the entire audio file, then the entire corpus (containing several audio files). We thereby gradually broaden the scope of the neural representation from a 'snippet-lect' (the speech present in an audio snippet[1]) to a 'doculect' (a linguistic variety as it is documented in a given resource [9]), then towards 'dialects' (other groupings could also be used: by sociolect, by speaking style/genre, etc.) and, beyond, entire languages.

In a set of exploratory experiments, we build neural representations of corpora of 11 dialects that belong to 5 under-resourced languages. We then use linguistic probes [10] (i.e. a multiclass classifier taking as input the frozen neural representation of an utterance and assigning it to a language, similarly to a language identification system) to assess the capacity of XLSR-53 to capture language information. Building on these first results, we propose to use our probe on languages not present in the training set and to use its decisions as a measure of similarity between two languages, following the intuition that if an audio segment of an unknown language is identified as being of language A, then the language in the audio segment is "close" to A.

Representations like XLSR-53 have already been used to develop language identification systems (e.g. [11, 12]), but their use in the context of under-resourced languages and linguistic fieldwork datasets raises many challenges. First, there is much less data available for training and testing these systems both in terms of number of hours of audio and number of speakers. For instance, VoxLingua [13], a dataset collected to train language identification models, contains $6,628$ hours of recordings in 107 languages, i.e. at least an order of magnitude more data per language than typical linguistic fieldwork corpora. Second, the languages considered in a language documentation context have not been used for (pre-)training speech representations and have linguistic characteristics that are potentially very different from the languages used for pre-training them (on consequences of narrow typological scope for Natural Language Processing research, see [14]). The ability of models such as XLSR-53 to correctly represent these languages remains an open question. We also aim to assess to what extent pre-trained models of speech can address these two challenges.

Similarity measurements between the 11 corpora bring out greatest closeness between those that are known to be dialects of the same language. Our findings suggest that dialect/language can emerge among the many parameters characterizing audio files as captured in XLSR-53 representations (which also include acoustic properties of the environment, technical characteristics of the recording equipment, speaker ID, speaker gender, age,

[1] 'Snippet-lect' is coined on the analogy of 'doculect' [9], to refer to the characteristics of a 5-second audio snippet.

social group, as well as style of speech: speaking rate, etc.), and that there is potential for arriving at useful estimates of phonetic/phonological closeness. The encouraging conclusion is that, even in the case of a little-resourced or fully unknown language, 'snippet-lects' and 'doculects' can be placed relative to other speech varieties in terms of their closeness.

An estimation of closeness between speech signals can have various applications. For computational language documentation [15, 16, 17, 8], there could be benefit in a tool for finding closest neighbours for a newly documented language (with a view to fine-tuning extant models for the newly documented variety, for instance), bypassing the need for explicit phoneme inventories, unlike in [18]. For dialectology, a discipline that traditionally relies on spatial models based on isogloss lines [19], neural representations of audio signals for cognate words allow for calculating a phonetic-phonological distance along a dialect continuum [3]; our work explores whether cross-dialect comparison of audio snippets containing *different utterances* also allows for significant generalizations. Last but not least, for the community of speech researchers, the task helps shed light on the type of information captured by neural representations of speech and how it can be extracted from these representations. This work is intended as a stepping-stone towards the mid-term goal of leveraging neural representations of speech to extract typological features from neural representations of speech signals: probing linguistic information in neural representations, to arrive at data-driven induction of typological knowledge [20]. Note that our work is speech-based, like [21, 22], and unlike text-based research predicting typological features (e.g. [23]).

This article is organized as follows. In Section 2 we introduce our system. In Section 3 we briefly review the languages used in our experiments. Finally, we report our main experimental findings in Section 4.

2. Probing Language Information in Neural Representations

Predicting the language of a spoken utterance can, formally, be seen as a multi-class classification task that aims at mapping an audio snippet represented by a feature vector to one of the language labels present in the training set. Our implementation of this principle is very simple: we use 5-second audio snippets and use, as feature vector, the representation of the audio signal built by `XLSR-53`, a cross-lingual speech representation that results from pre-training a single Transformer model from the raw waveform of speech in multiple languages [1]. `XLSR-53` is a sequence-to-sequence model that transforms an audio file (a sequence of real numbers along the time dimension) into a sequence of vectors of dimension 1,024 sampled at 47 Hz (i.e. it outputs 47 vectors for each second of audio). We use max-pooling to aggregate these vectors and map each audio snippet to a single vector. In all our experiments, we use a logistic regression (as implemented in the `sklearn` library [24]) as the multi-class classifier with ℓ_2 regularization.

Importantly, our language identification system uses the representations built by `XLSR-53` without ever modifying them and is therefore akin to a linguistic probe [10]. We do not carry out fine-tuning of a pre-trained model. Language identification is a well-established task in the speech community and has been the focus of much research; our work does not aim at developing a state-of-the-art language identification model, but at showing that neural representations encode language information, and that this information can be useful for language documentation and analysis. Said differently, we do not aim to leverage "emergent abilities" of large language models [25], but to explore one of their *latent* abilities.

Our experimental framework allows us to consider several questions of interest to linguists. We can use various sets of labels, e.g. language names, or any level of phylogenetic (diachronic) grouping, or again typological (synchronic) groupings. We can also vary the examples the classifier is trained on. Among the many possibilities, we consider three settings:

- a *dialect identification* setting in which the classifier is trained on recordings of N language varieties (dialects) and is then used (and evaluated) to recognize one of these;
- a *language identification* setting which differs from the previous setting only by the definition of the label to predict: the goal is now to identify languages, which constitute groups of dialects. Importantly, this classifier can be used to predict the language affiliation of a dialect that is not present in the train set, so that it can be used to predict, for instance, the language to which a hitherto unknown dialect belongs;
- a *similarity identification* setting which differs from the first setting only by the definition of the train set: in this setting, we use our model on utterances of a dialect that is not present in the train set. Since the classifier cannot predict the exact dialect (as its label is not available from within the train set), it seems intuitively likely to choose the label of a dialect with similar characteristics. Crucially, we believe that this setting will therefore allow to identify similarities between language varieties.

3. Information on Languages and Dialects

In all our experiments, we use datasets from the Pangloss Collection [26],[2] an open archive of (mostly) endangered languages.Our experiments focus on 11 dialects that belong to five languages:

- two dialects of Nepali: Achhami (Glottocode [27]: doty1234) and Dotyal (doty1234);
- two dialects of Lyngam (lyng1241): Langkma and Nongtrei;
- three varieties of Na-našu, a dialect of Shtokavian Serbo-Croatian (shto1241) spoken by Italian Croats;
- two dialects of War (khas1268): Amwi (warj1242) and Nongtalang (nong1246);
- two dialects of Na (yong1270): Lataddi Na (lata1234) and Yongning Na (yong1288).

We also consider two additional languages, Naxi (naxi1245) and Laze (laze1238), because of their closeness to Na [28].

For the sake of consistency in the experiments reported here, we use "dialect" as the lowest-level label, and "language" for the first higher level, as a convention. We are aware that the distance between "dialects" (and between "languages") varies significantly from one case to another. We do not assume that the distance between Achhami and Dotyal (dialects of Nepali) is (even approximately) the same as that between Langkma and Nongtrei (dialects of Lyngam), or between Lataddi Na and Yongning Na. The key assumption behind our use of terms is that language varieties referred to as "dialects" of the same language are close enough that it makes sense to assume that the degree of phonetic similarity between them can serve as a rule-of-thumb estimate for the distance that separates them, without requiring higher-level linguistic information (of the type used to train a language model).

In this preliminary study we have decided to focus on a small number of languages and to focus on qualitative analysis of our results, rather than running a large-scale experiment on dozens of

[2]Website: pangloss.cnrs.fr. A tool for bulk downloads and for tailoring reference corpora is available: OutilsPangloss.

	utterance split			file split		
	precision	recall	F_1	precision	recall	F_1
Achhami	0.98	0.91	0.95	0.88	0.93	0.90
Dotyal	1.00	0.98	0.99	1.00	0.30	0.46
Laze	0.96	0.98	0.97	0.80	0.96	0.87
Langkma	0.89	0.90	0.90	0.74	0.96	0.83
Nongtrei	1.00	1.00	1.00	1.00	1.00	1.00
Acquaviva Collecroce	0.93	0.88	0.90	0.80	0.95	0.87
Montemitro	0.92	0.91	0.92	0.94	0.80	0.87
San Felice del Molise	0.89	0.97	0.93	0.87	0.87	0.87
Naxi	0.99	0.96	0.97	0.85	0.95	0.90
Lataddi Na	0.97	0.98	0.97	0.93	0.96	0.94
Amwi	0.94	0.93	0.93	0.67	0.89	0.76
Nongtalang	0.93	0.95	0.94	0.90	0.84	0.87
Yongning Na	0.97	0.97	0.97	0.95	0.97	0.96
macro average	0.95	0.95	0.95	0.87	0.88	0.86

Table 1: *Result of our dialect identification experiments. "Utterance split" refers to the setting in which data from the same file can appear both in the train and test sets. "File split" corresponds to the setting in which we require that the files of the train and test sets are different.*

languages. The languages are chosen according to the size of the available corpora and specific properties. We favored continuous speech (we left aside corpora consisting solely of word lists or materials elicited sentence by sentence).

For each of these languages we extracted 2 to 50 files of variable length (from 33 seconds to 30 minutes).

4. Experiments

In all our experiments, we evaluate the capacity of our classifier to predict the correct language information (either the label of a specific dialect or the name of a language) using the usual metrics for multi-class classification, namely, precision, recall and their combination in the F_1 score.

Dialect Identification To test the ability of a classifier to recognize a dialect from the representations built by `XLSR-53`, we consider a classifier using the names of the 13 dialects or languages described in Section 3 as its label set. We try out two configurations. In the first one, all the utterances of a dialect are randomly divided into a test set (20% of utterances) and a training set (80%). In the second configuration, the training corpus is made up of 80% of the files of a dialect and the test corpus contains the remaining 20%. While the latter configuration is closer to the real conditions of use of our system (guaranteeing that the utterances of the test corpus come only from files that have not been seen at training), it is more difficult to control the size of the train and test sets, which makes the analysis less straightforward.

The results are reported in Table 1. They show that, in both configurations, a simple classifier is able to identify the correct dialect label for an utterance with high accuracy, showing that `XLSR-53` representations encode language information. Similar observations have already been reported (see, e.g., [29]), but to the best of our knowledge, our work is the first evaluation of the capacity of `XLSR-53` representations to identify under-documented language varieties whose characteristics are potentially very different from the languages seen at (pre-)training [8]. Interestingly, the quality of predictions does not seem to be influenced by the amount of training data (a similar paradox is reported in the evaluation of another large language model in multilingual learning: ChatGPT [30]).

The recordings considered in the experiment we have just described were all collected in the context of linguistic fieldwork, and thus have some peculiarities that may distort the conclusions

	precision	recall	F_1
Laze†	0.97	0.98	0.98
Lyngam	1.00	0.99	0.99
Na	0.96	0.99	0.97
Na-našu	0.99	0.99	0.99
Naxi†	0.89	0.98	0.93
Nepali	1.00	0.46	0.63
War	0.89	0.96	0.93
macro average	0.96	0.91	0.92

Table 2: *Performance of a classifier trained to predict languages (group of dialects). Languages consisting of a single dialect are indicated with a* †.

	precision	recall	F_1
Lyngam	0.59	0.81	0.68
Na	0.86	0.83	0.84
Na-našu	0.48	0.75	0.59
Nepali	0.09	0.09	0.09
War	0.74	0.60	0.66
macro average	0.55	0.62	0.57

Table 3: *Performance of a classifier trained to predict the language (group of dialects) of dialects not seen during training. Naxi and Laze have been left out as there is a single variety of these languages in our dataset.*

we have just drawn. In particular, most of the dialects we considered have recordings of a single speaker. Moreover, different dialects of the same language were often recorded by the same linguist, using the same recording setup (in particular, the same microphone). We therefore need to check whether our classifiers just learn to distinguish speakers (in many cases: one per dialect) or recording conditions. In order to rule out this possibility, we carried out a control experiment in which we tried to predict the file name (serving as proxy information for the speaker and the recording conditions). A logistic regression trained in the 80-20 condition described above achieved a macro F_1 score of 0.45, showing that the decision of the classifier is largely based on linguistic information, not solely on information about the recording conditions.

Language Identification In a second experiment, we test the ability of our classifier to identify languages (that is, groups of dialects). We consider, again, two conditions to train and evaluate our classifier. In a first condition, the train and test sets are randomly sampled from all the recordings we consider (with the usual 80%-20% split) without any condition being imposed on the files or languages. All dialects are therefore present in both the test and train sets. In a second condition, the test set is put together by selecting, for each language (group of dialects), all the recordings of a randomly chosen dialect. The test set is thus made up of 5 dialects that have not been seen at training.

Table 2 reports results in the first condition. The classifier succeeds in identifying the correct language in the vast majority of cases, a very logical result since the same languages are present in the train and test sets and the experiments reported in the previous paragraph proved that it is possible to identify dialects with good accuracy. To verify that the classifier was able to extract linguistic information rather than merely memorizing arbitrary associations between dialects, we performed a control experiment in which we divided the 13 dialects into 5 arbitrary

	Laze	Langkma	Nongtrei	Lataddi Na	Yongning Na	Acquaviva C.	Montemitro	San Felice dM	Naxi	Achhami	Dotyal	Amwi	Nongtalang
Laze													
Laze	—	0.3	0.0	1.7	**35.0**	9.2	6.6	0.0	17.9	0.1	0.1	2.4	26.7
Lyngam													
Langkma	6.2	—	0.0	0.6	11.9	0.0	0.0	3.7	4.2	2.5	0.0	5.4	**65.5**
Nongtrei	0.0	0.0	—	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0	0.0	**99.3**
Na													
Lataddi Na	0.6	0.4	0.0	—	**52.1**	1.3	0.5	0.8	28.5	1.6	0.2	3.5	10.6
Yongning Na	4.4	0.8	0.0	**72.4**	—	0.0	1.5	0.0	12.5	0.3	0.0	2.1	5.8
Na-našu													
Acquaviva Collecroce	0.5	0.5	0.0	4.0	0.2	—	**64.2**	4.2	8.6	0.0	0.0	3.0	14.8
Montemitro	0.7	0.0	0.0	0.0	0.7	**84.4**	—	6.4	4.7	0.0	0.0	0.0	3.1
San Felice del Molise	3.3	**28.2**	0.0	1.5	0.9	11.0	7.7	—	19.3	23.4	0.0	1.5	3.3
Naxi													
Naxi	16.5	1.4	0.2	13.0	8.3	0.3	5.7	5.0	—	0.6	1.3	4.0	**43.7**
Nepali													
Achhami	0.6	0.3	0.0	3.4	0.3	0.0	0.0	16.9	20.2	—	0.3	8.3	**49.7**
Dotyal	0.0	0.0	0.0	20.0	0.0	0.0	0.0	0.0	**35.0**	0.0	—	**35.0**	10.0
War													
Amwi	0.0	0.1	0.0	9.6	17.0	1.7	0.1	1.1	7.0	5.7	0.0	—	**57.5**
Nongtalang	5.7	3.2	2.8	8.6	8.4	5.4	1.4	5.0	21.2	5.8	0.1	**32.5**	—

Table 4: *Distribution of the labels predicted by a classifier trained on 12 dialects (in columns) and used on a 13th dialect (unseen at training). Thus a classifier trained on all except Yongning Na identifies 72.4% of Yongning Na utterances as Lataddi Na and 12.5% as Naxi.*

groups having the same size as the languages (dialect groups) considered in the previous experiment. A classifier considering these groups as labels achieves a macro F_1 score of 0.85. While this score is high, it is notably lower than the score obtained by predicting linguistic families, showing that the classifier decisions are, to a significant extent, based on linguistic criteria.

Table 3 shows the results for the second condition, in which we evaluate the capacity of a classifier to predict the language (dialect group) of a dialect that was not part of the train set. Scores vary greatly by language (group of dialects) and several factors make it difficult to interpret these results. First, removing a dialect completely from the train set can result in large variation in its size and the results of Table 3 are not necessarily comparable with those reported so far. Second, some confounders seem to cause particularly poor performance for certain groups of dialects. For example, recordings of Dotyal are mainly sung epic poetry, so it is not surprising that any generalization across the two dialects of Nepali is difficult. Gender seems to be another confounder: several corpora only contain recordings by speakers of the same gender, and a quick qualitative study seems to show that a model trained on a female speaker does not perform well on data by a male speaker (and conversely). Note, however, that our evaluation of the performance of the classifier puts it at a disadvantage since it is evaluated at the level of a 5-second snippet and not of an entire recording. It is not unlikely that the performance would be better if we predicted a single label for a whole recording (for example by taking the most frequent label among those of all snippets).

Similarity Identification Setting In our last experiment, we trained 12 classifiers, considering all dialects but one for training and looking at the distribution of predicted labels when the classifier had to identify snippets of the held-out language. As explained in Section 2, the classifier cannot predict the correct label (since the target language is not present in the training corpus) but might, we believe, pitch on a language with similar characteristics. Results of this experiment are reported in Table 4. They allow us to draw several interesting conclusions.

First, these results show that the classifier pitches consistently on one and the same label. In almost every case, the distribution of predicted labels is concentrated on a few labels. That means that the classifier typically identifies almost all snippets from an audio file as being in the same language. Second, in several cases (e.g. for dialects of Na, War or Na-našu), the classifier recognizes the unknown language as a dialect of the same group: for instance Yongning Na utterances are mainly labelled as Lataddi Na (the dialect of a nearby village). In addition to its interest for the automatic identification of dialect groups, this observation proves that `XLSR-53` uncovers representations that somehow generalize over small dialectal variation.

Further experiments are needed to understand the two cases where the output of the classifier disagrees with the gold-standard clustering: the San Felice del Molise dialect of Na-našu, and the two dialects of Lyngam (Langkma and Nongtrei). (For Nepali, a plausible confounder was mentioned above: data type – genre –, as the Dotyal corpus consists of sung epics.)

5. Conclusions

Our exploratory experiments exploring the capacity of `XLSR-53` to place audio signals in a language and dialect landscape confirm the interest of neural representations of speech as an exciting avenue of research. Further work is required to ensure that a dialect identification system bases its decisions on phenomena (detecting relevant phonetic-phonological structures), not on parameters such as recording conditions, speaker characteristics (gender, age...) and speech genre/style, which constitute confounders in a language identification task. In future work, we plan to reproduce the experiments on corpora of better-resourced languages, such as LibriVox or CommonVoice, for which it is easier to control recording conditions, speaker gender, and the amount of training data.

6. References

[1] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2426–2430. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-329

[2] M. Moradshahi, H. Palangi, M. S. Lam, P. Smolensky, and J. Gao, "HUBERT untangles BERT to improve transfer across NLP tasks," *CoRR*, vol. abs/1910.12647, 2019. [Online]. Available: http://arxiv.org/abs/1910.12647

[3] M. Bartelds, W. de Vries, F. Sanal, C. Richter, M. Liberman, and M. Wieling, "Neural representations for modeling variation in speech," *Journal of Phonetics*, vol. 92, pp. 101–137, 2022.

[4] R. K. Potter, G. A. Kopp, and H. C. Green, *Visible speech*. New York: D. Van Nostrand, 1947.

[5] S. A. Fulop, "The beginning of time-frequency analysis," *The Journal of the Acoustical Society of America*, vol. 152, no. 5, pp. R9–R10, 11 2022. [Online]. Available: https://doi.org/10.1121/10.0014987

[6] G. Fant, *Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations*. The Hague & Paris: Mouton, 1960.

[7] N. San, M. Bartelds, M. Browne, L. Clifford, F. Gibson, J. Mansfield, D. Nash, J. Simpson, M. Turpin, M. Vollmer, S. Wilmoth, and D. Jurafsky, "Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 1094–1101.

[8] S. Guillaume, G. Wisniewski, C. Macaire, G. Jacques, A. Michaud, B. Galliot, M. Coavoux, S. Rossato, M.-C. Nguyên, and M. Fily, "Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)," in *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 170–178. [Online]. Available: https://aclanthology.org/2022.computel-1.21

[9] J. Good and M. Cysouw, "Languoid, doculect, and glossonym: formalizing the notion 'language'," *Language Documentation & Conservation*, vol. 7, pp. 331–359, 2013. [Online]. Available: http://hdl.handle.net/10125/4606

[10] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=HJ4-rAVtl

[11] A. Tjandra, D. G. Choudhury, F. Zhang, K. Singh, A. Conneau, A. Baevski, A. Sela, Y. Saraf, and M. Auli, "Improved language identification through cross-lingual self-supervised learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*. IEEE, 2022, pp. 6877–6881. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9747667

[12] Z. Fan, M. Li, S. Zhou, and B. Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 1509–1513. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-1280

[13] J. Valk and T. Alumäe, "VoxLingua107: a dataset for spoken language recognition," in *Proc. IEEE SLT Workshop*, 2021.

[14] E. M. Bender, "Linguistically naïve!= language independent: Why NLP needs linguistic typology," in *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, 2009, pp. 26–32.

[15] A. Michaud, O. Adams, T. Cohn, G. Neubig, and S. Guillaume, "Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit," *Language Documentation & Conservation*, vol. 12, pp. 393–429, 2018.

[16] D. van Esch, B. Foley, and N. San, "Future directions in technological support for language documentation," in *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1, 2019.

[17] E. Prud'hommeaux, R. Jimerson, R. Hatcher, and K. Michelson, "Automatic speech recognition for supporting endangered language documentation," *Language documentation and conservation*, vol. 15, 2021.

[18] R. Cotterell and J. Eisner, "Probabilistic typology: Deep generative models of vowel inventories," *arXiv preprint arXiv:1705.01684*, 2017.

[19] C. Chagnaud, P. Garat, P.-A. Davoine, E. Carpitelli, and A. Vincent, "Shinydialect: a cartographic tool for spatial interpolation of geolinguistic data," in *Proceedings of the 1st ACM SIGSPATIAL workshop on Geospatial Humanities*, 2017, pp. 23–30.

[20] E. M. Ponti, H. O'horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen, "Modeling language variation and universals: A survey on typological linguistics for natural language processing," *Computational Linguistics*, vol. 45, no. 3, pp. 559–601, 2019.

[21] A. Suni, M. Wlodarczak, M. Vainio, and J. Simko, "Comparative analysis of prosodic characteristics using wavenet embeddings," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019)*. ISCA, 2019.

[22] M. De Seyssel, G. Wisniewski, E. Dupoux, and B. Ludusan, "Investigating the usefulness of i-vectors for automatic language characterization," in *Speech Prosody 2022-11th International Conference on Speech Prosody*, 2022.

[23] J. Bjerva and I. Augenstein, "Tracking typological traits of Uralic languages in distributed language representations," *arXiv preprint arXiv:1711.05468*, 2017.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] R. Schaeffer, B. Miranda, and S. Koyejo, "Are emergent abilities of Large Language Models a mirage?" *arXiv preprint arXiv:2304.15004*, 2023.

[26] B. Michailovsky, M. Mazaudon, A. Michaud, S. Guillaume, A. François, and E. Adamou, "Documenting and researching endangered languages: the Pangloss Collection," *Language Documentation & Conservation*, vol. 8, pp. 119–135, 2014. [Online]. Available: https://shs.hal.science/halshs-01003734

[27] H. Hammarström, "Glottolog: A free, online, comprehensive bibliography of the world's languages," in *3rd International Conference on Linguistic and Cultural Diversity in Cyberspace*. UNESCO, 2015, pp. 183–188.

[28] G. Jacques and A. Michaud, "Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze," *Diachronica*, vol. 28, no. 4, pp. 468–498, 2011. [Online]. Available: https://shs.hal.science/halshs-00537990

[29] M. de Seyssel, M. Lavechin, Y. Adi, E. Dupoux, and G. Wisniewski, "Probing phoneme, language and speaker information in unsupervised speech representations," in *Interspeech 2022 - 23rd INTERSPEECH Conference*, Incheon, South Korea, Sep. 2022. [Online]. Available: https://hal.inria.fr/hal-03830470

[30] V. D. Lai, N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen, "ChatGPT beyond English: Towards a comprehensive evaluation of Large Language Models in multilingual learning," *arXiv preprint arXiv:2304.05613*, 2023.