

Learning Sentence-Level Speech Representations

From Semantics to Multidimensional Speech Modeling

Yannick ESTÈVE

LIA - Computer Science Laboratory of Avignon
Speech and Language Group

Conversational AI Reading Group at Mila - Fall 2026 - 22 October 2026

One utterance contains several kinds of information

“The meeting has been moved to tomorrow.”

Meaning

What is being said?

Speaker

Who is speaking?

Language

Which language?

Emotion

How is it expressed?

Utterance-level representations summarize a speech segment into compact vectors.

A research programme built around the SENSE framework (available in SpeechBrain)

01 Semantic alignment

SAMU-XLSR, SONAR and SENSE

02 Cross-lingual summarization

Chaimae Chellaf · LIA / Lundi Matin



03 Language specialization

French and Arabic dialects · Salima Mdhaffar



04 Multiple attributes

Ongoing work · Maryem Bouziane

01 Semantic alignment

SENSE: Shared Embedding for N-lingual Speech and tExt

Mdhaffar, S., Elleuch, H., Chellaf, C., Nguyen, H., & Estève, Y. (2025). ***SENSE models: an open source solution for multilingual and multimodal semantic-based tasks***. In 2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8).

Word-level semantics

- A lot of works aims to join learn or to align continuous representations at the word level

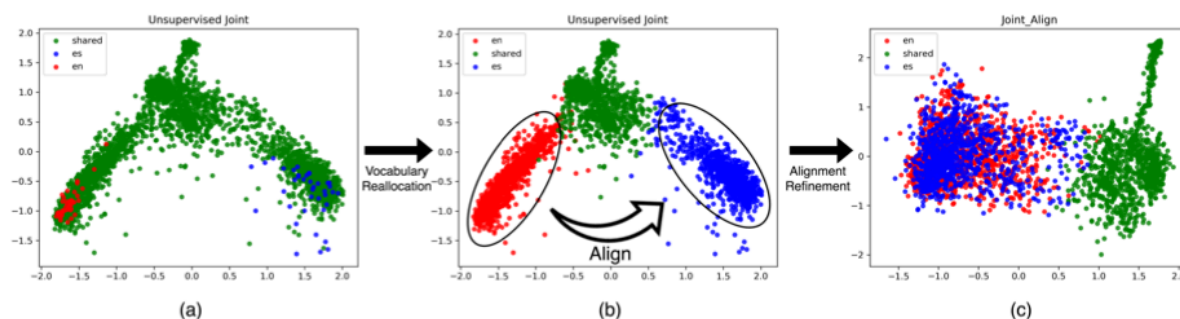


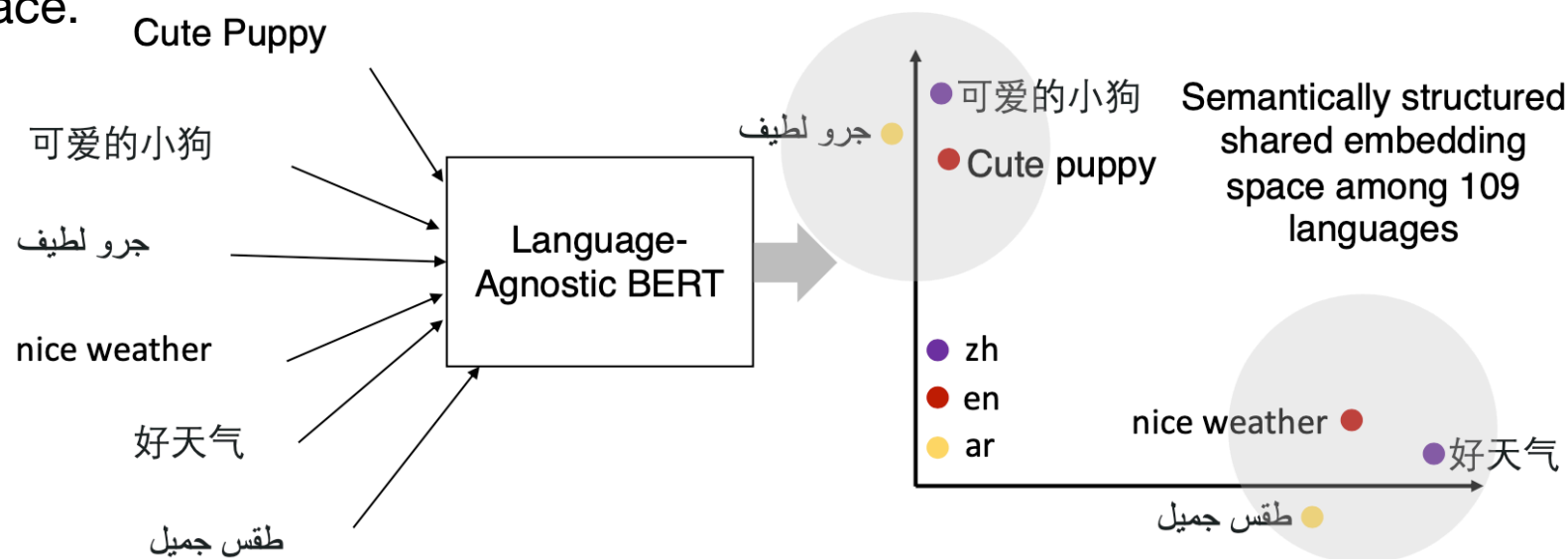
Figure 1: PCA visualization of English and Spanish embeddings learnt by unsupervised joint training as in Lample et al. (2018b). As shown by plots (a) and (b), most words are shared in the initial embedding space but not well-aligned, hence the oversharing problem. Plots (b) and (c) shows that the vocabulary reallocation step effectively mitigates oversharing while the alignment refinement step further improves the poorly aligned embeddings by projecting them into a close neighborhood.

Wang, Z., Xie, J., Xu, R., Yang, Y., Neubig, G., & Carbonell, J. G. (2019) **Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unified Framework**. In International Conference on Learning Representations.

Sentence-level semantics

LaBSE: Language-agnostic BERT Sentence Embedding generator

- BERT-like model trained using both masked and translation language modeling objective functions
- **LaBSE** embeds sentences from 109 languages into a shared semantically-aligned embedding vector space.



Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). *Language-agnostic BERT Sentence Embedding*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 878-891)

A frozen text teacher defines the semantic space

The idea: post-train an SSL speech encoder to project a utterance to the existing semantic space of a sentence-level embedding generator. *Text-speech segment pairs are needed*

SAMU-XLSR: the starting point (Khurana, Laurent & Glass, 2022)

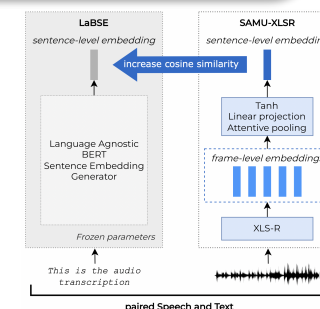
Speech student

XLS-R backbone
Attention pooling
Utterance-level output

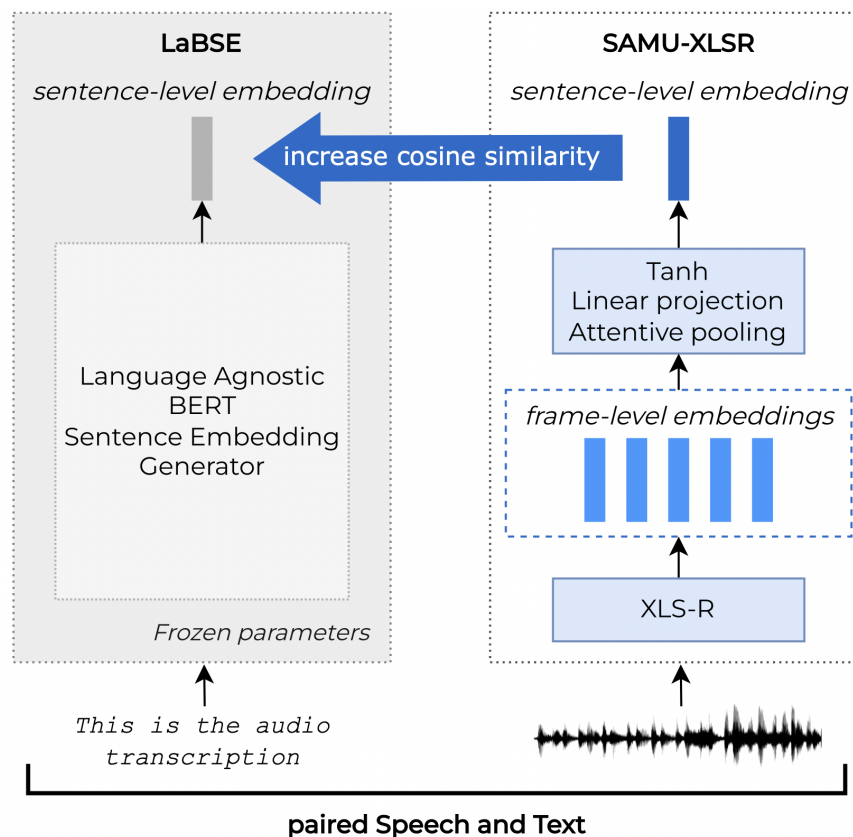
Semantic teacher

LaBSE sentence embeddings
Shared multilingual text space
Frozen teacher

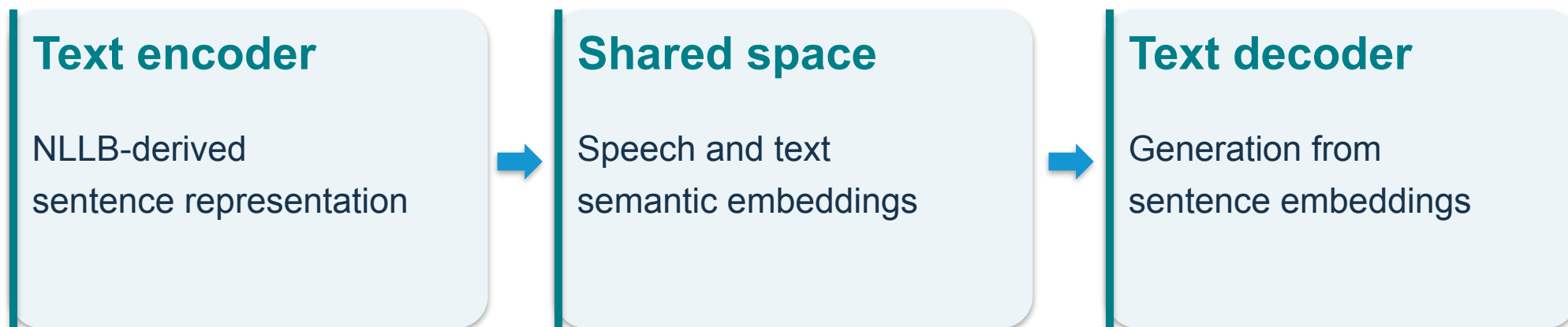
53 languages · Common Voice 8 · approximately 12.7K hours



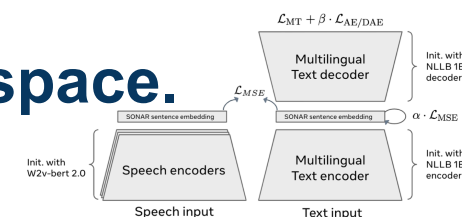
SAMU-XLSR: the starting point (Khurana, Laurent & Glass, 2022)



SONAR: a multilingual space for text and speech

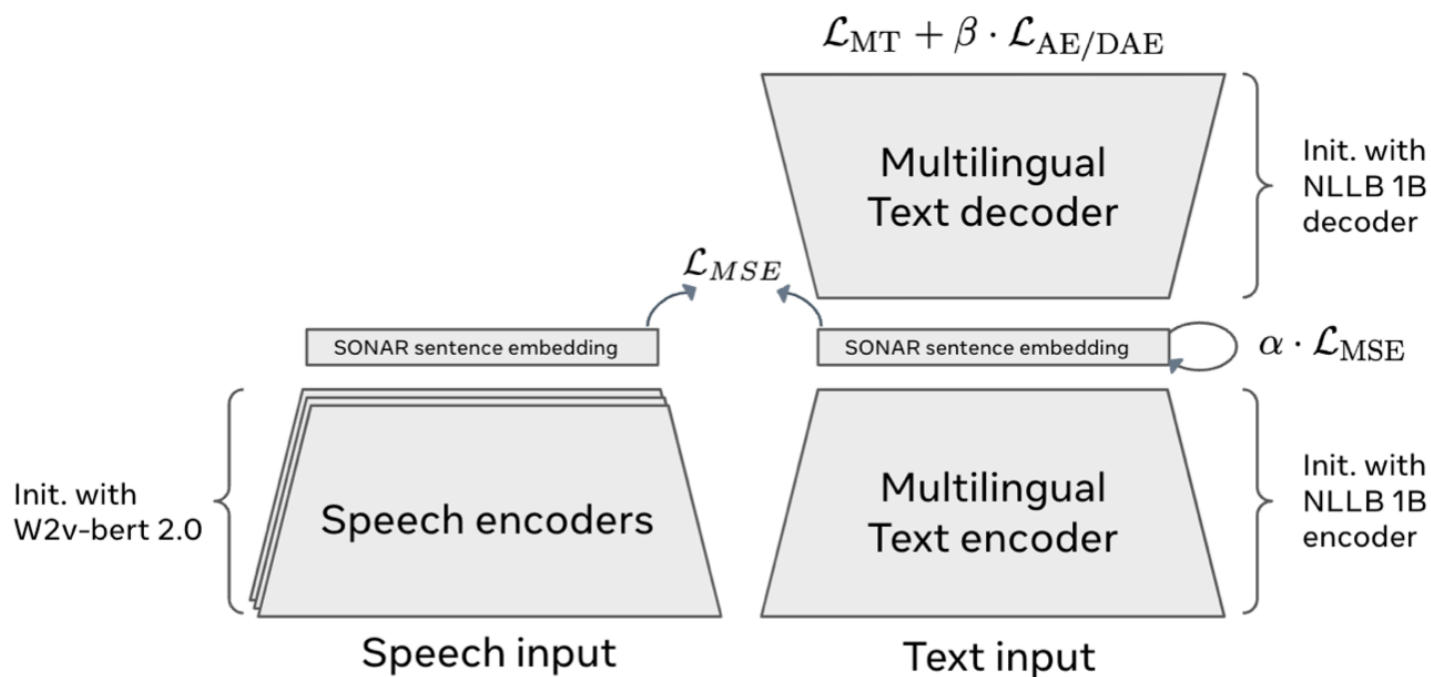


Speech encoders learn to enter the same multilingual space.



A comparable goal, with a different semantic teacher and representation space.

SONAR: a multilingual space for text and speech



SONAR architecture.

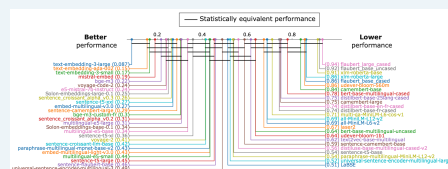
SENSE: an open framework and multilingual model

A speech backbone

w2v-BERT 2.0
Attention + projection

A stronger text teacher

BGE-M3
Multilingual semantic targets



An open implementation

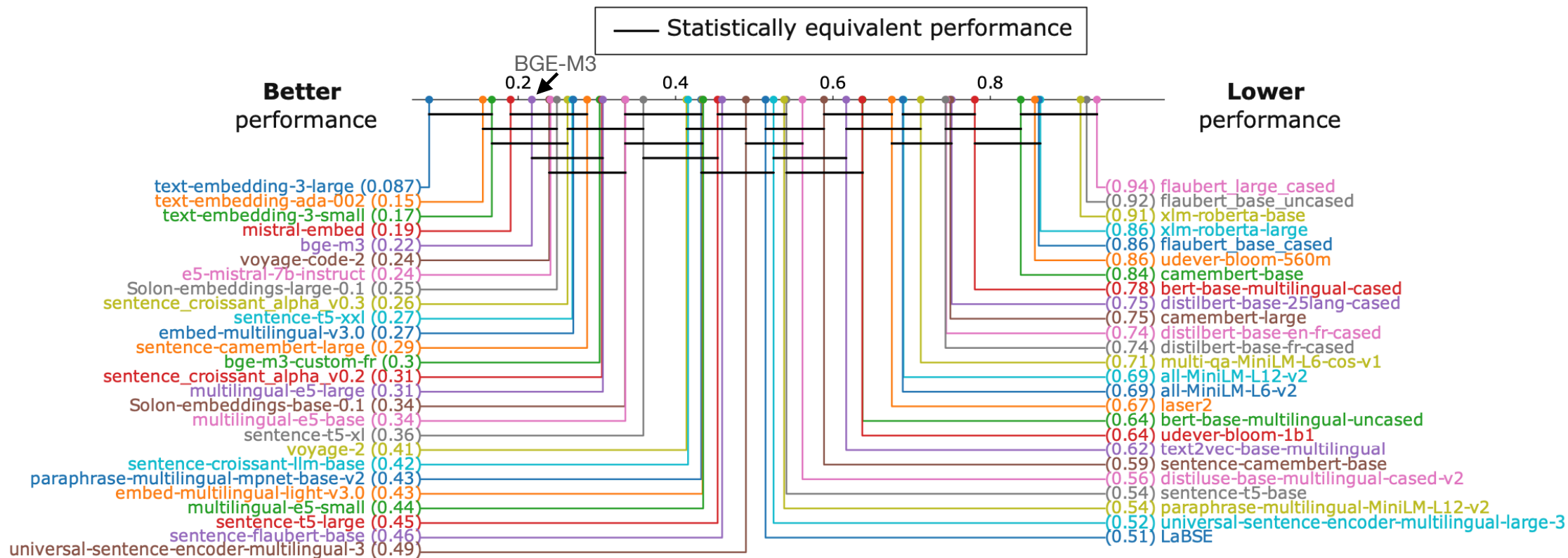
SpeechBrain recipes
Model on Hugging Face



Validated Common Voice 19 data · 83 languages

SENSE is a framework: the teacher, backbone and training data can evolve.

SENSE: an open framework and multilingual model

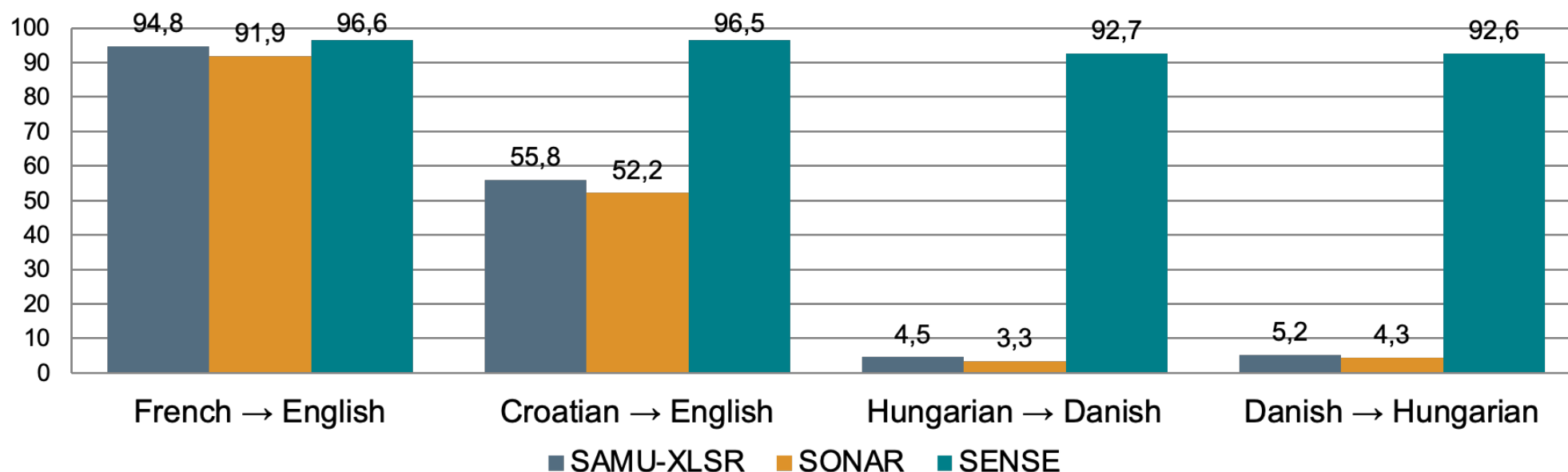


A stronger text teacher

Different teachers and backbones

Framework	Speech initialization	Semantic target
SAMU-XLSR	XLS-R	LaBSE
SONAR	w2v-BERT 2.0	SONAR text space (NLLB-derived)
First SENSE model	w2v-BERT 2.0	BGE-M3

SENSE is stronger, e.g. on speech-to-speech retrieval



VoxPopuli · Recall@1 (%) · selected pairs; the full table is in backup.

Semantic post-training also helps speech translation

Encoder initialization	FR → EN	DE → EN	AR → EN	NL → EN
SAMU-XLSR	20.56	19.01	14.80	6.29
w2v-BERT 2.0	21.02	20.40	0.65	3.69
SENSE	21.67	21.95	15.82	15.00

CoVoST-2 · BLEU ↑ · selected directions after task-specific fine-tuning

Two uses: extract semantic vectors, or *initialize an encoder for task-specific fine-tuning (here)*

3 people

Adding SENSE models (#2998)

aca5e41 · 8 months ago

History

Name	Last commit message	Last commit date
..		
hparams	Adding SENSE models (#2998)	8 months ago
README.md	Adding SENSE models (#2998)	8 months ago
common_voice_sense_prepare.py	Adding SENSE models (#2998)	8 months ago
train.py	Adding SENSE models (#2998)	8 months ago

Multilingual SENSE on Common Voice

This folder contains scripts to train **SENSE** models that align a self-supervised speech encoder (wav2vec-BERT) with multilingual text embeddings (BGE-M3) on the [Common Voice](#) corpus. The resulting speech embeddings live in a shared semantic space, similar in spirit to **SAMU-XLSR** and **Meta's SONAR** multilingual speech-text encoders. They follow the SENSE framework ("*SENSE models: an open source solution for multilingual and multimodal semantic-based tasks*") described in <https://arxiv.org/abs/2509.12093>.

A single multilingual model is trained on 90 Common Voice languages.

Two components are used:

1. A **student audio encoder**: wav2vec-BERT followed by attention pooling to obtain utterance-level embeddings.
2. A **teacher text encoder**: BGE-M3 sentence embeddings computed on-the-fly from the reference transcripts.

huggingface.co

 **Hugging Face**

Models

Datasets

Spaces

Buckets NEW

Docs

Pricing



LIA-AvignonUniversity/SENSE

Like

2

Following

Laboratoire Informa...

14

Safetensors

sense

custom_code

arxiv:2509.12093

Model card

Files and versions

xet

Community

Settings

Copy to bucket

NEW

Edit model card

YAML Metadata Warning:empty or missing yaml metadata in repo card

Check out the [documentation](#) for more information.

SENSE multilingual semantic speech embeddings

This repository provides a pretrained **SENSE** model for extracting multilingual semantic speech embeddings.

The model was trained on **90 Common Voice languages** following the SENSE framework described in:

SENSE models: an open source solution for multilingual and multimodal semantic-based tasks

<https://arxiv.org/abs/2509.12093>

Pipeline description

SENSE produces utterance-level semantic speech embeddings from 16 kHz audio.

Compute SENSE embeddings

```
import torch
from transformers import AutoFeatureExtractor, AutoModel
import soundfile as sf
```

Downloads last month

220

[View full history](#)

Safetensors

Model size

0.6B params

Tensor type

F32

Files info

Inference Providers

This model isn't deployed by any Inference Provider.

Ask for provider support

Paper for LIA-AvignonUniversity/SENSE

SENSE models: an open source solution for multilingual and multim...

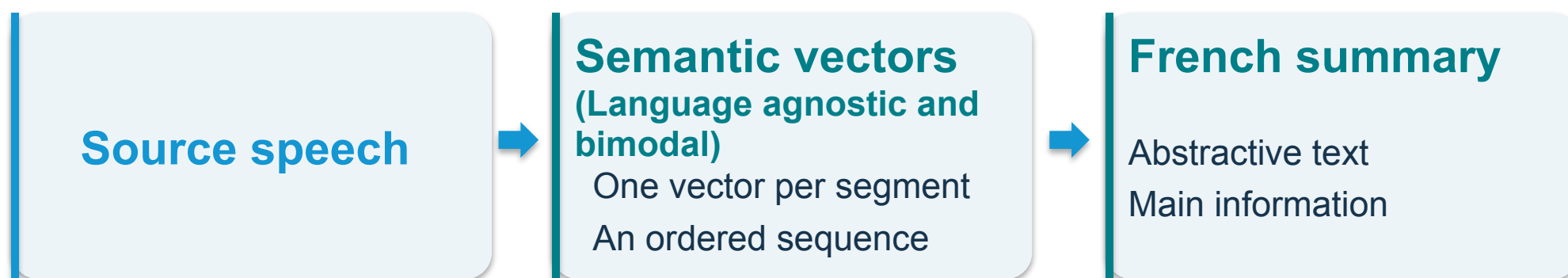
Paper • 2509.12093 • Published Dec 8, 2025

02 Cross-lingual summarization

Cross-lingual **and bimodal!**

Chellaf, C., Mdhaffar, S., Estève, Y., & Huet, S. (2026). ***Bridging Languages and Modalities: Lightweight Cross-Lingual Text and Speech Summarization for Low-Resource Scenarios***. In Interspeech 2026.

Can we summarize speech without transcribing it?



No ASR transcript or explicit translation step in this inference pipeline.

Can we summarize speech without transcribing it?

An first answer is given in: Barrault, L., Duquenne, P. A., Elbayad, M., Kozhevnikov, A., Alastruey, B., Andrews, P., ... & Schwenk, H. (2024). *Large concept models: Language modeling in a sentence representation space*. arXiv preprint arXiv:2412.08821.

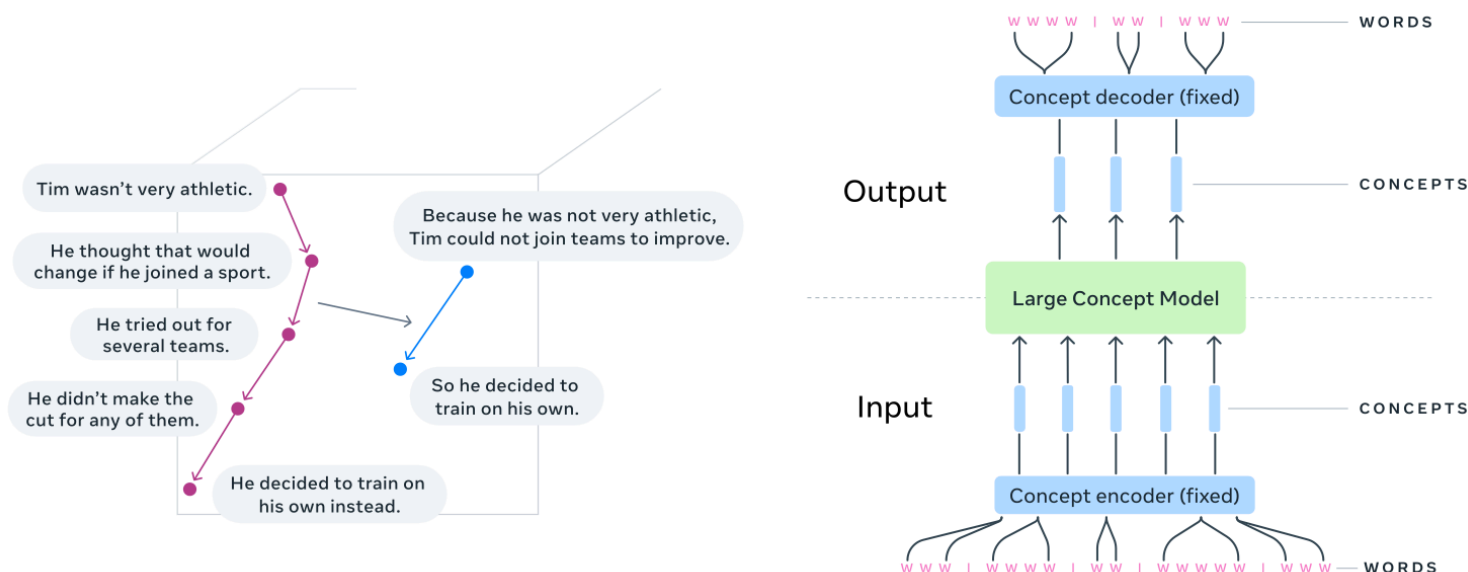
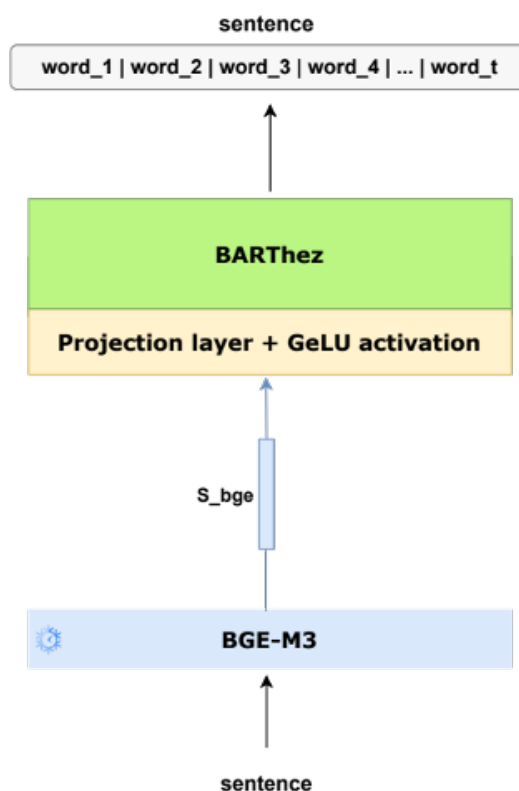


Figure 1 - Left: visualization of reasoning in an embedding space of concepts (task of summarization). Right: fundamental architecture of an LARGE CONCEPT MODEL (LCM).
 *: concept encoder and decoder are frozen.

SBARThez consumes a sequence of semantic vectors



We can use BGE embeddings (text) or SENSE (speech): they share the same bimodal and cross lingual space!

The input units (utterance-level semantic vectors) change while generation remains token-based.

Train on text, then transfer or adapt to speech

Stage 1 · Learn to summarize semantic vectors

French MLSUM · BGE-M3 sentence vectors → French summaries

Stage 2a · Text-only adaptation

Text vectors from translations

At test time: speech vectors

Stage 2b · Speech adaptation

SENSE speech vectors

At test time: speech vectors

Frozen embedding encoders · 140M trainable summarizer parameters in the manuscript.

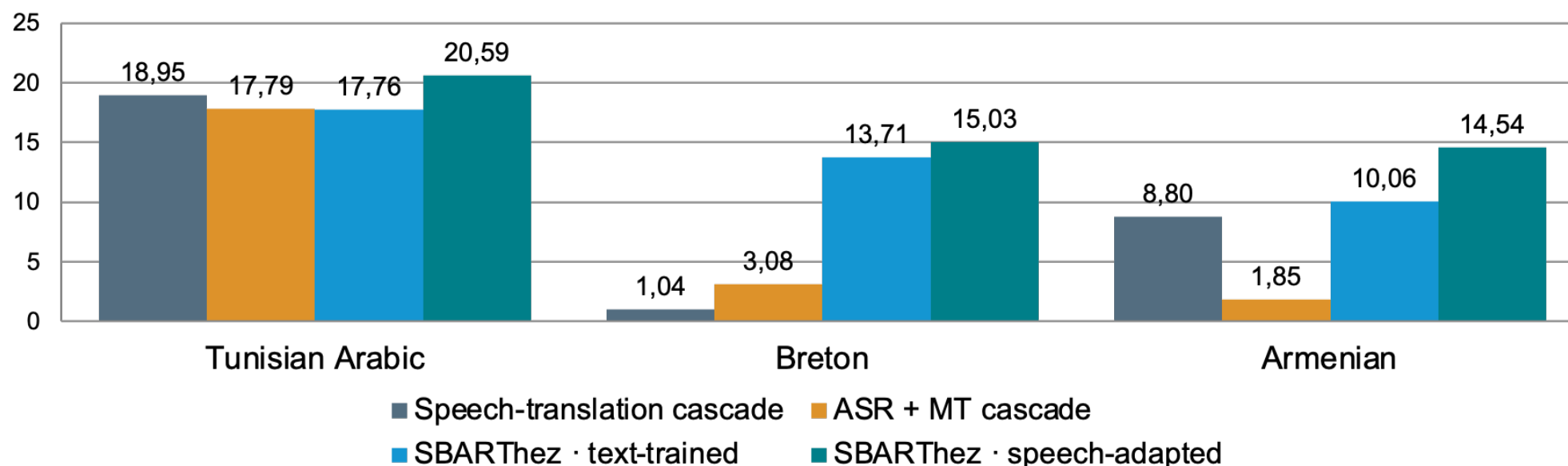
ABT-SpeechSUM: three low-resource speech settings

Source language	Domain	Total audio	Test recordings
Tunisian Arabic	TEDx talks	27 h 18 min	15
Breton	Tales / narratives	47 min	15
Armenian	Audiobooks / tales	6 h 41 min	10

French reference summaries

Generated from translations; test summaries manually checked.

Speech adaptation improves all three test settings



ABT-SpeechSUM · ROUGE-L ↑ · small test sets and translated reference construction.

03 Language specialization

For French and Arabic dialects

Salima Mdhaffar, Valentin Pelloin, Chaimae Chellaf, Yannick Estève. 2026. ***Semantic Speech Representation Learning from Large Unlabeled Monolingual Speech Collections***. Accepted at IEEE SLT 2026 (December)

French specialization: training a SENSE model from INA audio archives

INA archives

≈473K source hours
113 TV / radio channels

Selected speech

≈100K hours
≈12M 30-second chunks

French student

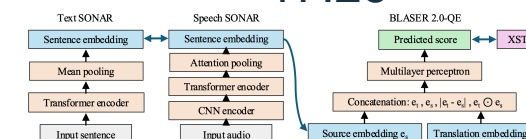
Pantagruel data2vec 2.0
300M-parameter backbone

Whisper pseudo-transcripts → BGE-M3 targets → semantic distillation

Learn semantic representations from unlabeled archives using automatic supervision.

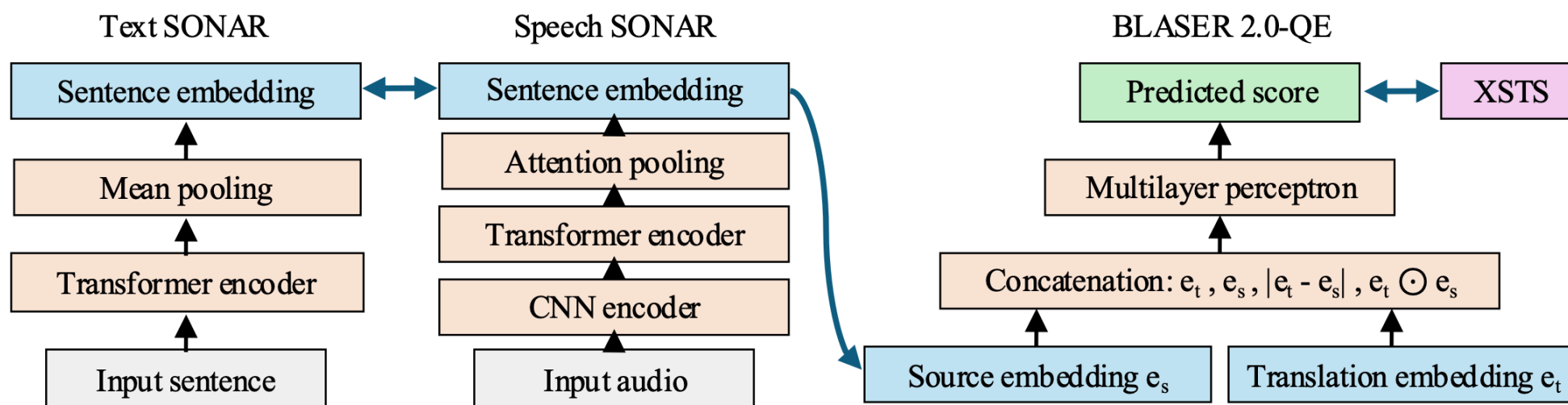
Semantic filtering helps

Training subset	FR → EN R@1 ↑	Topic micro-F1 ↑	Summary R-L ↑
100K hours: ALL	86.40	59.4	17.32
52K: score > 3.5	87.49	61.2	19.24
5K: best scores	87.49	60.8	17.96
5K: worst scores	68.16	27.3	17.25



Semantic filtering has really a positive impact (here we used BLASER 2.0)

Filtering helps more than keeping every pseudo-label



Left: SONAR architecture of SONAR.

Right: BLASER 2.0 architecture

A smaller French backbone remains competitive

Model / supervision	Backbone	FR → EN R@1 ↑	Topic micro-F1 ↑	Summary R-L ↑
Original SENSE	600M	87.01	57.1	21.51
French · 52K-aligned	300M	89.40	59.3	21.12
French · ALL-10s-noTRS	300M	87.73	57.6	22.80

Automatic text supervision and speech supervision offer different task trade-offs.

Arabic dialects: use English to build semantic targets

Dialectal transcript

Limited resources
Orthographic variation



English translation

Allam-7B
High-resource pivot



BGE-M3 target

Frozen semantic teacher
Speech student learns it

≈4,335 training hours - MSA + multiple dialects - rebalanced sampling

English is a training-time pivot; inference maps dialectal speech directly to vectors.

English-derived targets improve dialectal retrieval

Dialect	SONAR-ARB	SENSE AR targets	SENSE EN targets
Algerian	47.96	57.69	71.75
Egyptian	69.30	79.98	83.74
Jordanian	88.21	87.62	94.22
Mauritanian*	26.48	48.42	53.80
Moroccan	36.17	48.71	68.33
Palestinian	80.21	81.11	87.11
Emirati	57.14	75.96	85.82
Yemeni	64.63	77.58	81.57

Speech → English text · R@1 (%) · *unseen during semantic encoder training.

04 Multiple attributes

Learning Speech Attribute Representations via Multi-Teacher Distillation

Bouziane, M., Mdhaftar, S., & Estève, Y. (2026). *Learning Speech Attribute Representations via Multi-Teacher Distillation*. arXiv preprint arXiv:2610.21941.

Bouziane, M., Mdhaftar, S., & Estève, Y. (2026). *Learning Multiple Utterance-Level Attribute Representations with a Unified Speech Encoder*. Interspeech 2026

Beyond semantics: preserve complementary attributes

One shared speech encoder

Reuse the SSL backbone across all four attributes.

Four attribute-specific outputs

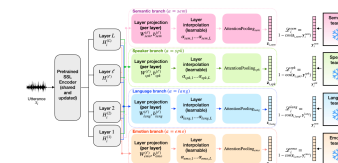
Semantics, speaker identity, language identity and emotion.

Four frozen teachers

Each teacher defines its own target representation space.

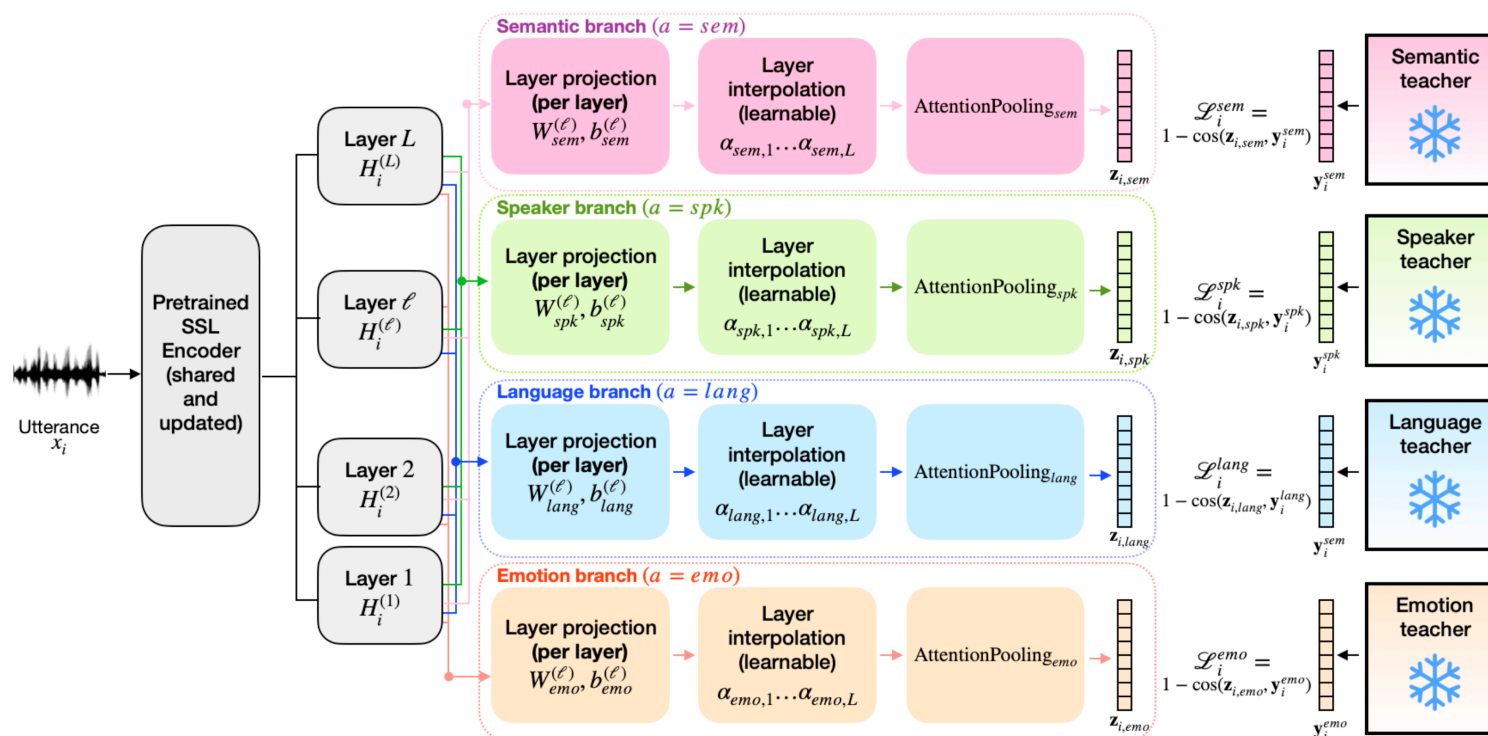
Share the backbone; retain a dedicated representation for each attribute.

Four teachers, four branches, one shared backbone



Each branch learns layer mixing + attention pooling + projection into its teacher space.

Four teachers, four branches, one shared backbone



Each branch learns layer mixing + attention pooling + projection into its teacher space.

Gain-based weighting strategy

Teachers differ in how much the student can learn from them.

1 Measure learning gains

Train one student per attribute.
Track normalized validation loss.

$$g_a = L_a^{\text{mono}}(1) - \min_e L_a^{\text{mono}}(e)$$

First-epoch loss minus the best validation loss

2 Normalize across attributes

Larger learning gain means a larger weight.

$$\lambda_a = g_a / \sum_k g_k \quad \sum_a \lambda_a = 1$$

3 Weight the joint distillation loss

Combine the four attribute objectives.

$$L = \sum_a \lambda_a [1 - \cos(z_a, y_a)]$$

Estimate weights once before joint training, then keep them fixed (NB: could be improved).

Joint training preserves useful attribute vectors

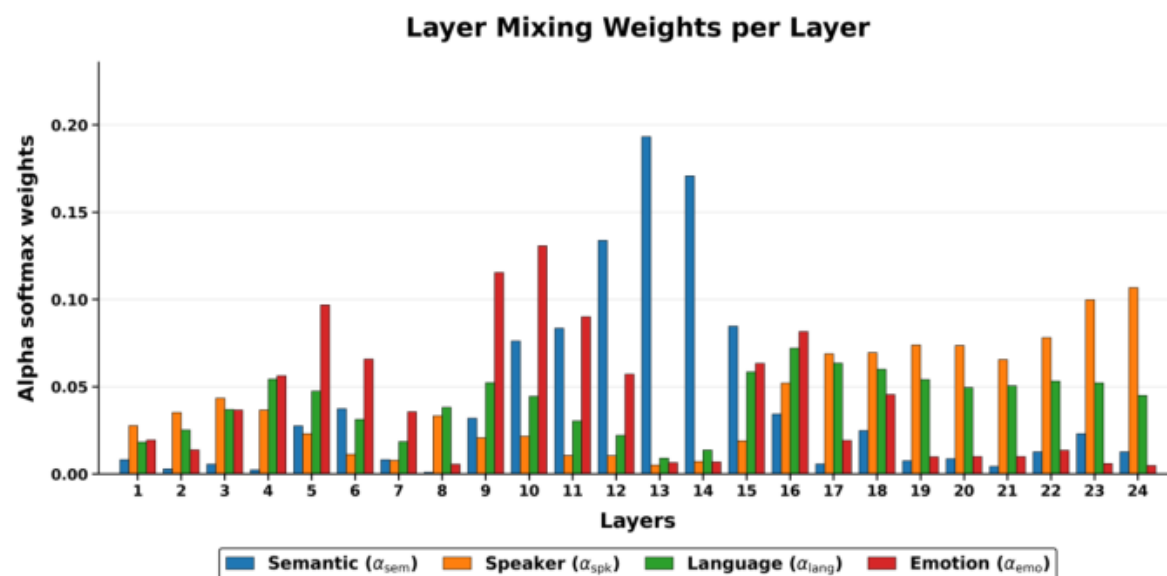
Evaluation	Teacher	Single-task	Multi-attribute
FR → EN retrieval R@1 ↑	96.55	94.05	95.19
Speaker verification EER ↓	0.899	0.899	0.915
Language ID error ↓	6.58	6.83	7.02

Emotion: competitive dataset-specific UA; full results in backup.

Complementary outputs can help

Semantic + language outputs: 6.39% language-ID error (VoxLingua107).

Which encoder layers matter for each speech attribute?



Learned layer mixing

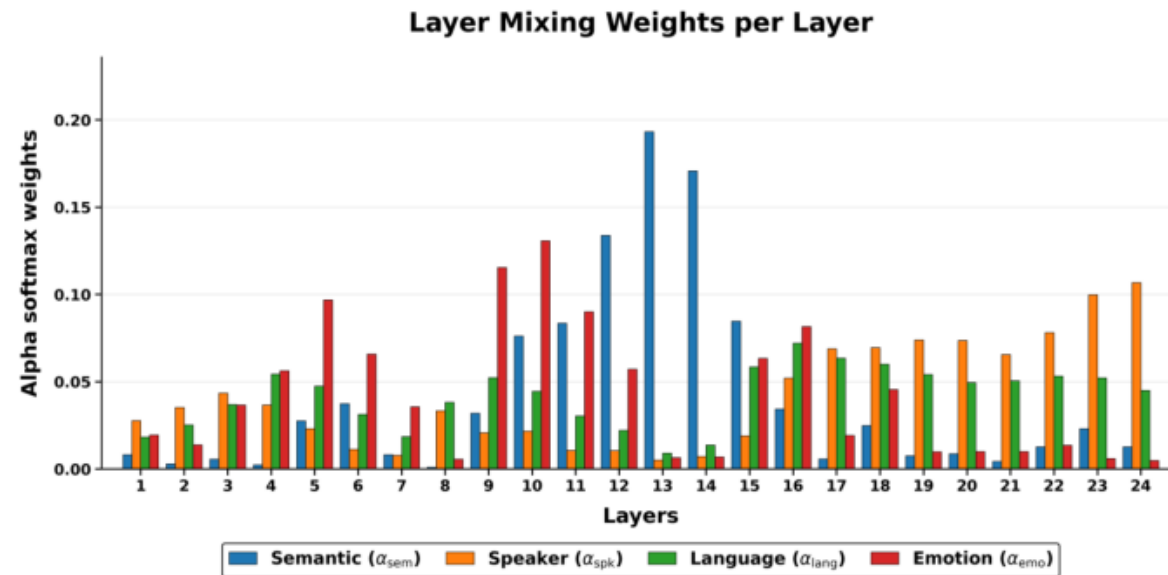
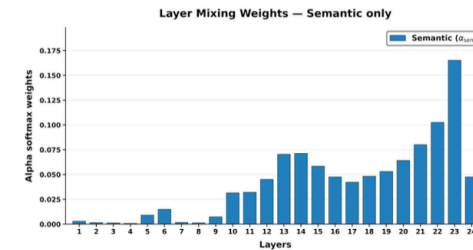
Semantics: middle layers

Emotion: early / middle

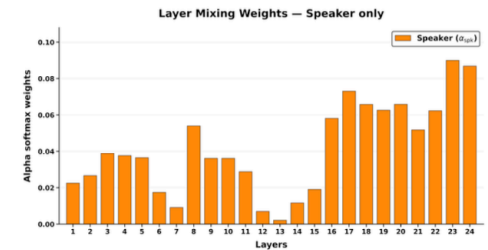
Speaker / language:
broader

Layer weights describe the learned readout; they do not establish causal localization.

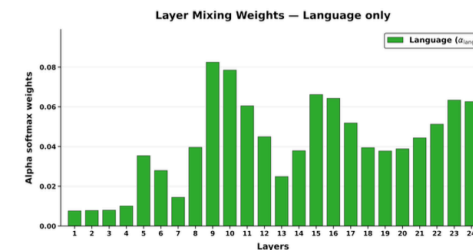
Different attributes draw on different encoder depths



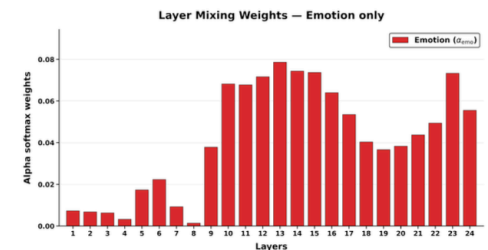
(a) SEM-only



(b) SPK-only



(c) LANG-only



(d) EMO-only

Three messages

Semantic vectors connect speech, text and languages

They support retrieval, summarization and task-specific adaptation.

Supervision design drives specialization

Filtering, alignment, speech teachers and pivot-language targets matter.

One backbone can support several attributes

Dedicated branches preserve useful, complementary utterance representations.

Thank you · discussion

From semantic alignment to richer speech representations

Chaimae Chellaf: cross-lingual summarization

Salima Mdhaftar: French and Arabic dialect models

Maryem Bouziane: ongoing multi-attribute work

SENSE in SpeechBrain - speechbrain.github.io - recipes/CommonVoice/SENSE

SBARThez - github.com/cchellaf/SBARThez

SENSE models - huggingface.co/LIA-AvignonUniversity

Emotion recognition across datasets

