24/11/2022

A brief (low-resource) speech processing adventure in the realm of self-supervised models

A collection of studies on the use of wav2vec models for low-resource and fair speech processing techniques

Marcely ZANON BOITO, PhD
Research Scientist @ **NAVER** Labs Europe

How did I even end up here?

2013 - 2017 Triple degree:
2x Computer Science (ENSIMAG + UFRGS) +
Master of Science (MOSIG UGA)

2017 - 2021 Thesis on Unsupervised Word Segmentation for Computational Language Documentation

2021 - 2022 Postdoc in (Multilingual, Low-resource) * Speech Translation

2022 - Postdoc in
(Multimodal, Simultaneous) * Speech Processing

The connection: Self-supervised Speech Processing blocks

These models emerged as popular foundation blocks for speech pipelines. They are trained using only audio examples and can reach the billions of parameters [1-5].

"Tons" of speech data

Rich and Contextualized Speech Representation (continuous or discrete)

The wav2vec family of models: wav2vec “1.0”

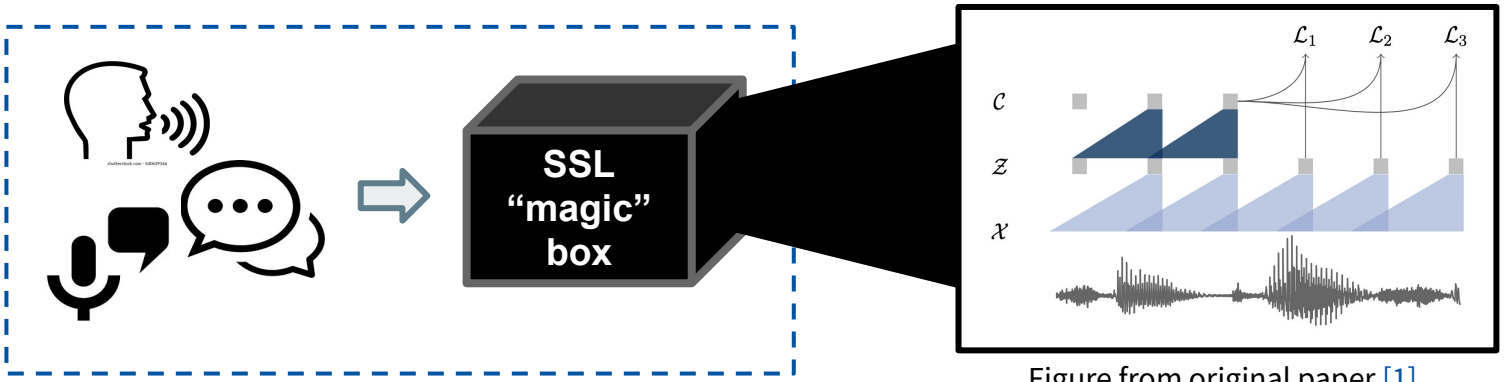


Figure from original paper [1]

Building a rich feature extractor for acoustic models in ASR

1. Encoder Network (CNN stack) f : X → Z
2. Context Network (CNN stack) g : Z → C
3. The network is trained on the **contrastive loss**:
   a. We sample negative examples from other audios, our set of *distractors*
   b. The network needs to correctly identify the future sample from the current example amidst this set

The wav2vec family of models: vq-wav2vec

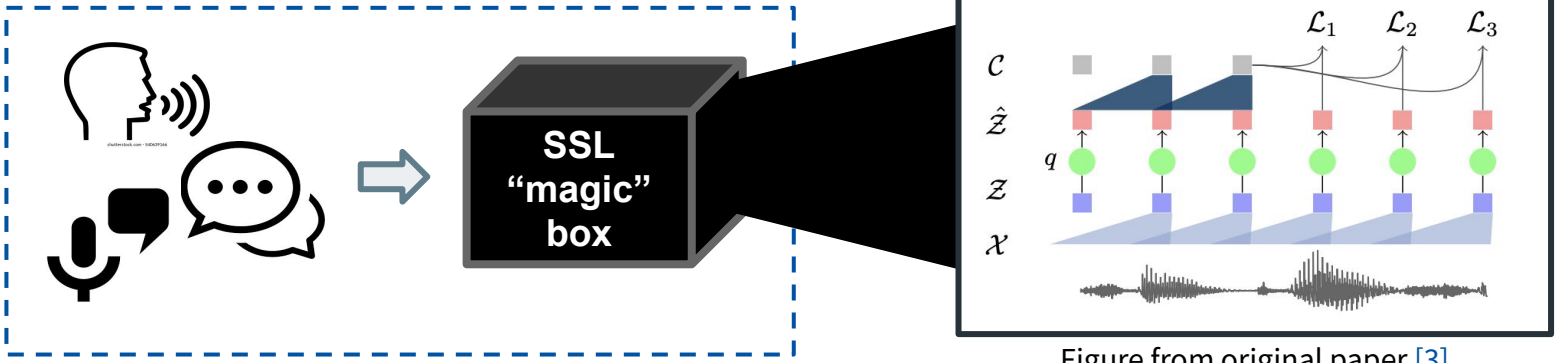


Figure from original paper [3]

Discretization allows us to “*directly plug*” it into NLP tasks

1. Encoder Network (CNN stack) $f : X \rightarrow Z$
2. Quantization Network (Vector Quantized Variational Autoencoder) $q : Z \rightarrow \hat{Z}$
3. Context Network (**still a CNN stack**) $q : \hat{Z} \rightarrow C$

The wav2vec family of models: wav2vec 2.0

Base: 12 transformer layers
Large: 24 transformer layers
XLarge: 48 Transformer layers

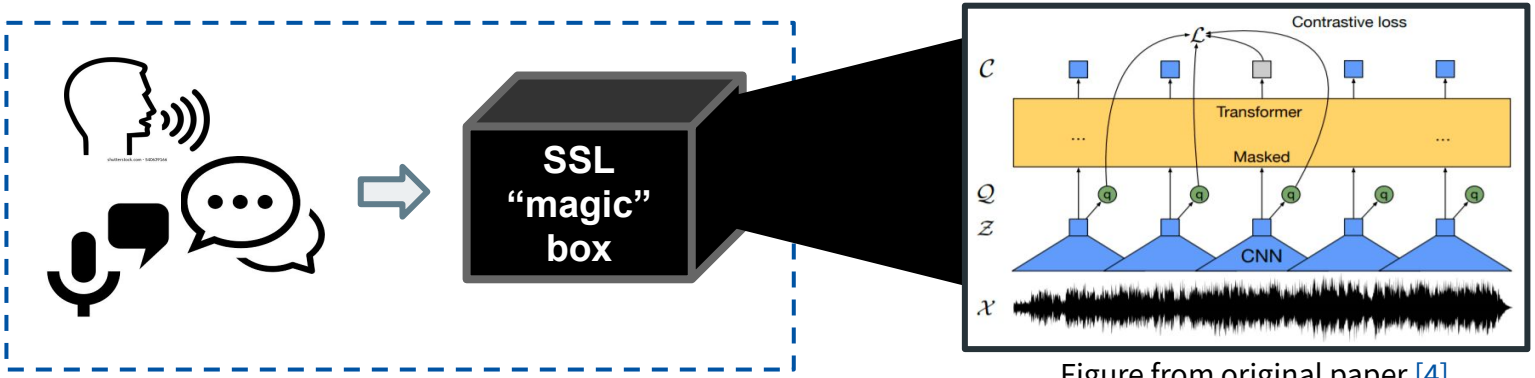


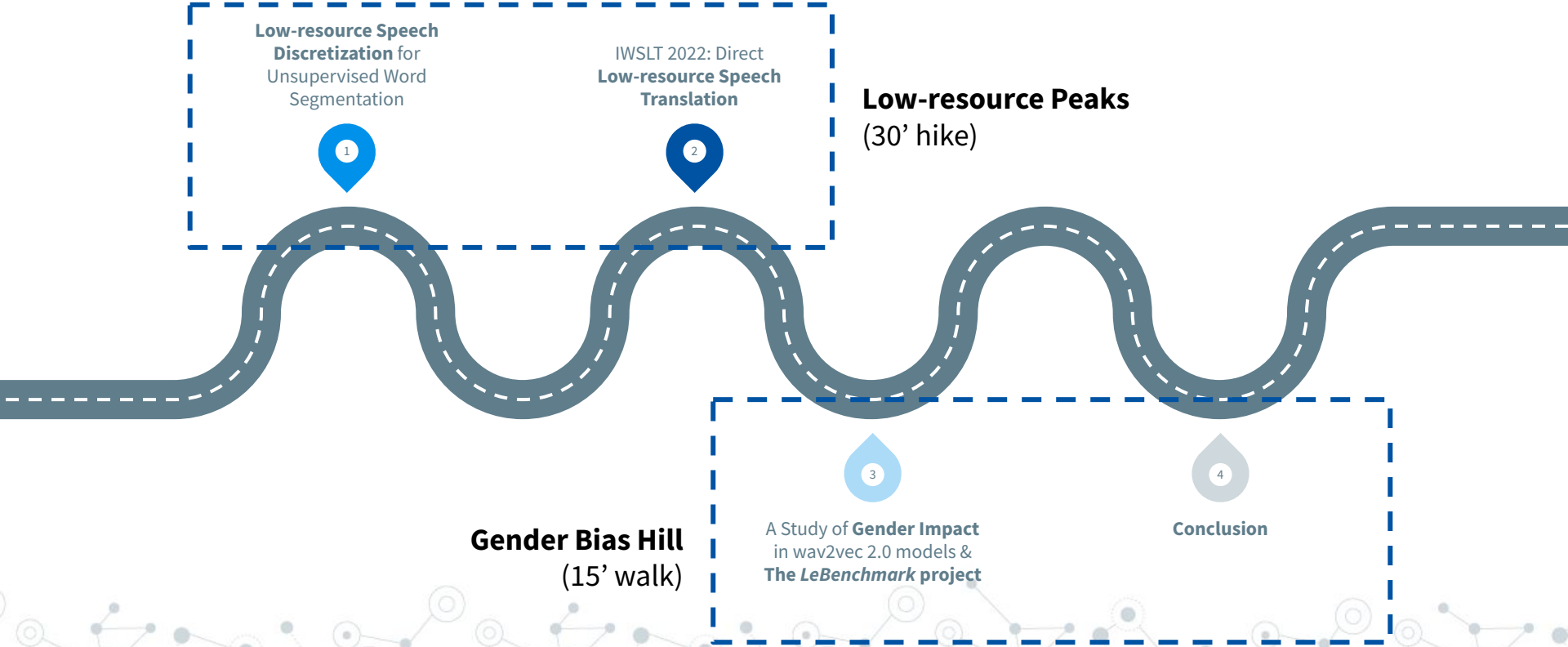
Figure from original paper [4]

Remember that Context Network? Throw a Transformer at it!

1. Encoder Network (CNN stack) $f : X \rightarrow Z$
2. Quantization Network (Vector Quantized Variational Autoencoder) $q : Z \rightarrow \hat{Z}$
3. **Context Network (Transformer Encoder stack)** $q : Z \rightarrow C$

The discrete representation is not the input of the Transformer stack,
it is instead used as objective by the loss over the learned continuous representation

A brief (low-resource) speech processing adventure in the realm of self-supervised models



1.

Low-Resource

Unsupervised Word Segmentation
Direct Speech Translation

Low-resource speech processing: Why should we care?

- Most of current speech technology is developed in a fraction of the existing languages and dialects (“high-resource languages”) [6]

- Pipelines based on text exclude oral languages
  - “Most of the world's languages are not actively written, even the ones with an official writing system” [7]

Unsupervised Word Segmentation for Computational Language Documentation

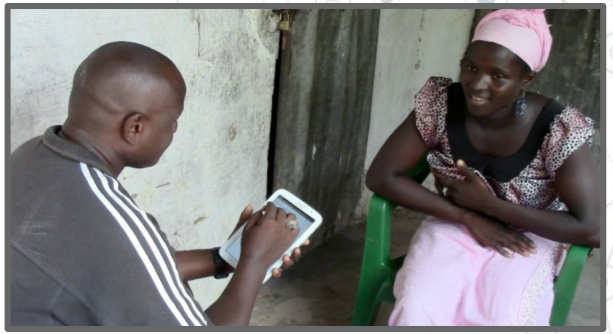


Figure: A field linguist recording utterances from a native speaker.

The nature of the data:

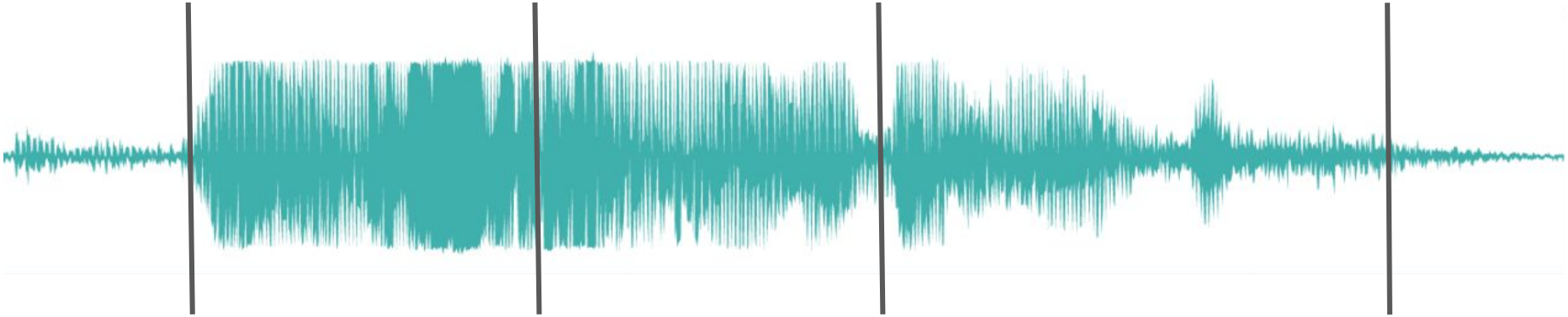
- ➔ Small size (difficult to collect)
- ➔ Often lack written form (oral-tradition languages)
- ➔ Parallel information (translations instead of transcriptions)

SPEECH

Translations
to a high-resource language [8]

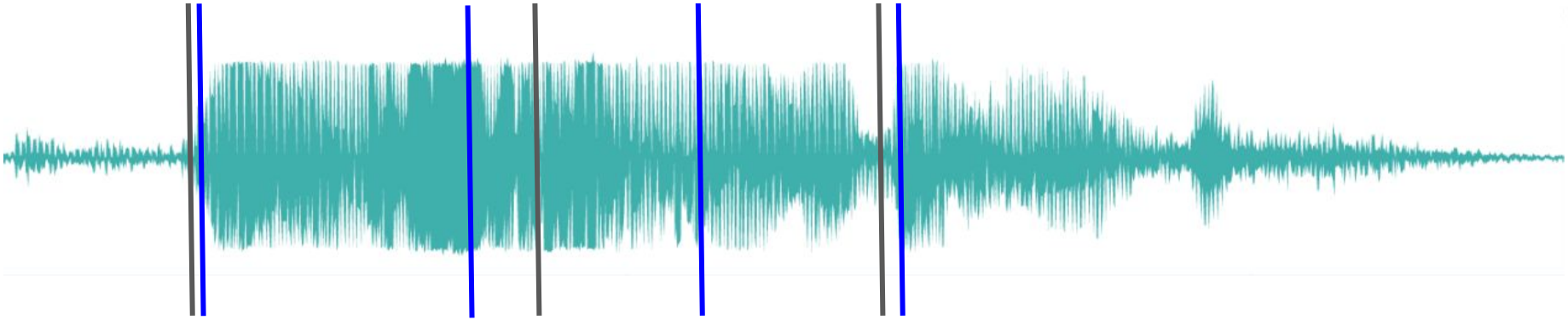
Unsupervised Word Segmentation (UWS) from speech

Example: Let’s imagine the speech utterance for “Hello my friend”.

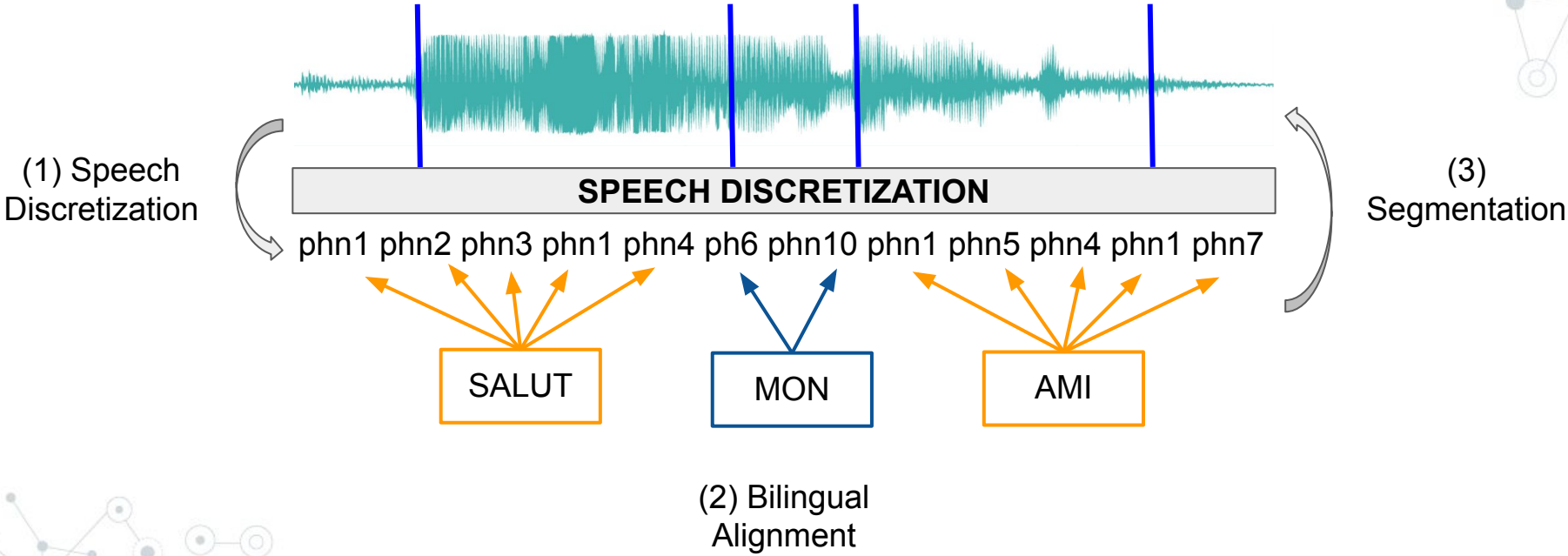


Unsupervised Word Segmentation (UWS) from speech

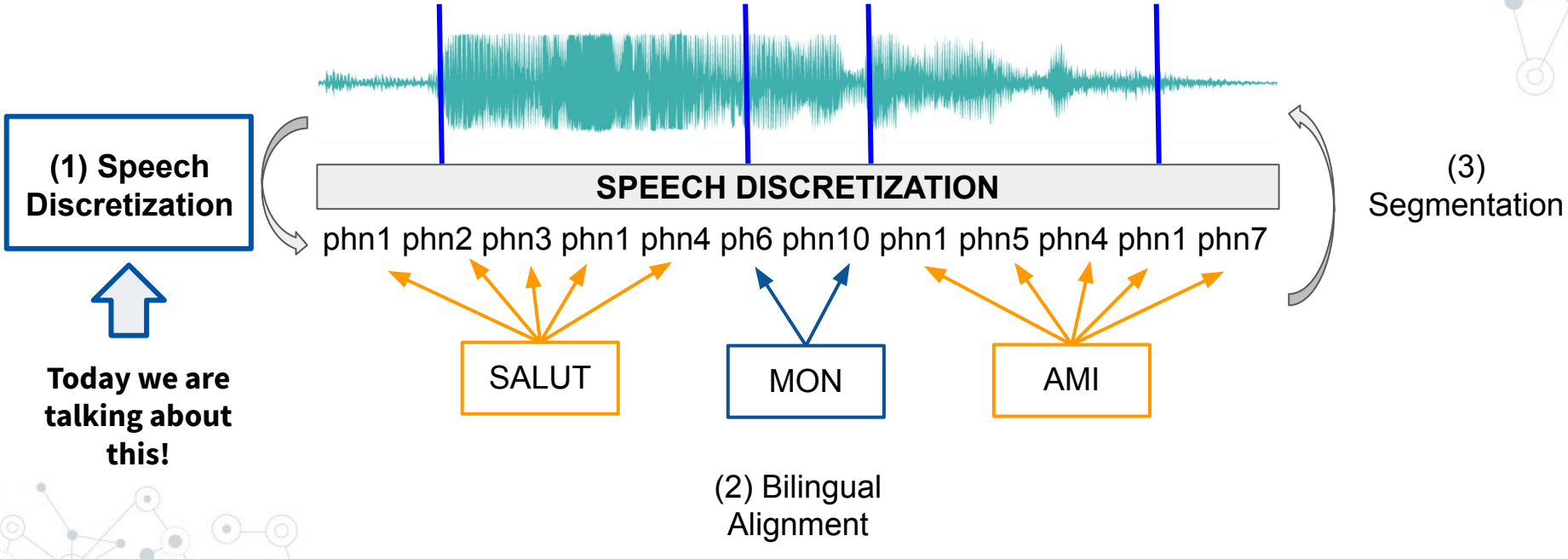
- ➔ We want a system which outputs time stamps corresponding to boundaries.



A pipeline for UWS in low-resource settings:
Discretizing first, segmenting later



A pipeline for UWS in low-resource settings:
Discretizing first, segmenting later



Speech Discretization in the Low-resource Land

GOAL: To discretize (represent, summarize) the input speech using a collection of **discrete speech units**

- Low-resource settings (4-5 hours of speech in Mboshi [14])
- **No access to transcription**

Speech Discretization in the Low-resource Land

- Vector Quantization (VQ) Approaches:
  1. **VQ-Variational Auto-Encoder (VAE)**: inspired by dimensionality reduction architectures [12]

  2. **VQ-WAV2VEC**: inspired by self-supervised models trained with a context-prediction loss [13]
     "Why not wav2vec 2.0?" Due to the diversity loss "issue"!

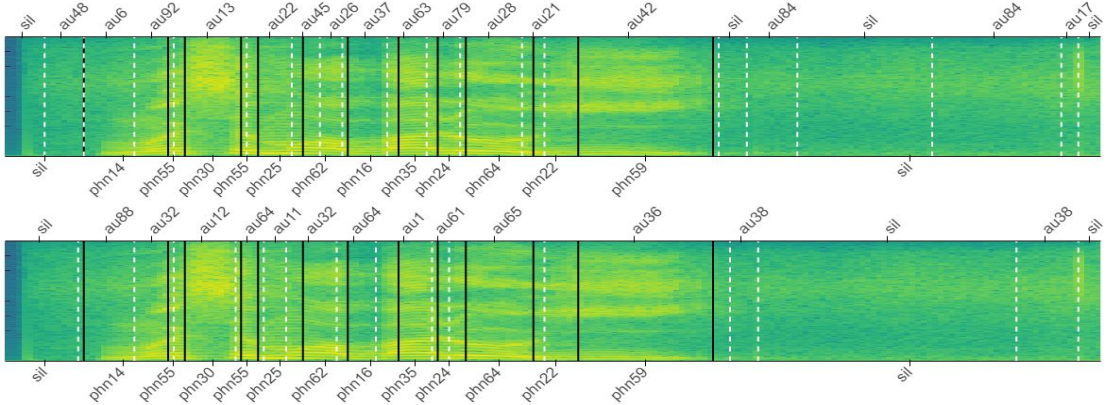
- Bayesian Generative Models (AUD):
  1. HMM/GMM **(HMM)**: Every possible sound can be a unit [9]

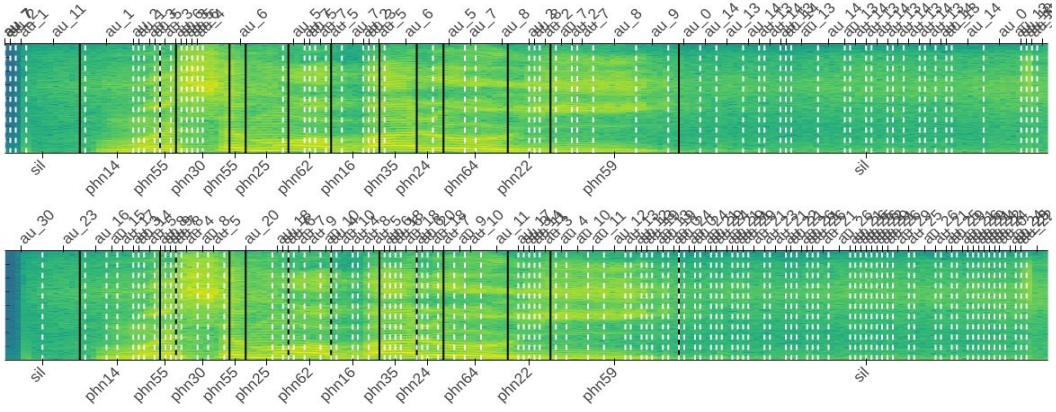
  2. Subspace HMM **(SHMM)**: Prior over a phonetic subspace [10]

  3. Hierarchical Subspace HMM **(H-SHMM)**: Subspace adaptation from different languages for unit prediction [11]

Well-informed versus self-supervised discretization:

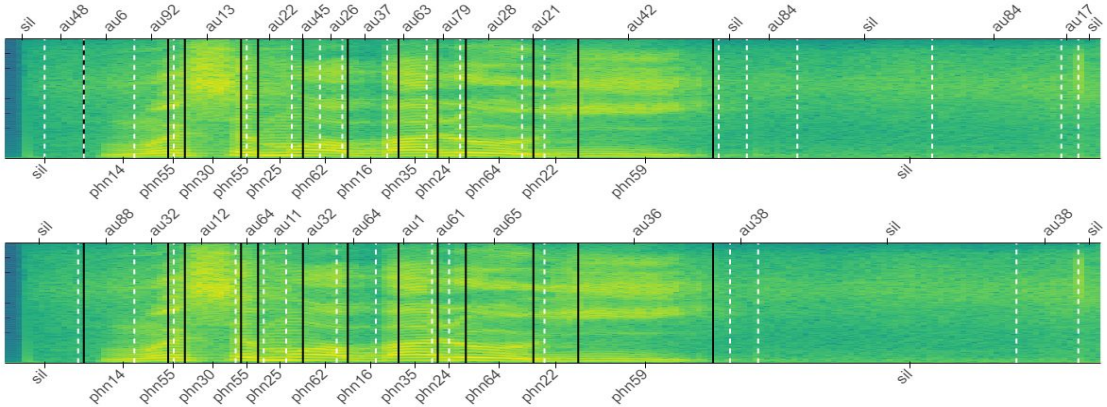
True Boundary
Output Boundary





Well-informed versus self-supervised discretization:

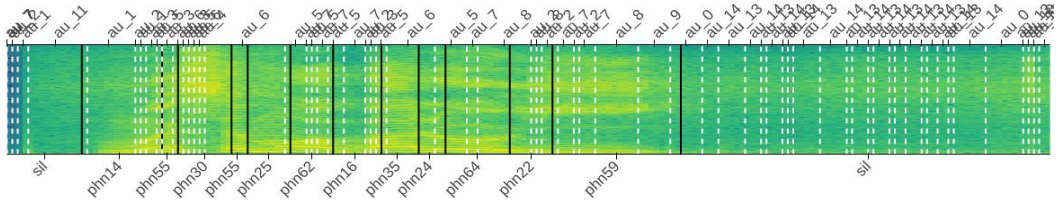
True Boundary
Output Boundary



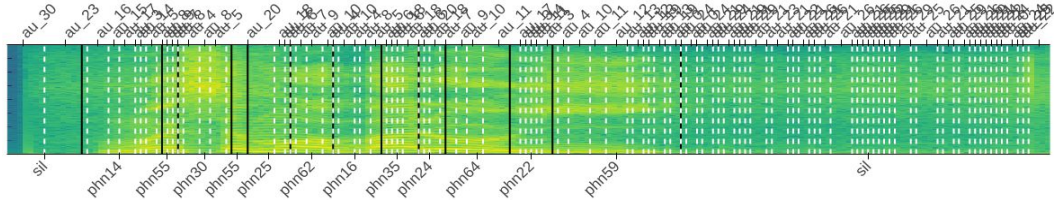
SHMM

H-SHMM

VQ-WAV2VEC
V = 16



VQ-WAV2VEC
V = 32



UWS results working from the discretization

Topline:

phonemic transcription

5 models, 6 setups

1. HMM
2. SHMM
3. H-SHMM
4. VQ-VAE
5. VQ-WAV2VEC **V=16**
6. VQ-WAV2VEC **V=36**

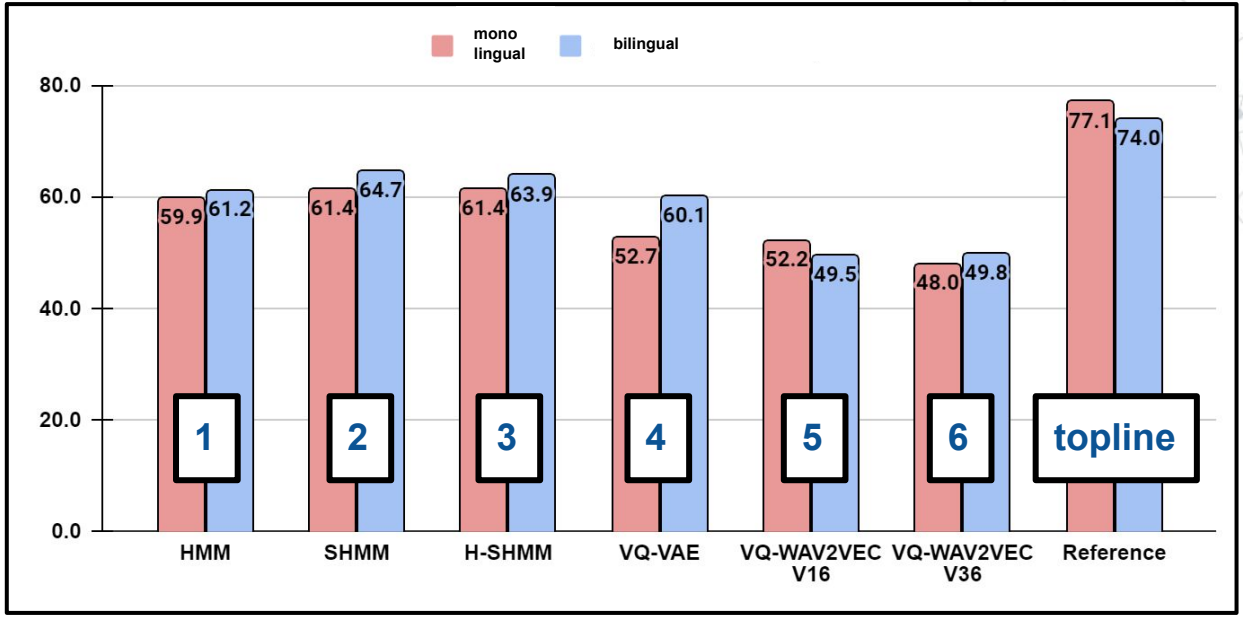


Figure: Boundary UWS F-score results for the different SD models, using the Mboshi-French dataset. The result is the average over 5 runs.

UWS results working from the discretization

Takeaways:

➔ **Bayesian models** are not only well informed, they are also less computationally expensive:

- Mboshi vq-wav2vec trained for **300h**
- Mboshi H-SHMM is trained in less than two days

➔ **Not a discretization silver bullet:** vq-wav2vec (and vq-vae) discretization is not “discrete enough” for our task

- Also verified recently in Kamper and Nieker [15]

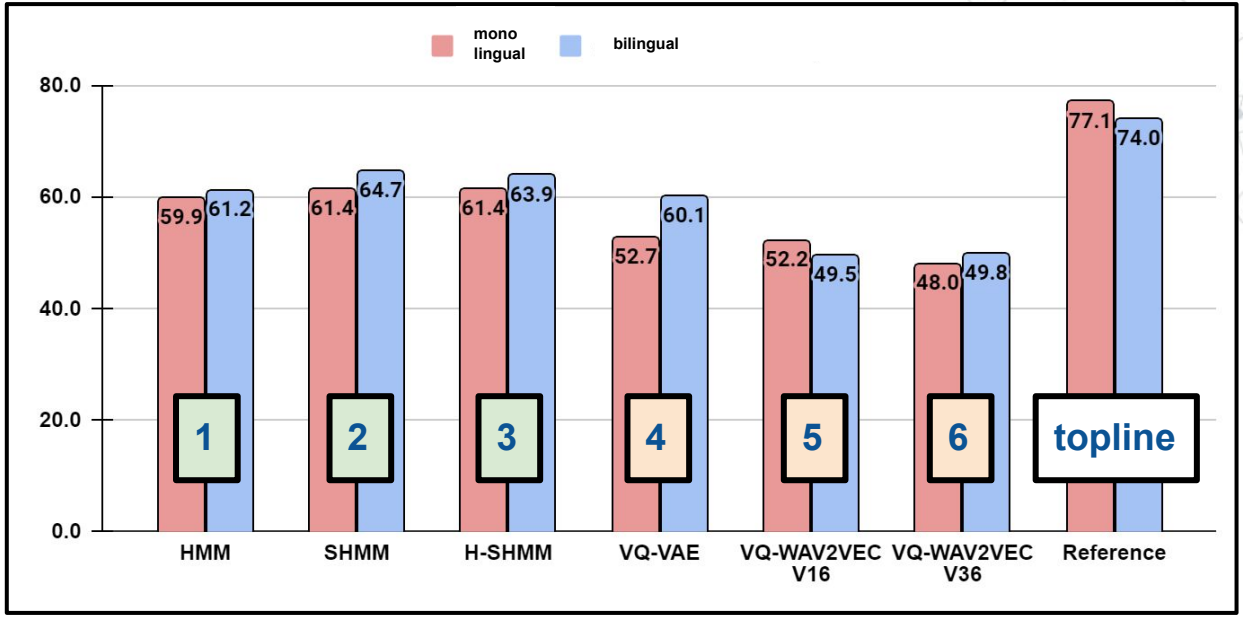
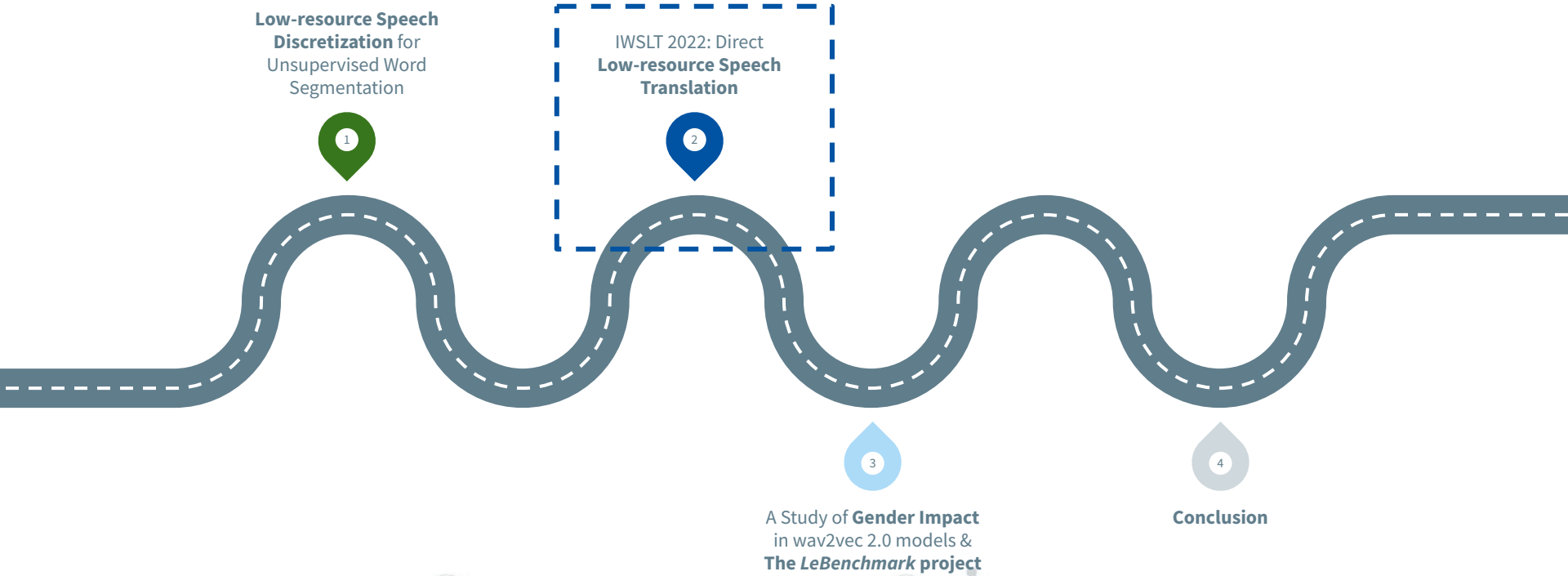
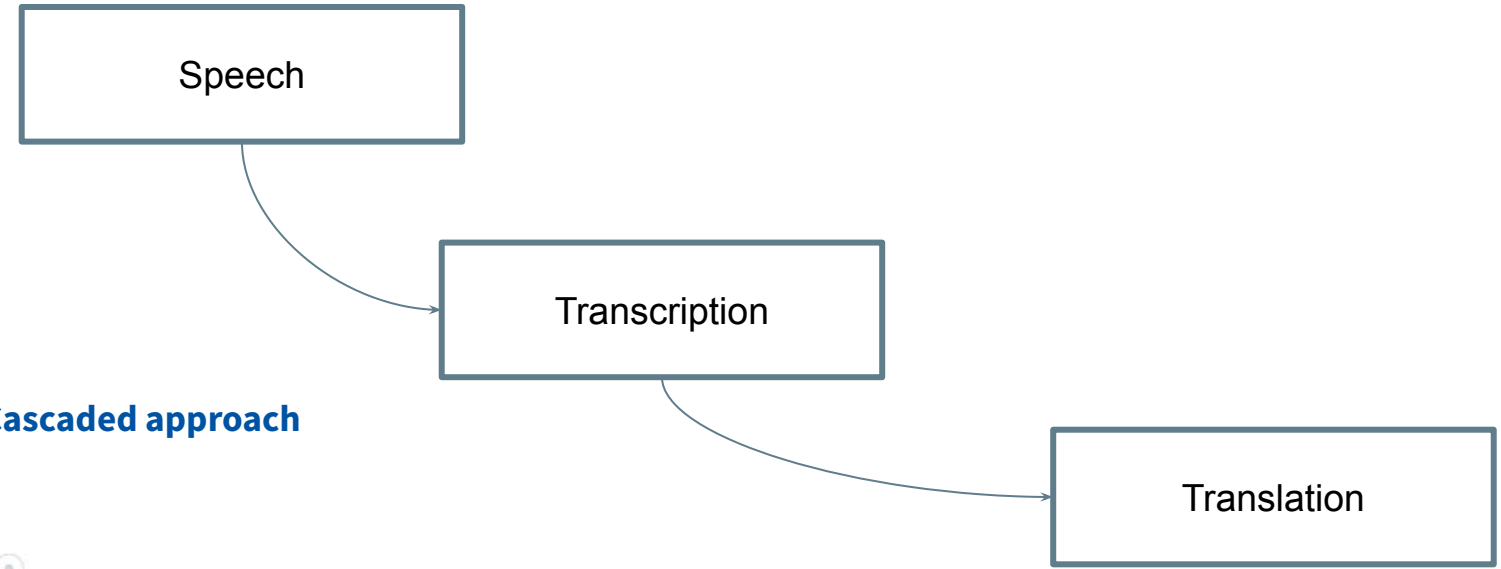


Figure: Boundary UWS F-score results for the different SD models, using the Mboshi-French dataset. The result is the average over 5 runs.

A brief (low-resource) speech processing adventure in the realm of self-supervised models

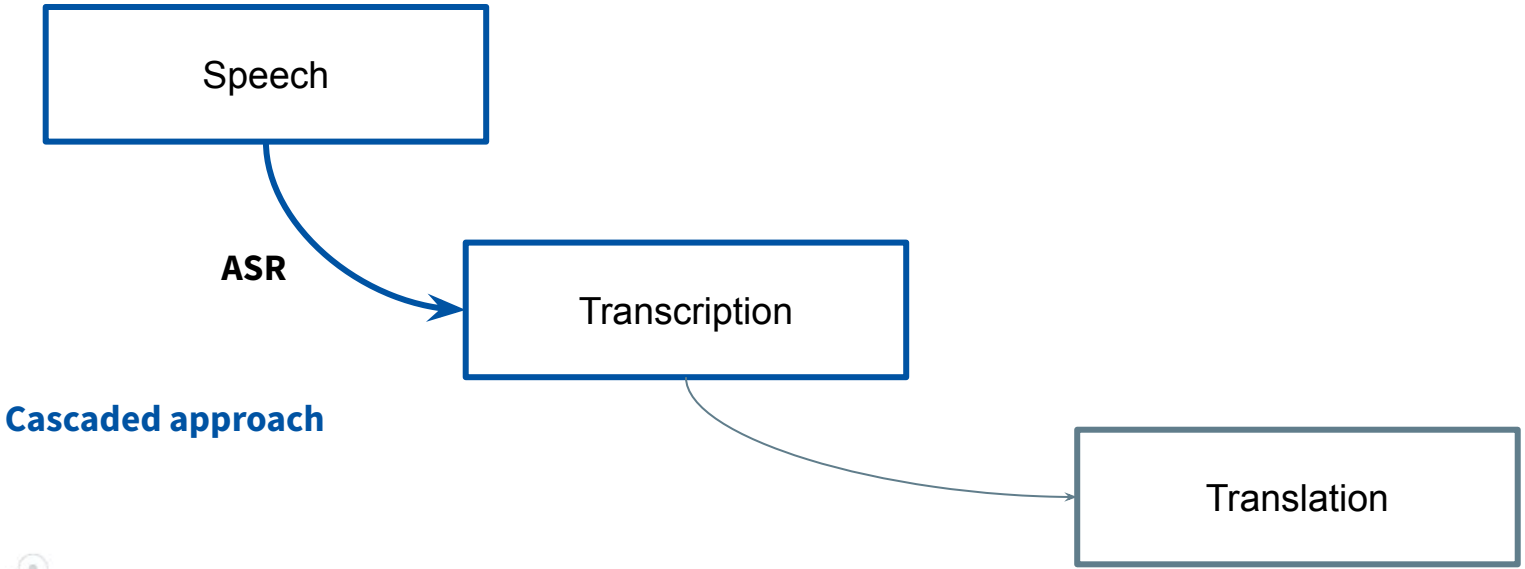


IWSLT 2022:
Direct Low-resource Speech Translation



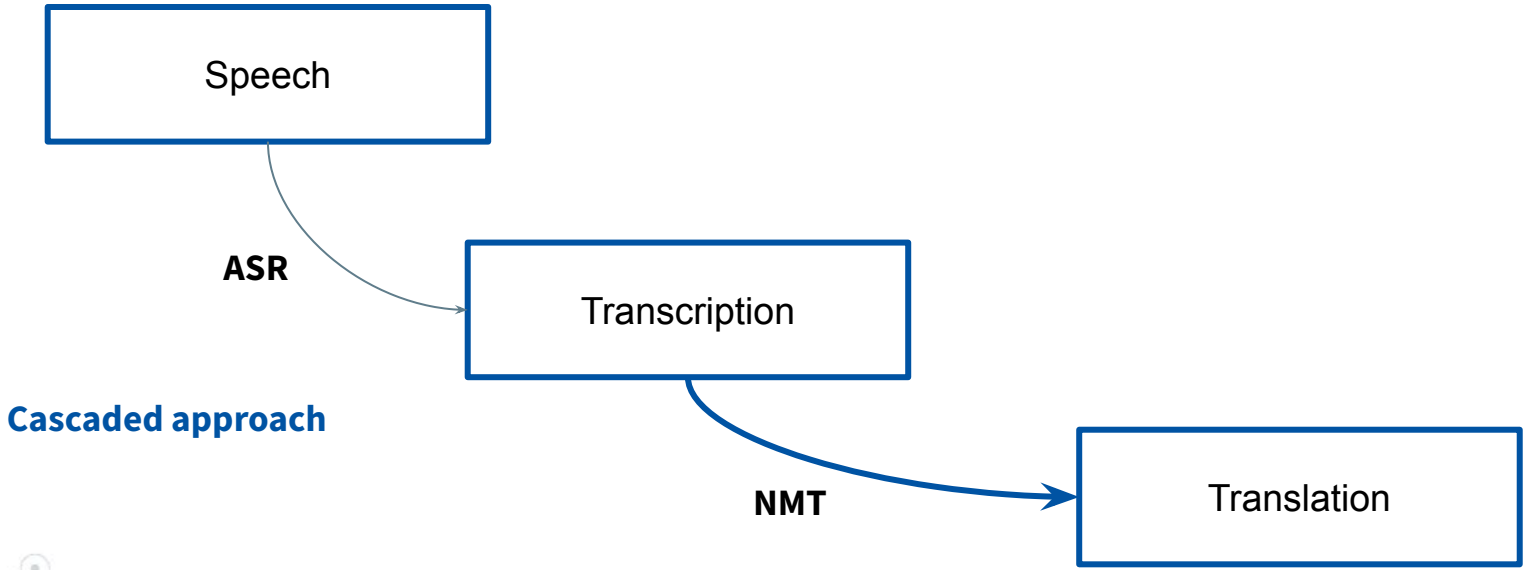
Cascaded approach

IWSLT 2022:
Direct Low-resource Speech Translation



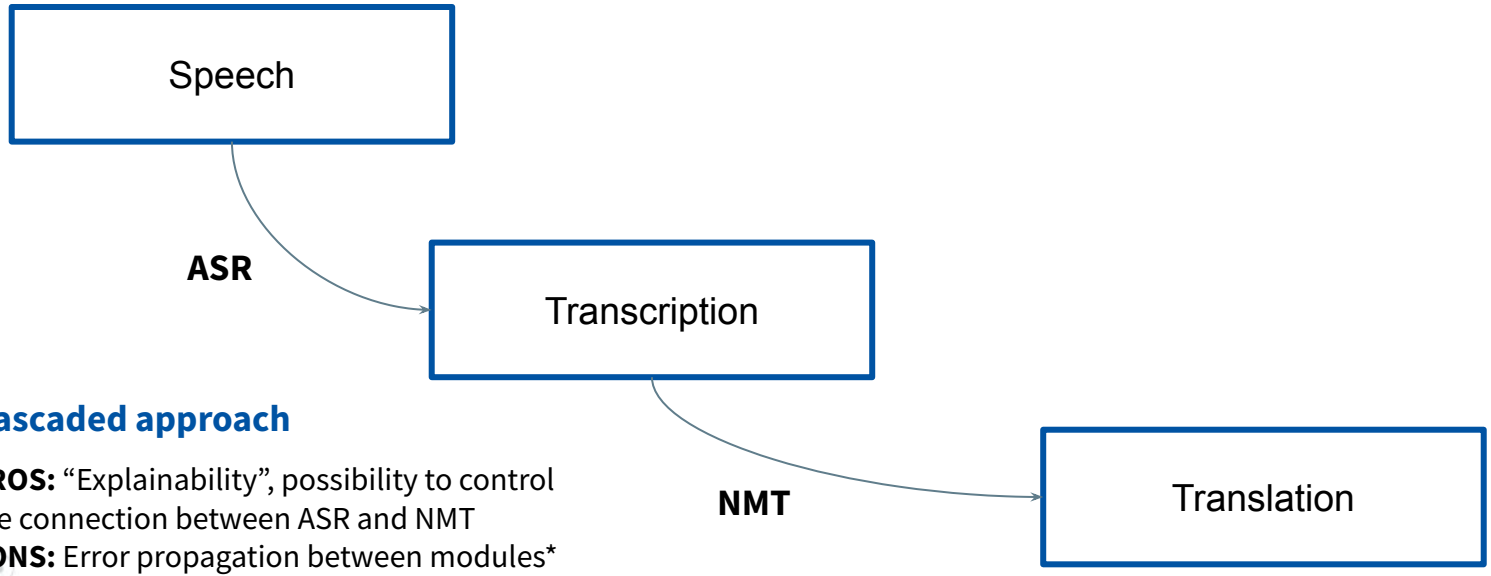
Cascaded approach

IWSLT 2022:
Direct Low-resource Speech Translation



Cascaded approach

IWSLT 2022:
Direct Low-resource Speech Translation



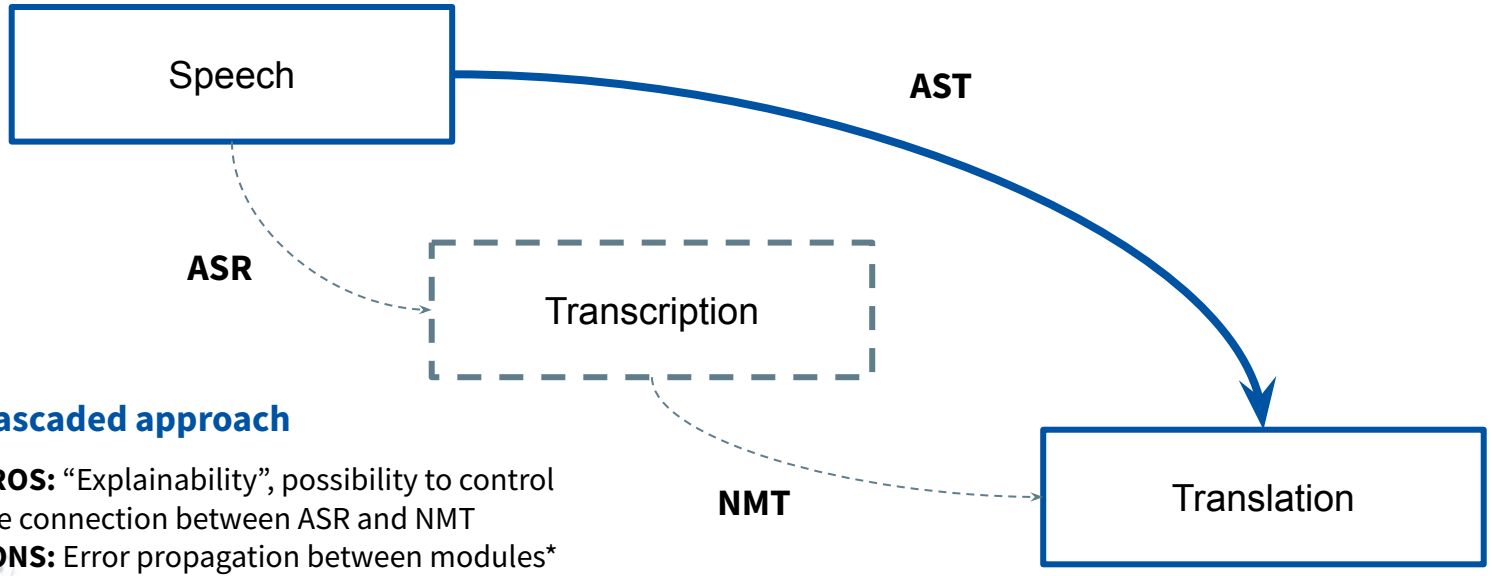
Cascaded approach

- **PROS:** “Explainability”, possibility to control the connection between ASR and NMT
- **CONS:** Error propagation between modules*

* Some IWSLTs ago, organizers claimed the apparent “victory” of end-to-end ST architectures over cascaded approaches, but in the last editions there was no clear conclusion about that, with both reaching similar results [16]

IWSLT 2022:
Direct Low-resource Speech Translation

Direct/end-to-end approach



Cascaded approach

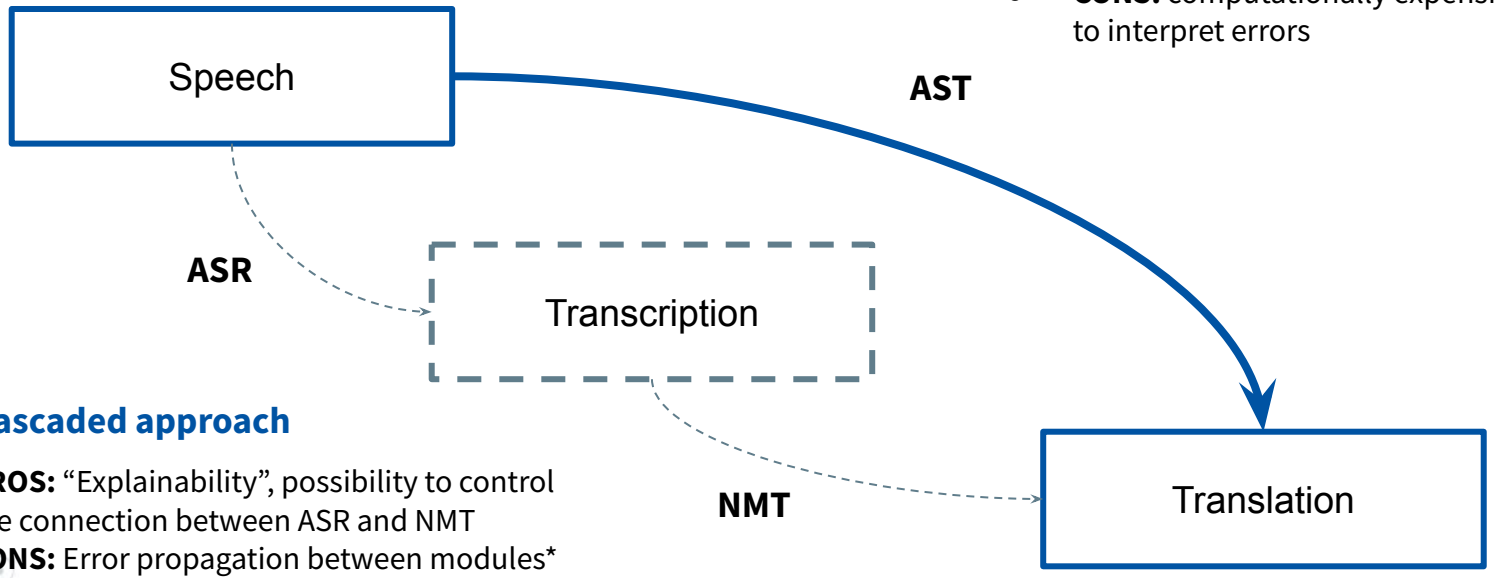
- **PROS:** “Explainability”, possibility to control the connection between ASR and NMT
- **CONS:** Error propagation between modules*

* Some IWSLTs ago, organizers claimed the apparent “victory” of end-to-end ST architectures over cascaded approaches, but in the last editions there was no clear conclusion about that, with both reaching similar results [16]

IWSLT 2022:
Direct Low-resource Speech Translation

Direct/end-to-end approach

- **PROS:** “cheaper” data annotation, inclusive of oral languages, no error propagation
- **CONS:** computationally expensive, difficult to interpret errors



Cascaded approach

- **PROS:** “Explainability”, possibility to control the connection between ASR and NMT
- **CONS:** Error propagation between modules*

* Some IWSLTs ago, organizers claimed the apparent “victory” of end-to-end ST architectures over cascaded approaches, but in the last editions there was no clear conclusion about that, with both reaching similar results [16]

IWSLT 2022:
Direct Low-resource Speech Translation Task

- We gave participants 17h of speech in Tamasheq, aligned to French translations [17]
- We also shared a collection of raw audio we webcrawled and segmented:
  - 224 hours of speech in Tamasheq
  - 417 hours in 4 other languages spoken in the same region (Fula/Fulfulde, French, Hausa, Zarma)
- **The data is challenging:** Radio recordings in Tamasheq with interviews, street noise, simultaneous translation over original speech, music…
- Participants could use **any** pre-trained models or extra data they could find

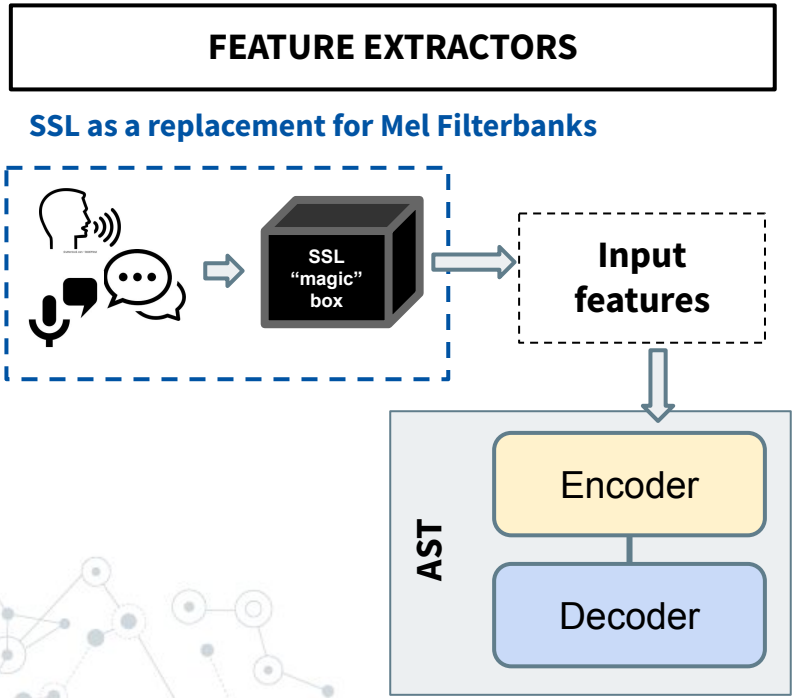
TOUTES LES VOIX DU MALI
STUDIO TAMANI
FONDATION HIRONDELLE | URTEL

AU RYTHME DU NIGER
STUDIO KALANGOU
FONDATION HIRONDELLE

FONDATION HIRONDELLE
media for peace & human dignity

IWSLT 2022:
Direct Low-resource Speech Translation Task

- As participants, our interest was in the application of **wav2vec 2.0** models as:



- ➔ Inexpensive in terms of GPU
- ➔ Multilingual models should be able to generalize to new languages
- ➔ Application of the features into a small AST (Transformer based) architecture

IWSLT 2022:
Our Methodology

- **Models tested:**
  - **LeBenchmark 7K model (large):** French model trained on **7,000h** of speech
  - **XLSR-53 model (large):** 53 languages, **56,000h** of speech
  - **Tamasheq-only (base): 243h** of Tamasheq
  - **Niger-Mali (base): 658h** of speech in Tamasheq + other languages (FU, FR, ZA, HA)
- **Task-agnostic Fine-tuning:**
  - For non-tamasheq-only models, we restart pre-training for 20k steps on Tamasheq data in order to *inform* the models

IWSLT 2022:
Feature Extraction (depressing) Results

Quick reminder: this is not WER!

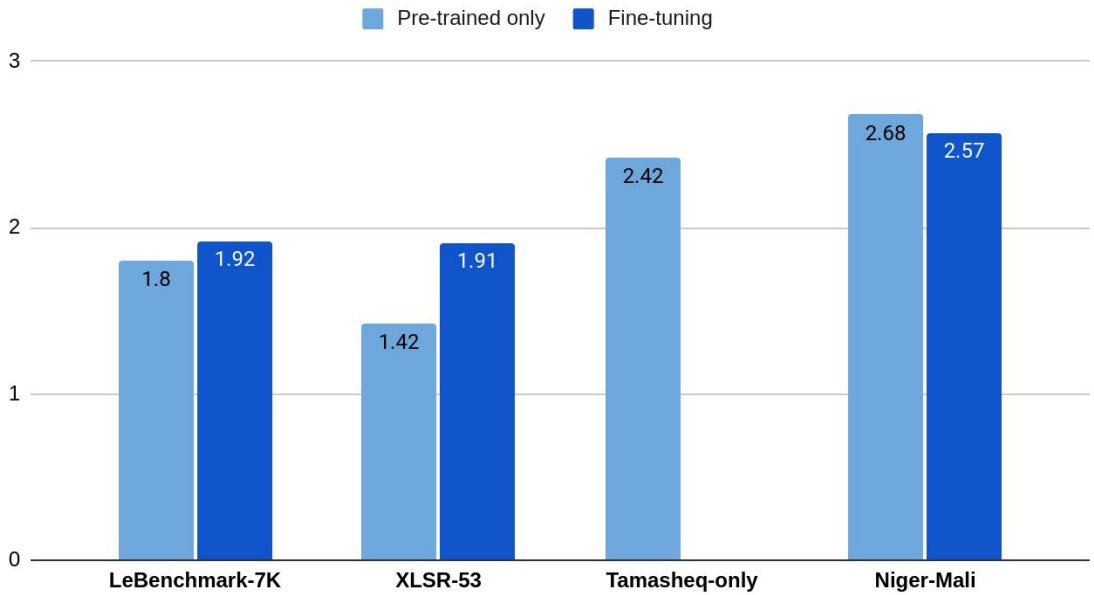


Figure: BLEU scores (test) for AST Tamasheq-French models using different wav2vec 2.0 models as feature extractors

IWSLT 2022: Feature Extraction (depressing) Results

Quick reminder: this is not WER!

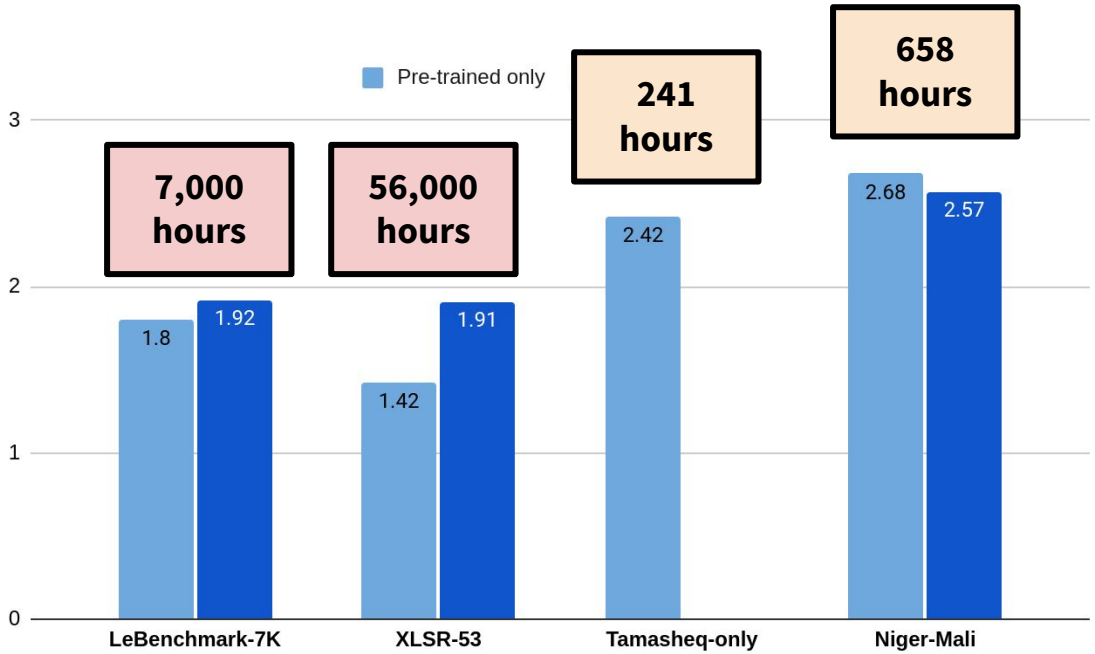
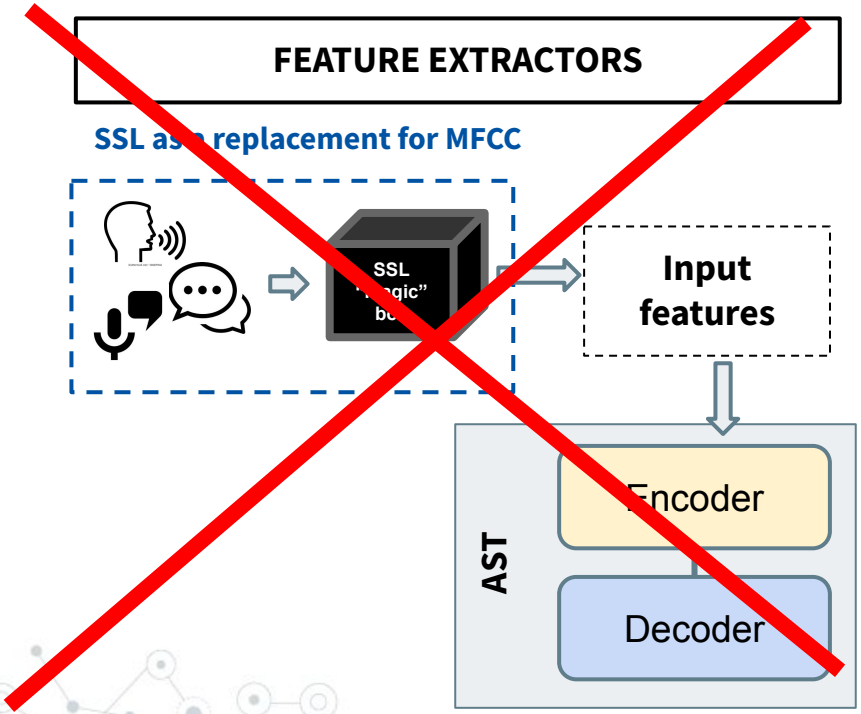


Figure: BLEU scores (test) for AST Tamasheq-French models using different wav2vec 2.0 models as feature extractors

Takeaways:

- Larger is not necessarily better!
- Fine-tuning for reducing domain shift was not enough!
- In-domain small models seem to be “more” effective, but results are ridiculous.
- **In general, it just doesn’t work!**

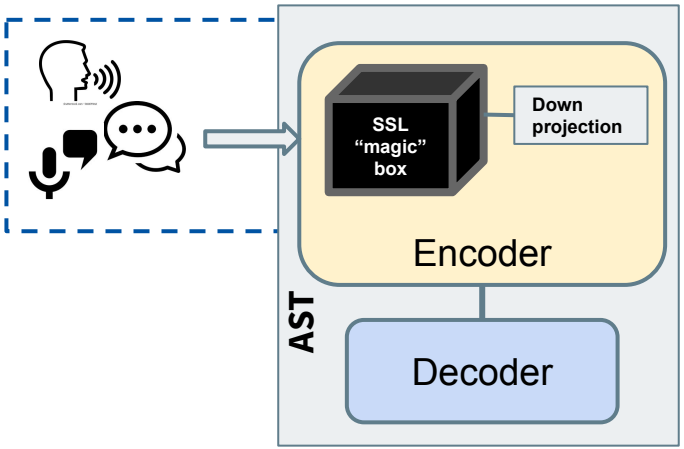
IWSLT 2022: Our Methodology



IWSLT 2022: Our Methodology 2.0

SPEECH ENCODER (Fine-tuning)

SSL as part of the AST network



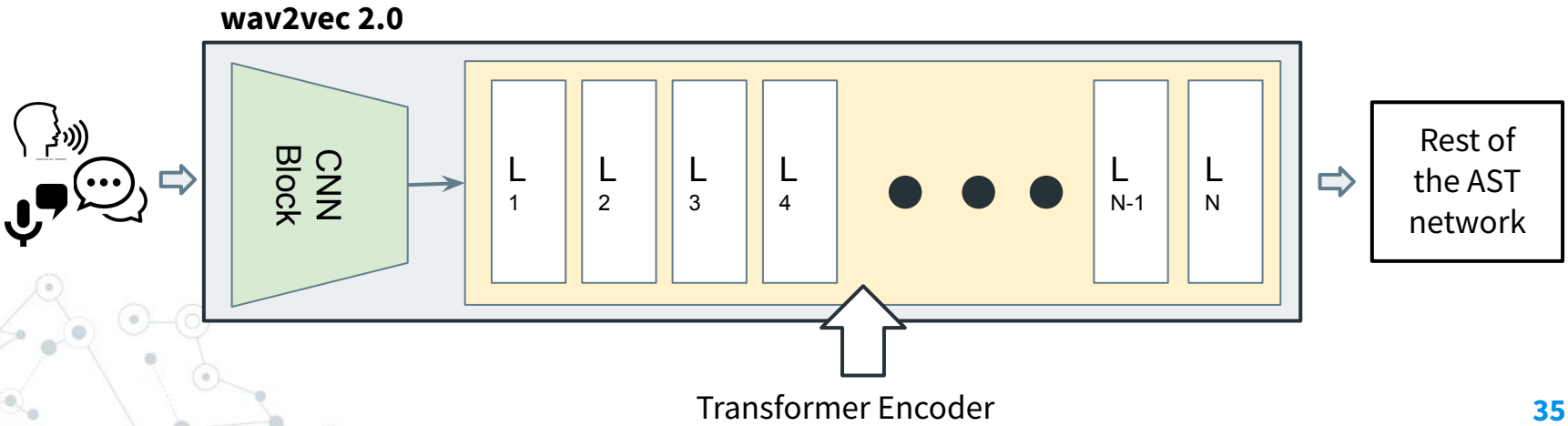
- ➔ Use the entire model inside the AST
- ➔ Fine-tuning everything is difficult:
  - The challenge of fitting new AST models in 32GB of GPU
  - The amount of supervised data is not a lot (large models have it worse)
- ➔ However, if SSL model is a frozen module, we get BLEUs of 0.00…1
- ➔ Results were overall bad, **so we decided to prune our wav2vec models** (inspired by [18])

IWSLT 2022:
How to chop your favorite wav2vec 2.0 model

Investigation in Pasad et al. [18] showed that:

- ➔ Due to the SSL training objective, the last layers of wav2vec 2.0 are quite “low-level”
- ➔ **It takes the task-specific model a lot to move these weights.** In their work they showed that randomly re-initializing the last 3 layers of wav2vec 2.0 models for fine-tuning resulted in better ASR.

But… We are in a low-resource setting.
If the middle layers are better informed… **why don’t we just chop the rest?**

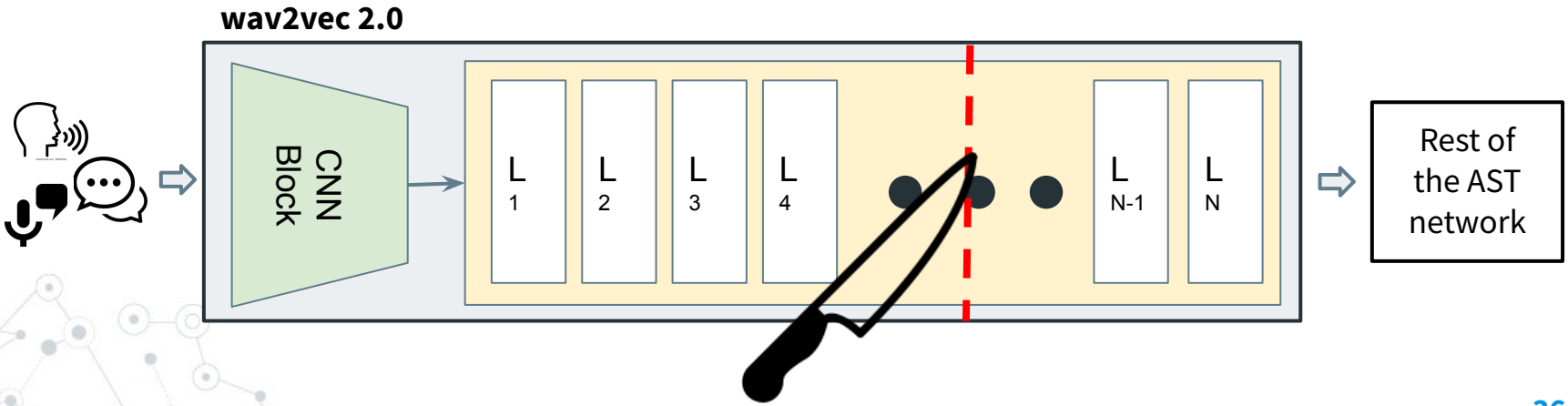


IWSLT 2022:
How to chop your favorite wav2vec 2.0 model

Investigation in Pasad et al. [18] showed that:

- ➔ Due to the SSL training objective, the last layers of wav2vec 2.0 are quite “low-level”
- ➔ **It takes the task-specific model a lot to move these weights.** In their work they showed that randomly re-initializing the last 3 layers of wav2vec 2.0 models for fine-tuning resulted in better ASR.

But… We are in a low-resource setting.
If the middle layers are better informed… **why don’t we just chop the rest?**

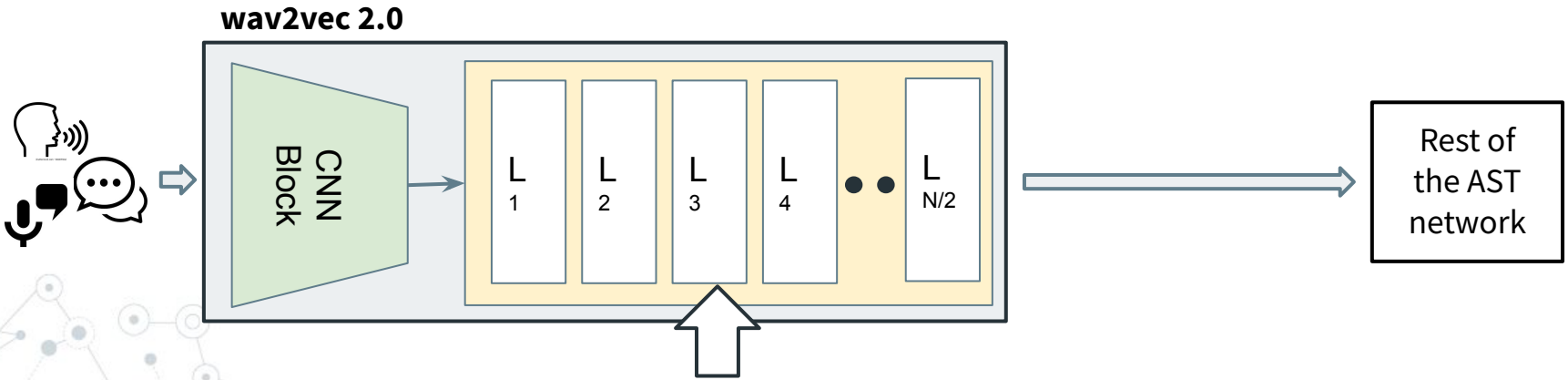


IWSLT 2022:
How to chop your favorite wav2vec 2.0 model

Investigation in Pasad et al. [18] showed that:

- ➔ Due to the SSL training objective, the last layers of wav2vec 2.0 are quite “low-level”
- ➔ **It takes the task-specific model a lot to move these weights.** In their work they showed that randomly re-initializing the last 3 layers of wav2vec 2.0 models for fine-tuning resulted in better ASR.

But… We are in a low-resource setting.
If the middle layers are better informed… **why don’t we just chop the rest?**



IWSLT 2022:
End-to-end Results (still depressing)

Quick reminder: this is still not WER!

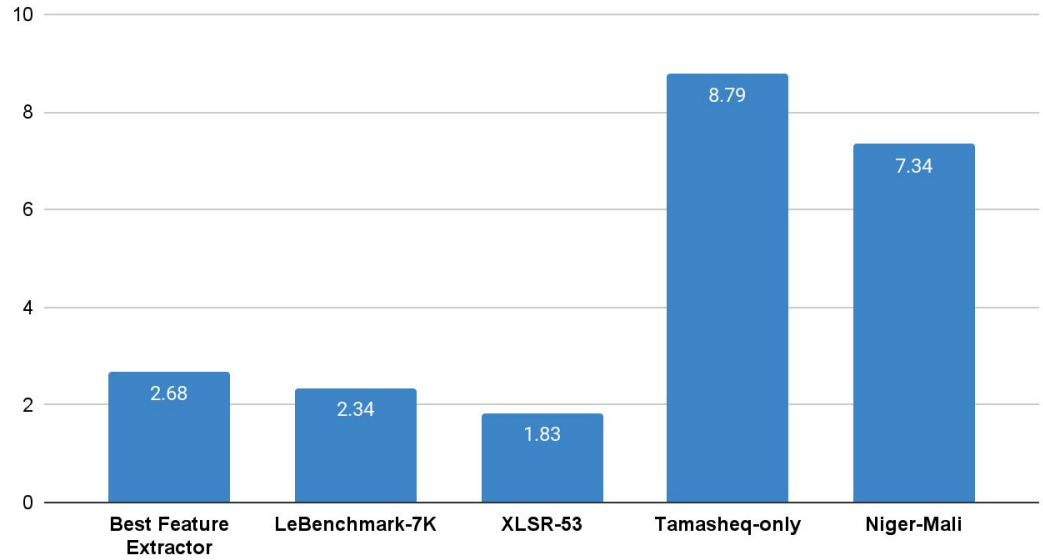


Figure: BLEU scores (dev) for AST Tamasheq-French models using **(HALF of)** different wav2vec 2.0 models

IWSLT 2022:
End-to-end Results (still depressing)

Quick reminder: this is still not WER!

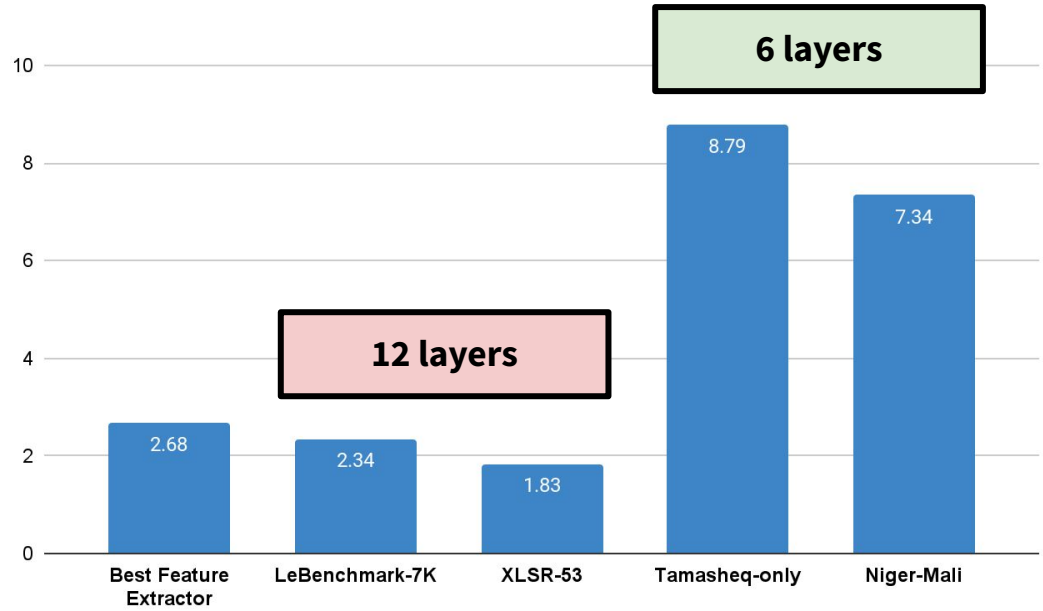


Figure: BLEU scores (dev) for AST Tamasheq-French models using **(HALF of)** different wav2vec 2.0 models

Takeaways:

- Half of a lot is still plenty
- Just pruning large models further resulted in very bad results (explanation in [18])
- **In-domain smaller models are the most effective**

IWSLT 2022: Where we are at in the land of Low-resource Speech Translation

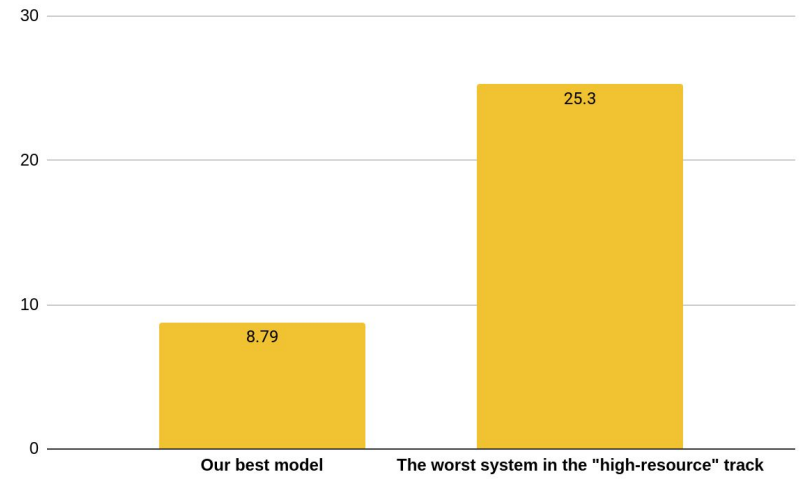


Figure: BLEU scores for our model (Tamasheq-French) and ALEXA-AI submission to Offline (English-Japanese)

Takeaways:

- **Speech Translation is not a solved problem**
  - Maybe *"very clean non-accented without noise English-centered mTEDx speech translation"* is already solved
- Depressingly we won the low-resource track with a model below 10 BLEU
- Total lack of interest from companies in our low-resource IWSLT track (the "just buy more data" mentality is still the norm)
- Small pre-trained models worked better than popular "general-purpose" models trained on thousand of hours
- Working with real data is difficult, there's no magic (box) solution!

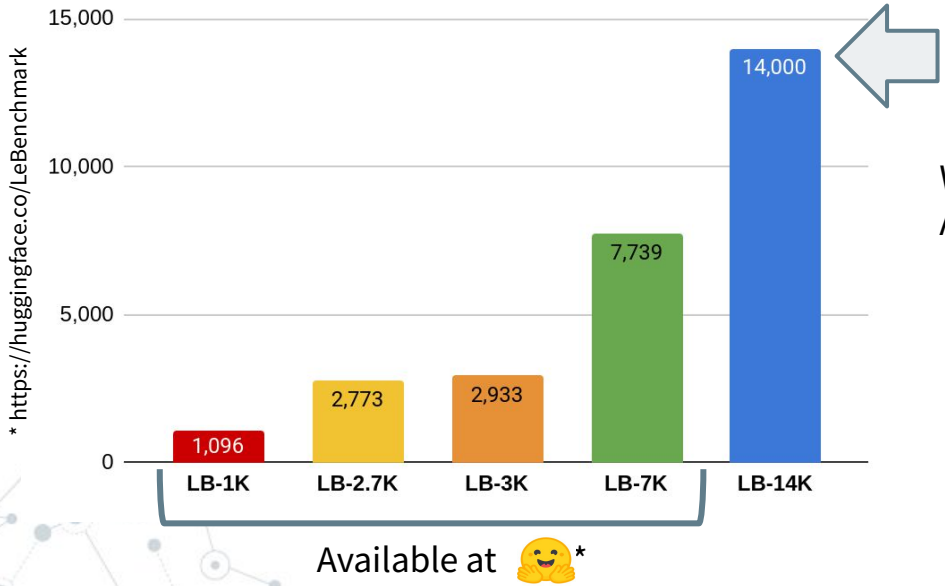
2.

Gender Bias

About huge models and bias in pre-training data

LeBenchmark project: Training and benchmarking wav2vec 2.0 models

- In 2022, we trained and released the *LeBenchmark* models [19, 20]: French wav2vec 2.0 models trained with a lot of diverse audio data



Soon to be released

We *benchmarked* them for ASR, AST, SLU and AER.

- Our results indicate that models trained with more hours of speech produce more robust feature extractors (ASR, AST, SLU, AER).
- They were also superior as speech encoders in ASR.

A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems

- During *LeBenchmark*, we collected rich metadata for the data used during pre-training
- The resulting models present different degrees of speech style and **gender balance**

Does gender distribution in the pre-training data affects ASR/AST models?

ALSO available at 🤗 *

Model	M%	F%	U%
F-1K-Large	-	100	-
M-1K-Large	100	-	-
LB-1K-Large	47.4	52.5	-
LB-3K-Large	62.2	35.2	2.5
LB-7K-Large	23.9	13.4	62.6

* https://huggingface.co/LeBenchmark

Results: wav2vec 2.0 as **feature extractor**

NOTE: Gender balanced mTEDx datasets

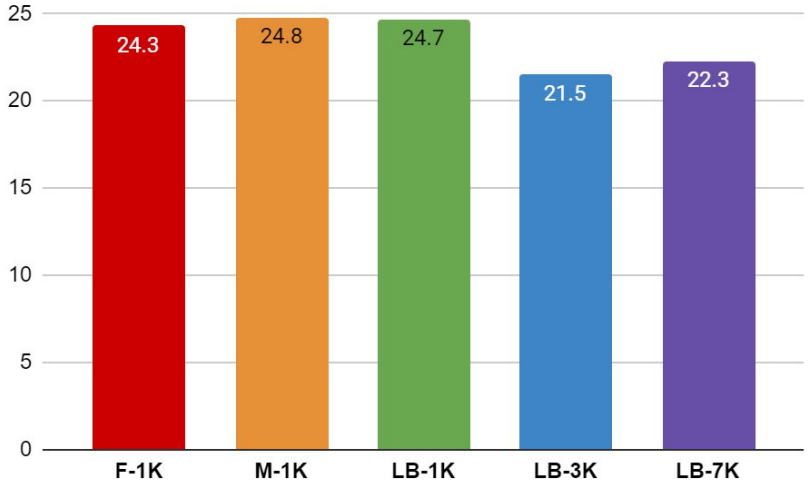


Figure: WER (↓) scores for **ASR** (mTEDx)

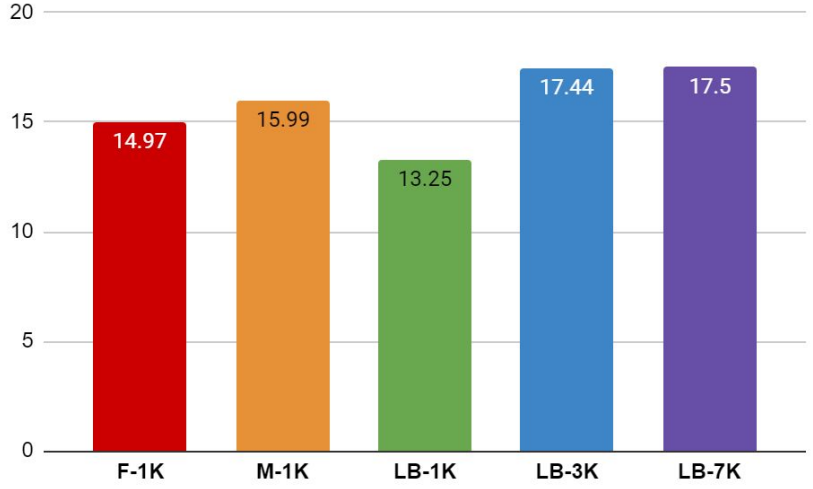


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

Results: wav2vec 2.0 as **feature extractor**

NOTE: Gender balanced mTEDx datasets

→ Gender-specific models are no different than the balanced model

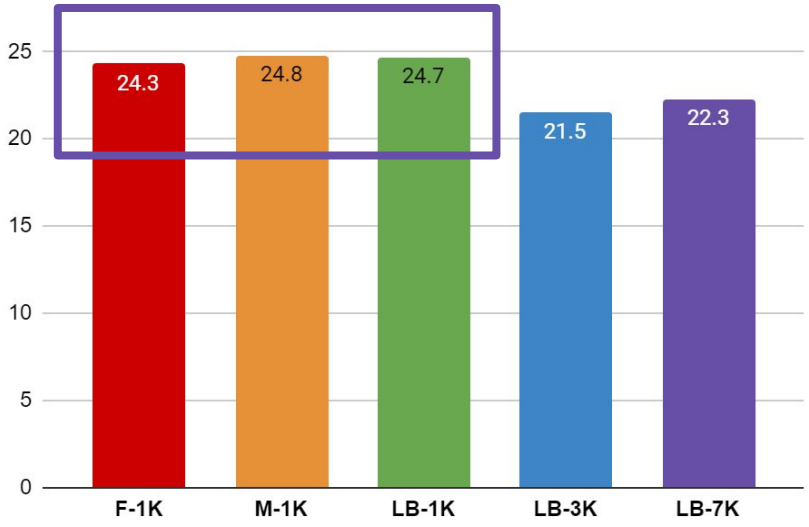


Figure: WER (↓) scores for **ASR** (mTEDx)

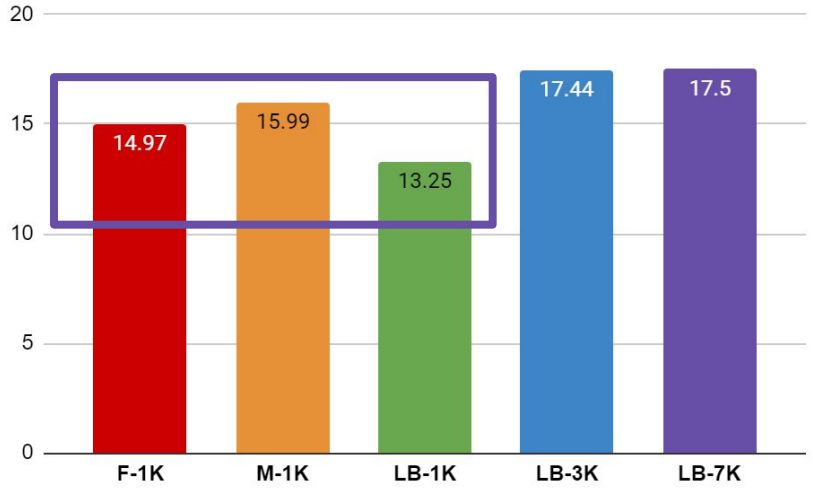


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

Results: wav2vec 2.0 as **feature extractor**

NOTE: Gender balanced mTEDx datasets

- Gender-specific models are no different than the balanced model
- Features seem robust to speaker information

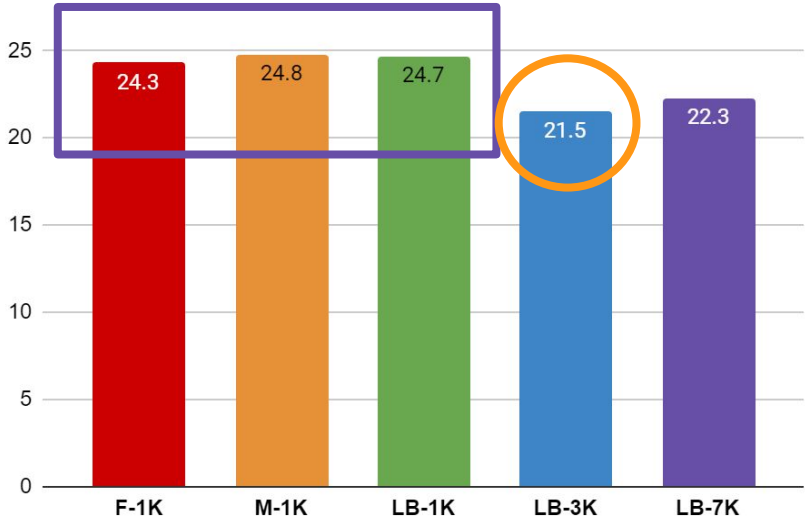


Figure: WER (↓) scores for **ASR** (mTEDx)

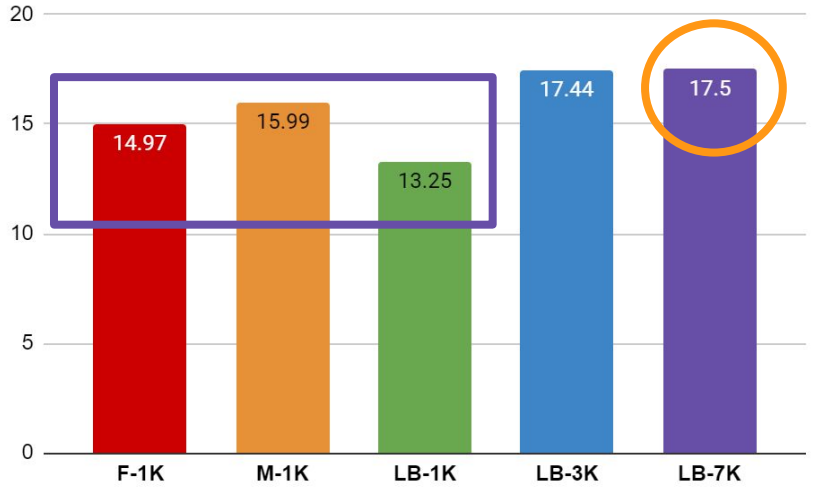


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

Results: wav2vec 2.0 as **speech encoder**

NOTE: Gender balanced mTEDx datasets

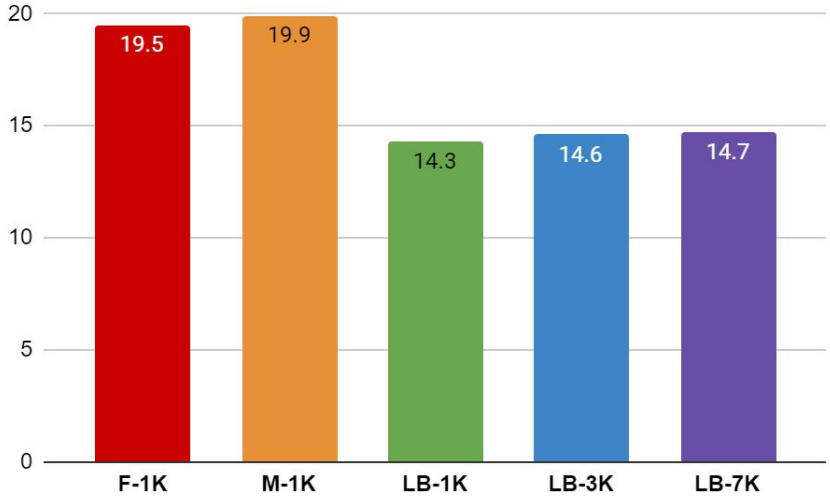


Figure: WER (↓) scores for **ASR** (mTEDx)

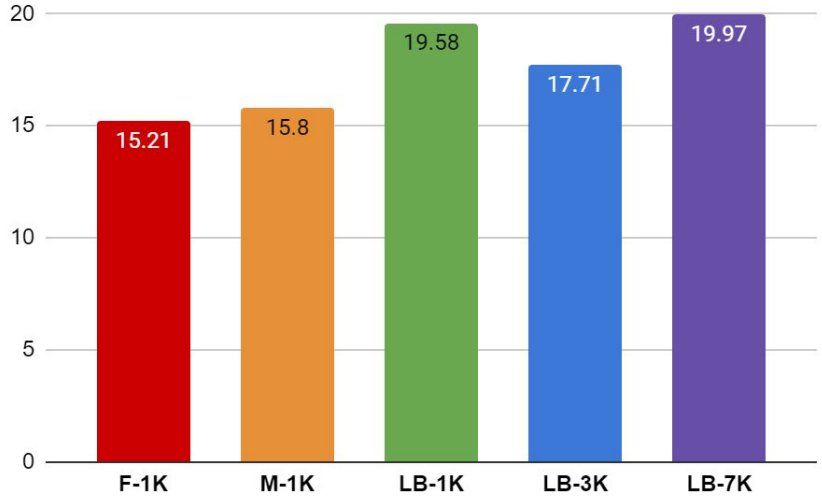


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

Results: wav2vec 2.0 as **speech encoder**

NOTE: Gender balanced mTEDx datasets

➔ Gender-specific Models are bad!

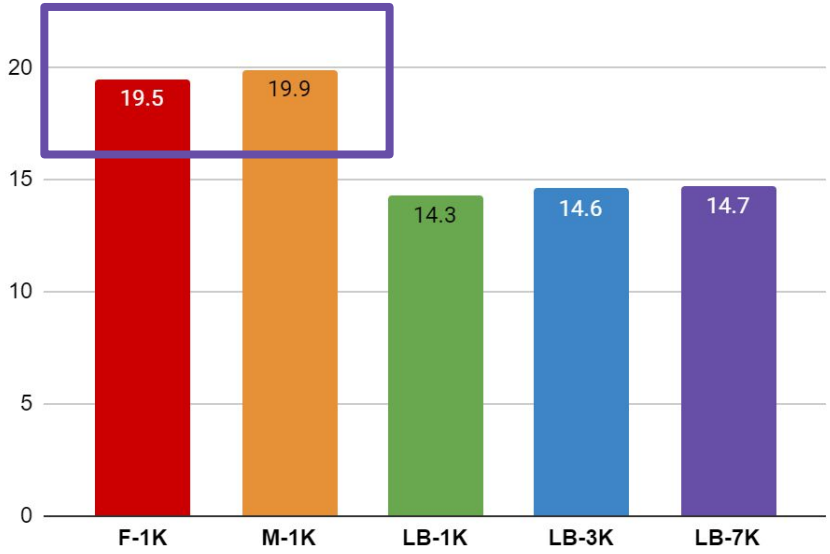


Figure: WER (↓) scores for **ASR** (mTEDx)

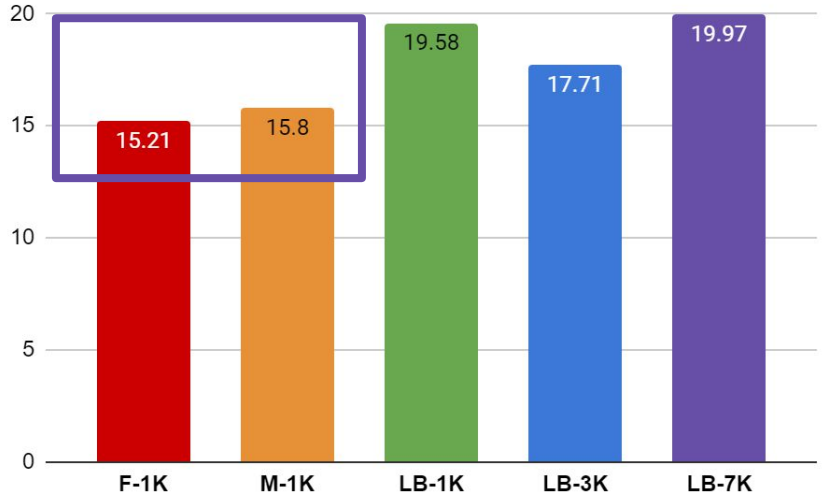


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

Results: wav2vec 2.0 as speech encoder

NOTE: Gender balanced mTEDx datasets

- Gender-specific Models are bad!
- Balanced model is competitive against models trained with 3-7x more data

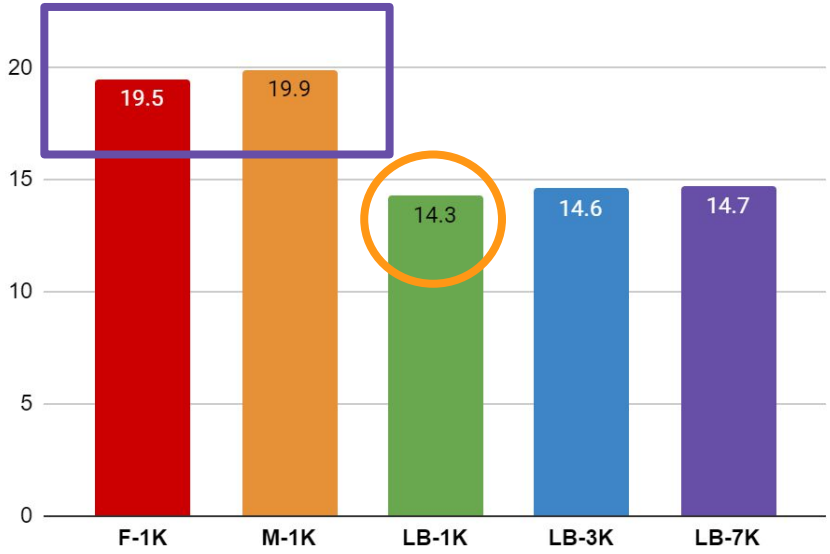


Figure: WER (↓) scores for **ASR** (mTEDx)

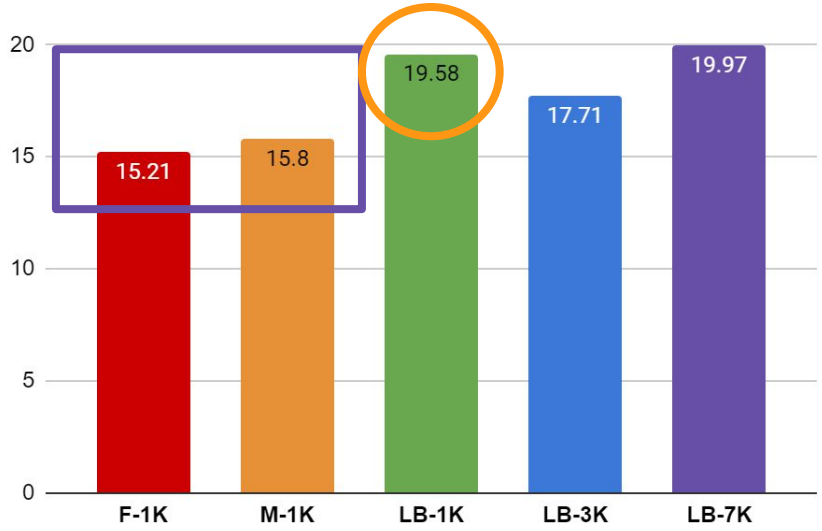


Figure: BLEU (↑) scores for **AST** (fr-en mTEDx)

That's a very steep (investigation) hill!

Takeaways:

- Limited investigation, much to be done still:
  - Better isolation of bias factors (speech style, speaker distribution…)
- Feature Extractors **seem** speaker independent
- Speech Encoders **seem** sensitive to this interference in speaker gender distribution

3.

Where are we again?

And where I would like to go next

TL;DR Before we leave

Low-resource speech discretization for UWS and speech translation:

- Pre-trained models are not silver bullets for every low-resource setting, informed smaller models are sometimes better

Gender Bias investigation:

- We (just barely) scratched the surface, but I’m generally cautious about just inserting “messy” data into huge models!

What I'm excited about right now @ UTTER Horizon Project

Multimodal contextualized speech processing for online meetings

- **Multimodal:** microphone information (who is speaking), video, audio, chat (text) history
- **Contextualized:** speaker information for disambiguation inside speech processing pipelines (Who am I speaking to? Who is this person?)
- **Speech processing:** simultaneous transcription and translation, summarization, automatic minuting

UTTER Horizon Project partners: Edinburgh University (UK), Amsterdam University (NL), NAVER LABS Europe (FR), Instituto de Telecomunicações (PT), Unbabel (PT)

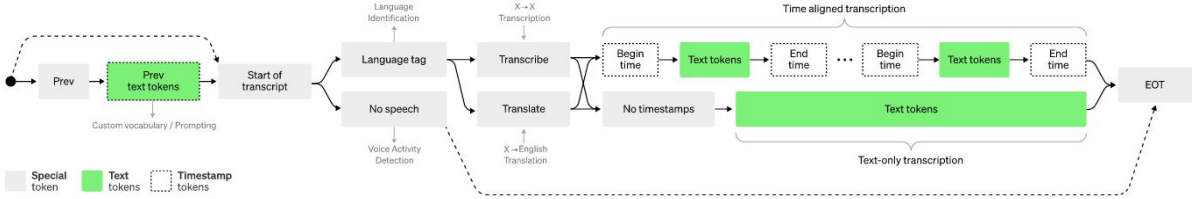
What I'm excited about right now in general

Multilingual and Multi-task training instead of* SSL pre-training

OpenAI

Introducing Whisper

We've trained and are open-sourcing a neural net called Whisper that approaches human level robustness and accuracy on English speech recognition.



https://openai.com/blog/whisper/

Can we guide "SSL" to make it more robust?

Dataset	wav2vec 2.0 Large 960h	Whisper Large	RER (%)
LibriSpeech test-clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.7**	72.7
Fleurs (English)	14.6	**4.6**	68.5
Common Voice	29.9	**9.5**	68.2
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.6**	61.1
WSJ	7.7	**3.1**	59.7
VoxPopuli (English)	17.9	**7.3**	59.2
AMI-IHM	37.0	**16.4**	55.7
CallHome	34.8	**15.8**	54.6
Switchboard	28.3	**13.1**	53.7
CORAAL	38.3	**19.4**	49.3
AMI-SDM1	67.6	**36.9**	45.4
LibriSpeech test-other	6.2	**5.6**	9.7
Average	29.5	**12.9**	55.4

Table 2. **Detailed comparison of robustness on various datasets.** Although both models perform equally well on LibriSpeech, a zero-shot Whisper model performs much better on other datasets than expected for its LibriSpeech performance and makes 55% less errors on average. Results reported in word error rate (WER) for both models after applying our text normalizer.

* but really, why not both!

Who is “we”?

Thanks to all my *hike buddies*!

- Speech Discretization for Unsupervised Word Segmentation [21]
  - UGA, Sheffield University, Brno University, LISN
- IWSLT 2022: Direct Speech Translation [22, 16]
  - ON-TRAC Consortium: LIA, UGA, LIUM, Airbus, ELYADATA
- *LeBenchmark* Project [19, 20]
  - UGA, LIA, Atos, ESPCI PSL, NAVER LABS Europe
- Gender Investigation [24]
  - NAVER LABS Europe, LIA

Thanks **JeanZay** for the 20,000 GPU hours I *burned* on your servers in 2021-2022!!!

Thanks!

Any questions?

You can find me at:

marcely.zanon-boito@naverlabs.com

BIBLIOGRAPHY

[1] S. Schneider et al., “wav2vec: Unsupervised pre-training for speech recognition,” arXiv preprint arXiv:1904.05862, 2019.
[2] W.-N. Hsu, B. Bolte et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021
[3] A. Baevski, M. Auli, and A. Mohamed, “Effectiveness of self-supervised pre-training for speech recognition,” arXiv preprint arXiv:1911.03912, 2019.
[4] A. Baevski, Y. Zhou et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, 2020.
[5] A. Babu, C. Wang et al., “Xls-r: Self-supervised cross lingual speech representation learning at scale,” arXiv preprint arXiv:2111.09296, 2021.
[6] Joshi, et al. The state and fate of linguistic diversity and inclusion in the NLP world. ACL 2020.
[7] S. Bird, Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. Linguistic Issues in Language Technology, vol. 6, no. 4, 2011
[8] Adda et al. Breaking the unwritten language barrier: The BULB project. SLTU 2016.
[9] Ondel et al. Variational inference for acoustic unit discovery. Procedia Computer Science 2016.
[10] Ondel et al. Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery. Interspeech 2019.
[11] Yusuf et al. A Hierarchical Subspace Model for Language-Attuned Acoustic Unit Discovery. ICASSP 2020.
[12] Oord et al. Neural Discrete Representation Learning. NeurIPS 2017.
[13] Baevski et al. vq-wav2vec: Self-supervised Learning of Discrete Speech Representations. arXiv, 2019.
[14] Godard et al. A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments. LREC 2018.
[15] Kamper and Nieker. Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks. arXiv, 2020.
[16]Anastasopoulos et al. Findings of the IWSLT 2022 Evaluation Campaign. IWSLT 2022.
[17] Boito et al. Speech resources in the tamasheq language. LREC 2022.
[18] Pasad et al. Layer-wise analysis of a self-supervised speech representation model. arXiv preprint arXiv:2107.04734.
[19] Evain et al. "Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech." Interspeech 2021.
[20] Evain et al. Task agnostic and task specific self-supervised learning from speech with LeBenchmark. NeurIPS 2022.
[21] Boito et al. Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings. SIGUL 2022.
[23] Boito et al. ON-TRAC Consortium Systems for the IWSLT 2022 Dialect and Low-resource Speech Translation Tasks. IWSLT 2022.
[24] Maison et al. Promises and Limitations of Self-supervised Learning for Automatic Speech Processing. CAID 2022.

IWSLT 2022:
End-to-End Results: Playing with the threshold

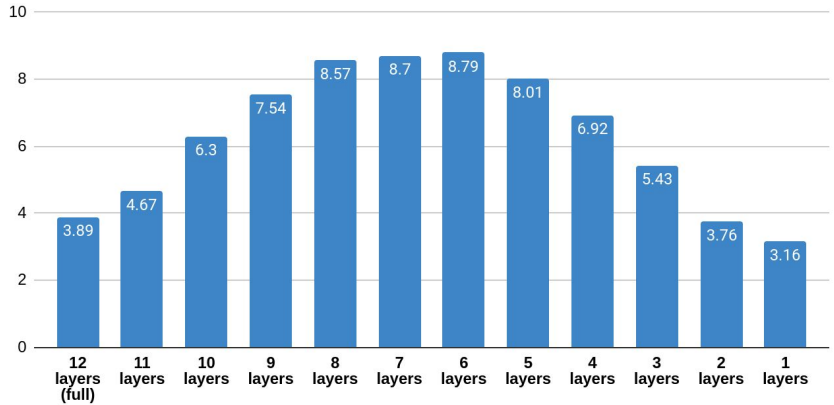


Figure: BLEU scores (dev) for AST Tamasheq-French models using different number of wav2vec 2.0 encoder layers (Tamasheq-only model)

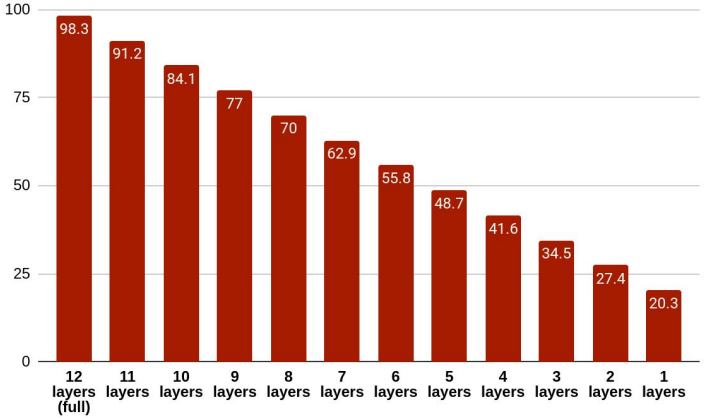


Figure: Number of parameters in millions for training a AST model with a wav2vec 2.0 base

IWSLT 2022:
End-to-End Results: Playing with the threshold

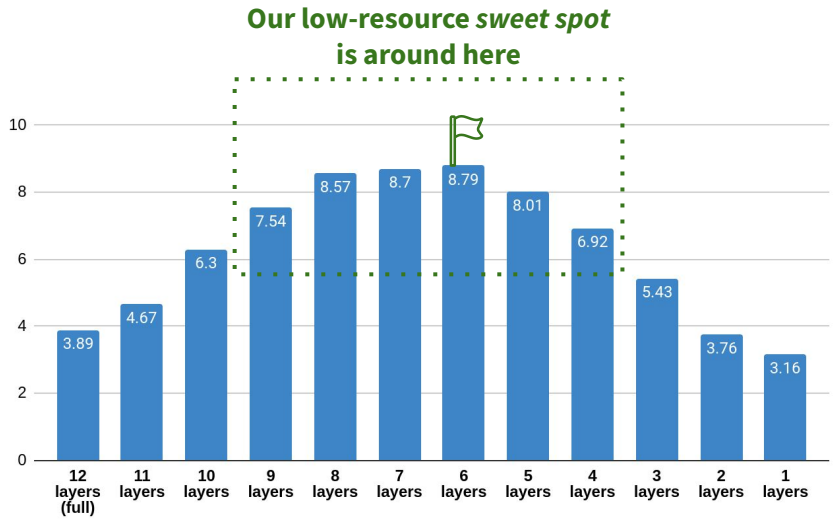


Figure: BLEU scores (dev) for AST Tamasheq-French models using different number of wav2vec 2.0 encoder layers (Tamasheq-only model)

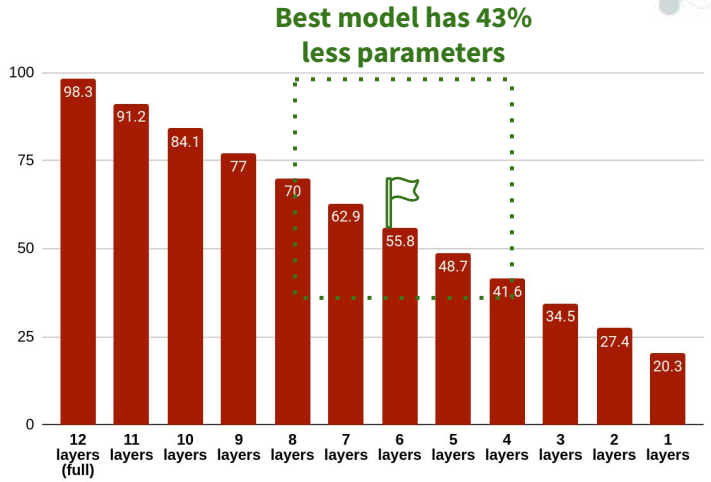


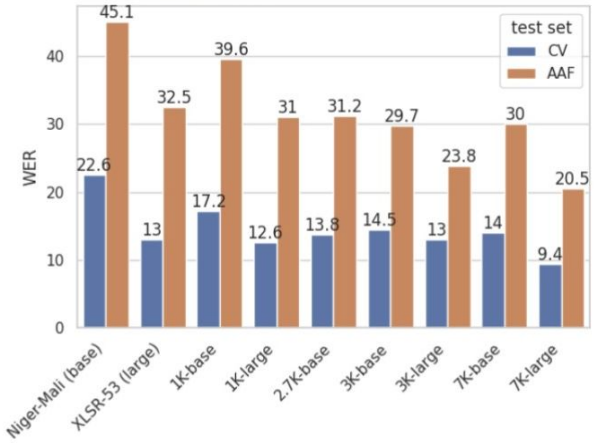
Figure: Number of parameters in millions for training a AST model with a wav2vec 2.0 base

Can we fix bias in pre-training?

Work from Lucas Maison on accented ASR from our CAID paper this year [24]

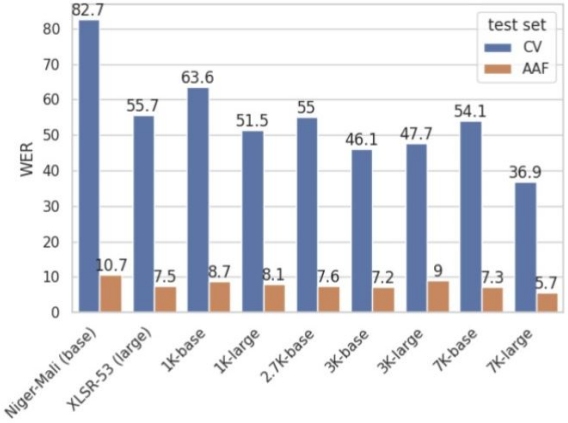
Non-accented fine-tuning

Fig. 2. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on CommonVoice.



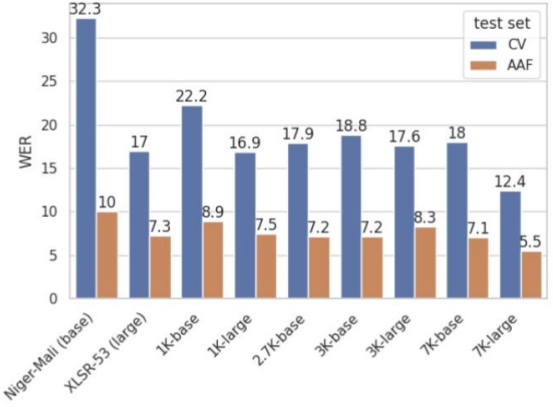
Accented fine-tuning

Fig. 3. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on African Accented French.



50/50 fine-tuning

Fig. 4. ASR results (WER, the lower the better) over the two test sets for models fine-tuned on a mixed dataset (CV+AAF).



BLUE: Standard French speech
ORANGE: African Accented speech

*The scale is a mess, but the message is:
50/50 fine-tuning helps recover the performance in non-accented speech, while increasing performance for accented speech

The connection: Self-supervised Speech Processing blocks

These models emerged as popular foundation blocks for speech pipelines. They are trained using only audio examples and can reach the billions of parameters [1-5].

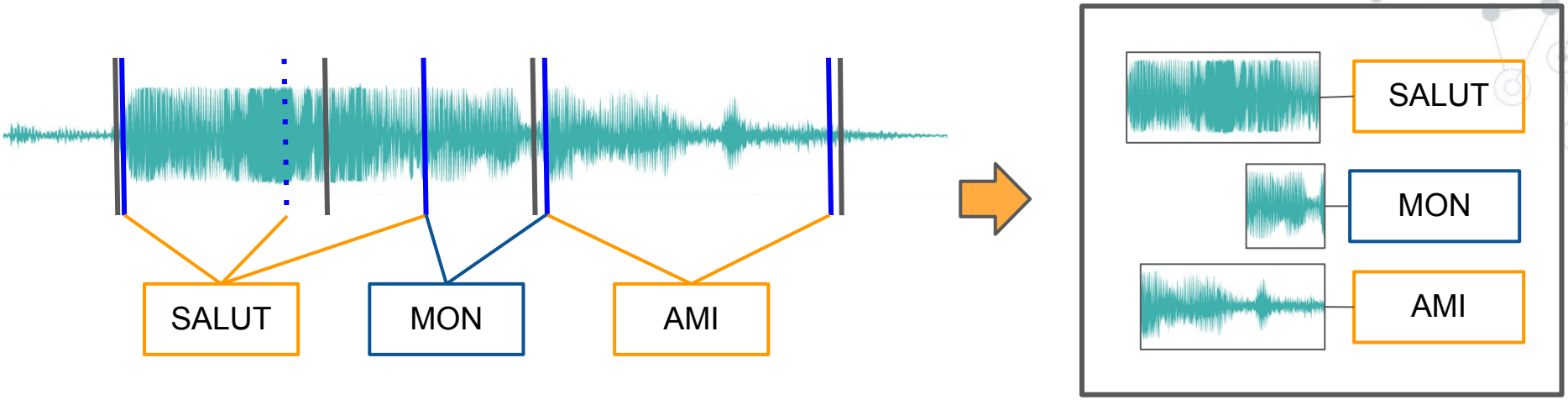
"Tons" of speech data

Rich and Contextualized Speech Representation (continuous or discrete)

1. Pre-training from scratch if:
   a. New data/domain/language/format
   b. You have between 16 and 48 A100 32GB to "waste" for a month

Otherwise, general purpose models available at HuggingFace 🤗 or GitHub

A pipeline for UWS in low-resource settings:
Grounding segmentation on translation



➔ In this setting, all our boundaries have an *annotation*: the aligned bilingual information.

Speech Discretization in the Low-resource Land

- Bayesian Generative Models (AUD):
    1. HMM/GMM **(HMM)**: Every possible sound can be a unit [9]
    2. Subspace HMM **(SHMM)**: Prior over a phonetic subspace [10]
    3. Hierarchical Subspace HMM **(H-SHMM)**: Subspace adaptation from different languages for unit prediction [11]