Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares

Journées scientifiques du Groupement de recherche LIFT - Grenoble

Cécile Macaire, Guillaume Wisniewski, Séverine Guillaume,
Benjamin Galliot, Guillaume Jacques, Alexis Michaud,
Solange Rossato, Minh-Châu Nguyên, Maxime Fily

7 décembre 2021

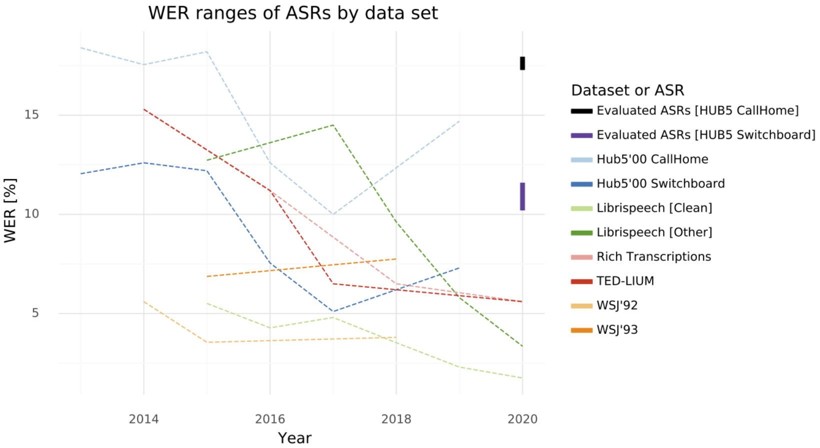
Context

Interdisciplinary collaboration

Field linguists ⇄ Computer scientists

- Relevant and beneficial collaboration for both.

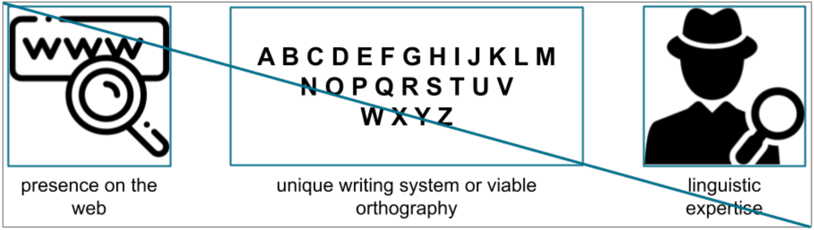
Automatic Speech Recognition (ASR): overview



ASR systems on benchmark datasets: ↘ 10% errors (WER) [1].

- What about low-resource languages ?

Under-resourced languages [2], [3]:



Two major interests in applying ASR systems on them:

1. **To document the world's declining linguistic diversity** for preservation and perpetuation.
2. **To reduce the workload** of field linguists and language workers (burden of repetitive tasks).

Automatic Speech Recognition for low-resource languages (2)

Spectacular results for under-resourced languages on the **phoneme-level** [4], [5].

Figure 1: Kaldi [6].

Figure 2: ESPnet [7].

→ using only ∼**10h of annotated data** [8], [9].

Objective

→ Towards the level of the **word**.

p æ ˩ ʈʂʰ ɯ ˥ ɖ ɯ ˧ ≠ pæ˩ ʈʂʰɯ˥ ɖɯ˧

For which purpose ?

S1

kɯɕɯŋgɯ kɯɕɯŋgɯ tɕe, tɤtɕɯ kɤndʑɯxtɤɣ χsɯm pjɤ-tú-nɯ, tɕendɤre nɤkínɯ,

kɯɕɯŋgɯ	**kɯɕɯŋgɯ**	**tɕe**	**tɤ-tɕɯ**	**kɤndʑɯxtɤɣ**	**χsɯm**	**pjɤ-tú-nɯ**	**tɕendɤre**	**nɤkínɯ**
autrefois	autrefois	\conj	\neu-garçon	frères\coll	trois	\med\ipf-avoir-\pl	\conj	cela

Il y a longtemps, il y avait trois frères,

Figure 3: First sentence of the “Le déluge” Japhug resource from the Pangloss Collection (*https://doi.org/10.24397/pangloss-0003359*).

Objective (2)

Demonstrate that a new neural approach based on the specialisation of a generic representation model (fine-tuning) can improve the quality of phonemic transcription, and automatically recognise higher-level entities, words.

Approach:

Use of supervised neural networks for ASR that have proven effective in low-resource settings.

→ *XLSR-53 wav2vec 2.0 model*

Fine-tuning XLSR-53 wav2vec 2.0 model

XLSR wav2vec 2.0 model

FB

Novel approach entitled **XLSR** introduced in Conneau et al. by Facebook AI, and based on wav2vec 2.0.

Competitive results compared to the most advanced ASR systems with self-supervised learning: (1) pre-training step, (2) fine-tuning on labelled speech data.

Release of the Transformers v4.3.0 library[1] by HuggingFace[2].
→ added the first automatic speech recognition model to the library: **Wav2Vec2** by Facebook AI [10].

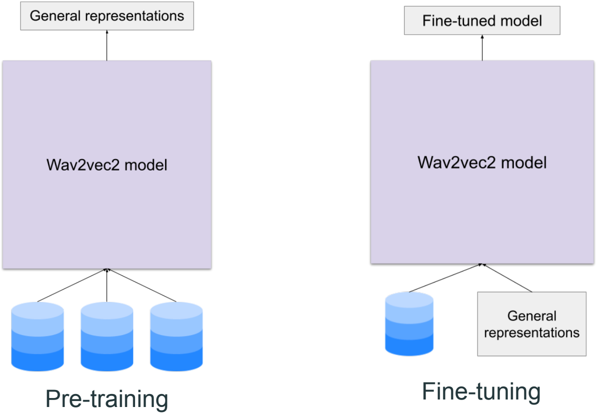
Transformers

build passing | license Apache-2.0 | website online | release v2.0.0

[1] *https://huggingface.co/transformers/*
[2] *https://huggingface.co/*

Self-supervised learning

1. Pre-training: use large amounts of unlabeled data to learn robust representations on audio recordings.
2. Fine-tuning: use these representations to fine-tune a model for a specific language on a small amount of labeled data.



Experiments

Experiments: Fine-tuning of the XLSR wav2vec 2.0 model pre-trained on 53 languages (**multilingual**).
→ *Dutch, English, French, German, Italian, Polish, Portuguese, Spanish, Arabic, Basque, Breton, Chinese (CN), Persian, Portuguese, Russian, ...*

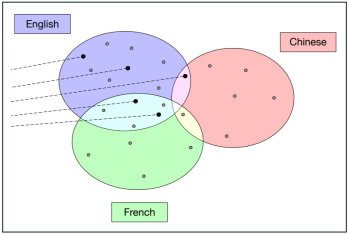


Figure 4: Multilingual quantized latent speech representations, taken from [11].

Input: vocabulary in C classes, labeled data.

Connectionist Temporal Classification (CTC) — classifier on top of the model representing the output vocabulary, trained on labeled data.

Datasets

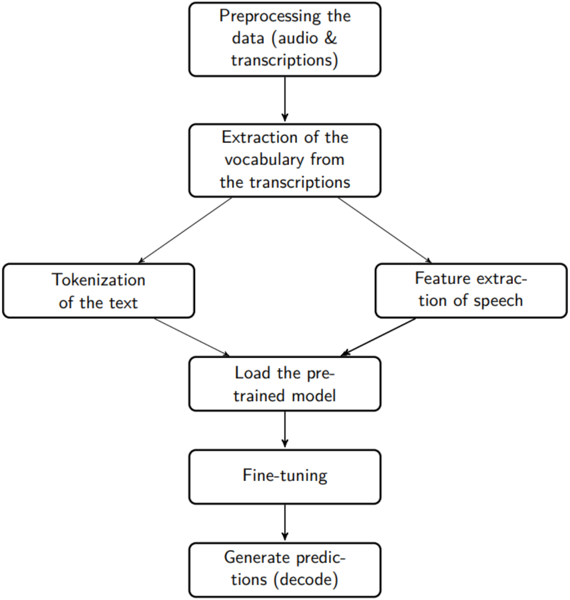
2 corpora from the Pangloss Collection[3]: **Yongning Na** & **Japhug**.

Corpus	**Yongning Na**	**Japhug**
Number of files	57 <audio, xml>	357 <audio, xml>
Number of sentences	2,484	31,864
Total duration (in minutes)	209.52 (≈ 3h30)	1907.57 (≈ 31h47)
Number of speakers	1 female speaker	2 male and 2 female speakers

- IPA-based transcriptions.

[3] *https://pangloss.cnrs.fr/*

Pipeline


Preprocessing the data (audio & transcriptions)
Extraction of the vocabulary from the transcriptions
Tokenization of the text
Feature extraction of speech
Load the pre-trained model
Fine-tuning
Generate predictions (decode)

Pipeline: preprocessing

Preprocessing the data (audio & transcriptions):

- Cutting each audio file according to the corresponding sentence segments in the transcription, which creates a .tsv file.

path	sentence
hist-14-tApitaRi_S001.wav	api stu kɯwxti nɯnɯ tɕheme ŋu tɕe jidɤm mtshu rmi
hist-14-tApitaRi_S002.wav	tɕe jidɤm mtshu rmi tɕe
hist-14-tApitaRi_S003.wav	iʑo kɤndʑɯʁi ɯŋgɯz stu kɯwxti pɯŋu (ŋu)
hist-14-tApitaRi_S004.wav	tɕendɤre amu awa ni, ndʑircɯrca tɕe iʑora kɯnɤ thɯ́wɣtɕɤti ɕti ma
hist-14-tApitaRi_S005.wav	tɕe ɯʑo pɯwxti qhe, ʑatsa thɯcha qhe

- Splitting the data into train, validation, and test sets (respectively with 70, 15, and 15% ratio),

- Cleaning of the transcriptions (deletions or substitutions of specific characters (punctuation, etc.) and conversion of the audio files (WAV format in mono, 16kHz sampling rate).

Ref: ʈʂʰɯ˧ne˧-ʝi˥ | tʰi˧-tɕɯ˧-ɲi˥-tsɯ˩ ◊ -mv̩˩. |

Ref_processed: ʈʂʰɯ˧ne˧ʝi˥|tʰi˧tɕɯ˧ɲi˥tsɯ˩|mv̩˩

Definition of a **vocabulary** from the list of symbols (tokens).
→ character units.

ʈʂʰ ↦ ʈ, ʂ, ʰ

Special characters:

- Space token: pipe symbol ‘|’.
- [PAD]: padding token.
- [UNK]: unknown token.

Generated vocabulary from the Na corpus:

“ɔ”: 0, “æ”: 1, “ɭ”: 2, “˩”: 3, “ŋ”: 4, “ɣ”: 5, “ĩ”: 6, “ɬ”: 7, “e”: 8, “b”: 9, “t”: 10, “ʈ”: 11, “p”: 12, “ˌ”: 13, “k”: 14, “ʁ”: 15, “ç”: 16, “ɛ”: 17, “ʰ”: 18, “ɤ”: 19, “s”: 20, “…”: 21, “ẽ”: 22, “h”: 23, “w”: 24, “z”: 25, “l”: 26, “d”: 27, “f”: 28, “q”: 29, “v”: 30, “˞”: 31, ...

Pipeline: tokenization & feature extraction

Tokenizer's goal: converts the text into the corresponding token IDs.

Feature extractor's goal: transforms the speech signal into the model's input format.

Example:	stu kɯwxti chondɤre nɯ ɯpa nɯ tɯlɤt ni wuma ʑo pjɤçqraʁndʑi
Tokenizer:	[25, 11, 15, 47, 20, 34, 23, 5, 11, 26, ...]
Feature extractor:	sequence of vectors of floats

Results on the test sets: quantitative analysis

Model	Training size	WER (%)	CER (%)
xlsr-na-180	180 mn	41.51	7.97
xlsr-jya-600	600 mn	18.56	7.44

Table 1: WER and CER on the Na and the Japhug test sets when training on low-resource labeled data setups of 180 minutes and 600 minutes respectively.

Few examples

Ref: tɤmu kɤtsa ci pjɤtundʑi tɕe

Hyp: tɤmɯ kɤtsa ci pjɤtu tʑɕe tɕe

Ref: ʑi˧ʁæ˧ dzi˧˥ ʑi˧ʁæ˧ dzi˧pi˥ zo˩no˧ne˩ʝi˩zo˩ əəə···tʰɑ˧v̩˥ tʰææ̃˧ mɤ˧di˥

Hyp: ʑi˩ʁæ˩ dzi˩˥ ʑi˧ʁæ˧dzɯ˧ pi˥ əzo˩no˧ni˩zo˩ əə···tʰɑ˩v̩˧ ʈʰææ̃˧ mɤ˧di˧

Results on the test sets: qualitative analysis

Main observations:

- Incorrect predictions of word boundaries for both language predictions.

  Japhug: pjɤtundʑi ↦ pjɤtu˽tʑɕe

  Na: ʑi˧ʁæ˧˽dzi˧pi˥ ↦ ʑi˧ʁæ˧dzɯ˧˽pi˥

- Main incorrect predictions for the Na come from the tones (uni tones and bi tones).

  ˧ ⇄ ˩, ˩˥ ↦ ˩, ˧˥ ↦ ˩˥, ˧˥ ↦ ˧, ...

- Wholly mistaken assumptions of Japhug reference sentences, meaning that the audio does not match the reference sentence.

  Ref: cai ɯjwaʁ ɯtaʁ ri ɲɯβze ɲɯŋu

  Hyp: bɣɤʑu qhe ʑɯrɯʑɤri

Model	Test size (words)	WER (%)	CER (%)
xlsr-na-180	71	38.5	5.7
xlsr-jya-600	236	5.4	1.3

Table 2: WER and CER of the predictions by the xlsr-na-180 and the xlsr-jya-600 models of unseen speech files.

Ref:	ə˧ʝi˧-ʂɯ˥ʝi˩-dʑo˩, ə˩-gi˩, zo˩no˥, hĩ˧ ʈʂʰɯ˧-dʑo˩, əəə... ɖʐwæ˧ ɖʐwæ˥-hwɤ˩ hwɤ˩, mmm... pi˧-dʑo˩, ʈʂɯ˧ʈʂɯ˩ ɻæ˧ɻæ˧ tʰv̩˧, ɖʐwæ˧ ɖʐwæ˥-hwɤ˩ hwɤ˩ tʰv̩˩ pi˧-kv̩˩ mæ˩,
Hyp:	ə˧ʝi˧ʂɯ˥ʝi˩dʑo˩ ə˩gi˩ zo˩no˥ hĩ˧ʈʂʰɯ˧dʑo˩ əə... ɖʐwæ˧ ɖʐwæ˥hoɤ˩ mə... pi˧dʑo˩ ʈʂɯ˧ʈʂɯ˩ ɻæ˧ɻæ˧tʰv̩˧ ɖʐwæ˧ ɖʐwæ˥hwɤ˩hɤ˩ tʰv̩˩ pi˧kv̩˩mæ˩

Ref:	tɕendɤre nɯ ɯqhu tɕe tɕendɤre kɯki @zhangxiaobing nɯnɯ @henan nɯtɕu lorɤʑi qhe
Hyp:	tɕendɤre nɯ ɯqhu tɕe tɕendɤre kɯki @zhangxiaobin nɯnɯ @huolan nɯtɕu lorɤʑi qhe

Discussion

Discussion

What to remember ?

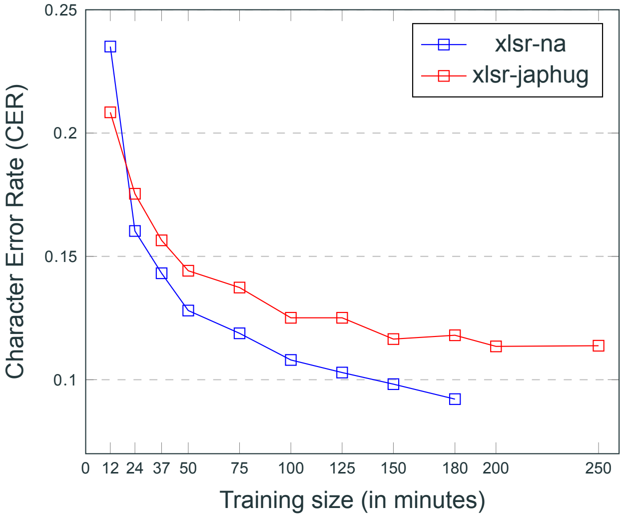
(1) **recognizing entities** from a higher level, here **words**.
(2) **dealing with a scarce-resource context**, where labeled data are only available in small amounts.

By fine-tuning XLSR wav2vec 2.0:

- **Fulfilled the task** of predicting word sequences.
- Qualitative and quantitative analysis.
  ★ Importance of **interdisciplinary collaboration** between field linguists and computer scientists.

Discussion (2)

- How many training data ?



Future prospects

1. Use the carried experiments on other low-resource languages.
   $\rightarrow$ multi-speaker, multilinguality, ...

Language Name	Iso code	city	audio/minutes	transcribed/number of minutes
Japhug	jya	Sichuan	3502	2486
Ersu	ers	China	2075	2030
Duoxu	ers	China	1509	1163
Phong Nha dialect	vie	Quảng Bình	978	978
Yongning Na	nru	Yongning Township	2306	931
Xârâcùù	ane	Nakéty	1117	787
Northern Raglai	rog	Ninh Thuận	348	714
Mường	mtq	tỉnh Phú Thọ	1524	444
Kakabe	kke	Guinea	21	390
Nepali	nep	Surkhet	362	362
Vatlongos	tvk	Mele Maat	53	342
Chru	cje	Lâm Đồng	306	306
Mwotlap	mlv	Motalava	3279	257
Dotyal	nep	Doti District	254	254
Naxi	nxq	Yunnan	672	250
Chrau	crw	BR-VT	247	247
Xumi	sxg	China	572	229

Future prospects (2)

2. Explore the newly XLS-R pretrained on half a million hours of audio data in 128 languages.
(see `https://ai.facebook.com/blog/xls-r-self-supervised-speech-processing-for-128-languages/`).

XLS-R: SELF-SUPERVISED CROSS-LINGUAL SPEECH REPRESENTATION LEARNING AT SCALE

Arun Babu[△*], **Changhan Wang**[△*], **Andros Tjandra**[△], **Kushal Lakhotia**[◊†], **Qiantong Xu**[△], **Naman Goyal**[△], **Kritika Singh**[△], **Patrick von Platen**[♣], **Yatharth Saraf**[△], **Juan Pino**[△], **Alexei Baevski**[△], **Alexis Conneau**[□‡], **Michael Auli**[△‡]
△ Meta AI □ Google AI ◊ Outreach ♣ Hugging Face

ABSTRACT

This paper presents XLS-R, a large-scale model for cross-lingual speech representation learning based on wav2vec 2.0. We train models with up to 2B parameters on nearly half a million hours of publicly available speech audio in 128 languages, an order of magnitude more public data than the largest known prior work. Our evaluation covers a wide range of tasks, domains, data regimes and languages, both high and low-resource. On the CoVoST-2 speech translation benchmark, we improve the previous state of the art by an average of 7.4 BLEU over 21 translation directions

References

[1] P. Szymański, P. Żelasko, M. Morzy, *et al.*, “Wer we are and wer we think we are,” *arXiv preprint arXiv:2010.03432*, 2020.

[2] S. Krauwer, “The basic language resource kit (blark) as the first milestone for the language resources roadmap,” in *Proceedings of SPECOM*, vol. 2003, 2003, pp. 8–15.

[3] V. Berment, “Méthodes pour informatiser les langues et les groupes de langues «peu dotées»,” Ph.D. dissertation, Université Joseph-Fourier-Grenoble I, 2004.

[4] 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.

[5] 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.

[6] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 6465–6469.

[7] S. Watanabe, T. Hori, S. Karita, *et al.*, "Espnet: End-to-end speech processing toolkit," *arXiv preprint arXiv:1804.00015*, 2018.

[8] A. Michaud, O. Adams, C. Cox, and S. Guillaume, "Phonetic lessons from automatic phonemic transcription: Preliminary reflections on na (sino-tibetan) and tsuut' ina (dene) data," in *ICPhS XIX (19th International Congress of Phonetic Sciences)*, 2019.

[9] G. Wisniewski, A. Michaud, and S. Guillaume, "Phonemic transcription of low-resource languages: To what extent can preprocessing be automated?" In *1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, European Language Resources Association (ELRA), 2020, pp. 306–315.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.

[11] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” *arXiv preprint arXiv:2006.13979*, 2020.

Thank you for your attention.

Any questions?