Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings

Marcely Zanon Boito, Bolaji Yusuf, Lucas Ondel,
Aline Villavicencio, Laurent Besacier

Université Grenoble Alpes
Avignon Université
Laboratoire Informatique d'Avignon
Brno University of Technology
LISN Laboratoire Interdisciplinaire des Sciences du Numérique
The University Of Sheffield.
Naver Labs Europe

INTRODUCTION

Speech Technologies for Low-resource Languages

- Most of current speech technology is developed in a fraction of the existing languages and dialects (“high-resource languages”) [1]

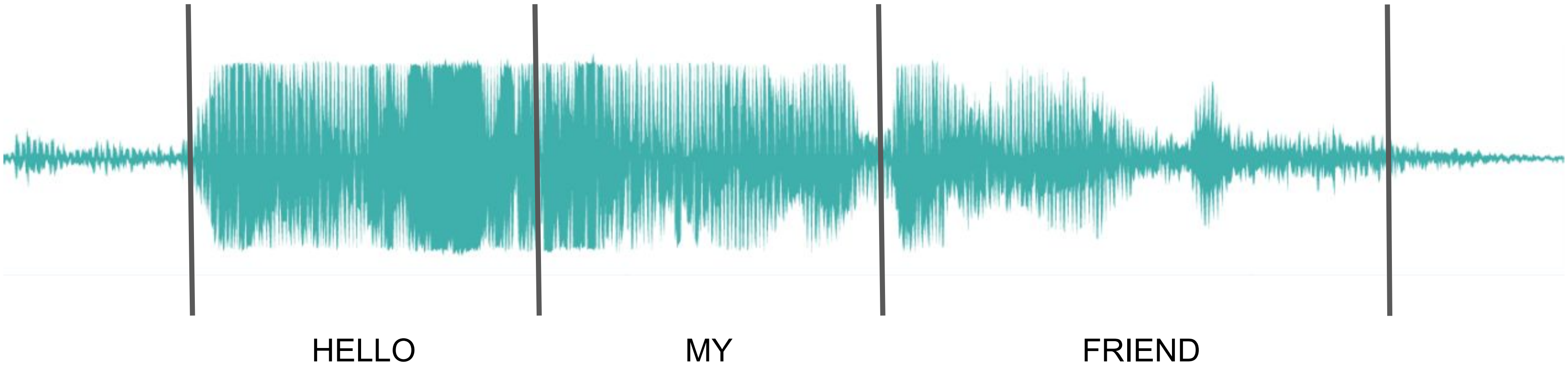
- Pipelines based on text exclude oral languages
  - “Most of the world's languages are not actively written, even the ones with an official writing system” [15]

- This work focuses on **low-resource speech processing:**
  - **Our goal:** performing unsupervised word segmentation from speech



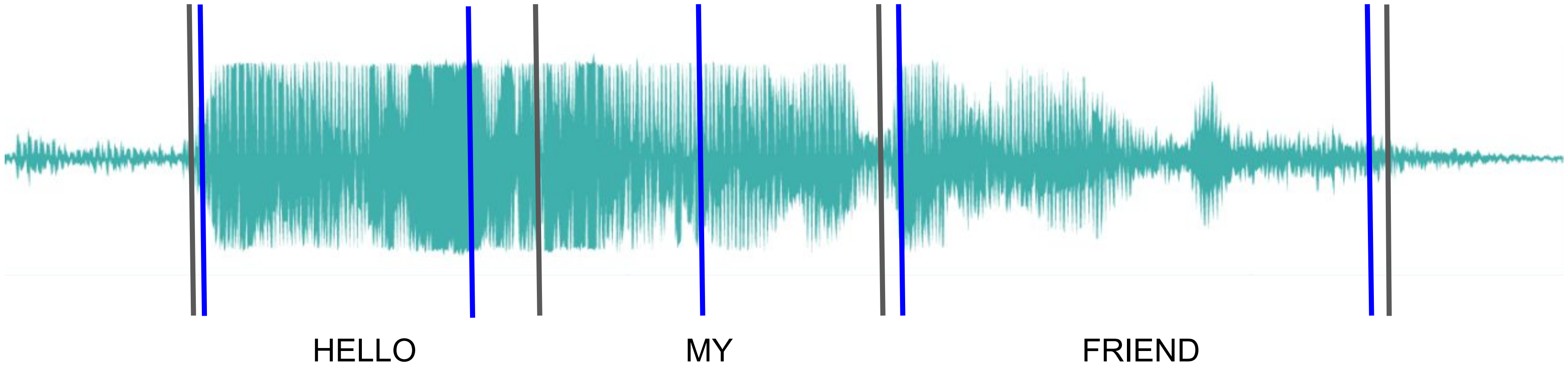
Unsupervised Word Segmentation (UWS) from speech

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.



UWS for Language Documentation

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

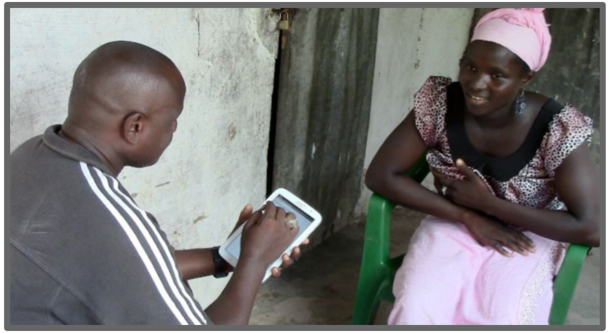


Figure: A field linguist recording utterances from a native speaker.

SPEECH

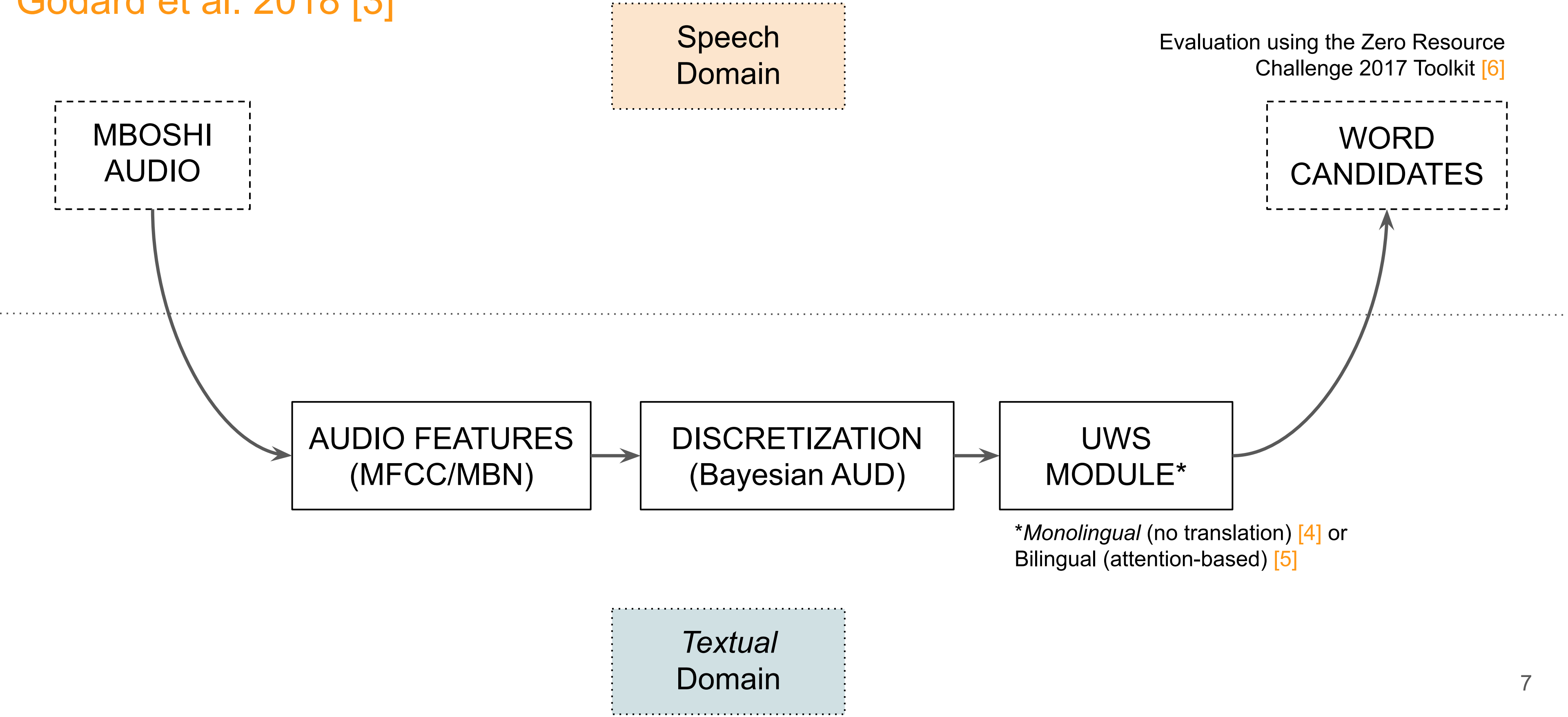
+

Translations
to a high-resource language [2]

Unsupervised Word Segmentation from Speech with Attention

Godard et al. 2018 [3]

Evaluation using the Zero Resource Challenge 2017 Toolkit [6]

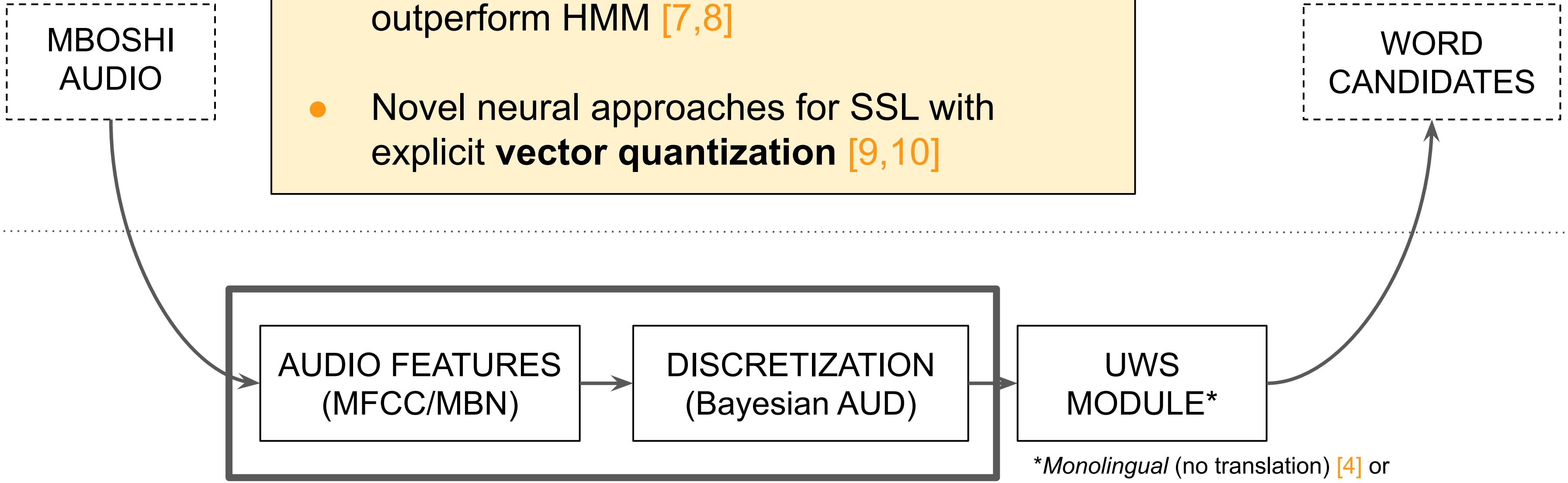


**Monolingual* (no translation) [4] or Bilingual (attention-based) [5]

Since then…

- Efficient Bayesian models for AUD now outperform HMM [7,8]
- Novel neural approaches for SSL with explicit **vector quantization** [9,10]

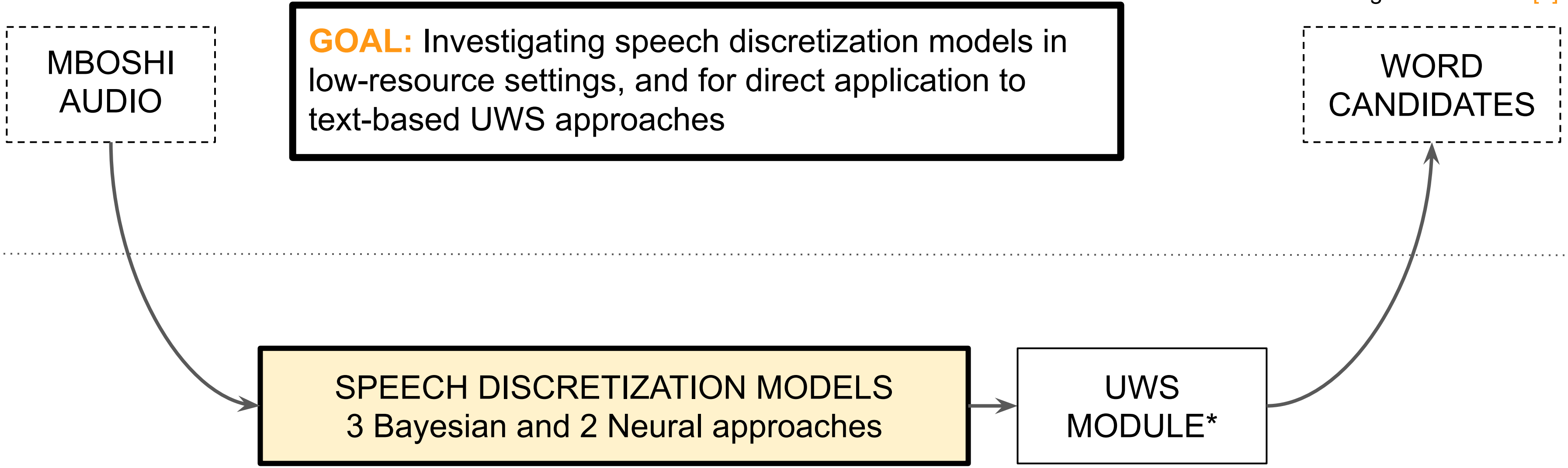
Evaluation using the Zero Resource Challenge 2017 Toolkit [6]



**Monolingual* (no translation) [4] or Bilingual (attention-based) [5]

This work: Revising the Pipeline

Evaluation using the Zero Resource Challenge 2017 Toolkit [6]

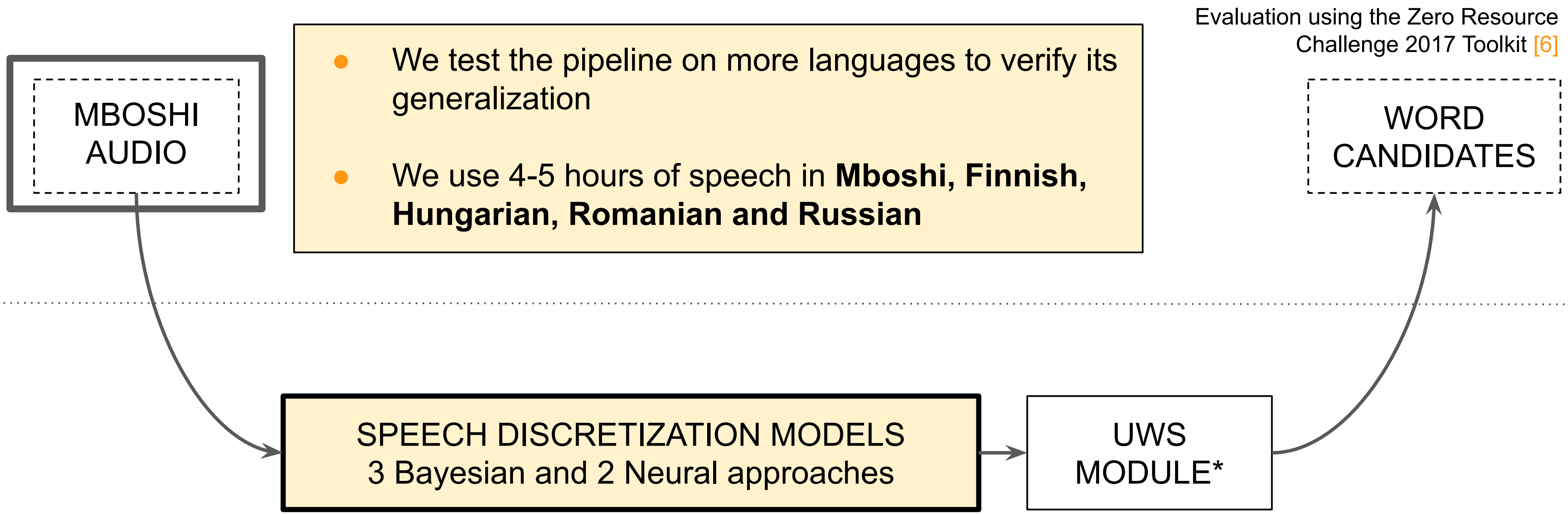


Bayesian AUD models: HMM [11], SHMM [7], H-SHMM [8]
Vector Quantization Approaches: VQ-VAE [9], vq-wa2vec [10]

**Monolingual* (no translation) [4] or Bilingual (attention-based) [5]

This work: A Revision of this Pipeline

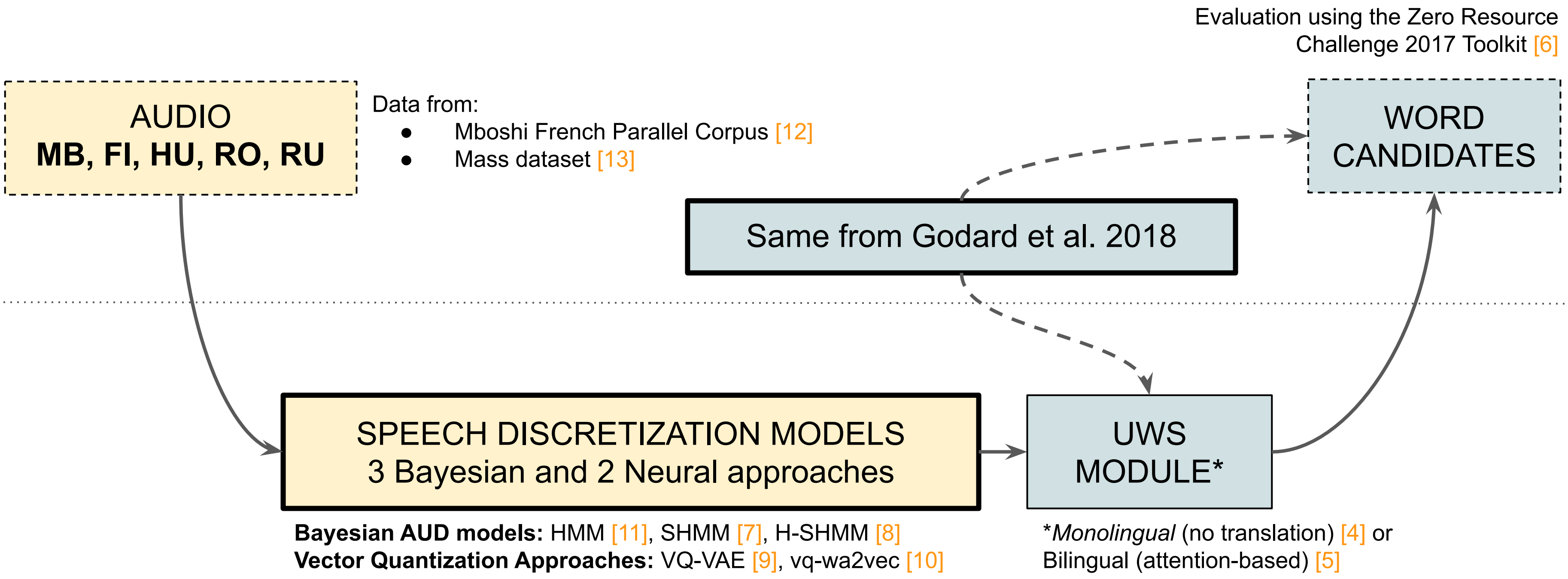
- We test the pipeline on more languages to verify its generalization
- We use 4-5 hours of speech in **Mboshi, Finnish, Hungarian, Romanian and Russian**



Bayesian AUD models: HMM [11], SHMM [7], H-SHMM [8]
Vector Quantization Approaches: VQ-VAE [9], vq-wa2vec [10]

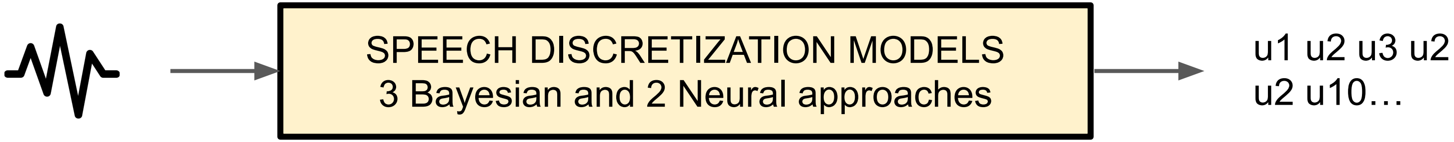
**Monolingual* (no translation) [4] or Bilingual (attention-based) [5]

This work: A Revision of this Pipeline



SPEECH DISCRETIZATION (SD)

Starting point: Producing Discrete Speech Units



Bayesian AUD models: HMM [11], SHMM [7], H-SHMM [8]
Vector Quantization Approaches: VQ-VAE [9], vq-wa2vec [10]

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

- Low-resource settings (4-5 hours of speech)
- No access to transcription

Speech Discretization (SD) Models

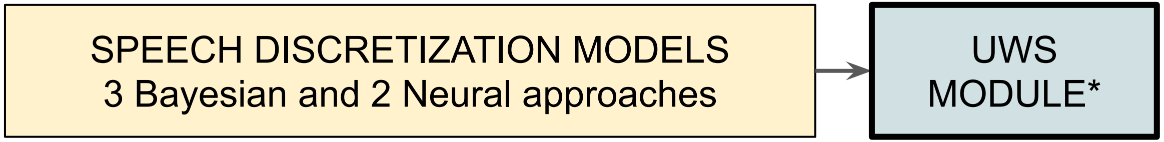
- Bayesian Generative Models (AUD):
    1. HMM/GMM (**HMM**): Every possible sound can be a unit [11]
    2. Subspace HMM (**SHMM**): Prior over a phonetic subspace [7]
    3. Hierarchical Subspace HMM (**H-SHMM**): Subspace adaptation from different languages for unit prediction [8]

Speech Discretization (SD) Models

- Bayesian Generative Models (AUD):
    1. HMM/GMM (**HMM**): Every possible sound can be a unit [11]
    2. Subspace HMM (**SHMM**): Prior over a phonetic subspace [7]
    3. Hierarchical Subspace HMM (**H-SHMM**): Subspace adaptation from different languages for unit prediction [8]

- Vector Quantization (VQ) Approaches:
    1. **VQ-Variational Auto-Encoder (VAE)**: inspired by dimensionality reduction architectures [9]
    2. **VQ-WAV2VEC**: inspired by self-supervised models trained with a context-prediction loss [10]

Next Step: Apply Segmentation!



Bayesian AUD models: HMM [11], SHMM [7], H-SHMM [8]
Vector Quantization Approaches: VQ-VAE [9], vq-wa2vec [10]

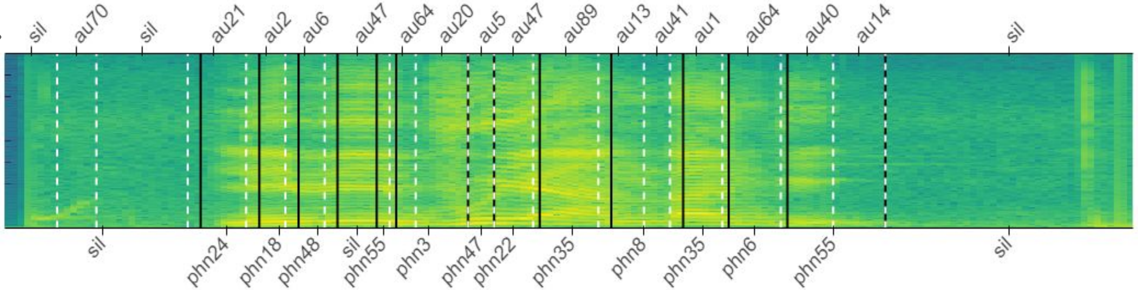
**Monolingual* (no translation) [4] or
Bilingual (attention-based) [5]

Studying the SD Representation

Example: The same sentence, two approaches

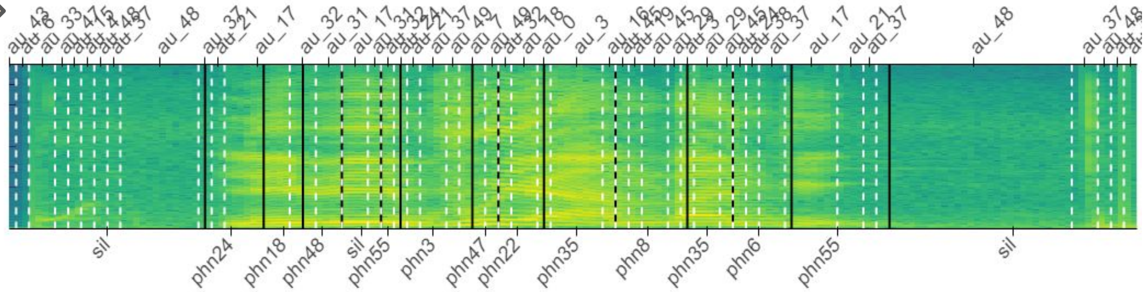
True Boundary
Output Boundary

H-SHMM output (Bayesian)



Reference

VQ-VAE output (Neural)



Reference

UWS RESULTS

Results for Mboshi

- **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

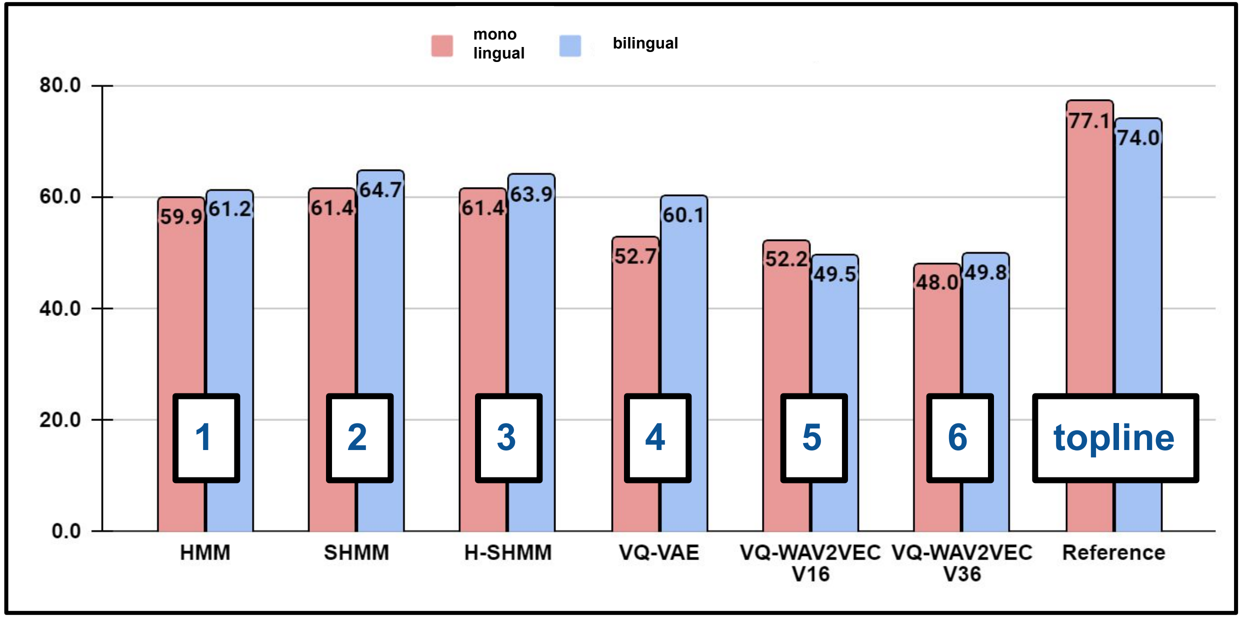


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Results for Mboshi

- We notice a drop in performance, **but we still successfully generate segmentation**
- Bilingual UWS is competitive against Monolingual UWS
- All languages tested followed the **same trend**

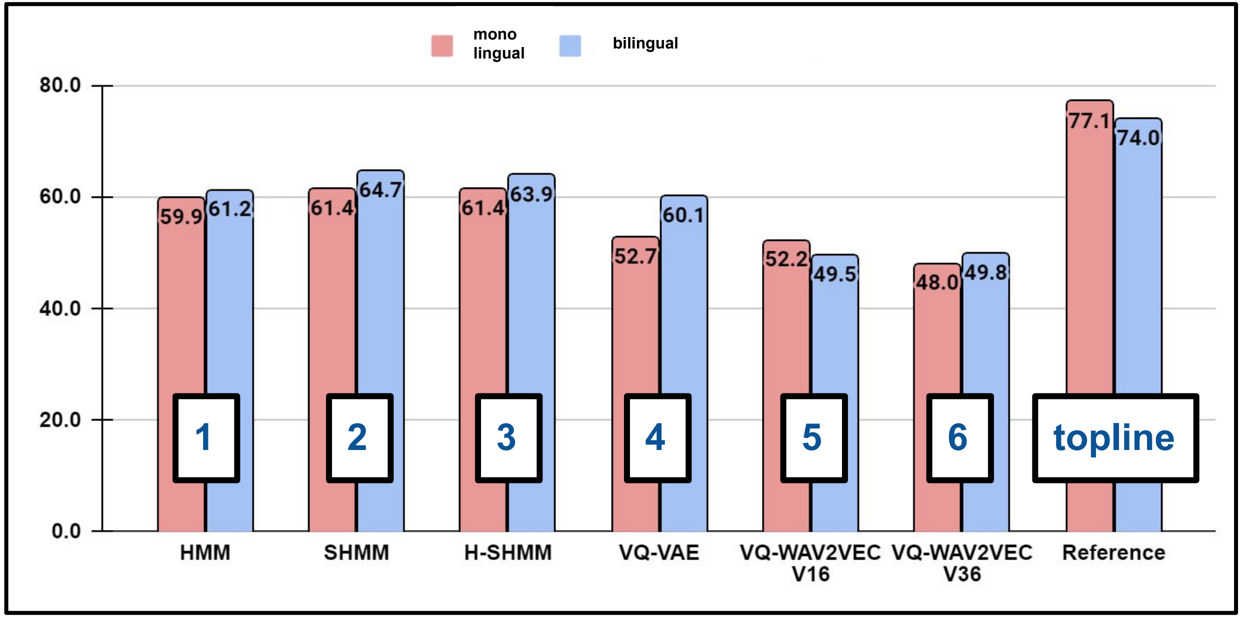


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Results for Mboshi

- **Bayesian models** are the most exploitable, in special SHMM and H-SHMM
- **VQ-models** are difficult to directly exploit for our task
  - ➔ Also verified recently in Kamper and Nieker [14]
  - ➔ An extra step of post-treatment might be necessary

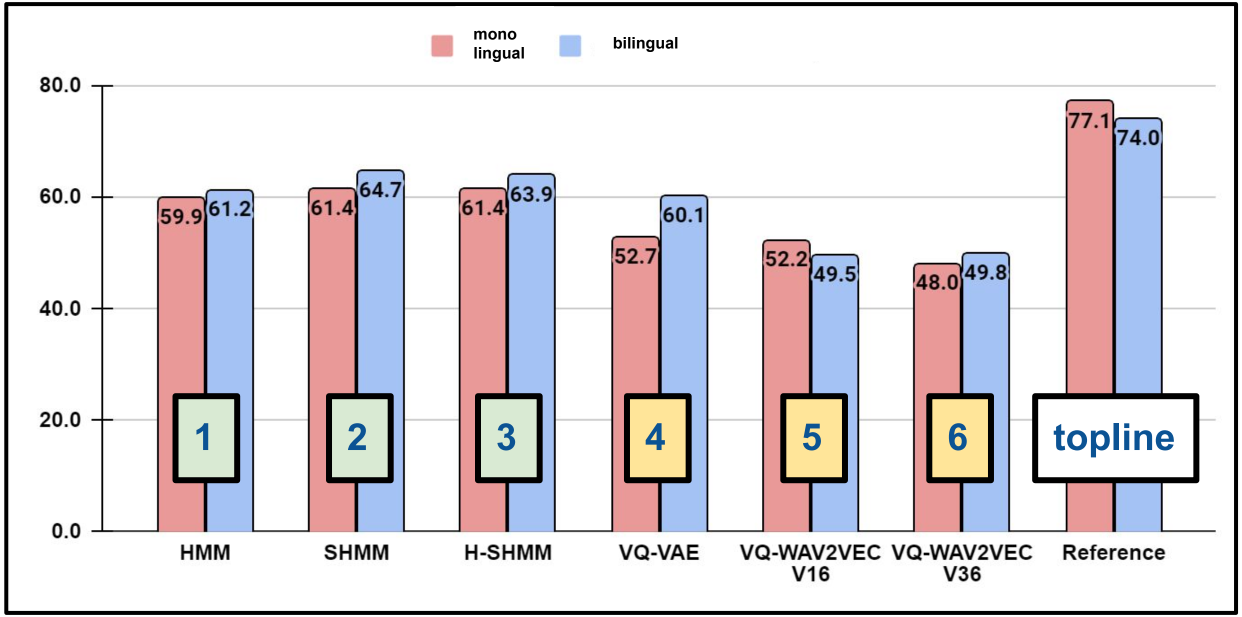


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Results for the MASS Languages (FI, HU, RO, RU)

- Results only for Bayesian SD due to the **excessive output discretization length for neural**
- Results follow the same trend from the Mboshi language: **Bilingual UWS** is competitive against **Monolingual UWS**.

	FI	HU	RO	RU
HMM	45.6 | **53.4**	49.9 | **51.2**	53.5 | **56.6**	47.1 | **54.9**
SHMM	49.0 | **56.0**	52.3 | **53.9**	53.5 | **57.7**	50.5 | **57.7**
H-SHMM	50.5 | **56.1**	52.9 | **53.3**	58.0 | **59.6**	52.9 | **56.0**

Table: Boundary UWS F-score results for the different SD models, using the MASS dataset (dpseg/**attention-based**). The result is the average over 5 runs.

CONCLUSIONS

Concluding…

- We update our pipeline for unsupervised word segmentation (UWS) from speech
    - We test in more languages, and we reach higher scores for Mboshi
    - We explore novel approaches for speech discretization

- **Neural speech discretization approaches do not perform well** in our pipeline
    - They produce inconsistent representation, difficult for downstream text-based approaches

- **Extra annotation can be beneficial when the input is noisy!**
    - The bilingual UWS model (access to translations) consistently outperforms monolingual UWS

Thank you!

Questions?

Bibliography

[1] Joshi, et al. ***The state and fate of linguistic diversity and inclusion in the NLP world.*** ACL 2020.

[2] Adda et al. ***Breaking the unwritten language barrier: The BULB project.*** SLTU 2016.

[3] Godard et al. ***Unsupervised word segmentation from speech with attention.*** Interspeech 2018.

[4] Goldwater et al. ***A Bayesian framework for word segmentation: Exploring the effects of context. Cognition.*** 2009.

[5] Boito et al. ***Unwritten languages demand attention too! word discovery with encoder-decoder models.*** ASRU 2017.

[6] Dunbar, Ewan, et al. ***The zero resource speech challenge 2017.*** ASRU 2017.

[7] Ondel et al. ***Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery.*** Interspeech 2019.

[8] Yusuf et al. ***A Hierarchical Subspace Model for Language-Attuned Acoustic Unit Discovery.*** ICASSP 2020.

[9] Oord et al. ***Neural Discrete Representation Learning.*** NeurIPS 2017.

[10] Baevski et al. ***vq-wav2vec: Self-supervised Learning of Discrete Speech Representations.*** arXiv, 2019.

[11] Ondel et al. ***Variational inference for acoustic unit discovery.*** Procedia Computer Science 2016.

[12] Godard et al. ***A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments.*** LREC 2018.

[13] Boito et al. ***MaSS: A large and Clean Multilingual Corpus of Sentence-aligned Spoken Utterances Extracted from the Bible.*** LREC 2020.

[14] Kamper and Nieker. **T*owards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks.*** arXiv, 2020.

[15] 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