Plugging a neural phoneme recognizer into a simple language model: a workflow for low-resource settings

Séverine Guillaume[1], *Guillaume Wisniewski*[2], *Benjamin Galliot*[1], *Minh-Châu Nguyên*[3], *Maxime Fily*[1, 4], *Guillaume Jacques*[5], *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, 75 013 Paris, France
[3] Laboratoire d'Informatique de Grenoble (LIG), CNRS – Université Grenoble Alpes – Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)
[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
[5] Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), CNRS – École des Hautes Études en Sciences Sociales – Institut National des Langues et Civilisations Orientales (INALCO)

severine.guillaume@cnrs.fr, guillaume.wisniewski@u-paris.fr, b.g01lyon@gmail.com, minhchau.ntm@gmail.com, maxime.fily@gmail.com, rgyalrongskad@gmail.com, alexis.michaud@cnrs.fr

Abstract

Recently, several works have shown that fine-tuning a multilingual model of speech representation (typically XLS-R) with very small amounts of annotated data allows for the development of phonemic transcription systems of sufficient quality to help field linguists in their efforts to document the languages of the world. In this work, we explain how the quality of these systems can be improved by a very simple method, namely integrating them with a language model. Our experiments on an endangered language, Japhug (Trans-Himalayan/Tibeto-Burman), show that this approach can significantly reduce the WER, reaching the stage of automatic recognition of entire words.

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

1. Introduction

Fine-tuning a multilingual language representation (be it written or spoken) such as the one uncovered by `wav2vec` [1] from raw speech or by `mBERT` [2] from texts, is today at the core of many workflows in speech recognition (and, more broadly, in Natural Language Processing). It is widely considered to be the most promising way to develop NLP and speech systems beyond the thirty or so languages (representing only 0.5 % of the world's linguistic diversity) for which there are large amounts of annotated data [3].

In this work we explain how this approach can be used to develop a phonemic transcription system for the documentation of "under-documented", "under-resourced" languages (mostly endangered and unwritten), taking as examples Japhug, an endangered language of the Trans-Himalayan (Sino-Tibetan) family. Our approach is part of a recent trend [4, 5] to develop transcription systems by using neural network methods to help field linguists in their language documentation work.

The method proposed here is very simple: it consists in plugging a neural phoneme recognizer, built by fine-tuning the representation of `XSL-R` [1], a multilingual `wav2vec` model into a simple count-based language model. The use of a language model provides soft constraints that help correct the predictions of a neural phoneme recognizer and thus overcome the small volume of data available in the context of language documentation. Using an approach that relies on the combination of two models (one of which is not even based on a neural network) may seem to go against the grain, as the current trend is more towards end-to-end approaches. However, the specificities of the task motivating this work (and in particular the size of the corpus involved) justifies the use of methods that can be learned from very little data.

The contribution of this work is threefold:

1. we evaluate the relevance of approaches consisting in fine-tuning a generic speech representation to develop speech recognition systems on Japhug, which, as we explain in § 2, offers a particularly challenging and interesting testbed both to evaluate the capabilities of these representations and to better understand their inner working;

2. our experiments show that it is possible to develop speech recognition systems of sufficient quality to help fieldworkers documenting endangered languages and to be of use to the speaker communities (the users of minority languages);

3. by measuring the impact of integrating a count-based language model with a neural network trained on corpora of different sizes, we are able to better understand the information that a neural network is able to extract from raw signal, a task that motivates a lot of work [6].

The rest of this article is organized as follows. In Section 2 we introduce the task motivating this work: the development of automatic transcription systems to help linguists working on the documentation of endangered languages. We aim to convey a feel for the interest of this task for the speech recognition community at large. In Section 3 we describe a first transcription system that relies on fine-tuning a `wav2vec` model and we highlight the limitations of this approach when the volume of data is low. Finally, in Section 4, to overcome these limits, we introduce an extension of this approach in which the phoneme recognizer is plugged into a simple language model.

2. Language documentation: a task that presents major challenges for NLP

Automatic Speech Recognition of "minority", "under-resourced" languages is not only extremely important for the field of language documentation [7, 8, 9]: it also raises various scientific challenges [10]. Specifically, this area constitutes a particularly interesting test bed for evaluating and analyzing the properties of unsupervised language representations uncovered by neural networks such as `wav2vec`. In particular:

- the amount of data available for such languages is very small: for instance, of the 197 languages in the Pangloss Collection [11], which hosts audio recordings in various languages of the world (most of them endangered), only 44 corpora contain more than one hour of recordings. There is therefore a need for speech recognition methods that require as little training data as possible.
- Endangered/little-described languages have structural features of their own, which may be widely different from those of the languages routinely taken into account in the work of the speech processing community. It has even been argued that highly elaborate linguistic structures and typological oddities are more likely to be found in minority languages, for sociolinguistic reasons [12, 13]. For example, Japhug has a degree of morphosyntactic complexity that is particularly impressive, especially in view of its areal context [14]. Knowing whether models pre-trained on the most common languages (and more generally the neural architectures used for these languages) are also able to correctly represent languages in all their diversity would (i) help out with the modeling of these languages and (ii) improve our understanding of these models (e.g. by identifying which features of the signal are captured and which are not).
- Speakers of minority languages frequently use words (or multi-word expressions, or even entire sentences) from other languages — typically the majority language of the country, or of the area [15, 16]. The presence of various loanwords, as well as cases of code-switching in the recordings, are a challenge for the automatic transcription of linguistic fieldwork data.

3. Using `wav2vec` as an acoustic model

To build our first phonemic transcription system we followed the (now well-established) approach of fine-tuning a generic speech representation to our task. We will start by describing the generic representation we used, then proceed to set out the fine-tuning stage in some detail.

The `wav2vec` model `wav2vec` [1] is a transformer-based model that allows to automatically build a language-independent, 'generic' representation of the signal from raw audio files (i.e. without any supervision information). Since the model only uses information coming from the signal, it falls into the category of "accoustic models". The model allows to associate each signal frame to a vector capturing relevant information from the signal, which can then be used as input to another neural network to perform a task (e.g. speech recognition or speaker identification). These representations can be learned from a corpus containing data in several languages in order to learn a multilingual representation that can encode data from any language. The multilingual aspect of these representations is particularly appealing because it opens the possibility of developing speech recognition systems for the languages that constitute the focus of fieldwork language documentation tasks: languages for which there is no available corpus of sufficient size to train a neural representation from scratch.

In our experiments, we used the `XSL-R` multilingual model[1] trained in an unsupervised way on a corpus of 56,000 hours of recordings in 53 languages, and we used the HuggingFace API [17] to use and fine-tune it.

Application to phoneme prediction To develop a phonemic transcription system, we simply fine-tune a `wav2vec` model on a corpus made of speech utterances associated to their phonemic transcription. The use of CTC loss function spares us from having to explicitly specify an alignment at the phonemic level and we can directly use the annotations that are collected to document a language and that are accessible in open archives such as the Pangloss Collection, with time codes at the level of the sentence (each typically corresponding to less than 10 seconds of audio signal).

More precisely, as `wav2vec` makes prediction at the character level, the sequences to be predicted are made up directly of the phonemic transcriptions, including the spaces separating words as well as the punctuation marks. We apply a very simple preprocessing step consisting only in deleting some comments of the field worker who annotated the data (that are all between brackets) and in adding spaces before the punctuation marks. This last step allows a more reliable evaluation of the prediction quality and, more importantly, the integration of a language model, as we will see in Section 4.

Considering word boundaries as a label that has to be predicted by the transcription system is a novelty of our work and aims at allowing the system to recognize words directly. In state-of-the-art systems, the labels considered by the system consist solely of the phonemes of the language, so that the system predicts a continuous flow of phonemes without trying to identify word boundaries (which are removed during training, and are thus unavailable to the model). Word boundaries are independently predicted later on by applying a system specifically designed for this task [18]. Our approach makes it possible to dispense with this second step.

Note that our system predicts the characters that make up the phonemes, rather than predicting directly the phoneme as a unit. Thus, the phoneme /tsʰ/, noted by a trigraph t+s+ʰ, corresponds to three different predictions. As shown by our previous experiments [5], this method obviates the need for an explicit list of all the phonemes of a language (without affecting the quality of predictions).

3.1. A phonemic transcription system for Japhug

We use the approach described in the previous paragraph to train a phonemic transcription system for Japhug. More precisely, we extract from the Pangloss Collection [11] a corpus containing 3 hours of audio recordings manually transcribed and segmented into utterances.[2] We use 90% of this corpus as a train-

[1] This model is named `facebook/wav2vec2-large-xlsr-53` in Hugging Face API.

[2] To facilitate replication of the experiments, the Japhug corpus is made available as a Huggingface dataset: `https://huggingface.co/datasets/BenjaminGalliot/pangloss` It is also available from Zenodo: `https://doi.org/10.5281/zenodo.5521111`

train set size	CER		WER	
	full	no punct.	full	no punct.
1h30	18.2 $\pm$ 2.6	13.6 $\pm$ 2.2	50.1 $\pm$ 5.9	41.3 $\pm$ 3.7
3h	13.6 $\pm$ 3.5	8.6 $\pm$ 2.4	35.7 $\pm$ 8.8	26.7 $\pm$ 4.6
6h	13.0 $\pm$ 3.4	7.7 $\pm$ 2.3	33.0 $\pm$ 3.7	22.0 $\pm$ 8.5

Table 1: *CER and WER averaged over all the stories of our test set for different train set sizes.*

ing set and the remaining 10% as a validation set. This corpus is much larger than the corpora usually gathered for documenting a language (Japhug is an exception in this respect since nearly 32 hours of data have been collected, annotated and are freely available[3]). We use the validation set to tune the various hyperparameters (in particular: the learning rate, the learning rate schedule, the batch size, and the various drop-out parameters). All the results reported in this section have been achieved by the model with the best CER on the validation set. We also consider, as a point of comparison, a corpus of 1.5 hour (corresponding to a more realistic size for linguistic documentation tasks) and a corpus of 6 hours that provides a *topline*.

The performance of our model is evaluated on 11 held-out narratives (stories) corresponding to 45 minutes of audio recording. To get as close as possible to "real" evaluation conditions, we consider, as input, only the raw audio files and not their manual segmentation into sentences. These files were segmented using a voice activity detection algorithm[4] that detects silences of 100 ms using a threshold of 60 dB. In the same vein, we also decided to predict all the information annotated by the field linguist including word boundaries, punctuation marks and a special symbol used to single out Chinese loanwords in romanized transcription *(Pinyin)*.

This experimental setting significantly departs from the usual evaluation of machine learning systems in which test sets are randomly selected at the utterance level. Choosing this setting complicates the task of a system based on machine learning since test examples may different from training examples. Each story bears on a different topic, so that the blend of lexical items differs notably from one story to another. The recording conditions may also be different from story to story. (Not to mention the theoretical possibility that the annotations produced by the linguist could vary over the years, adding to the – relative – inner complexity of the corpus.) However, we believe that this way of evaluating the system is closer to the use that will be made of the system than an evaluation based on a test corpus sampled from the same corpus as the training data. We therefore consider that 'story-fold' splitting of data yields a more reliable picture of the quality of the system.

Table 1 reports the quality of the automatic phonemic transcription predicted by `wav2vec` on the narratives considered in our test set, as evaluated by both the character error rate (CER) and the word error rate (WER). To make the comparison with earlier results easier, we have systematically computed these two metrics (i) directly on the outputs of our systems (which, unlike many state-of-the-art systems, predict punctuation signs and word boundaries) and also (ii) after removing all punctuation marks from references and predictions.

Results reported in Table 1 show that our phoneme recognizer already achieves very good performance. Even though the quality of the automatic transcriptions varies greatly from one text to another, the number of corrections required is on average very low. However, it should be noted that punctuation is very badly predicted. It appears safe to hypothesize that the poor results on punctuation have to do with the 'hard' chunking of audio files. We use silent pauses to cut the audio signal into chunks fed into the speech recognizer, thereby obfuscating information that is relevant for predicting punctuation.

A clear trend emerges from the results reported in Table 1: the quality of predictions at the character level (i.e. the CER) is significantly better than at the word level (i.e. the WER), especially when the punctuation is not taken into account. This observation does not come as a surprise: the number of characters is much higher than the number of words and the impact of a mispredicted character (which systematically results in a word-level error) is therefore much stronger on the WER than on the CER. The results suggest, however, that most errors are very "local" (i.e. it is rare for a word to have more than one mispredicted character) and that it is possible to correct `wav2vec` predictions, which are made at the character level, by including higher-level constraints on the words: by building on a dictionary or a language model, for instance. This observation, confirmed by a qualitative analysis of the prediction, is the motivation for the method we describe in the next section.

Impact of the training set size Results reported so far are achieved with a model trained on a corpus of 3 hours of annotated data. To quantify the impact of the training set size, we have also reported in Table 1 the performances (evaluated in the same conditions as before) achieved by models trained on corpora of 1.5 hour and 6 hours. 1.5 hour is a realistic size for linguistic documentation tasks. As expected, the performance improves with training corpus size, but the gains achieved by adding 3 hours of training data may seem small in view of the corresponding annotation effort.[5]

4. Integration with a language model

Approach The idea of combining the predictions of an acoustic model with higher level constraints (typically a language model) is almost as old as the concept of speech recognition itself. Instantiating this principle in the case of an acoustic model estimated by `wav2vec` is straightforward since this model predicts a phoneme lattice defining for each frame of the input signal a probability distribution over the set of labels it predicts. Even if this combination raises some technical issues (mainly related to the handling of the blank token ϵ introduced by CTC to account for frames not producing a label and for repeated characters in the output sequence), there are now several implementations that allow for plugging a neural acoustic model into a language model. In this work we have used `pyctcdecode`,[6] a CTC beam search decoder that finds the sequence of phonemes that maximizes a linear combination of the language model score and the acoustic model score.[7]

In all our experiments we consider a standard n-gram language model estimated from phonemic transcription using modified Kneser-Ney smoothing as implemented in `KenLM` [20].

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

[4] We used the VAD algorithm implemented in `libROSA` [19].

[5] As a rule of thumb: producing 3 hours of transcriptions in an endangered language can take skilled transcribers and their consultants about 80 to 300 hours of work, depending on familiarity with the topic and level of detail of annotation.

[6] `https://github.com/kensho-technologies/pyctcdecode`

[7] The weights of this linear combination are chosen to minimize the WER on the validation set.

train set size	CER		WER	
	full	no punct.	full	no punct.
No LM				
1h30	18.2%	13.6%	50.1%	41.3%
3h	13.6%	8.6%	35.7%	26.7%
6h	13.0%	7.7%	33.0%	22.0%
small LM				
1h30	21.3%	13.0%	47.2%	35.7%
3h	16.2%	8.1%	33.0%	23.5%
6h	16.2%	7.2%	31.1%	19.4%
large LM				
1h30	20.6%	12.7%	46.4%	31.0%
3h	17.1%	8.3%	34.1%	22.9%
6h	16.5%	7.3%	31.9%	19.2%

Table 2: *CER and WER averaged over all the stories of our test set when decoding with a Language Model.*

We train two language models considering as training data transcriptions pre-processed in the same way as the training data of the acoustic model. The first one, referred to as `small` in what follows, is trained on the same exact data as the acoustic model. The second one, `large`, is trained on a corpus corresponding to 10 hours of audio recordings (8,400 sentences and 94,082 tokens) containing the data used for the training of the acoustic model. The purpose of the `large` model is to quantify the usefulness of including additional data, which are intuitively easier to obtain because these sentences do not necessarily have to be aligned with the audio signal, and need not even be transcribed from extant audio (but could, if necessary, be produced through automatic text generation or various data augmentation techniques [21]). The `small` model will allow us to put to the test our intuition about the usefulness of adding a language model.

Results Table 2 reports the performance achieved by our approach for the various language models and acoustic models considered. These results show that, in line with our intuition, the inclusion of a language model systematically improves the quality of the automatic transcriptions at the word-level. This improvement is even very strong (up to 10 points) when the acoustic model is trained on a small corpus. It should however be noted that the results in terms of CER are worse with a language model than without. A likely explanation is that cases where the language model mis-identifies a word result in several character substitutions which are then reflected in a stark increase in CER. Another finding is that, the larger the training set for the acoustic model, the smaller the improvements gained by using a language model down the line.

To sum up, inclusion of a language model (even of small size) seems useful for corpora of small sizes, such as those typically used for newly documented minority languages.

Our experiments also show that, counter-intuitively, representations uncovered by Transformer-based models fail to capture all the contextual information that could be extracted from training data: despite the self-attention mechanism, which makes it possible to build contextualized representations, `wav2vec` representations are not able to capture the information of neighboring letters, which can be easily taken into account by soft-constraints such as the ones provides by a simple count-based language model.

5. Conclusion

We have shown, in this work, how the performance of a neural phonemic recognizer built by fine-tuning a `wav2vec` model can be improved by plugging it into a simple count-based language model, even if the latter is estimated on a very small corpus. Our experiments on an endangered language, Japhug (Trans-Himalayan/Tibeto-Burman), show that this approach can significantly reduce the WER, reaching the stage of automatic recognition of entire words. If the quality of the transcriptions predicted by our system is already sufficient to help linguists in their language documentation work, the performances (in terms of WER) are still far from those achieved for languages with large corpora. Whether these transcriptions are of sufficient quality to be of use to the speaker communities is still an open question we plan to address in our future work.

6. References

[1] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," *CoRR*, vol. abs/2006.13979, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NaACL*. Minneapolis, Minnesota: ACL, Jun. 2019, pp. 4171–4186.

[3] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[4] O. Adams, T. Cohn, G. Neubig, H. Cruz, S. Bird, and A. Michaud, "Evaluating phonemic transcription of low-resource tonal languages for language documentation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1530

[5] G. Wisniewski, S. Guillaume, and A. Michaud, "Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?" in *SLTU*, D. Beermann, L. Besacier, S. Sakti, and C. Soria, Eds., Marseille, France, 2020, pp. 306–315.

[6] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] 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 and Conservation*, vol. 12, pp. 393–429, 2018, http://hdl.handle.net/10125/24793.

[8] N. Partanen, M. Hämäläinen, and T. Klooster, "Speech recognition for endangered and extinct Samoyedic languages," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 2020. [Online]. Available: https://arxiv.org/abs/2012.05331

[9] E. Prud'hommeaux, R. Jimerson, R. Hatcher, and K. Michelson, "Automatic speech recognition for supporting endangered language documentation," *Language Documentation & Conservation*, vol. 15, pp. 491–513, 2021.

[10] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol. 56, pp. 85–100, 2014.

[11] B. Michailovsky, M. Mazaudon, A. Michaud, S. Guillaume, A. François, and E. Adamou, "Documenting and researching endangered languages: the Pangloss Collection," *Language Documentation and Conservation*, vol. 8, pp. 119–135, 2014.

[12] A.-G. Haudricourt, "Number of phonemes and number of speakers [translation of: *Richesse en phonèmes et richesse en locuteurs*]," *L'Homme*, vol. 1, no. 1, 2017 [original publication: 1961].

[13] P. Trudgill, *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford: Oxford University Press, 2011.

[14] G. Jacques, *A grammar of Japhug*, ser. Comprehensive Grammar Library. Berlin: Language Science Press, 2021, no. 1, https://langsci-press.org/catalog/book/295.

[15] P. Moore, "Re-valuing code-switching: lessons from kaska narrative performances," in *Activating the heart: Storytelling, knowledge sharing, and relationship*, J. Christensen, C. Cox, and L. Szabo-Jones, Eds. Wilfrid Laurier Univ. Press, 2018.

[16] A. Aikhenvald, "Language contact and endangered languages," *The Oxford handbook of language contact*, pp. 241–260, 2020.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6

[18] P. Godard, M. Zanon Boito, L. Ondel, A. Berard, F. Yvon, A. Villavicencio, and L. Besacier, "Unsupervised Word Segmentation from Speech with Attention," in *Interspeech 2018*, Hyderabad, India, Sep. 2018. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01818092

[19] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[20] K. Heafield, "KenLM: Faster and smaller language model queries," in *WMT*. ACL, Jul. 2011, pp. 187–197.

[21] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for NLP," 2021.