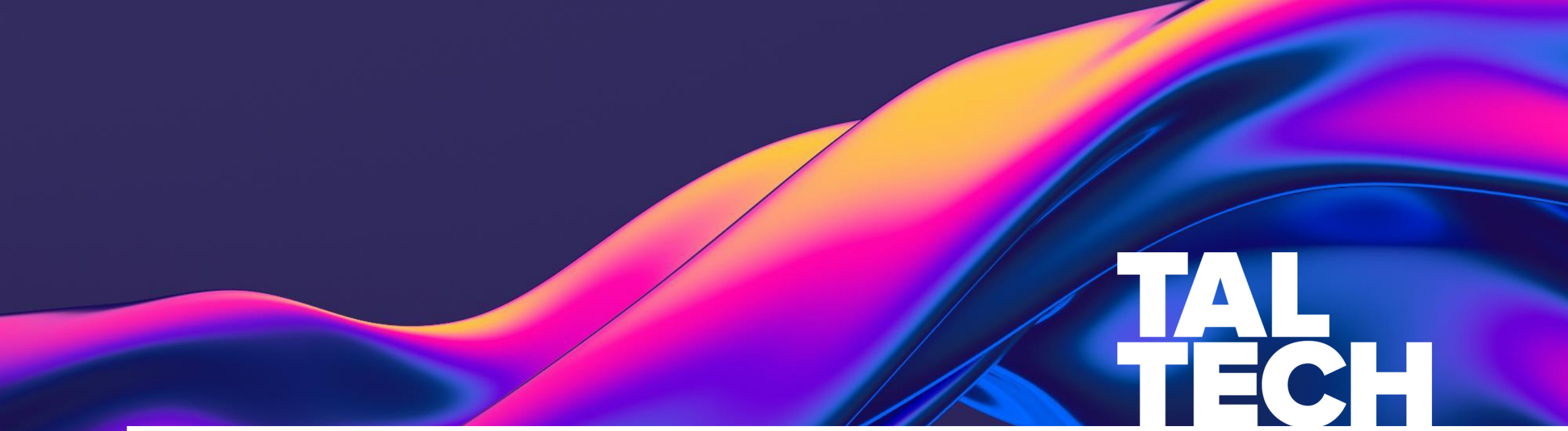



# IMPROVING SPOKEN LANGUAGE IDENTIFICATION FOR NON-NATIVE SPEECH (NOT ONLY)

Tanel Alumäe

April 23, 2026: MILA Conversational AI Reading Group

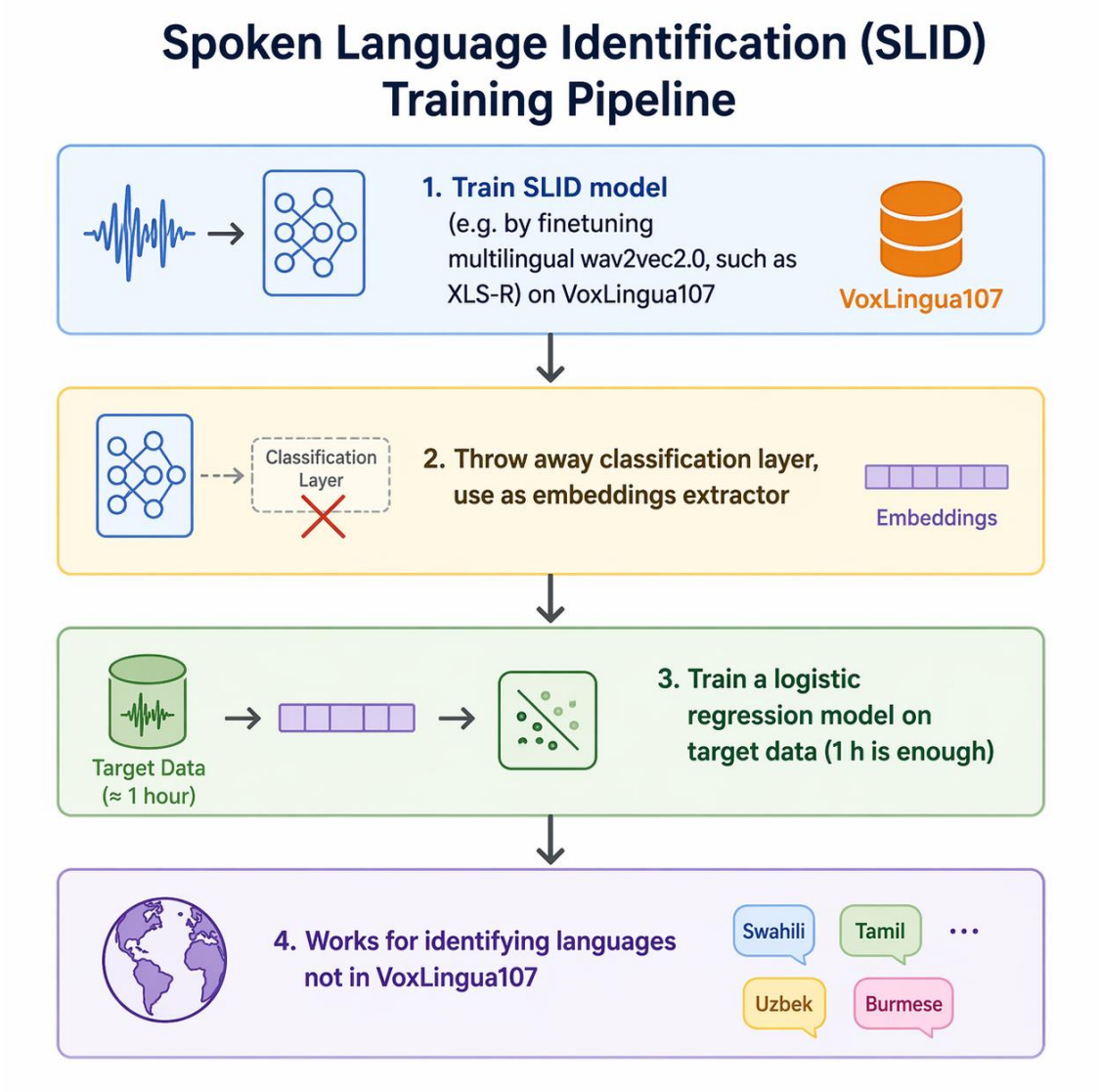
TALLINN UNIVERSITY OF TECHNOLOGY

# SPOKEN LANGUAGE IDENTIFICATION (SLID)

- Why?
  - Preprocessing step in dialogue systems
  - Even when the downstream system is highly multilingual
    - E.g., to decide which language to respond in
  - To understand whether our (multilingual) system supports user's language
  - For indexing and annotating large audio archives

TAL TECH

# STANDARD APPROACH




TAL TECH

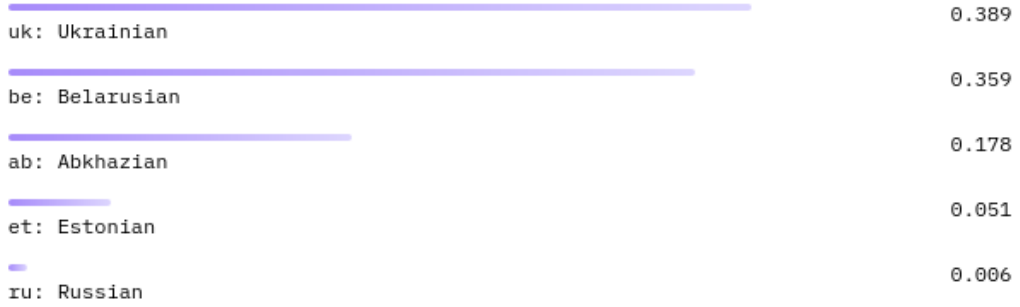
# DEMO: NATIVE SPEECH

- Some examples
  - Estonian
  - Hungarian
  - Swahili

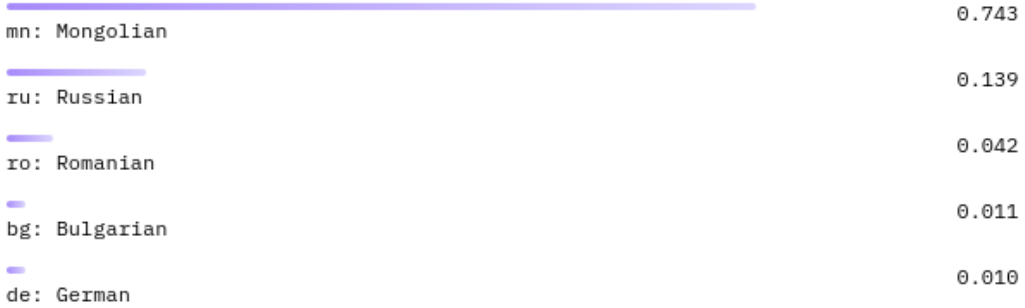
# DEMO: NON-NATIVE SPEECH

- Estonian, German accent


- Estonian, Russian accent




- German, Estonian accent




TAL TECH

# ON NON-NATIVE ACCENTED SPEECH, ACCURACY DROPS DRAMATICALLY

# ERROR RATES:

| | VL107 dev | Estonian L1 | Arabic L1 | Mixed L1 | Russian L1 | Estonian L2 | African French | Arabic L2 | Mixed L2 | Russian L2 | Mixed L2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Wav2vec2.0-BERT | 4.3 | 0.6 | 1.6 | 0.0 | 27.4 | 38.1 | 49.3 | 56.7 | 21.2 | 72.2 | 60.4 |

Also in practice: ChatGPT voice mode answers my (accented) English spoken queries sometimes in Finnish (probably because my English is misclassified as Finnish)

TAL TECH

# WHY THIS HAPPENS

- Why such dramatic drop in accuracy on non-native speech?
- Training data (e.g. Voxlingua107) contains mostly non-native speech
- Models learn to  utilize very short (phoneme-like) features, because they are highly discriminative on native speech
- Non-native speakers' pronunciation is **biased** by their native language learned phoneme inventory
  - This confuses the acoustic LID, hence very odd predictions
- OK, what about adding non-native speech to training data?
  - Our experiments show that this only improves accuracy on the exact language/accent combinations, and not on unseen accents and languages with no non-native training data
  - It is difficult to collect data that would cover all language/accent combinations

TAL TECH

# WHAT TO DO?

- How do humans solve this problem?
- When we hear somebody talk with a heavy non-native accent, we first don't realize it's a language we understand
- Then we start hearing words
  - *And words in a combination that makes sense!*
- What about using ASR system to "test" a language?
  - We can apply text-based LID to ASR produced hypothesis
- ASR system's quality also degrades with non-native speech, but usually not dramatically

# EXPERIMENT: ASR

- Setup: let's transcribe Voxlingua107 with Estonian and English ASR systems
  - CTC-based models to avoid hallucination and accidental translation
  - Train 2 Naïve Bayes classifiers for classifying all 107 languages, using character 4-grams of the Estonian/English ASR transcripts
  - Test resulting model on various native and non-native datasets

| Model | CMU Arctic Native | L2 Arctic Non-native | CSLU FAE Non-native | CSLU 22 en Native | Estonian Native | Estonian Non-native | V107 dev Native |
|---|---|---|---|---|---|---|---|
| XLS-R 300M | 87.6 | 74.6 | 79.5 | 71.9 | 99.6 | 51.8 | 95.3 |
| NB on Estonian ASR char 4-grams | 83.8 | 79.8 | 84.7 | 57.1 | 20.2 | 21.0 | 54.6 |
| NB on English ASR char 4-grams | 81.5 | 74.6 | 45.7 | 34.3 | 71.0 | 65.3 | 48.7 |
| Fusion of NB systems | 90.1 | 86.0 | 83.4 | 53.8 | 69.2 | 63.7 | 75.3 |
| Fusion of XLS-R and ASR-based systems | **95.5** | **90.5** | **88.2** | **72.6** | 99.5 | **69.5** | **95.3** |

Kukk, K., Alumäe, T. (2022) *Improving Language Identification of Accented Speech.* Proc. Interspeech 2022

TAL TECH

# OBSERVATIONS

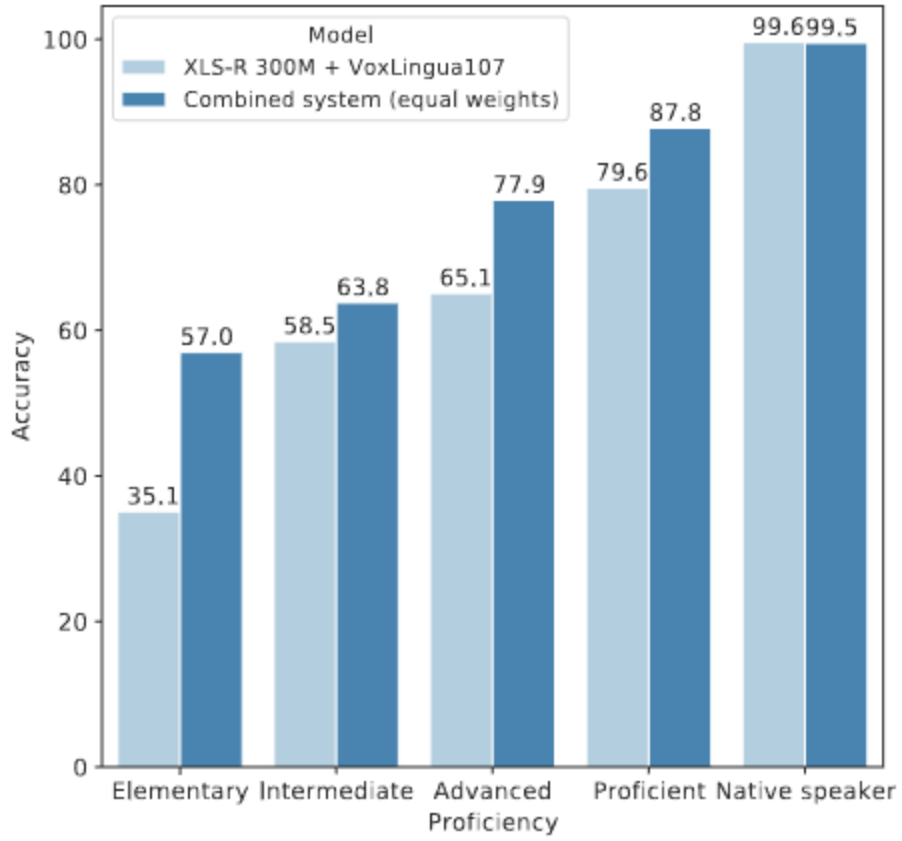
- ASR-based LID is worse than wav2vec-based LID on native speech
- But it beats wav2vec2 on non-native speech!
  - But only when using the ASR system of the corresponding language
  - E.g., on Estonian non-native speech, English ASR transcripts only give 21% recall, but Estonian ASR transcripts 65% recall (vs 52% recall of the wav2vec-based LID)
- Fusion of wav2vec2 and ASR-based systems give the best across all datasets

| Model | CMU Arctic Native | L2 Arctic Non-native | CSLU FAE Non-native | CSLU 22 en Native | Estonian Native | Estonian Non-native | V107 dev Native |
|---|---|---|---|---|---|---|---|
| XLS-R 300M wav2vec | 87.6 | 74.6 | 79.5 | 71.9 | 99.6 | 51.8 | 95.3 |
| NB on English ASR char 4-grams | 83.8 | 79.8 | 84.7 | 57.1 | 20.2 | 21.0 | 54.6 |
| NB on Estonian ASR char 4-grams | 81.5 | 74.6 | 45.7 | 34.3 | 71.0 | 65.3 | 48.7 |
| Fusion of NB systems | 90.1 | 86.0 | 83.4 | 53.8 | 69.2 | 63.7 | 75.3 |
| Fusion of XLS-R and ASR-based systems | **95.5** | **90.5** | **88.2** | **72.6** | 99.5 | **69.5** | **95.3** |

TAL TECH

# SPOKEN LID VS LANGUAGE PROFICIENCY

- Estonian non-native data has speakers' self-reported language proficiency
- There is a clear correlations between proficiency and spoken LID accuracy
- Hybrid approach improves accuracy across all proficiency levels



# HOW MANY ASR SYSTEMS TO USE?

- We only tested with two languages (Estonian and English)
- We saw that to improve the recall of a certain language, we need transcripts from the corresponding language's ASR model
- But what if we have more languages?
  - E.g., do we need 107 ASR systems to do LID on 107 languages?
  - It quickly gets expensive to run and to maintain
- **How to make it scalable?**

- Can we use a multilingual ASR system?
  - Not really, as multilingual ASR models (implicitly or explicitly) also do LID, and also tend to fall into the "non-native" trap

TAL TECH

# ML-SUPERB 2.0 CHALLENGE

- Interspeech 2025 Challenge
- Goal: LID + ASR
- 153 languages, 200+ language varieties (dialects, accents)
- Training data given for 141 of the 153 languages
- Using external data/models was allowed
- Evaluation performed on a server
  - Participants had to upload a containerized solution
- **Perfect testing ground for scalable and robust spoken LID!**

TAL TECH

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

## The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties

*William Chen*[1], *Chutong Meng*[2], *Jiatong Shi*[1], *Martijn Bartelds*[3], *Shih-Heng Wang*[1], *Hsiu-Hsuan Wang*[4], *Rafael Mosquera*[5,8], *Sara Hincapie*[5,8], *Dan Jurafsky*[3], *Antonis Anastasopoulos*[2,7], *Hung-yi Lee*[4], *Karen Livescu*[6], *Shinji Watanabe*[1]

[1]Carnegie Mellon University, [2]George Mason University, [3]Stanford University, [4]National Taiwan University, [5]ML Commons, [6]TTI-Chicago, [7]Archimedes, Athena RC, [8]Factored AI,

williamchen@cmu.edu

### Abstract

Recent improvements in multilingual ASR have not been equally distributed across languages and language varieties. To advance state-of-the-art (SOTA) ASR models, we present the Interspeech 2025 ML-SUPERB 2.0 Challenge. We construct a new test suite that consists of data from 200+ languages, accents, and dialects to evaluate SOTA multilingual speech models. The challenge also introduces an online evaluation server based on DynaBench, allowing for flexibility in model design and architecture for participants. The challenge received 5 submissions from 3 teams, all of which outperformed our baselines. The best-performing submission achieved an absolute improvement in LID accuracy of 23% and a reduction in CER of 18% when compared to the best baseline on a general multilingual test set. On accented and dialectal data, the best submission obtained 30.2% lower CER and 15.7% higher LID accuracy, showing the importance of community challenges in making speech technologies more inclusive.

**Index Terms**: multilingual, speech recognition

### 1. Introduction

In the past decade, studies on scaling end-to-end neural networks have led to dramatic improvements in models for Automatic Speech Recognition (ASR) [1, 2]. Importantly, ASR systems are no longer limited to solely the English language: state-of-the-art (SOTA) models achieve strong performance on over 50 languages [3–5]

by having all submissions adhere to an API. As such, the test set remains fully hidden to participants, preventing benchmark overfitting.

The focus on robustness to *both* different languages and language varieties is novel, as research in these areas has often been disjoint. Many works aim to solely increase the amount of languages a system can handle, but ignore variation that occurs within languages [5, 17, 18]. On the other hand, research on improving ASR performance on different accents or dialects often focuses on variation within single languages [6, 19–21]. This benchmark acts as a first step towards better unifying the fields of *multilingual* and *fair* speech processing. We hope such cross-disciplinary interactions can lead to both a more inclusive speech processing community and more inclusive speech processing models. This paper is outlined as follows:

- Section 2 details the collection and cleaning process for the data used in the challenge, while discussing the difficulties in developing a benchmark with such a wide coverage of languages and language varieties.
- Section 3 outlines the challenge rules and design decisions.
- Section 4 presents baseline results with SOTA self-supervised and supervised ASR models, and compares them to the submitted systems.

Our contributions can thus be summarized as:

1. We introduce a new challenge that evaluates multilingual ASR performance across 149 languages and 93 language varieties, representing the broadest coverage of any speech benchmark to date.

# ML-SUPERB 2.0 CHALLENGE: OUR SLID SOLUTION

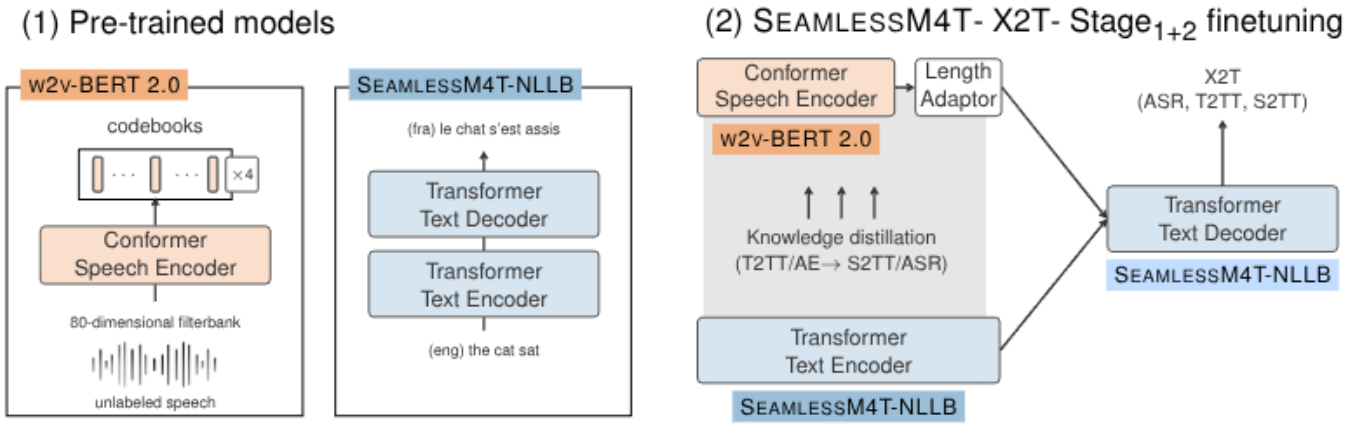



**Figure 5: Overview of the SeamlessM4T X2T model.** (1) describes the main two building blocks: w2v-BERT 2.0 and SeamlessM4T-NLLB. (2) describes the training of the X2T model. In $\text{Stage}_1$, the model is trained on X–eng directions and in $\text{Stage}_2$, eng–X directions are added.

- Our LID system used a uniform interpolation of 2 model's outputs
  - Acoustic and ASR-based
- Acoustic LID:
  - Instead of "raw" SSL-trained wav2vec2 model, we extracted the encoder module of the SeamlessM4T multilingual speech recognition/translation model
  - W2V-BERT-2.0 model, but already finetuned on highly multilingual data (as part of the encoder- decoder model), after being pre-trained on vast amounts of unlabelled speech
  - We used weighted outputs from W2V layers + attention pooling to turn this into LID model
    - Trained on VoxLingua107
  - Finally used it as feature (embeddings) extractor on ML-SUPERB 2.0 training data (provided 141 languages + web-scraped for data for the remaining languages with no given data)

TAL TECH

# ASR-BASED LID: ZERO-SHOT MMS MULTILINGUAL MODEL

- Our ASR-based LID model is based on MMS Zero-shot
- Multilingual ASR model
- Uses a romanized ("*uroman*") character vocabulary
- Trained on 1078 language data

TAL TECH

## Scaling A Simple Approach to Zero-Shot Speech Recognition

Jinming Zhao*
*Monash University*
Melbourne, Australia
jinming.zhao@monash.edu

Vineel Pratap
*Meta FAIR*
Menlo Park, USA
vineelkpratap@meta.com

Michael Auli
*Meta FAIR*
Menlo Park, USA
michaelauli@meta.com

***Abstract*—Despite rapid progress in increasing the language coverage of automatic speech recognition, the field is still far from covering all languages with a known writing script. Recent work showed promising results with a zero-shot approach requiring only a small amount of text data, however, accuracy heavily depends on the quality of the used phonemizer which is often weak for unseen languages. In this paper, we present MMS Zero-shot, a conceptually simpler approach based on romanization and an acoustic model trained on data in 1,078 different languages or three orders of magnitude more than prior art. MMS Zero-shot reduces the average character error rate by a relative 46% over 100 unseen languages compared to the best previous work. Moreover, the error rate of our approach is only 2.5x higher than in-domain supervised baselines, while MMS Zero-shot uses no labeled data for the evaluation languages at all.**

***Index Terms*—zero-shot, speech recognition, unsupervised learning, transfer learning.**

### I. Introduction

There has been significant work [1–4] in enabling automatic speech recognition (ASR) for more of the over 7,000 known languages spoken around the world [5]. One approach has been to perform self-supervised learning [6–9] followed by fine-tuning on labeled data to build models supporting between 100 and 1,000 languages [10, 11]. Another line of work focuses on traditional supervised learning using large amounts of labeled data [12] which is only available for a small subset of languages. However, despite rapid progress in language expansion, it appears unlikely that it is possible to obtain a reasonable amount of labeled data for all languages with a writing script. An alternative is to do away with labeled data altogether via unsupervised ASR methods [13–15] but a drawback of these methods is the requirement for both and that the G2P phonemizer has poor performance for many unseen languages[1].

To sidestep these challenges, we do away with allophones and phonemes and use a single intermediate text representation by romanizing text in different languages which standardizes them to a common Latin-script [20]. Our acoustic model is trained on labeled data in 1,078 different languages and outputs romanized text. For inference, we map model outputs to words by employing a simple lexicon based on a romanized encoding of a modest amount of supplied text in a new language ([20]; see Figure 1).

We do not require language independent phonemizers which we find to result in poor accuracy for many languages and experiments show that this simple approach can lead to large accuracy improvements on 100 unseen MMS-lab and FLEURS languages [11, 21].

The MMS Zero-shot models are available at https://github.com/facebookresearch/fairseq/tree/main/examples/mms/zero_shot. We also provide a HuggingFace demo at https://hf.co/spaces/mms-meta/mms-zeroshot.

### II. Method

#### A. Universal Acoustic Model

We address the challenge of standardizing various writing scripts by converting text to a single writing script [22]. Specifically, we use uroman [20], a universal romanizer, which performs the mapping from text in any language to the standard Latin alphabet using a set of heuristics. After converting all the transcripts to their romanized version, we finetune XLS-R [23], a pretrained multilingual wav2vec 2.0 [7] model,

# ROMANIZATION

- So-called uroman characters: basic romanized (ASCII) letters
- Training data transcripts for all languages converted to uroman using the *uroman* library that covers **almost all** written languages
  - `uroman.romanize_string("Tere päevast!", "est") -> 'Tere paeevast!'`
  - `uroman.romanize_string("Ντέιβις Καπ", "ell") -> 'Deivis Kap'`
  - `uroman.romanize_string("নমস্কাৰ", "asm") -> 'namaskaar'`

README | License

## uroman

*uroman* is a *universal romanizer*. It converts text in any script to the standard Latin alphabet.
Example (Greek): Νεπάλ → Nepal
Example (Hindi): नेपाल → nepaal
Example (Urdu): نیپال → nypal
Example (Chinese): 三万一 → 31000

- *uroman* enables the application of string-similarity metrics to texts from different scripts without the need and complexity of an intermediate phonetic representation.
- *uroman* converts digital numbers in various scripts to Western Arabic numerals.
- *uroman* uses m-to-n character mappings, context, and a user-provided language code (optional), i.e. *uroman* does not just replace characters one by one.
- *uroman* expects all input to be encoded in UTF-8.

New Python version: 1.3.1.1 (released on June 27, 2024)
Last Perl version: 1.2.8 (released on April 23, 2021)
Author: Ulf Hermjakob, USC Information Sciences Institute
Quick links (inside this doc): uroman CLI, import uroman, Old Perl version, change history, reversibility, limitations

### (New) Python version

**Installation**

```
python3 -m pip install uroman
```

### Command Line Interface (CLI)

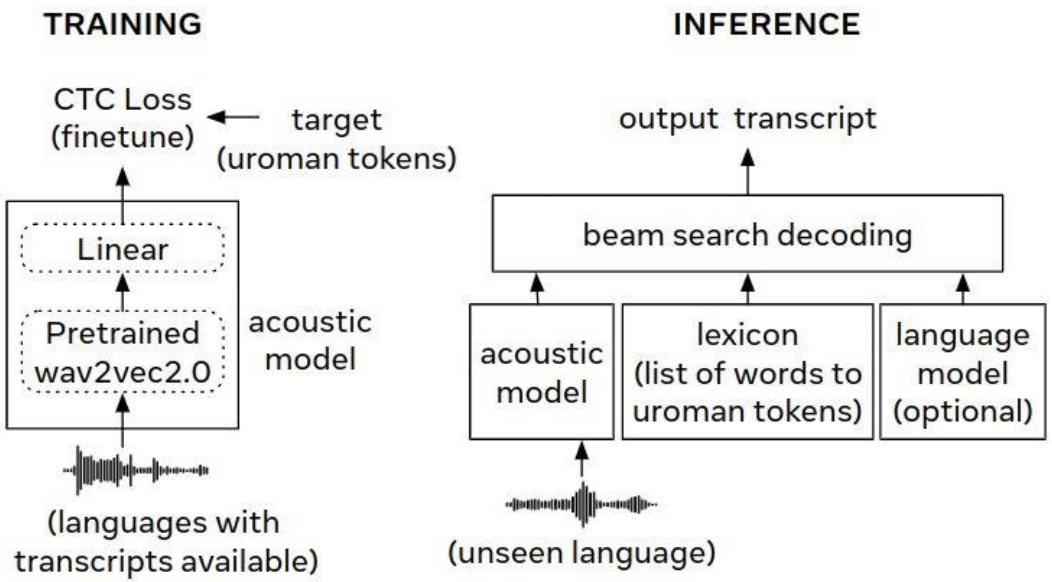
**Examples**

```
python3 -m uroman "Игорь Стравинский"
python3 -m uroman Игорь -l ukr
python3 -m uroman Ντέιβις Καπ -l ell
python3 -m uroman "\u03C0\u03B9" -d
python3 -m uroman -l hin -i mini-test/hin.txt
python3 -m uroman -l fas -i mini-test/fas.txt -o mini-test/fas-rom.jsonl -f edges
python3 -m uroman < mini-test/multi-script.txt > mini-test/multi-script.uroman.txt
python3 -m uroman -h
```

TAL TECH

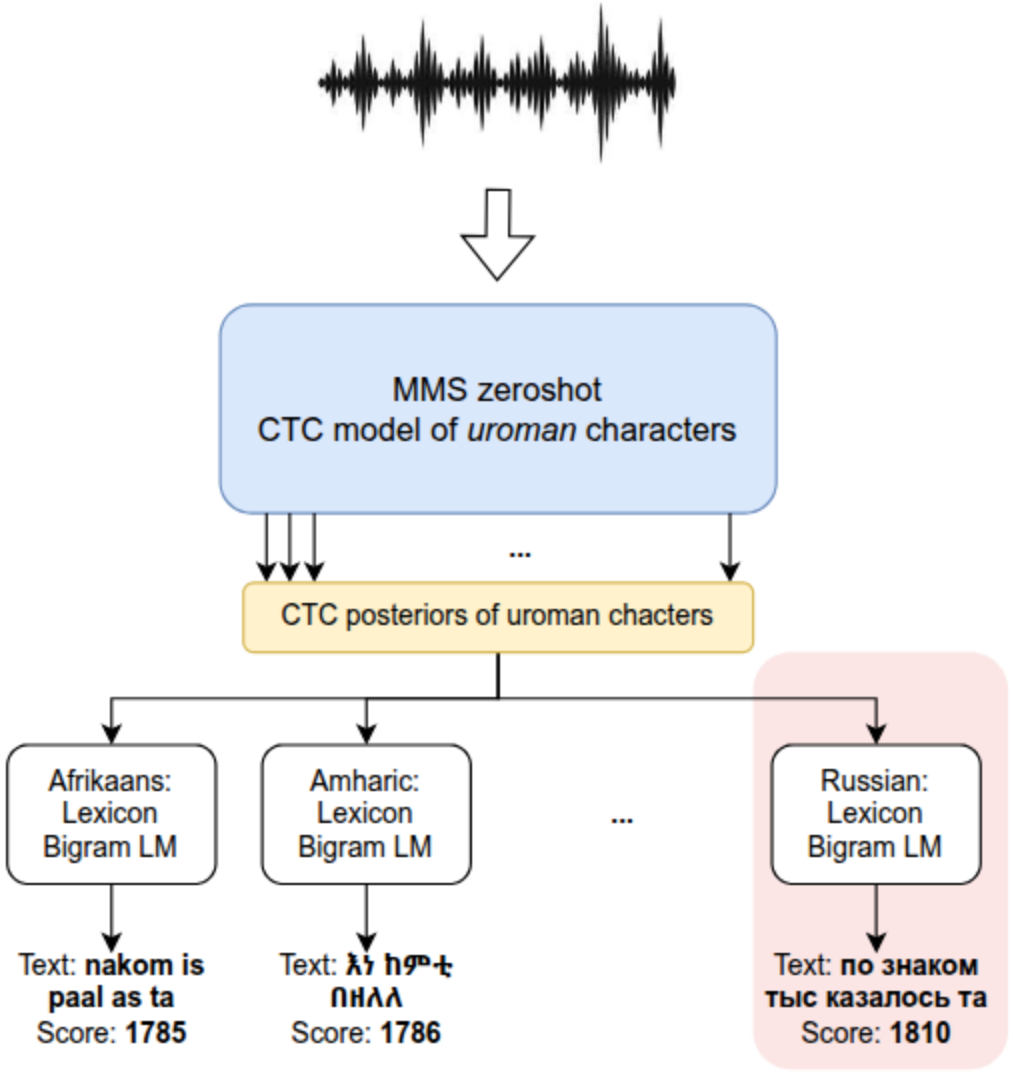
# ZERO-SHOT MMS MODEL

- But MMS model can actually produce transcripts in the original script!
- How:
  - Use a language model (e.g. trigram) of words (written in the original script)
  - And a lexicon that maps words to uroman "pronunciation"
  - Beam search combines CTC probabilities from MMS model with the LM and lexicon
  - Why important: now, given a multilingual uroman model, we can transcribe any language, given only text data and a romanizer



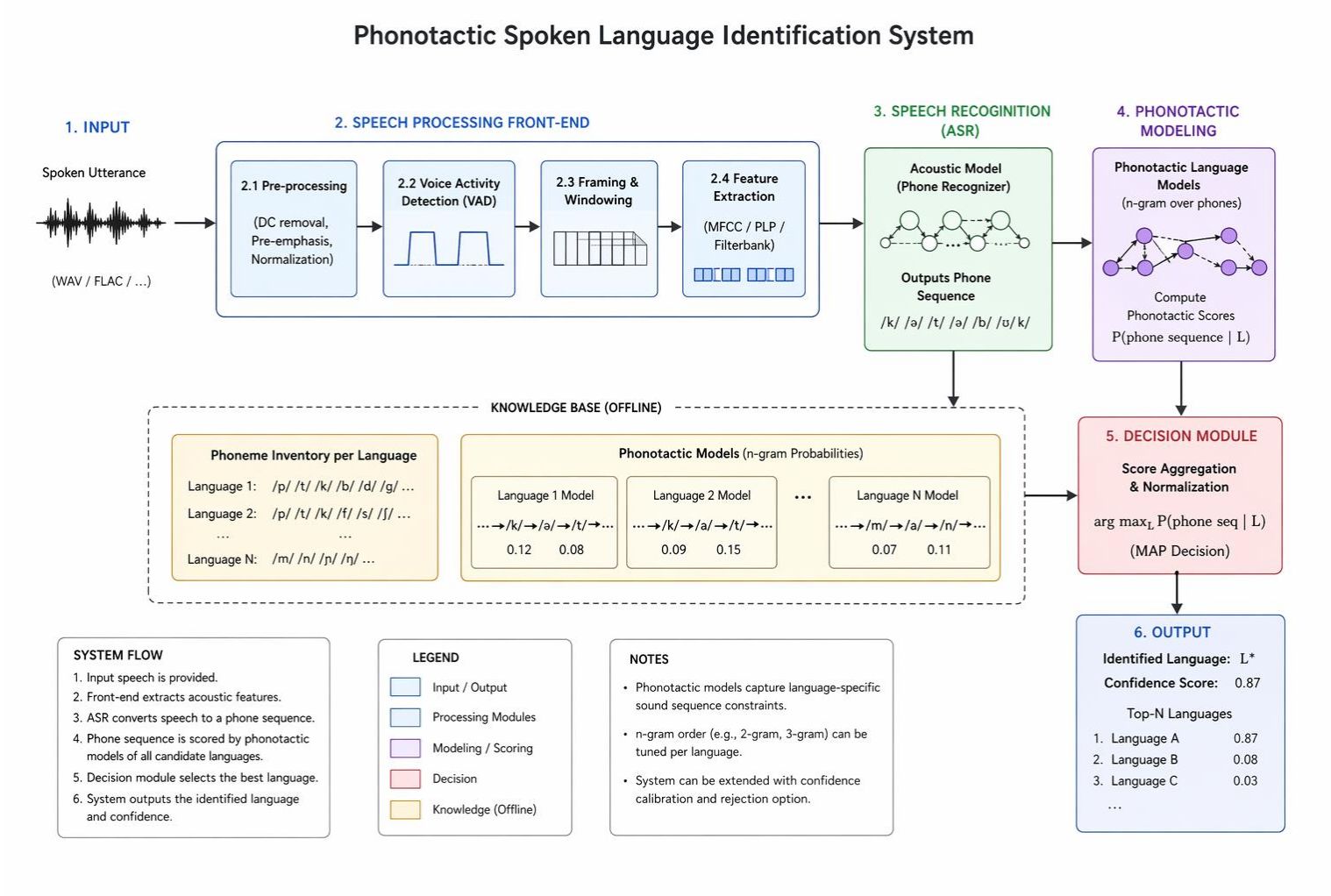
# MMS-ZEROSHOT AS SPOKEN LID MODEL

- MMS-zeroshot model, combined with LMs and lexicons in different languages, acts as a very decent language ID system
- Method:
  - Generate CTC posteriors using MMS-zeroshot model (requires GPU)
  - Decode to words using each language's LM+lexicon
  - Check, for which language the probability of the best decode is the highest
  - **Only requires textual training data!**
- Is it scalable? Requires decoding for 153 languages, no?
  - **But the expensive encoding step is run only once**
  - Decoding can use bigram LM, highly parallelizable on CPUs
  - In practice, it takes ~1 second per ~10 second utterance (when decoding is parallelized across ~16 CPUs)



# RELATIONSHIP TO PHONOTACTIC SLID

- MMS-Zeroshot based LID is related to phonotactic SLID models
- However:
  - Phonotactic models require a G2P converter or speech-based training data
  - Since they operate on phonemes, they are also vulnerable to non-native "traps"



# MMS-ZEROSHOT AS SPOKEN LID MODEL: DOES IT WORK?

- First, we experimented with this model on the 107 language LID task
- Textual training data for MMS-zeroshot based model from the GlotLID corpus
- Preliminary experiments on several languages (both L1 (native) and L2 (non-native)) show that:
  - LID accuracy of MMS-zeroshot is not as good as the best audio-based model (for native speech)
  - However, interpolating audio-based model's predictions with MMS-zeroshot gives nice improvements for all L2 datasets!


| | VL107 dev | Arabic L1 | Russian L1 | Estonian L2 | African French | Arabic L2 | Mixed L2 | Russian L2 | Mixed L2 |
|---|---|---|---|---|---|---|---|---|---|
| Wav2vec2.0 -BERT | **4.3** | **1.6** | 27.4 | 38.1 | 49.3 | 56.7 | 21.2 | 72.2 | 60.4 |
| MMS zero-shot 10k bigram | 17.0 | 11.5 | 35.3 | 13.9 | 48.9 | 65.1 | 0.9 | 59.3 | 30.9 |
| A+B, optimized | **4.3** | **1.6** | **25.8** | **13.1** | **34.4** | **51.6** | **0.9** | **56.9** | **29.4** |

TAL TECH

# MMS-ZEROSHOT LID FOR ML-SUPERB 2.0

- For ML-SUPERB 2.0, we scaled the MMS-Zeroshot based LID model to 153 languages
  - Used word-piece vocabulary, as some languages do not have "words"
- Surprisingly decent accuracy (taking into account 153 output classes)
- Gives a 5%-point accuracy boost on both "normal" and dialectal data, when combined with acoustic LID (simple linear interpolation of posterior probabilities)
- We obtained the best LID scores in the challenge

Table 2: *LID accuracies (%) of the two models and their uniform interpolation on two development sets.*

| | Embeddings | Gener. (MMS-zeroshot) | Inter-polated |
|---|---|---|---|
| Dev | 85.3 | 70.7 | **89.9** |
| $\text{Dev}_{dialects}$ | 80.5 | 73.2 | **84.9** |

TAL TECH

# TIDYLANG 2026 CHALLENGE

- Speaker Odyssey 2026 Challenge
- Speaker-controlled and zero-shot language recognition
- Track 1: closed-set language ID with 35 languages, but training and dev data contain native and non-native speech from same speakers
  - Each speaker contributes utterances in 2–10 languages
  - Tests speaker/language distanglement
- Track 2: very-few shot language ID with unknown languages
  - Enrollment-based (20–65 s per enrollment ID, compare with test utterance)


TAL TECH

# TIDYLANG 2026 CHALLENGE: TRACK 1

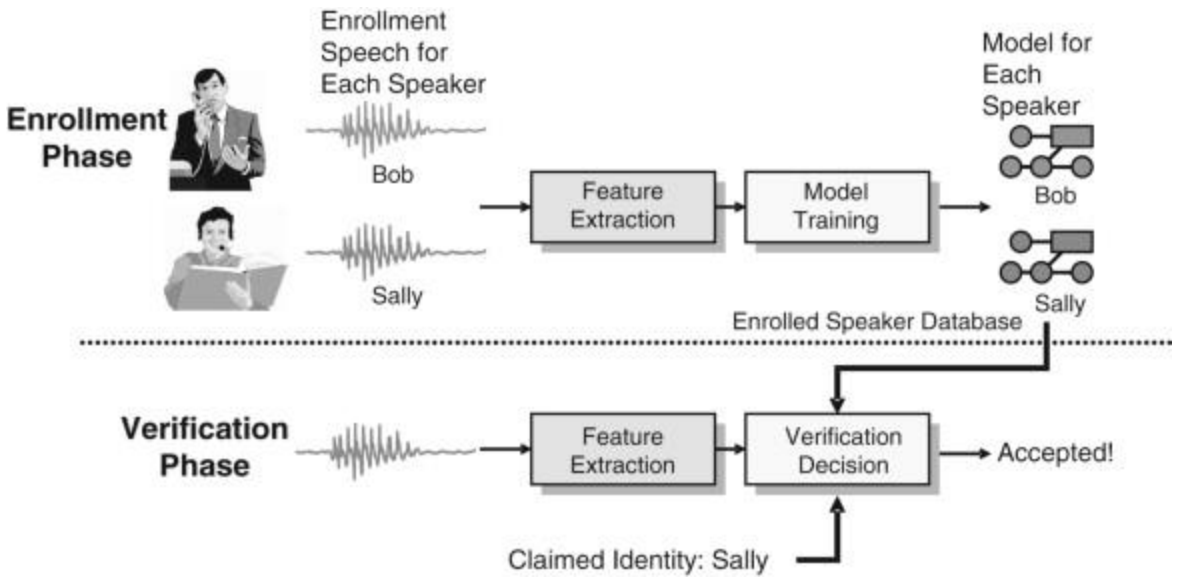
- We experimented with 3 models:
  - W2V-BERT-based model
  - MMS-Zeroshot
  - Qwen/Qwen3-Omni-30B-A3B-Instruct, finetuned for language identification using LoRA
    - To obtain a 35-class posterior distribution from the generative model, we extract the logits at the first generated token position and compute a softmax over the token IDs corresponding to each language code.
    - This is possible because all language codes map to distinct tokens in the model's vocabulary.

| Model / Setup | Dev (Macro-acc) | Dev-cross-lingual (Macro-acc) | Eval (Macro-acc) |
|---|---|---|---|
| Qwen3-Omni (zero-shot) | 61.0 | | |
| W2V-BERT | 93.1 | 99.0 | 88.5 |
| MMS-zeroshot | 84.7 | 87.4 | — |
| Qwen3-Omni (finetuned) | 94.0 | 98.1 | — |
| W2V-BERT + MMS-zeroshot | 95.1 | 99.3 | — |
| W2V-BERT + Qwen3-Omni | 94.0 | 99.5 | — |
| MMS-zeroshot + Qwen3-Omni | 92.9 | 98.6 | — |
| All 3 models | **95.2** | 99.4 | **96.8** |

TAL TECH

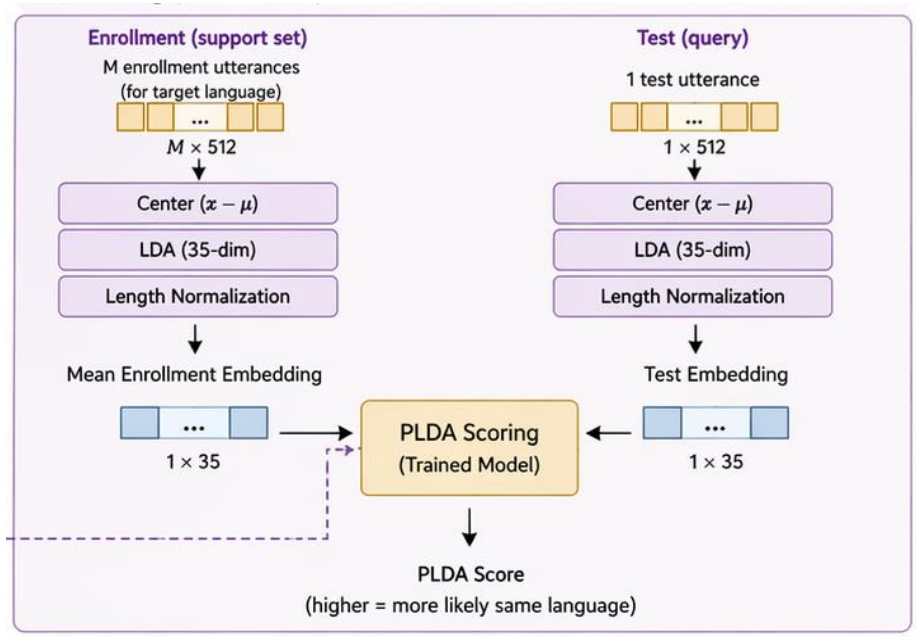
# TIDYLANG 2026: TRACK 2

- Task: Language verification
- Similar to speaker verification:
  - Is test utterance in the same language, given enrollment utterances (7–15 utterances -- 20–65 seconds of audio)
- Note: is it an ill-posed problem? What is a "*language*"?
  - Distinction between language and dialect is often political/cultural
  - E.g. NIST LID evaluations usually treat different varieties of Arabic as unique languages
  - If enrollment contains one dialect but test another dialect, should we accept it as verified?
  - Anyway, dev and test data contain only well-established languages



# TRACK 2: OUR BASELINE

- Our "baseline": use W2V-BERT2.0 based model trained for Track 1 as embedding extractor
  - Train LDA/PLDA classifier on Track 1 data
  - Use this to evaluate enrollment/test match
  - Standard approach from speaker verification
  - EER: 2.5% on dev, but 7.9% on test
- Can we do better?



# TRACK 2: ASR-BASED APPROACH

- ASR-based verification pipeline
  - Transcribe enrollment and test utterances using a massively multilingual ASR model (Meta's Omnilingual ASR: omniASR-CTC-7B)
  - Apply text-based LID (GlotLID)
  - Verify if enrollment utterances and test utterance is in the same language
- Easy!
- Does it work? Not very well...
- OmniASR and GlotLID not very robust on smaller languages and non-native speech
- Need to bring in some intelligence

**Enrollment (ASR output):**
*This is an enrollment  sentence*
*This anher enrollement tense*

**Test (ASR output)**
*Dis iz a dest sentenz wid akzent*

**GlotLID text-based LID**
Enrollment: English, English
Test: German

-> NOT same language

TAL TECH

# TRACK 2: USE LLM FOR ADDITIONAL INTELLIGENCE

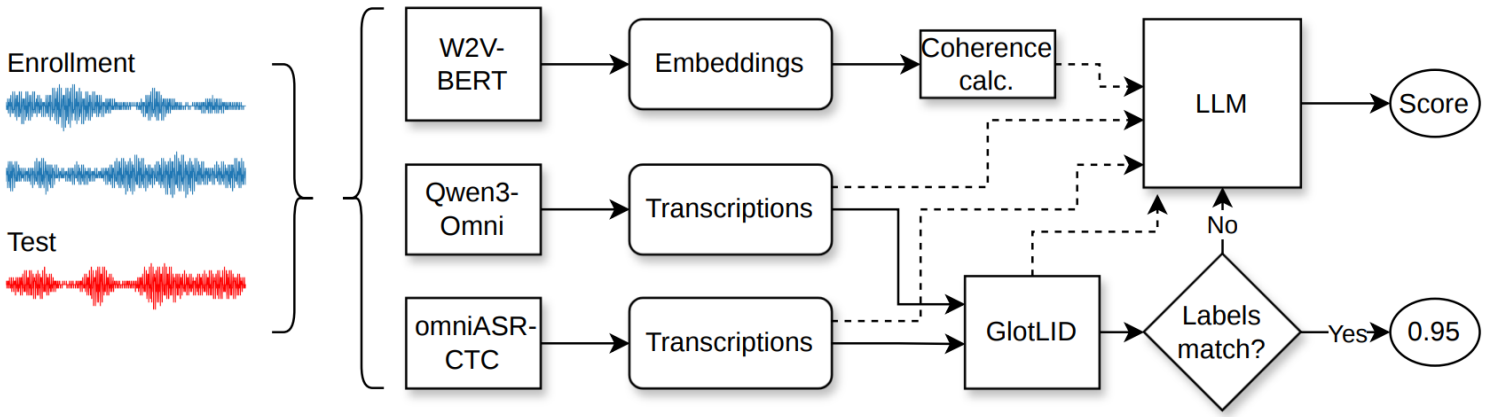
- Use 2 separate ASR systems to obtain transcripts (Qwen3—Omni and omniASR-CTC-7B)
- Apply GlotLID on the transcripts
- Ask Qwen/Qwen3.5-397B-A17B (not finetuned) if enrollment matches with test, given transcripts and GlotLID scores
  - LLM not only relies on character n-grams (as GlotLID) but can also judge whether the transcripts "make sense" (e.g. contain phrases that mean something)
- To improve speed, we did not use thinking mode

```
System:  You are a language verification
expert.  Determine if utterances are in
the same language.
Output ONLY: SCORE: <float>
where 0.0 = definitely different language,
1.0 = definitely same language.

User:  ## Enrollment utterances (speaker's
known language(s)):
Qwen3-Omni transcription 1:  "All of the
elders working on the
project were men."
OmniASR transcription 1:  "All of the
elders working on the project
were men." [GlotLID: eng_Latn, conf:
0.98]
Qwen3-Omni transcription 2:  "The
committee met again before the
final report was submitted."
OmniASR transcription 2:  "The committee
met again before the final
report was submitted." [GlotLID: eng_Latn,
conf:  0.97]
## Test utterance:
Qwen3-Omni transcription:  "There is also
a large Buddhist Taoist
monastery built near the cemetery."
OmniASR transcription:  "There is also a
large Buddhist Taoist
monastery built near the cemetery."
[GlotLID: eng_Latn, conf:  0.99]
## Task
Determine if the test utterance is in the
same language as the
enrollment utterances.  Consider script,
morphology, vocabulary.
If ASR produced gibberish, compare the
gibberish patterns.
```

# COHERENCE-BASED PIPELINE

- If GlotLID labels agree on all enrollment utterances and the test utterance, LLM check is skipped
- Furthermore, LLM prompt is varied based on the GlotLID "coherence"
  - If coherence (GlotLID label agreement) is low, LLM is warned that speakers have strong accents and ASR transcripts contain errors
  - This improves EER on dev data from 2.6% to 1.9%



# TRACK 2: RESULTS

- LLM-based language verification system outperforms acoustic LID, despite operating on text transcriptions and metadata rather than direct acoustic features
- Large improvements when the two systems' predictions fused together
- Weakness: computational performance
  - LLM-based system required 2 days to process 4M evaluation trials
  - Acoustic LID did it in less than an hour

| System | Dev | Eval |
|---|---|---|
| W2V-BERT + LDA/PLDA | 2.50 | 7.90 |
| LLM-based (standalone) | 1.89 | — |
| Fusion ($\alpha = 0.60$) | **0.90** | **3.06** |

TAL TECH

# FINAL TIDYLANG RESULTS

- We obtained top scores in both tracks
- Large improvements over baseline and over other submissions

| Rank | Team | Task 1 | | Task 2 | |
|---|---|---|---|---|---|
| | | Macro | Micro | EER | minDCF |
| 1 | Ours | **96.78** | **97.49** | **3.06** | **0.41** |
| 2 | Team 2 | 92.46 | 96.02 | 5.43 | 0.74 |
| 3 | Team 3 | 85.95 | 90.96 | 17.08 | 1.00 |

TAL TECH

# CONCLUSION

- Robust spoken language recognition depends on combining heterogeneous evidence sources with different failure modes: acoustic embeddings, generative decoding, transcript-derived features, and LLM-based reasoning

- Open problems:
  - Can LLM-based pipelines be made cheaper while keeping the complementarity benefit?
  - Would tighter joint training outperform late fusion?
    - Or will the model "shortcut" to use only phoneme-like features, given that most of the training data is native speech?
  - What to do with heavy accents combined with heavy code-mixing?

TAL TECH

# Thank you for attention!

**We are looking for motivated PhD students and postdocs!**

- https://taltech.ee/en/laboratory-language-technology
- Contact: tanel.alumae@taltech.ee