Gender and language identification in Multilingual Models of Speech: exploring the genericity and robustness of speech representations

Séverine Guillaume, Maxime Fily, Alexis Michaud, Guillaume Wisniewski

Interspeech 2024
1-5 September 2024, Kos, Greece

Gender and Language Identification in Multilingual Models of Speech: Exploring the Genericity and Robustness of Speech Representations

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

[1] LACITO, CNRS, Université Sorbonne Nouvelle, F-94800, Villejuif, France
[2] LLF, CNRS, Université Paris-Cité, F-75013, Paris, France

severine.guillaume@cnrs.fr, maxime.fily@gmail.com, alexis.michaud@cnrs.fr, guillaume.wisniewski@u-paris.fr

Abstract

Models such as XLS-R and UniSpeech have proven effective in speech processing across diverse languages, even with limited annotated data, enabling, for instance, the development of transcription systems for some under-documented languages. This work aims to test the hypothesis that these models can build "generic" representations of an audio snippet that do not depend on characteristics that are irrelevant to understanding the message conveyed. Through two sets of experiments, we assess their ability to abstract away from speaker-specific details and distill core informational contents — in an informational-communicational sense to be refined further: *all the information contained in the audio signal that contributes evidence on the speaker's communicative intent*. The results of our experiments show that pre-trained models of speech such as XLS-R do not necessarily encode information in the same way, depending on the speaker's gender.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Pre-trained multilingual models such as XLS-R [1] or UniSpeech [2] can build vector representations of an utterance spoken in any language. This capability has been used, with undisputed success, to develop speech processing models for an open range of languages and/or domains even when little annotated data is available. The ability of these models to selectively extract relevant information from the audio signal has even made it possible to develop transcription systems for very low-resource languages [3, 4, 5], for which very little data and few transcriptions are available and which have very different characteristics from the languages used to (pre-)train the speech model.

The success of pre-trained models raises the issue whether they have ability to uncover generic, universal speech representations that abstract certain features from the audio signal and capture the essentials needed to model the content of an audio utterance. Thus, to develop an accurate transcription system by fine-tuning pre-trained representations with only a few minutes of annotated data,[1] it is crucial that pre-training goes beyond learning a mere conversion of the audio signal representation (from sound pressure variations to vectors), and instead distills the core message conveyed by an audio snippet, somehow separating it from paralinguistic or nonlinguistic information: those pieces of information in the acoustic signal that reflect characteristics of the speaker, the environment, the microphone, etc.

This intuition seems to receive support from mind-boggling conclusions drawn from experiments on multilingual models of text [7]: these models are able to match words from different languages that refer to the same objects and/or concepts, even though they are not given any information about the translation relationships between words (e.g. in the form of parallel sentences or bilingual lexicons), which suggests that Transformers trained on multilingual data are capable of learning cross-lingual generalizations and constructing abstract representations that do not depend on the surface form of words [8].

To verify this intuition, we investigate, in this paper, to what extent the representations built by a pre-trained speech model are robust to domain changes. Specifically, we use a linguistic probe — a simple linear classifier — to predict certain features of a recording, namely the speaker's language and gender.[2] Using carefully constructed test sets, we evaluate the performance of these probes in an out-of-domain scenario. Linguistic probes have several advantages over the evaluation of the performance of models trained under various conditions on downstream tasks (e.g. transcription tasks). Not only does this reduce the computational cost of our experiments, enabling us to carry out a greater number of tests (it is cheaper, in all respects, to learn a probe than a complete transcription model): crucially, it enables us to test our hypotheses on corpora of newly-documented languages for which we do not necessarily have sufficient annotated data to train (or even just fine-tune) a complete transcription system.

Our results show that, contrary to our initial intuition, the representations built by pre-trained speech models do not actually abstract away from the audio signal: in addition to the core message conveyed by an audio snippet, they also encapsulate speaker-specific information. More importantly, we show that this information can play the role of confounding variables and affect the ability of fine-tuned models to generalize. These results have important consequences: they show that it is not feasible to fine-tune pre-trained representations of speech for any task using a training set containing only one speaker, or speak-

[1] For example, [6] reports a WER of 4.8 on the Librispeech test set after tuning a pre-trained model on only 10 minutes of annotated data (but considering a large n-gram language model); [5] succeeds in learning a high-quality phonemic transcription system for Japhug, a newly-documented Sino-Tibetan language and shows that in the case of languages with transparent orthography, transcription performance was already good without a language model.

[2] In this work, we utilize "gender" — the term used in the metadata of the CommonVoice corpus — as a simplified representation for the biological differences typically categorized as female and male, focusing on anatomical factors that are known to have a statistical bearing on phonetic realizations (see [9, p. 5] and references therein). Considering a binary gender system raises ethical questions [10], but on this point, we are limited by the metadata collected in CommonVoice. While we recognize that phonetic differences among gender groups stem in no small part from social influences beyond this binary framework, the present piece of research does not delve into these social aspects.

10.21437/Interspeech.2024-953

ers of the same gender. This is a vital consideration in low-resource scenarios, such as language documentation and/or language revitalization. In these situations, researchers often have access to only small corpora, which may come from a single speaker, and with no possibility of collecting more data. Our results also shed new light on the possible limitations of debiasing methods, particularly that of gender neutralization (e.g. [11]). This method, which essentially consists of modifying speech embeddings to remove certain potentially prejudicial information, could lead to a drop in performance. By demonstrating that gender information is encoded in a language-independent way, we suggest that it is fully plausible that debiasing methods developed for a given language can be successfully applied to recordings from other languages.

The rest of the article is organized as follows. In Section 2 we present the data used and our experimental protocol. In Section 3, we set out our experimental results, which show that pre-trained speech models such as `XLS-R` do not necessarily encode information in the same way, depending on the speaker's gender.

2. Probing Language and Gender Information in Speech Embeddings

Linguistic Probes The goal of our experiments is to find out whether information about the language of an audio snippet and the gender of its speaker is encoded in the representations that a neural network has induced from raw audio during pre-training.

For this purpose, we use a linguistic probe (see [12] for an overview): we consider audio snippets of 5 s or of 2 s and construct a vector representation of the audio signal using `XLSR-53`, a cross-lingual speech model that results from pre-training a single Transformer model from the raw waveform of speech in multiple languages [1].[3] Note that [13] has shown that the `wav2vec2` architecture on which `XLSR-53` is based is sensitive to the gender distribution in the training data: depending on the task under consideration and the way in which the models are used, not having a gender-balanced corpus can hurt performance. We, however, do not explore this dimension here and limit ourselves to studying the properties of a vanilla `XLSR-53` model as it is used in many works.

Several recent studies (e.g. [14]) have also shown that, counterintuitively, `XLSR-53` representations in the last layers are not necessarily those that capture the most linguistic information, probably because they are specialized for the task of reconstructing the masked part of the signal (the task considered during pre-training). In fact, our preliminary experiments (on an independent validation set) showed that the best performance on the language prediction task was achieved when using representations from the 21st layer. In all our experiments we have used the audio embeddings extracted from this layer.

All our experiments rely on a very simple framework: the vector representation built by the neural network is used as the feature vector to train a linear classifier to predict either the gender of the speaker or the language of the snippet. We use a logistic regression with ℓ_2 regularization as the multi-class classifier.[4] Note that the classifier is trained on "frozen" representations computed by the pre-trained model: unlike what is sometimes done when fine-tuning a model, the neural representations are not modified when learning our classifier. We are therefore evaluating the ability of a vanilla `XLSR-53` model to learn, during pre-training, how to encode information about languages and speakers: it is not the probe training that causes this information to appear in the representations. The choice of a linear classifier also avoids some of the drawbacks of linguistic probes: the limited capacity of the classifier limits the risk that the target information (in our case, gender or language) is captured by the probe and not by the representation [16].

Language identification is a well-established task and has been the focus of much research (see e.g. [17] for recent work using a pre-trained model, as we do). However, the goal of our work is not to develop a state-of-the-art model for language identification, but rather to assess the ability of pre-trained models to generalize to new speaker and/or languages.

Corpus We performed all our experiments on corpora from the `CommonVoice` project, a collection of audio recordings and their transcriptions in a wide variety of languages [18]. The corpus contains recordings of sentences associated with a broad range of metadata: transcription, unique speaker identifier, speaker gender[5] and age, language of recording (down to the level of dialects/'accents' where appropriate). Not all metadata fields are filled in for all audio files.

In our experiments, we consider[6] five languages: three Indo-European languages, English (`en`), French (`fr`) and Spanish (`es`) and two Bantu languages, Luganda (`lg`) and Swahili (`sw`). For each language, we sampled training corpora containing between 2 hours and 20 hours of audio recordings and test corpora containing 20 minutes, with two constraints: no speaker from the training set appears in the test set; and all the corpora are perfectly gender balanced. We then combine the data sets of equal size of the 5 languages to obtain our final training set, ensuring that they are also equally balanced in language. Note that English, French and Spanish are part of the training set of `XLSR-53`, whereas Luganda and Swahili are not.

To assess the capacity of our linguistic probe to generalize beyond the languages seen at the stage of the pre-training of the model, we also extract a test set from the Pangloss Collection [19],[7] an open archive of audio recordings in various languages of the world (most of them endangered). We selected from the Pangloss Collection the languages for which information on speaker gender was available,[8] and built a corpus of 20 languages spanning over 7 language families:[9] Oto-Manguean (Mixtec (`mix`)); Northern Berber (Tasahlit (`mis`)); Sino-Tibetan (Yongning Na (`nru`)); Indo-European (Na-našu

[3] We used max-pooling to build a single vector from the output of the Transformer model.

[4] We used the implementation of logistic regression provided by the `sklearn` library [15].

[5] `CommonVoice` metadata considers 3 different genders ("male", "female", and "other"). This information is optional and is therefore not always provided. There are not enough records where the speaker's gender is "other" for us to include them in this work.

[6] We had to limit ourselves to languages for which the `CommonVoice` corpus contained sufficient data for both genders. In our experiments we consider version 15 of the `CommonVoice` corpus.

[7] `https://pangloss.cnrs.fr/?lang=en&mode=pro`

[8] The gender of speakers in recordings from the Pangloss Collection is not encoded as such in the metadata — a shortcoming shared with major archives [20]: There is no "Gender" field in the OLAC metadata standard, and from fieldworkers' perspective, the information may seem self-evident from hints such as given names, speaking styles and so on. We have therefore selected only those recordings for which the researcher explicitly specified the gender of the speaker, for example when providing its name or describing the record.

[9] Our datasets are available for download at `https://nakala.fr/10.34847/nkl.bf2e8mgi`.

(svm), Bulgarian-Macedonian (bul), Nashta (mkd)); Austronesian (Cèmuhî (cam), Chru (cje), Xârâcùù (ane), Mwotlap (mlv)); Austro-Asiatic: Tampuan (tpu), Chong Heup (cog), Chong Tratt (cog), Chong Lo (cog), War (aml), Pear (pcb), Cardamom Khmer (khm), Mường (mtq)); Tai-Kadai (Tai Yo (tyj)). Unlike the other corpora used, this one is not perfectly gender-balanced: 54.7% of utterances are pronounced by women.

Experimental Setting We conducted two types of experiments to evaluate the ability of pre-trained representations of speech to capture gender and language information. The first set of experiments falls within the methodological framework of linguistic probes and consists in training a linear classifier (a probe) to evaluate whether specific information is encoded in XLSR-53 representations. Achieving high accuracy in this task implies that this information is encoded in the representation.

In the second set of experiments, we aimed to determine the probe's ability to generalize beyond its training data. To achieve this, we consider test and training sets that differ on specific criteria. More specifically, we are interested in two research scenarios. In the first scenario, we train a classifier to predict the gender of an audio snippet by considering only recordings in a single language and we evaluate its ability to predict the gender in other languages. In the second scenario, we address the task of predicting the language of an audio snippet and we test the ability of a model to generalize from one gender to another. More specifically, we train a classifier on recordings of one gender only, and we measure its ability to predict the language of a recording produced by a speaker of another gender.

This out-of-distribution evaluation helped us to assess the robustness of the representations uncovered by pre-trained models, and to determine whether these models could produce representations that abstract away from certain characteristics of the audio signal. Using a classifier on non-identically distributed data runs counter to the very theoretical foundations of machine learning, but it makes sense to entertain the hypothesis that the representations learned to model a very extensive and diverse corpus will be sufficiently generic to allow generalization from one domain to another. Crucially, the representations of the audio signal we are working with are learned from very large corpora (56,000 hours in the case of XLSR-53) that include recordings from various languages, speakers, and recording conditions.

3. Experimental Results

3.1. In-Domain Evaluation

We begin by evaluating the ability of a linguistic probe to predict gender and language information when the test and training sets are similar.

Table 1 reports the accuracy[10] of a classifier trained on data sets of increasing sizes to predict either the language of an audio snippet or the gender of its speaker. As explained in Section 2, the test and training sets were chosen so that they do not contain the same speakers. These results clearly show that both gender and language can be predicted with high accuracy even when the classifier is learned on a small amount of data, revealing that both types of information are encoded in the representations built by XLSR-53.

[10]Recall that our test sets are built to be perfectly balanced (in gender and language). Therefore, there is no need to consider recall and precision to evaluate our classifiers.

	gender		language	
snippet size →	2s	5s	2s	5s
↓ train size				
2.5 hours	86.2	88.7	70.3	75.2
10 hours	87.4	90.3	72.1	78.2
20 hours	88.6	88.8	71.1	78.4
40 hours	90.3	90.8	73.4	82.5
60 hours	90.4	90.5	71.9	82.7
80 hours	90.2	91.0	72.1	82.7
100 hours	90.2	90.6	71.9	83.3

Table 1: *Accuracy (in %) of our linguistic probe predicting either gender (2 labels) or language information (5 labels).*

The results in Table 1 also show that, while performance for gender prediction is essentially the same for the two snippet sizes we are considering (with a systematic but very small improvement from 2 s to 5 s), there is a substantial difference according to snippet length when we predict language. This result is in line with our intuition: language prediction requires more contextual information (a broader time window), whereas gender can be determined from local information.

Figure 1 shows the confusion matrix of our language-identifying classifier trained on 20 hours of data. This matrix clearly brings out the two language families we are considering: confusions between two Bantu languages (Luganda and Swahili) are much more frequent than those between a Bantu language and an Indo-European language (and vice versa). This observation provides a piece of anecdotal evidence that analyzing the errors of a language identification system can yield information about the closeness between languages, and possibly about language families, an observation similar to the conclusions of [21].

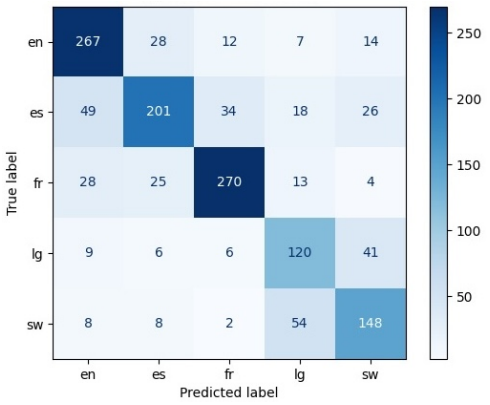

Figure 1: *Confusion matrix for language identification systems.*

3.2. Out-Domain Evaluation

We now consider the probe's ability to generalize beyond its learning data. First, we test the ability of our classifier to identify the gender of an utterance and learn cross-lingual generalization. Specifically, we learned a gender-identifying classifier on a corpus containing only recordings from one language, and tested its performance on other languages. The results (Table 2)

snippet size → ↓ train size	2 s	5 s
30 minutes	85.7	86.2
8 hours	86.9	87.3
20 hours	88.6	88.0

Table 2: *Average of the accuracies (in %) of a linguistic probe trained to predict gender on a corpus containing only recordings of one language and tested on the four other languages.*

	2 s snippet		5 s snippet	
gender in test → ↓ train set size	female	male	female	male
training on "female" snippets only				
1.25 hours	66.0	63.7	66.8	60.1
20 hours	72.1	69.6	79.3	74.1
50 hours	71.7	69.0	79.7	75.1
training on "male" snippets only				
1.25 hours	61.2	64.1	71.9	68.3
20 hours	68.4	72.5	77.3	81.3
50 hours	68.2	71.5	78.7	82.6

Table 3: *Accuracies (in %) of a probe identifying languages on a dataset containing only speakers of a given gender.*

show that applying the linguistic probe to languages that were not considered at training (but were present in the data set used at pre-training) does not harm performance: the performances are on a par with those obtained on an equal amount of data including the languages of the test set. Detailed results by language (Table 4) show, however, that performance varies widely from one language to another: performance on Luganda (one of the two languages not present in the XLSR-53 learning corpus) is very poor. We come back to this point in § 3.3.

We then considered the case where a classifier is trained to predict the language of an audio snippet when the train and test sets only contain speakers of a given gender. The results (Table 3) show that, while a probe trained and tested only on recordings spoken by males yields similar performance to another probe trained on a corpus containing both genders, the reverse is not true. Considering only recordings spoken by females harms performance. This result is surprising, because the probe's training set is perfectly balanced in terms of gender. Our interpretation is that it is the representations themselves that are the cause of the observed difference in performance.

Another interesting finding made when testing the classifier is that it has difficulty generalizing its predictions to a new gender: out-domain performance (when the classifier is tested on speakers of a gender it did not see during training) is systematically lower than in-domain performance. It seems worth reporting that the difference between in-domain and out-domain performance does not depend on the size of the corpus on which the probe was trained. That suffices to demonstrate that it is actually the neural representations of the audio recordings that are different, and that they do not encode language information in the same way for recordings by female and male speakers.

3.3. Quality of Gender Prediction on New Languages

The ability of `XLSR-53` to capture information on the gender of speakers, highlighted by the experiments we have just presented, may be overestimated: indeed, all the languages used in our experiments were seen when (pre-)training `XLSR-53`. In a final series of experiments, we aim at testing the ability of `XLSR-53` representations to generalize to languages that have never been seen (neither during training the linguistic probe, nor during pre-training). These languages (see §2), collected within the framework of documentary linguistics, present us with a variety of linguistic features very different from those usually considered in NLP experiments, and with a wide variety of recording conditions (unlike for `CommonVoice`). A linear classifier trained on a corpus extracted from `CommonVoice` and composed of 20 hours of the 5 languages considered in this work (i.e. a total of 100 hours) and used to predict the gender of the speaker of recordings extracted from the Pangloss Collection (cf. § 2 for a detailed list of the languages of the train and test set) obtains an accuracy of 82.7% when snippets of 2 s are considered and of 87.0 % for snippets of 5 s.

snippet size → ↓ Language	2 s	5 s
English	94.2	91.8
Spanish	89.3	90.0
French	93.8	93.8
Luganda	78.6	84.4
Swahili	91.8	90.2

Table 4: *Results of training a linguistic probe on all languages of the `CommonVoice` corpus but one and testing on the language that has been discarded.*

These performances are comparable with those obtained by performing a similar experiment on our `CommonVoice` corpus (Table 2). But as shown in Table 4, these comparisons are biased by the results for Luganda. Be that as it may, the results for the Pangloss languages are clearly inferior, suggesting that `XLSR-53` is specialized on the languages seen at pre-training.

4. Conclusion

In this paper we described several experiments showing that speech representations built by a pre-trained multilingual model encapsulate information about the gender of the speaker and the language of the snippet, as these pieces of information can be retrieved with good accuracy. Our results are in line with previous experiments showing that `wav2vec2` representations encode much more information than the sole linguistic content [22]. Using out-domain settings, we also observed that these representations do not necessarily encode information (at least information about the language of the snippet) in the same way depending on the gender of the speaker. This result is particularly important in situations where we do not have the possibility to tailor data sets to suit our exact needs — as is typically the case in the low-resource setting of computational linguistic documentation — or in work that relies on speech modeling (e.g. to understand how infants learn language), such as the reverse engineering approach of [23].

Our observations raise several fundamental questions about the use of pre-trained speech models. In particular, we need to understand why representations differ between genders, how these differences affect downstream tasks, and to what extent it is possible to construct gender-independent representations.

5. Acknowledgments

This work was partially funded by the DIAGNOSTIC project supported by the French Defense Innovation Agency (grant no. 2022 65 007) and the DEEPTYPO project supported by the French National Research Agency (ANR-23-CE38-0003-01).

We would like to state here our gratitude to the designers of the `CommonVoice` corpus, to the linguists and language workers who deposited data in the Pangloss Collection at a time when academic policies did not yet make it an institutional requirement, and to the speakers (of all genders) who generously agreed to have their vocal productions distributed under a license allowing use in research.

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] C. Wang, Y. Wu, Y. Qian, K. Kumatani, S. Liu, F. Wei, M. Zeng, and X. Huang, "Unispeech: Unified speech representation learning with labeled and unlabeled data," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 10 937–10 947. [Online]. Available: https://proceedings.mlr.press/v139/wang21y.html

[3] R. Jimerson, Z. Liu, and E. Prud'hommeaux, "An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1008–1016. [Online]. Available: https://aclanthology.org/2023.acl-short.87

[4] D. Liu, Z. Liu, Q. Yang, Y. Huang, and E. Prud'hommeaux, "Evaluating the performance of transformer-based language models for neuroatypical language," in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3412–3419. [Online]. Available: https://aclanthology.org/2022.coling-1.301

[5] S. Guillaume, G. Wisniewski, B. Galliot, M.-C. Nguyen, M. Fily, G. Jacques, and A. Michaud, "Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings," in *Interspeech 2022 - 23rd Annual Conference of the International Speech Communication Association*, ser. Proceedings of Interspeech 2022. Incheon, South Korea: International Speech Communication Association, Sep. 2022, pp. 4905–4909. [Online]. Available: https://shs.hal.science/halshs-03625581

[6] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[7] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. [Online]. Available: https://aclanthology.org/P19-1493

[8] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *CoRR*, vol. abs/1910.11856, 2019. [Online]. Available: http://arxiv.org/abs/1910.11856

[9] O. Niebuhr and A. Michaud, "Speech data acquisition: the underestimated challenge," *KALIPHO - Kieler Arbeiten zur Linguistik und Phonetik*, vol. 3, pp. 1–42, 2015.

[10] B. Larson, "Gender as a variable in natural-language processing: Ethical considerations," in *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–11.

[11] D. Fucci, M. Gaido, M. Negri, M. Cettolo, and L. Bentivogli, "No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation," 2023.

[12] Y. Belinkov and J. Glass, "Analysis methods in neural language processing: A survey," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 49–72, 2019.

[13] M. Zanon Boito, L. Besacier, N. Tomashenko, and Y. Estève, "A study of gender impact in self-supervised models for Speech-to-Text systems," in *Proc. Interspeech 2022*, 2022, pp. 1278–1282.

[14] F. Bordes, R. Balestriero, Q. Garrido, A. Bardes, and P. Vincent, "Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=ZgXfXSz51n

[15] 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.

[16] T. Pimentel, N. Saphra, A. Williams, and R. Cotterell, "Pareto probing: Trading off accuracy for complexity," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 3138–3153. [Online]. Available: https://aclanthology.org/2020.emnlp-main.254

[17] 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

[18] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[19] 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.

[20] M. Burke and O. L. Zavalina, "Descriptive richness of free-text metadata: A comparative analysis of three language archives," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, p. e429, 2020.

[21] S. Guillaume, G. Wisniewski, and A. Michaud, "From 'snippetlects' to doculects and dialects: Leveraging neural representations of speech for placing audio signals in a language landscape," in *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, 2023, pp. 29–33.

[22] M. Fily, G. Wisniewski, S. Guillaume, G. Adda, and A. Michaud, "Establishing degrees of closeness between audio recordings along different dimensions using large-scale cross-lingual models," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2332–2341.

[23] E. Dupoux, "Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner," *Cognition*, vol. 173, pp. 43–59, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010027717303013