Models and Resources for Attention-based Unsupervised Word Segmentation

An Application to Computational Language Documentation

Marcely Zanon Boito

September 30, 2021

Laboratory: LIG, University Grenoble Alpes (2017-2021)

Supervisors: Laurent **Besacier** (UGA and Naver Labs Europe)
Aline **Villavicencio** (Sheffield University, UK)

Before that…

Master Degree in Artificial Intelligence and Web (MoSIG, FR);
Double Degree in Computer Science (UFRGS, BR) and Information Systems Engineering (ENSIMAG, FR).

And now…

Postdoctoral Researcher in Low-resource Speech-to-Text Translation
SELMA Project - LIA Lab - Avignon University

This presentation agenda:

1. **Introduction:**

   The field: Computational Language Documentation
   The task: Unsupervised Word Segmentation (UWS) from speech

2. **Our Contribution:** An *attention-based* pipeline for UWS from speech

   PART 1: *Attention* for segmentation
   PART 2: *Speech discretization* in low-resource settings

3. **Conclusion**

Introduction

Language Documentation

- 50 to 90% of the currently spoken languages will go extinct before 2100 [1]
- Manually documenting all these languages is infeasible

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

Computational Language Documentation (CLD)

- 50 to 90% of the currently spoken languages will go extinct before 2100 [1]
- Manually documenting all these languages is infeasible

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

GOAL: to automatically retrieve information about language structures to speed up language documentation

Approaches for CLD: Documentation Corpora

- 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 [2]

s

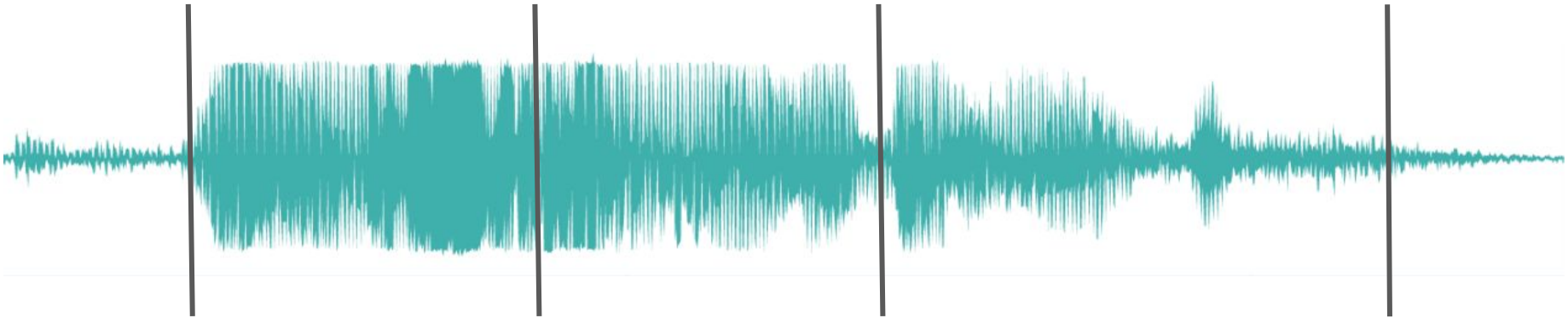
Approaches for CLD: Documentation Corpora

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

Therefore, CLD approaches need to...

1. Deal with speech
2. Be robust to low-resource
3. Incorporate bilingual (or multilingual) annotations

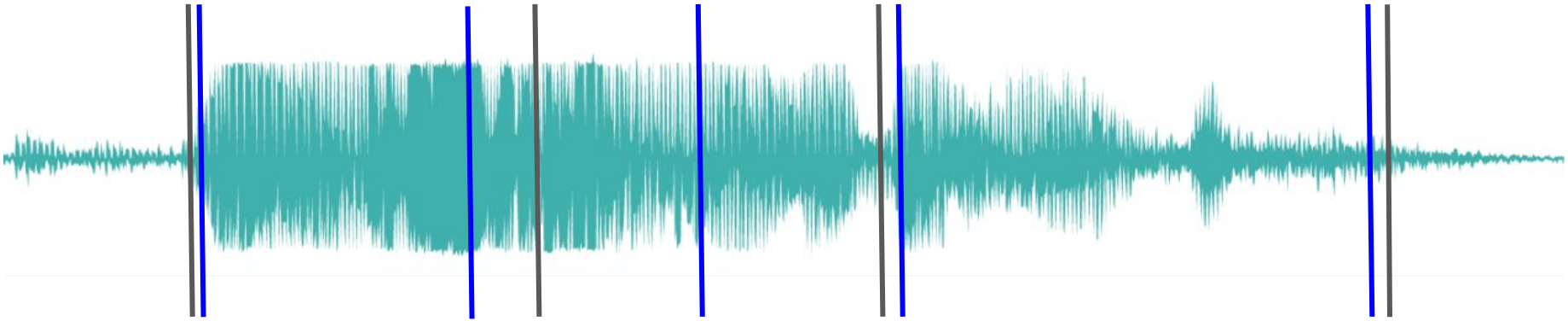
UNSUPERVISED WORD SEGMENTATION (UWS) from speech



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.



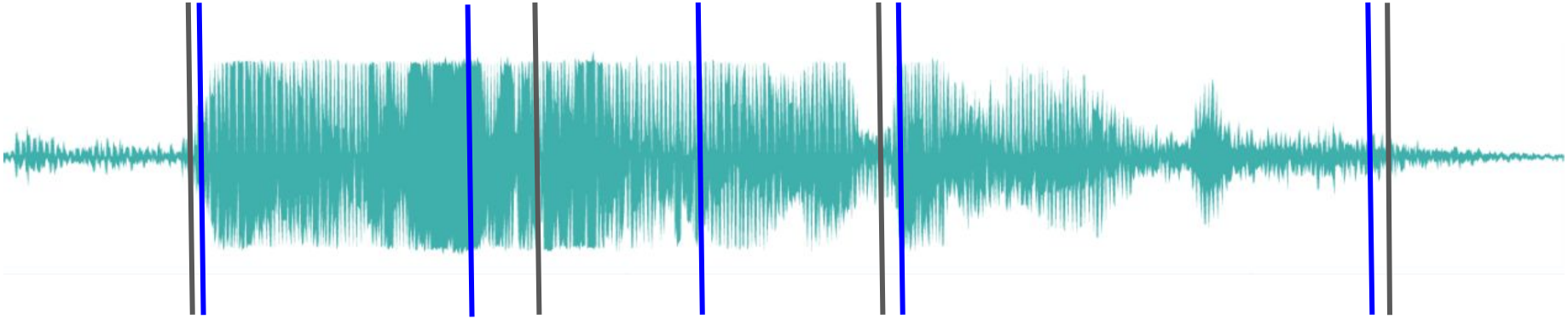
Literature in (monolingual) UWS

- The UWS task is more often solved in the symbolic domain (grapheme or phonemes) [3,4,5,6]
  - **Transcribing one minute of audio takes on average one hour and a half of work from a trained linguist** [38]

- For speech, there's mostly research on *Unsupervised Term Discovery*, which produces a partial segmentation of the speech signal [7-9]
  - Focus of Zero Resource Speech Challenge these last years [36,37]

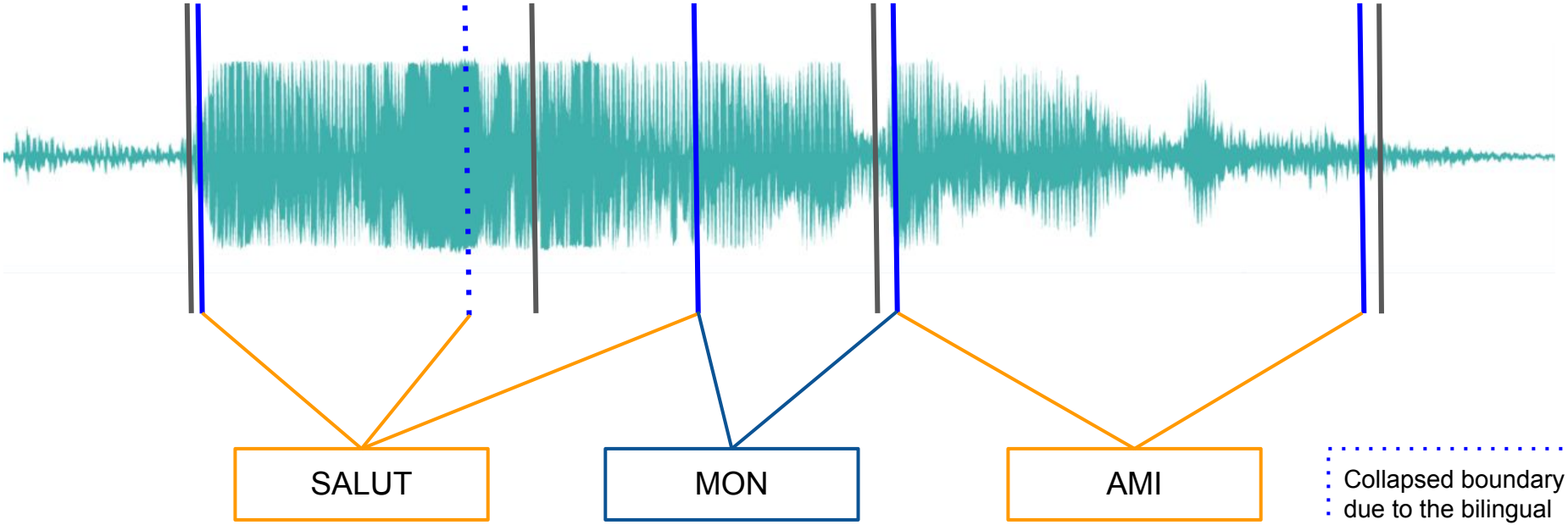
What we propose: Grounding Segmentation on Translations

Our system outputs segmentation based on…

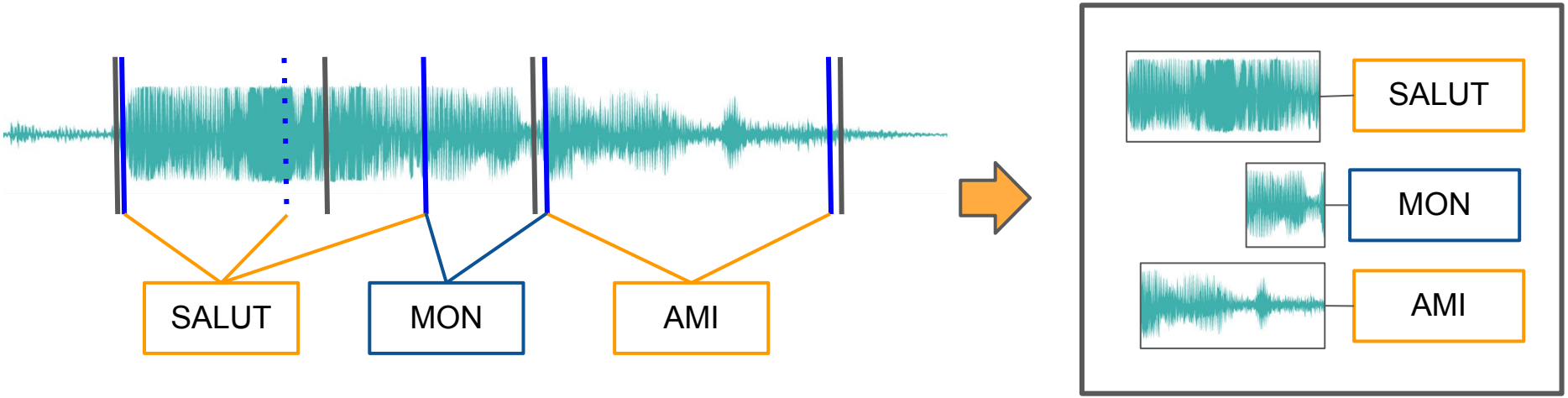


What we propose: Grounding Segmentation on Translations

Our system outputs segmentation based on… **sentence-level translations.**



Grounding Segmentation on Translations



➔ In this setting, all our boundaries have an *annotation*: the bilingual information aligned.[1]

[1]Using phonemes (textual domain), this bilingual segmentation setting was studied in [10, 11].

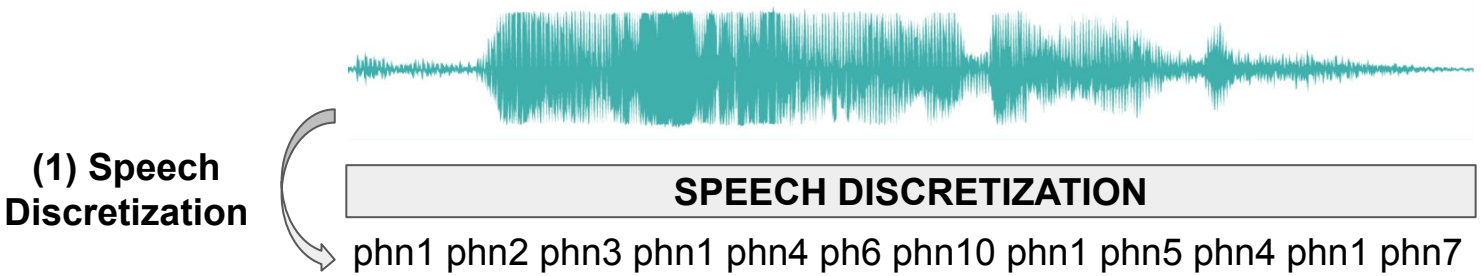
Bilingual UWS from Speech in Low-resource Settings

SALUT

MON

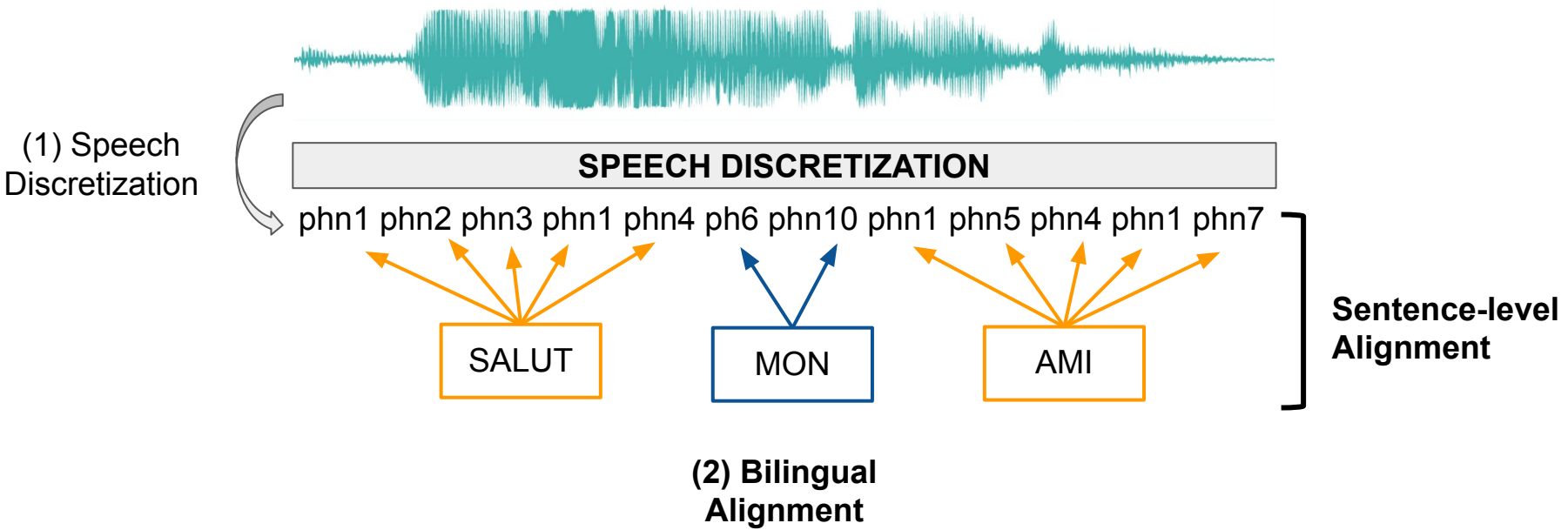
AMI

Bilingual UWS from Speech in Low-resource Settings

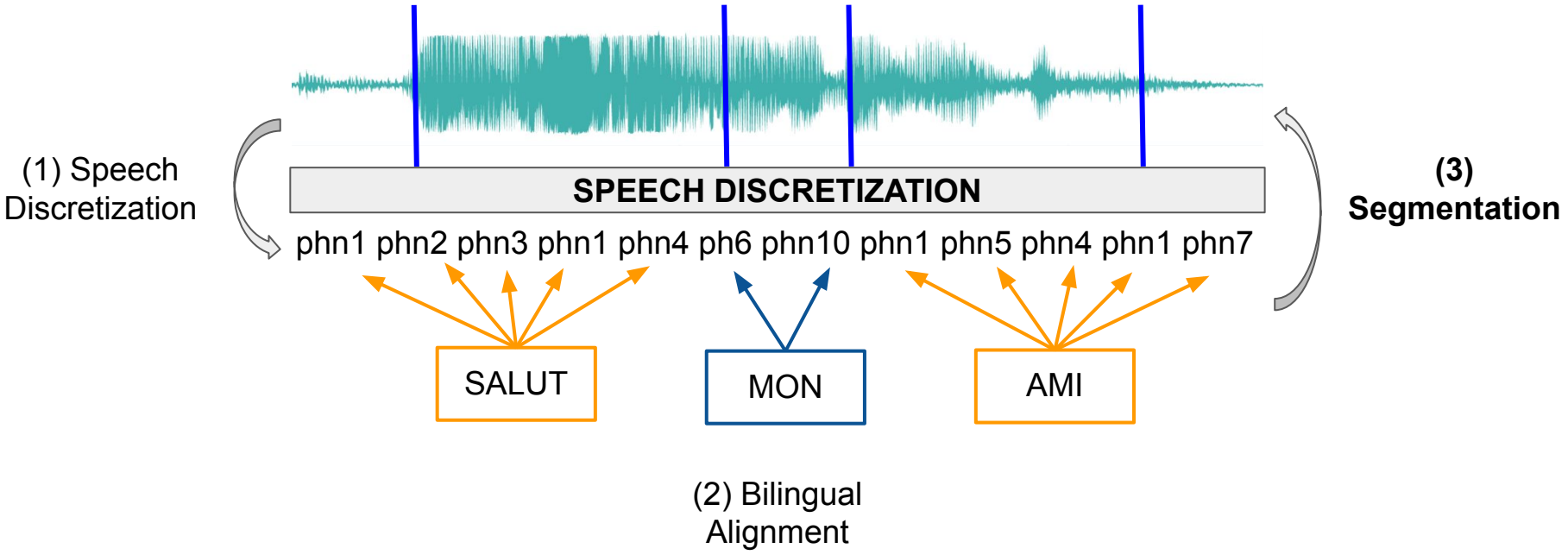


➔ Accommodates the challenge of processing speech in low-resource settings by first creating an unsupervised discretization of the signal

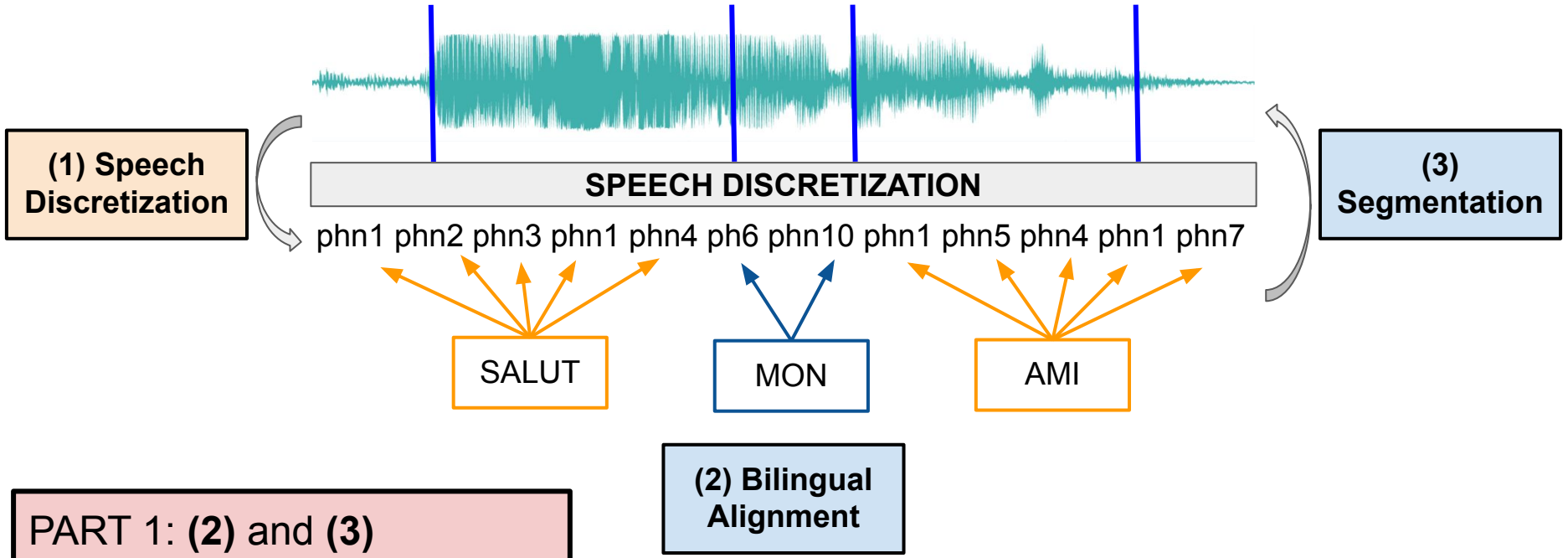
Bilingual UWS from Speech in Low-resource Settings



Bilingual UWS from Speech in Low-resource Settings



Bilingual UWS from Speech in Low-resource Settings



PART 1: **(2)** and **(3)**
PART 2: **(1)** and final results

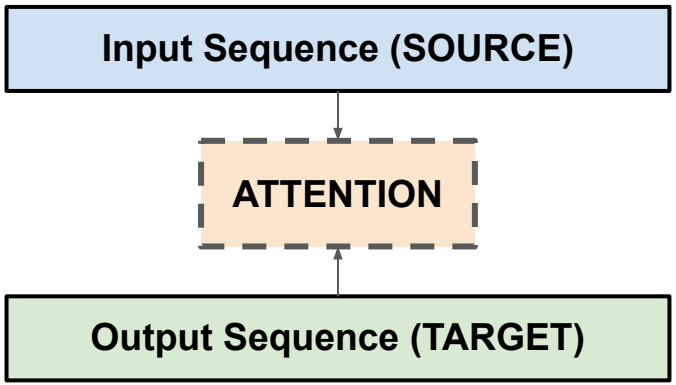
PART 1
A Bilingual Attention-based UWS Model

Corresponding publications:

- **Empirical Evaluation of Sequence-to-Sequence Models for Word Discovery in Low-resource Settings.** Boito et al. **INTERSPEECH 2019.**
- **Investigating Alignment Interpretability for low-resource NMT.** Boito et al. **Machine Translation Journal: Special Issue on Machine Translation for Low-resource Languages. Springer Netherlands 2021.**

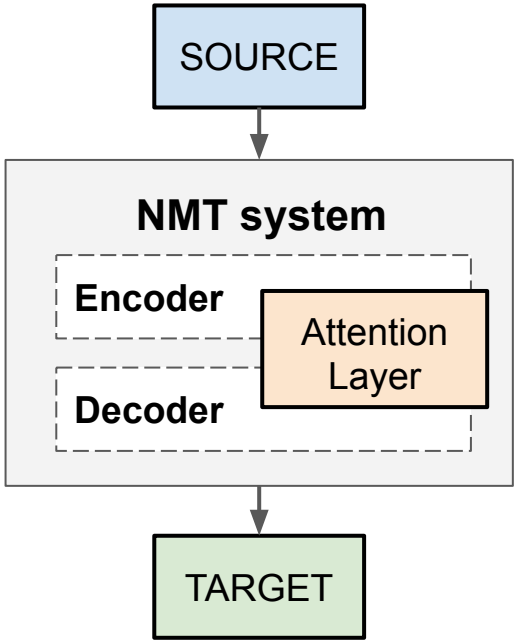
Towards Bilingual Supervision

- Sequence-to-sequence (seq2seq) models interfaced with **attention** emerged as popular solutions for a variety of NLP tasks:
  - Automatic Speech Recognition [24,25] (Source: speech, Target: text)
  - Text-to-Speech Synthesis [22,23] (Source: text, Target: speech)
  - **Neural Machine Translation** [12,15,16] (Source: text/speech, Target: text)



Towards Bilingual Supervision

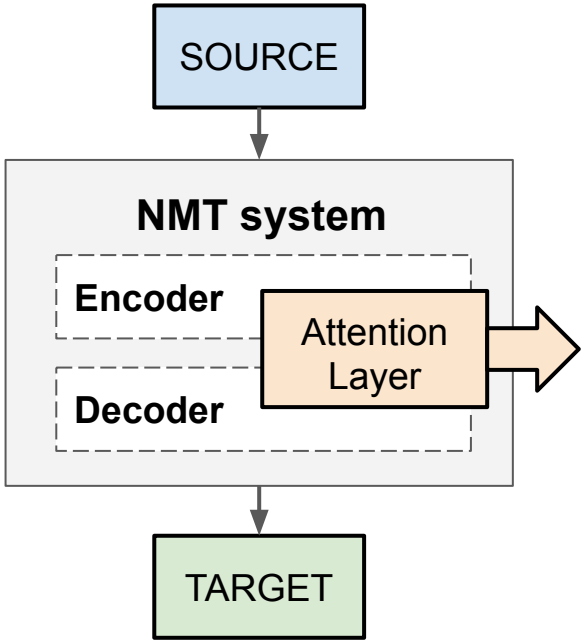
Neural Machine Translation (NMT) models



- Trained with bilingual datasets
- Attention Layer captures the *importance* of source tokens for generating each target token
- Posterior to training, the output of this layer can be visualized

Towards Bilingual Supervision

Neural Machine Translation (NMT) models



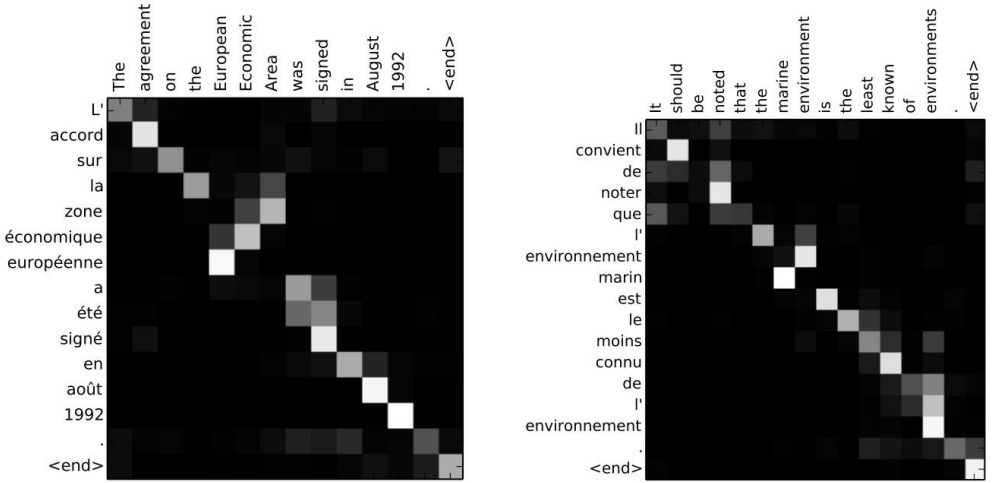
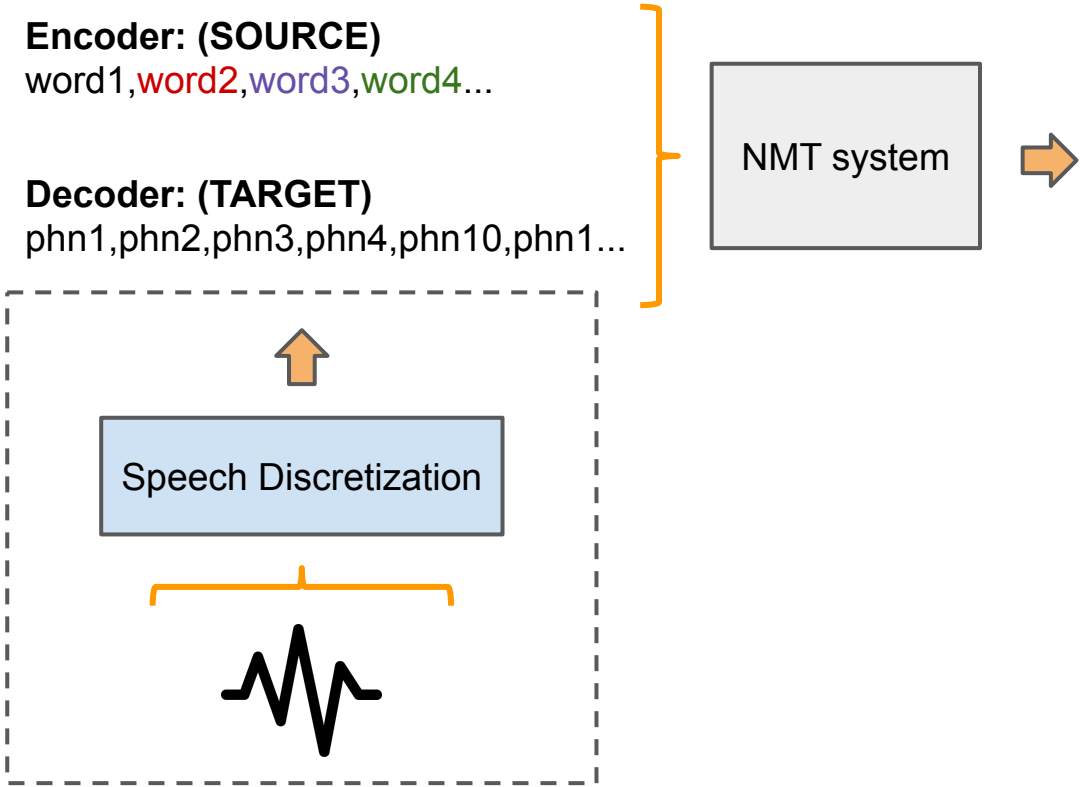


Figure: soft-alignment heatmaps from Bahdanau et al. 2015 [12]

Producing Bilingual Alignment and Segmentation

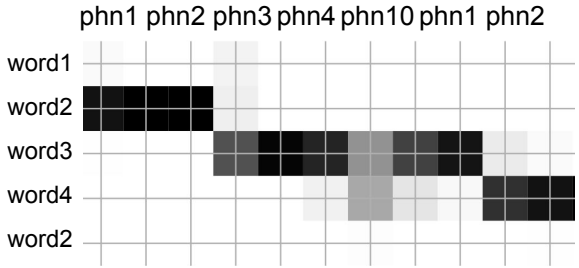


Producing Bilingual Alignment and Segmentation

Encoder: (SOURCE)
word1,word2,word3,word4...

Decoder: (TARGET)
phn1,phn2,phn3,phn4,phn10,phn1...

NMT system



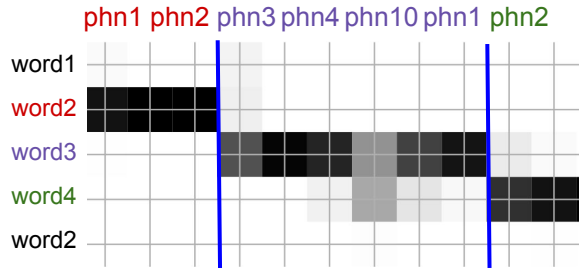
Speech Discretization

Producing Bilingual Alignment and Segmentation

Encoder: (SOURCE)
word1,word2,word3,word4...

Decoder: (TARGET)
phn1,phn2,phn3,phn4,phn10,phn1...

NMT system



Speech Discretization

Segmentation:
phn1phn2, phn3phn4phn10phn1, phn2

Alignment:
(phn1phn2, word2);
(phn3phn4phn10phn1, word3);
(phn2, word4)

Bilingual UWS: Research Questions

R1. Can we use the **soft-alignment probability matrices** learned during NMT training for segmentation in low-resource settings?

R2. What is the impact of the **type of attention mechanism**?

R3. What is the impact of **dataset** size?

R4. What is the **language** impact?

Not presented here, but investigated in Boito et al. [13] (Chapter 4, Section 2)

Experimental Settings

R1. Can we use the soft-alignment probability matrices learned during NMT training for segmentation in low-resource settings?

- ➔ We start from the **topline performance** expected for a speech discretization model: the true phones in the target language.
- ➔ We compare our model against a strong (monolingual) baseline **dpseg**[1][3]. This baseline is a *monolingual* approach for UWS, very robust in low-resource.

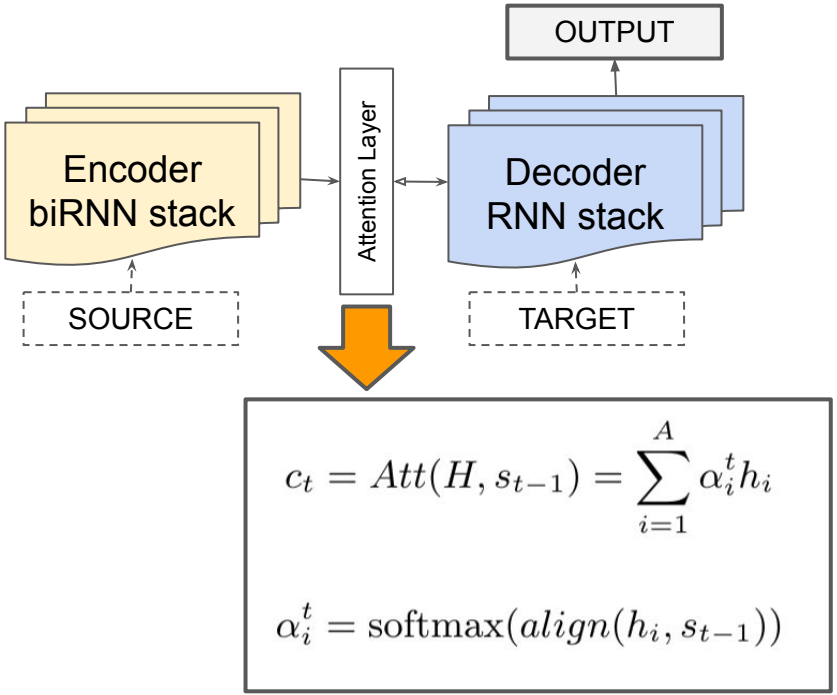
[1] Available at https://homepages.inf.ed.ac.uk/sgwater/, parameters from **[14]**.

Experimental Settings: 3 different NMT models

R2. What is the impact of the **type of attention mechanism**?

NMT Models (1): RNN

Global attention from Bahdanau et al. 2015 [12]

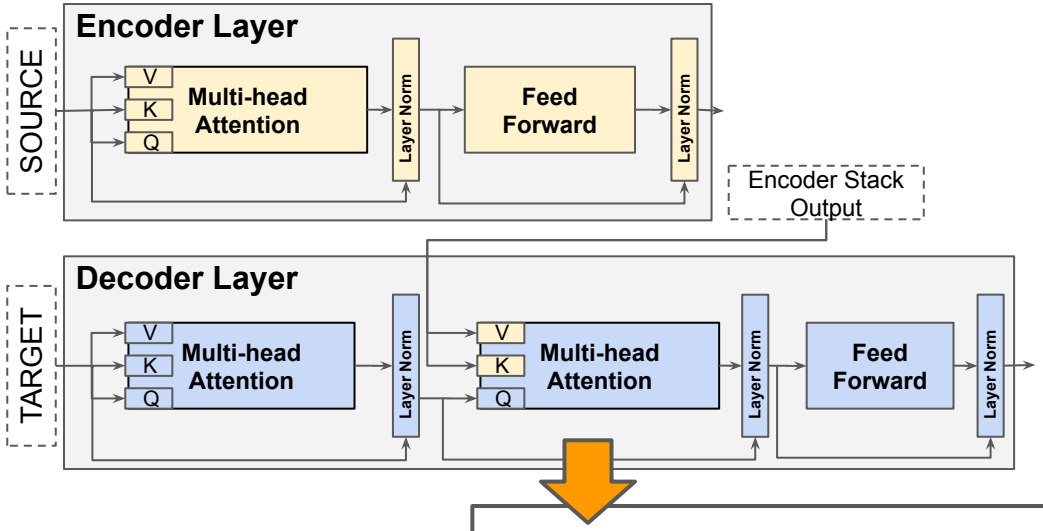


Attention appears in the form of **context vectors** for each decoder step *t*.

Computed using the set of source annotations H and the last state of the decoder network s_{t-1} (translation context).

The align layer is a feed-forward neural network trained jointly.

NMT Models (2): Transformer



$$MultiHead(V, K, Q) = f(Concat(H))$$

$$h_i = Att(f_i(V), f_i(K), f_i(Q))$$

$$Att(V, K, Q) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Multi-head attention from Vaswani et al. 2017 [15]

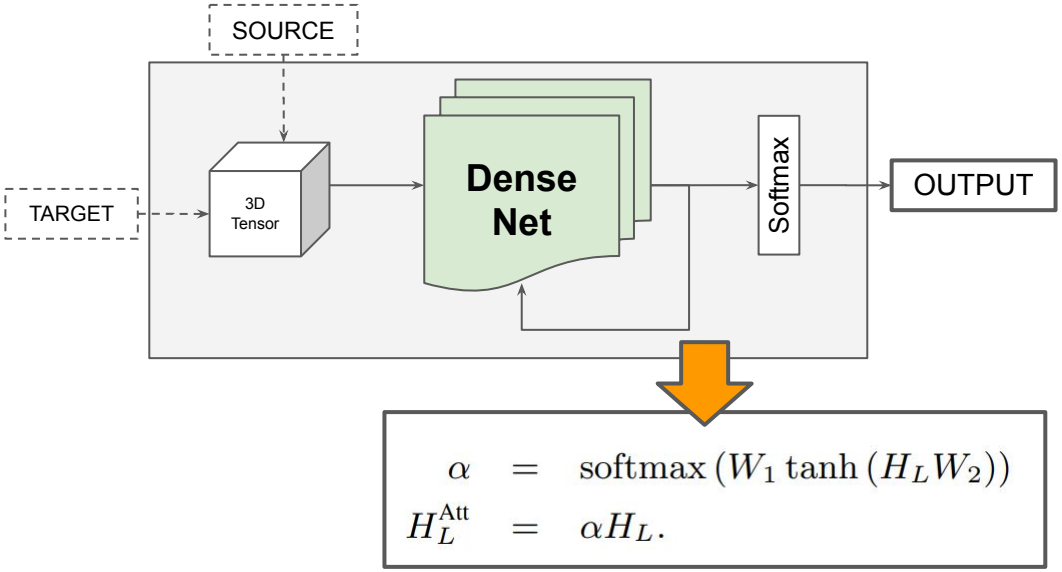
From a pair of key-value vectors and a query vector, the **attention layer** produces the weighted sum.

Weights computed by **Scaled dot-product (SDP) Attention** for each head.

Multi-head attention: SDP for several heads.

NMT Models (3): 2D-CNN

Pervasive attention from Elbayad et al. 2018 [16]



$$\alpha = \text{softmax}\left(W_1 \tanh\left(H_L W_2\right)\right)$$
$$H_L^{\text{Att}} = \alpha H_L.$$

Source and target sequences are encoded jointly. This acts as an **attention-like mechanism,** since individual source elements are re-encoded as the output is generated.

Attention weight tensor α is computed from the last activation tensor H_L, to pool the elements of the same tensor along the source dimension.

Experimental Settings: 3 different NMT models

R2. What is the impact of the **type of attention mechanism**?

➔ **RNN:** Global Attention [12]

The attention layer creates context vectors for weighting each target token.

➔ **Transformer:** Multi-head Attention [15]

Multiple *attentions* in parallel (heads) capture different equivalence functions between sequences.

➔ **2D-CNN:** Pervasive Attention [16]

Joint encoding acts as an attention-like mechanism. Source elements are re-encoded as the output is generated.

Experimental Settings: 3 datasets

R3. What is the impact of **dataset** size?

(MB-FR) Mboshi-French parallel corpus [17]
documentation dataset; tailored sentences

5,130 sentences (4h of speech) from the documentation of Mboshi, an unwritten language spoken in Congo-Brazzaville.[1]

[1]For comparison, the original Transformer NMT model was trained on 4.5 million parallel sentences

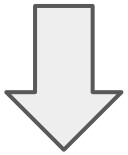
Experimental Settings: 3 datasets

R3. What is the impact of **dataset** size?

(MB-FR) Mboshi-French parallel corpus [17]
documentation dataset; tailored sentences

(EN-FR) English-French parallel corpus [18]
librispeech augmentation in French; noisy aligned information (filtered)

Data impact analysis

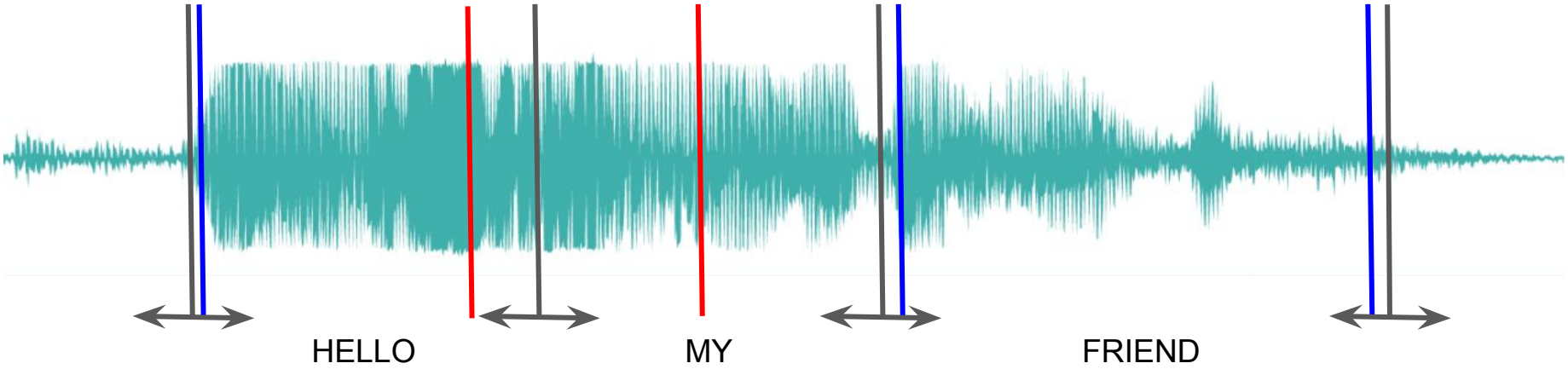


33K	**EN-FR (1)**	
5K	**EN-FR (2)**	**MB-FR (3)**

Experimental Settings: Evaluation

We evaluate it using tolerance windows.[1]

Precision (P):	**Recall (R):**	**F-score (F):**
Predicted & correct/predicted = 3/5	Predicted & correct / true = 3/4	2*(P*R)/(P+R) = 2/3



Inside the tolerance: **a hit.**
Outside the tolerance: **a miss.**

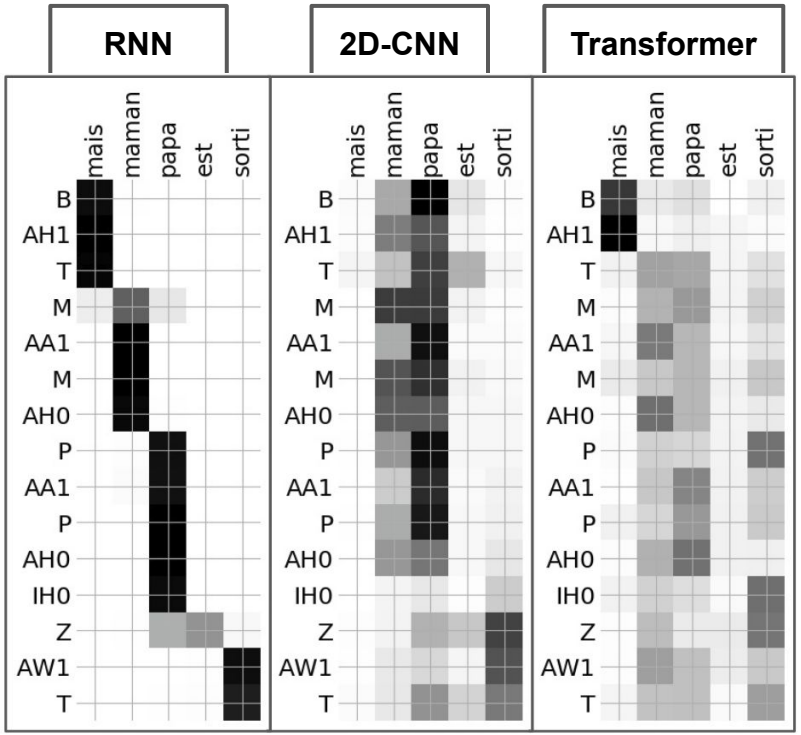
[1]The tolerance window we use is defined on the **Zero Resource Challenge 2017 Track 2.**

Experimental Settings: Evaluation

What about the alignment quality?

How do we evaluate this without having gold (word-level) alignment information?

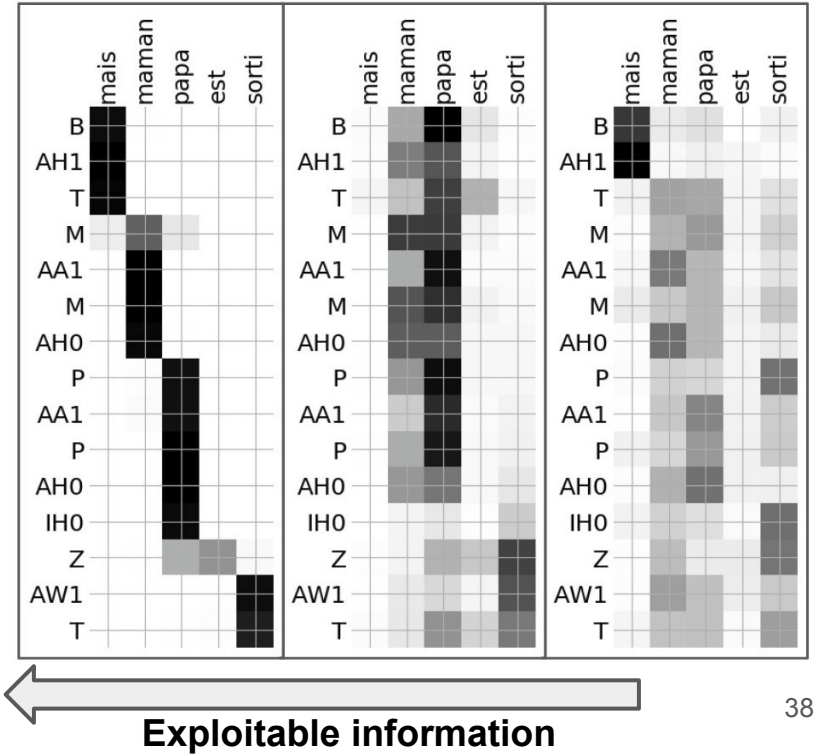
In a more practical sense:
Are all three of these good for our task?



Experimental Settings: Evaluation

Alignment Assessment with AVERAGE NORMALIZED ENTROPY (ANE)

- **Intuition:** sharper alignments are more informative.
- **Soft-alignment probability matrix:** one probability distribution per line (target symbol)

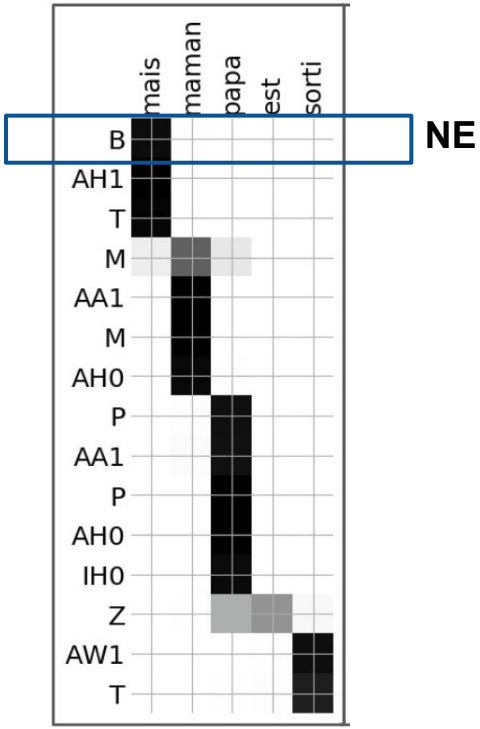


Experimental Settings: Evaluation

Alignment Assessment with AVERAGE NORMALIZED ENTROPY (ANE)

For every line in the matrix we compute normalized entropy (NE).

$$NE(t_i, s) = -\sum_{j=1}^{|s|} P(t_i, s_j) \cdot \log_{|s|}(P(t_i, s_j))$$



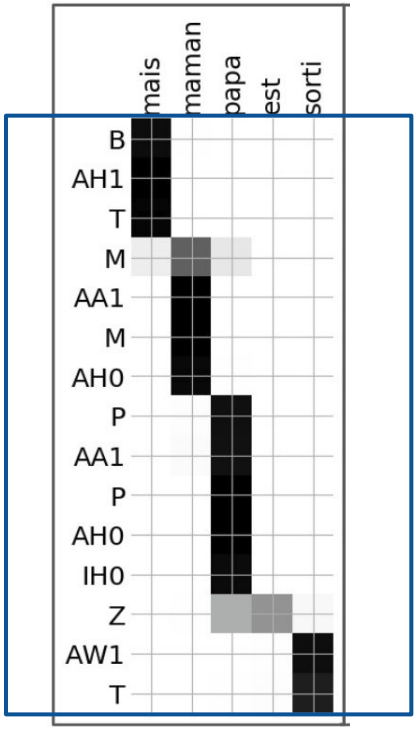
Experimental Settings: Evaluation

Alignment Assessment with AVERAGE NORMALIZED ENTROPY (ANE)

For every line in the matrix we compute normalized entropy (NE). We average over sets of distributions.

$$NE(t_i, s) = -\sum_{j=1}^{|s|} P(t_i, s_j) \cdot \log_{|s|}(P(t_i, s_j))$$

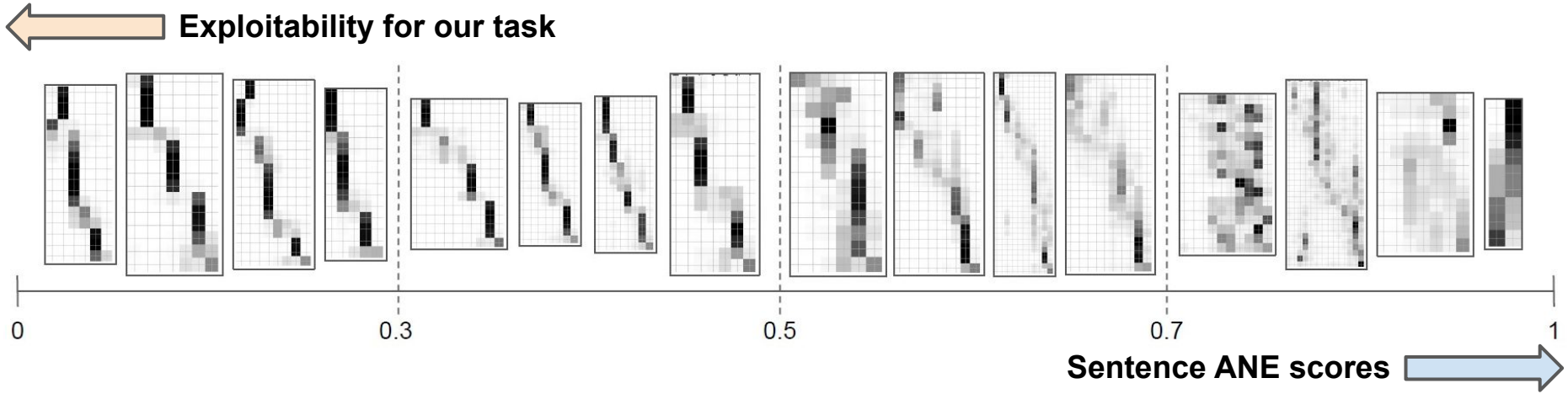
$$ANE(t, s) = \frac{\sum_{i=1}^{|t|} NE(t_i, s)}{|t|}$$



Sentence
ANE

Experimental Settings: Evaluation

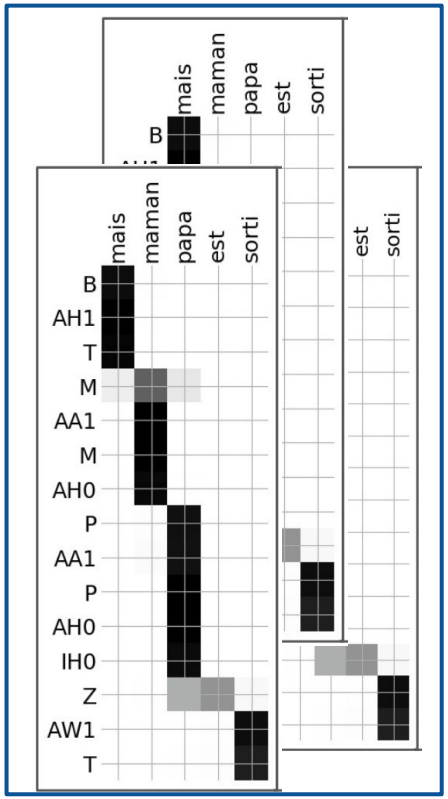
Alignment Assessment with AVERAGE NORMALIZED ENTROPY (ANE)



Experimental Settings: Evaluation

Alignment Assessment with AVERAGE NORMALIZED ENTROPY (ANE)

➔ To summarize the quality of the soft-alignment probability matrices produced by a given NMT model using a given dataset



Corpus
ANE

UWS Results

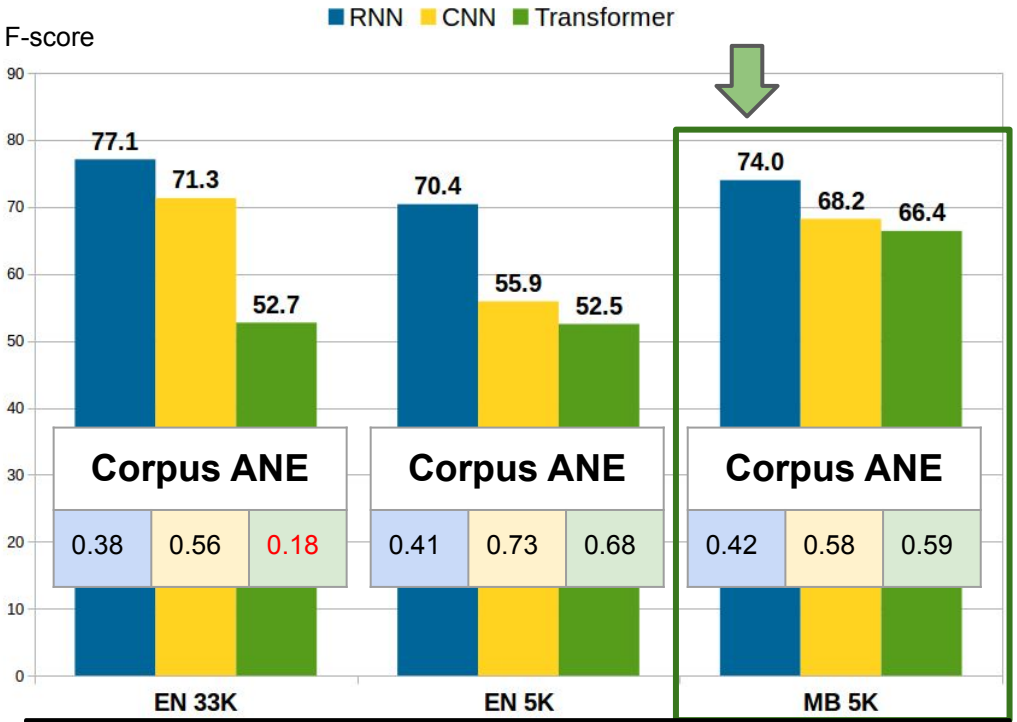


Figure: Boundary F-score Results averaged over 5 runs [19]

→ We are able to train models in very low-resource settings, scoring some points behind the **dpseg baseline** (77.1 for MB). **(R1)**

→ The **RNN-based model** performed the best in our setting. **(R2)**

Experimental settings and corpora: https://gitlab.com/mzboito/attention_study

UWS Results

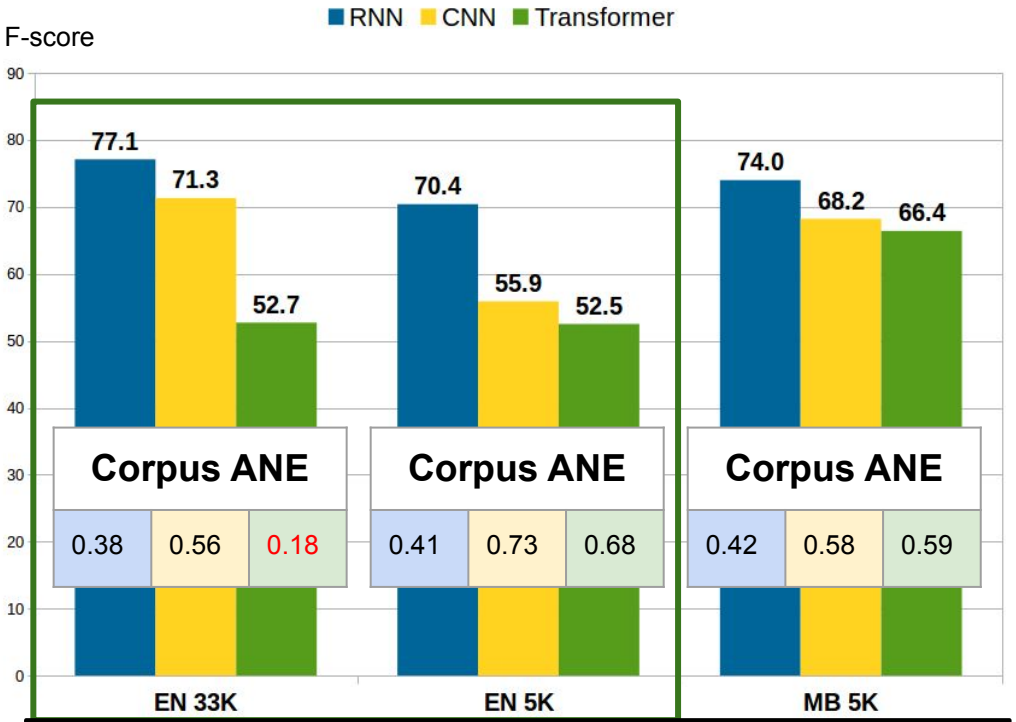


Figure: Boundary F-score Results averaged over 5 runs [19]

- We can see the impact of data reduction, but some models are more sensitive to it than others. **(R3)**
- Models with lower **Corpus ANE** reached better segmentation results (negative Pearson's correlation relationship). **(alignment assessment)**

Experimental settings and corpora: https://gitlab.com/mzboito/attention_study

UWS Results

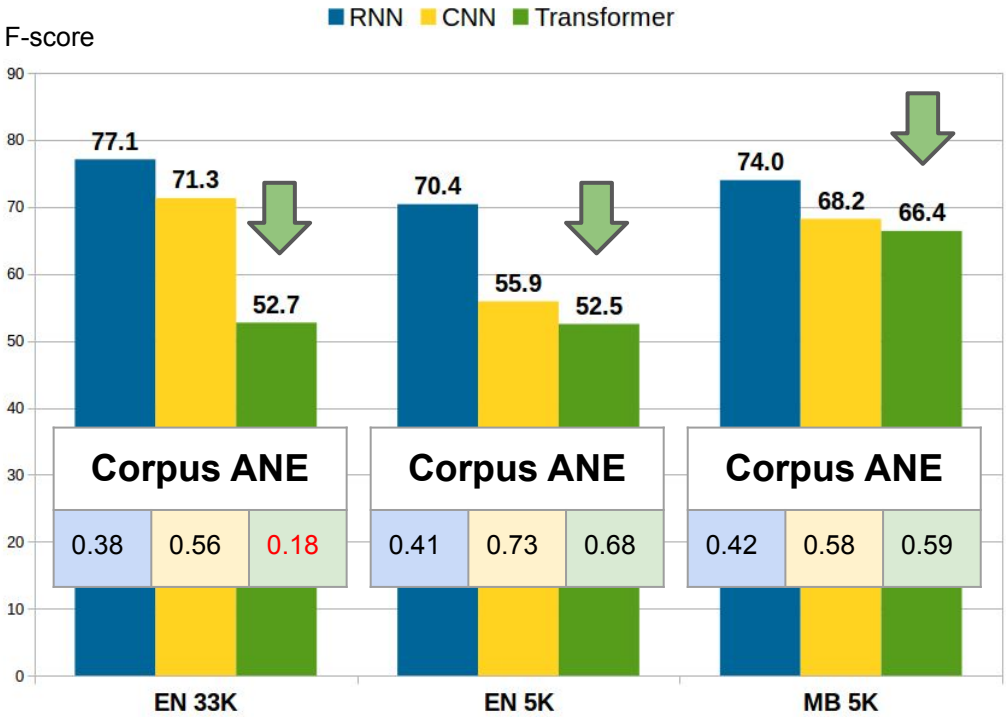


Figure: Boundary F-score Results averaged over 5 runs [19]

- How to choose a head from Transformer? [20,21]
- We reported results using corpus ANE for selecting the head.

We also experimented with:

- Models from 1 to 3 layers
- 1, 2 and 4 heads
- Intra- and inter-layer averaging

Experimental settings and corpora:
https://gitlab.com/mzboito/attention_study

UWS Results

➔ We showed that we are able to apply this pipeline for bilingual segmentation **starting from a perfect discretization for the speech**

We now focus on generating real speech discretization in low-resource settings

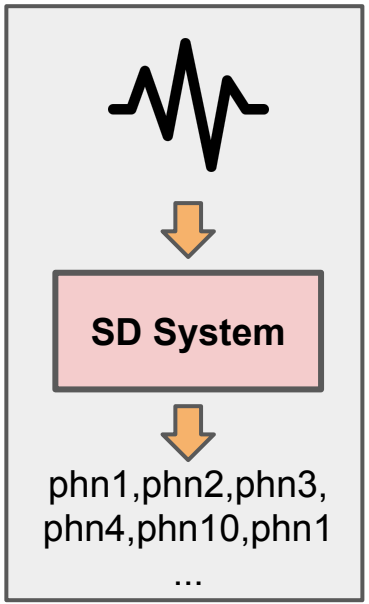
PART 2
Speech Discretization for UWS

Corresponding publications:

- **Unsupervised Word Segmentation from Speech With Attention.** Boito et al. **INTERSPEECH 2018.**
- **Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings.** Boito et al. **ArXiv 2021.**

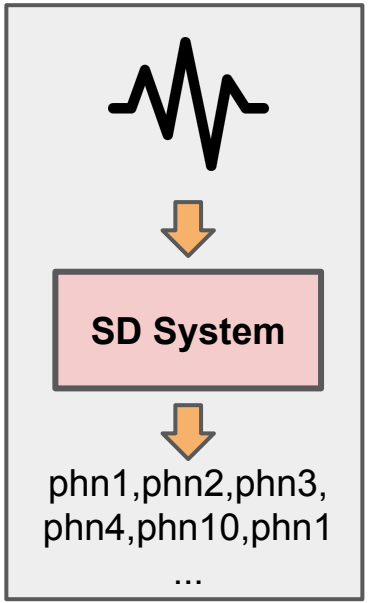
Exploitable SD models for Low-Resource UWS

➔ Speech Discretization (SD) models produce a sequence of **discrete speech units** representing input utterances with **no access to transcriptions** [26-30]



Exploitable SD models for Low-Resource UWS

➔ Speech Discretization (SD) models produce a sequence of **discrete speech units** representing input utterances with **no access to transcriptions** [26-30]



What do we expect from our discretization process?

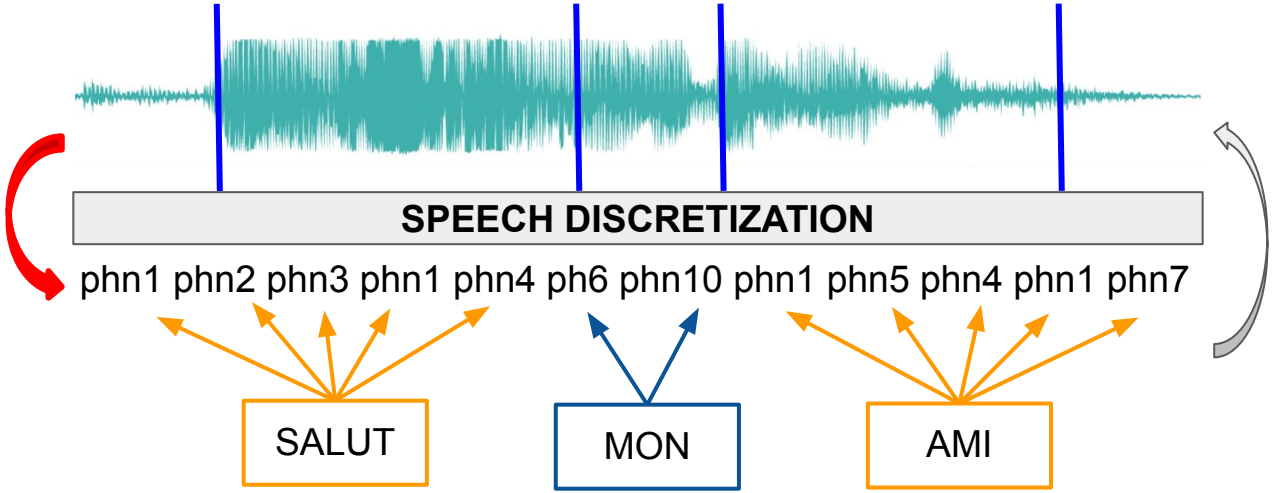
➔ The model needs to work well in low-resource.

➔ The model needs to output a **concise representation**:

- The baseline dpseg cannot deal with sequences longer than 350 units
- Our models can accommodate longer sequences, but it impacts performance (challenging alignment)

SD for Bilingual UWS: Research Question

R5. Can we directly use the output of SD models as input for our bilingual UWS approach in low-resource settings?



Speech Discretization Models: Bayesian Generative Models

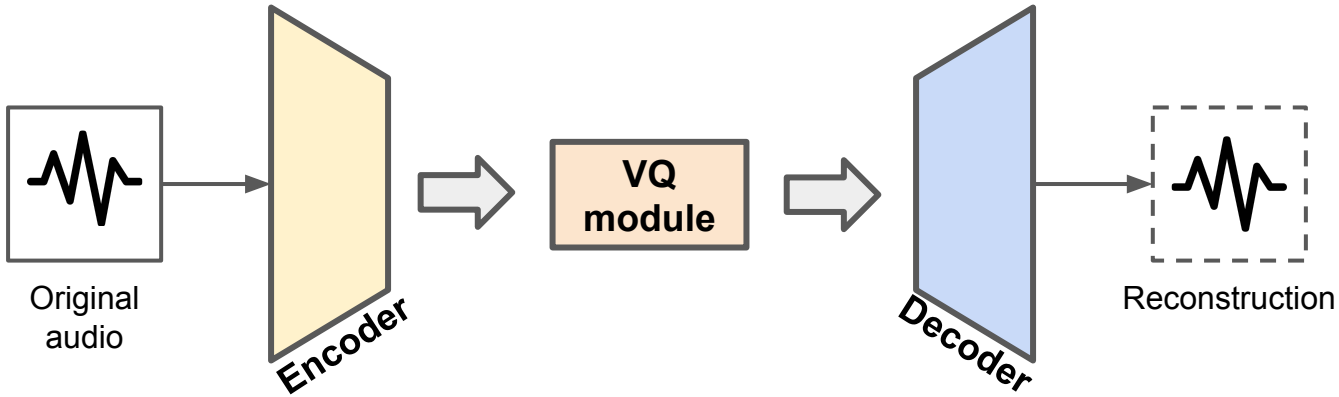
- Very efficient in low-resource settings

- Similar to a phone-loop model:
  - Each unit is modeled by an HMM/GMM
  - The prior distribution over all HMMs is modeled by a Dirichlet Process

- Models:
  1. HMM/GMM (**HMM**) [26]: Every possible sound can be a unit
  2. Subspace HMM (**SHMM**) [27]: Prior over a phonetic subspace
  3. Hierarchical Subspace HMM (**H-SHMM**) [28]: Subspace adaptation from different languages for phone prediction

Speech Discretization Models: Vector Quantization (VQ) Models

- Novel approaches for speech processing, popular in high-resource settings.
- Models:
  1. **VQ-Variational Auto-Encoder (VAE)** [29]: inspired by dimensionality reduction architectures



Speech Discretization Models: Vector Quantization (VQ) Models

➔ Models:

1. **VQ-VAE** [29]: inspired by input dimensionality reduction architectures

2. **VQ-WAV2VEC** [30]: inspired by self-supervised models trained with a context-prediction loss

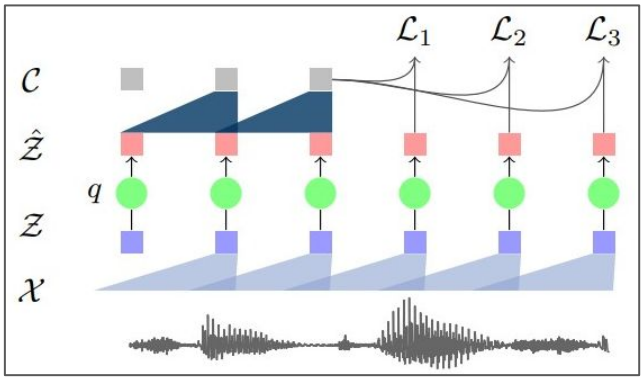


Figure: The vq-wav2vec architecture. Figure taken from the original paper [30]

1. Encoder (X→Z)
2. Quantizer (Z→Z')
3. Aggregator (Z'→C)

Experimental Settings

➔ We train all models with only 4 hours of speech. We focus on generating **concise representations**.

- Bayesian Models
  - HMM/GMM (**HMM**)
  - Subspace HMM (**SHMM**)
  - Hierarchical Subspace HMM (**H-SHMM**)
- VQ Neural Models
  - **VQ-VAE**
  - **VQ-WAV2VEC** (V16)
  - **VQ-WAV2VEC** (V36)

Trained on 4 hours of Mboshi data!

Statistics Over the Produced Sequences

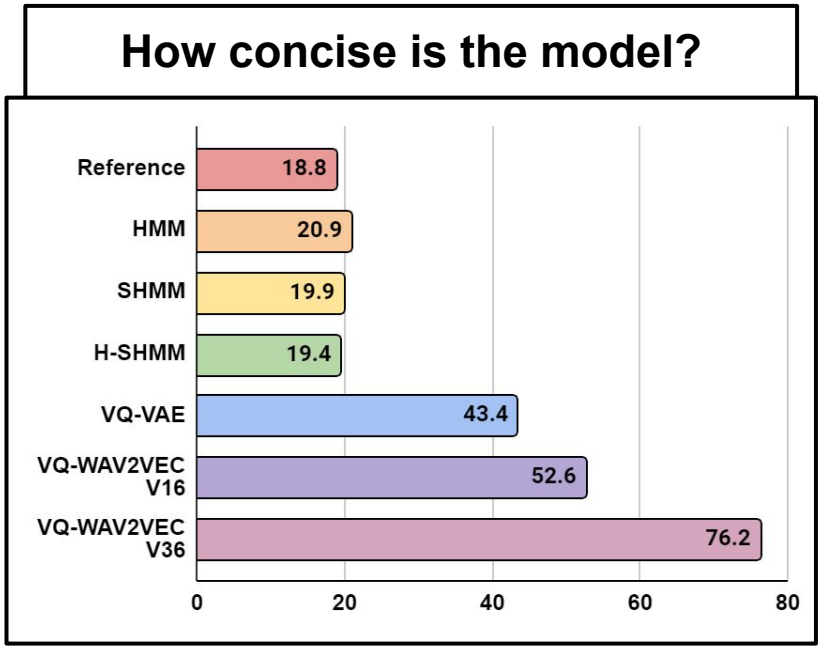


Figure: Average Sequence Length for SD models

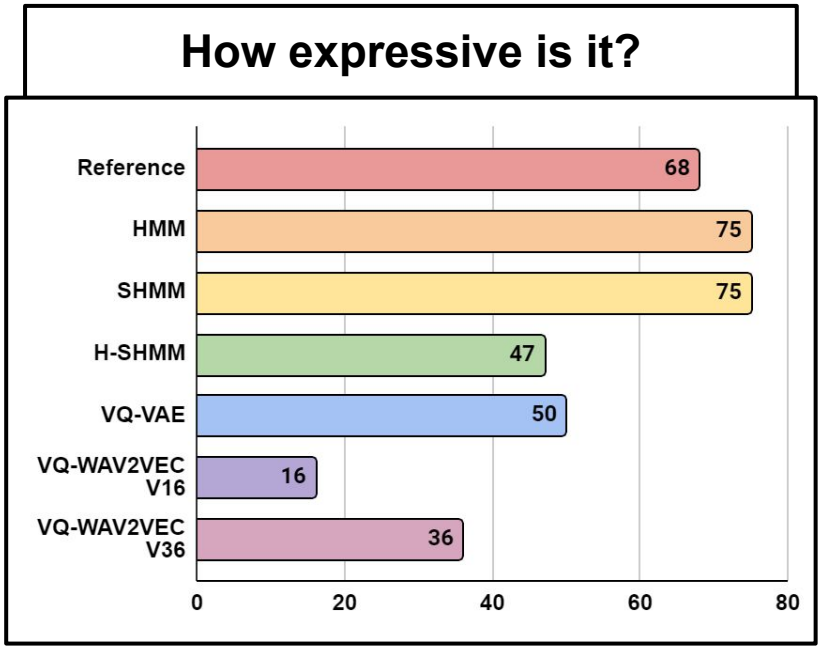


Figure: Vocabulary (# units) for SD models

Statistics Over the Produced Sequences: Bayesian Models

- ➔ The Bayesian models produce a more concise output, closer to the reference
- ➔ They also produce a similar number of units (excluding H-SHMM)

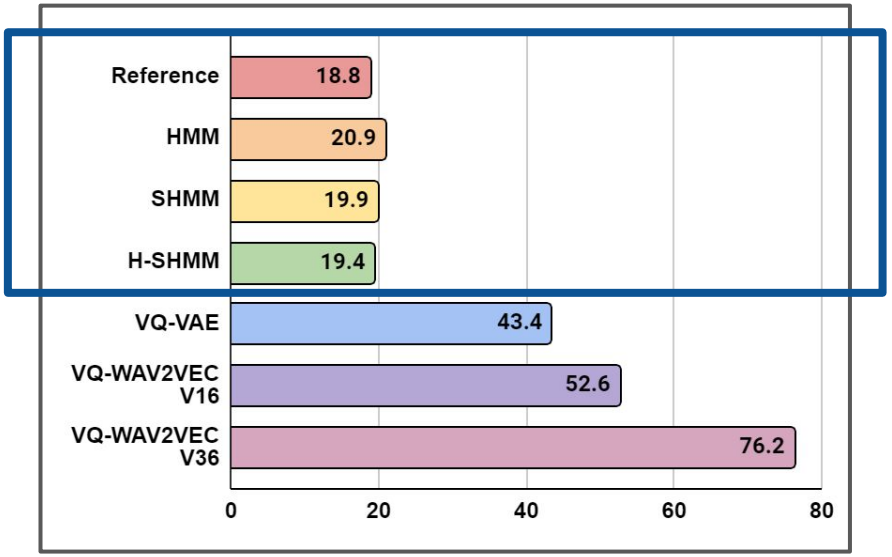


Figure: Average Sequence Length for SD models

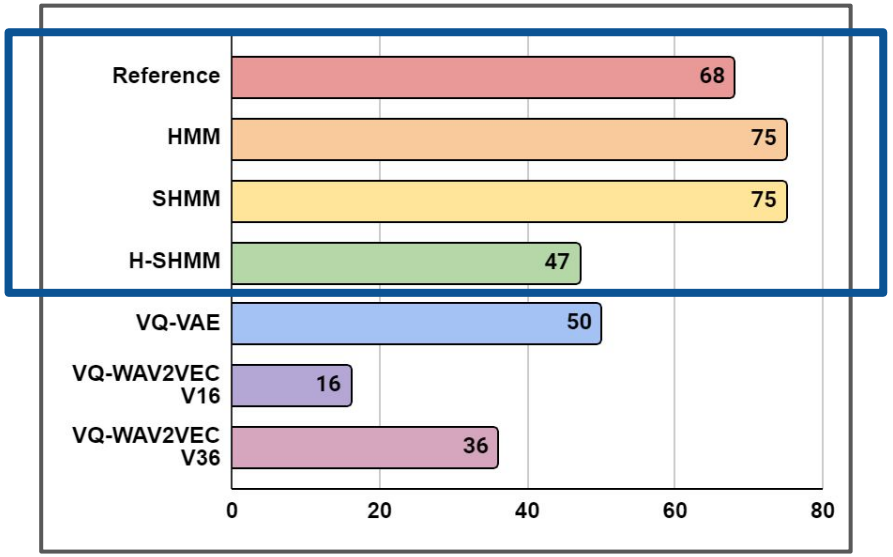


Figure: Vocabulary (# units) for SD models

Statistics Over the Produced Sequences: VQ Neural Models

➔ In order to reduce the length of the representation generated by VQ-based models, we are forced to also reduce the phone vocabulary.

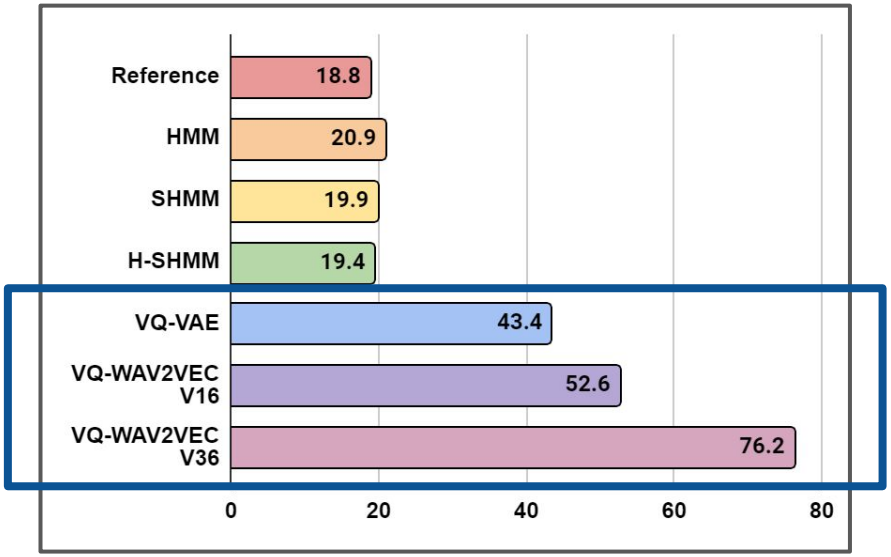


Figure: Average Sequence Length for SD models

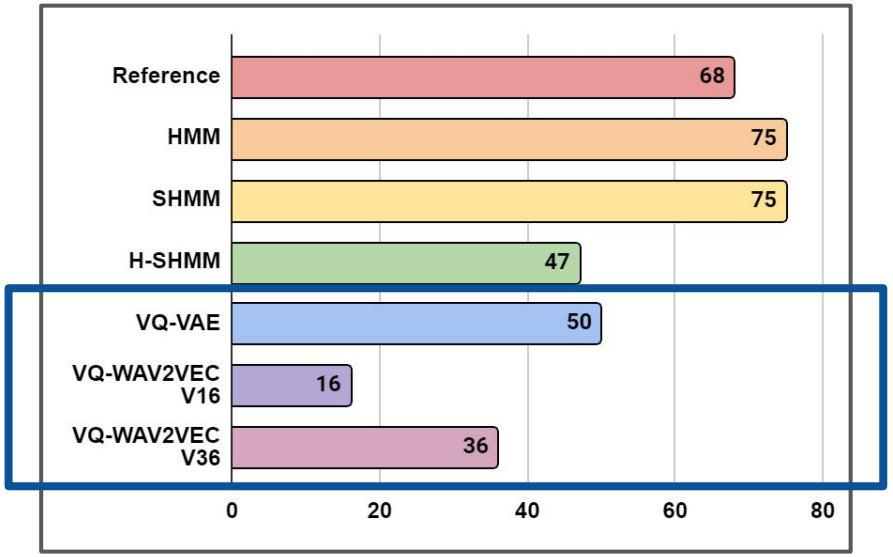


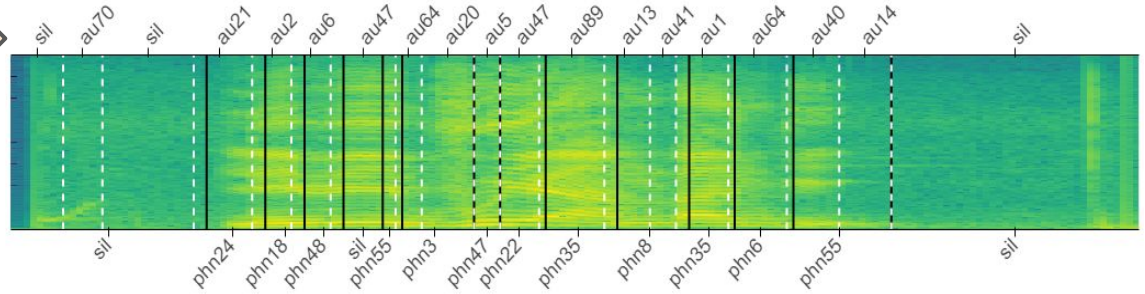
Figure: Vocabulary (# units) for SD models

Studying the SD Representation

Example: The same sentence, two approaches

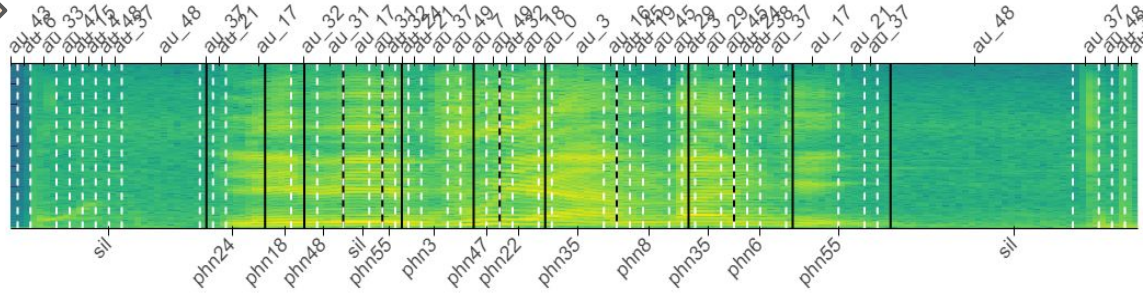
True Boundary
Output Boundary

H-SHMM output



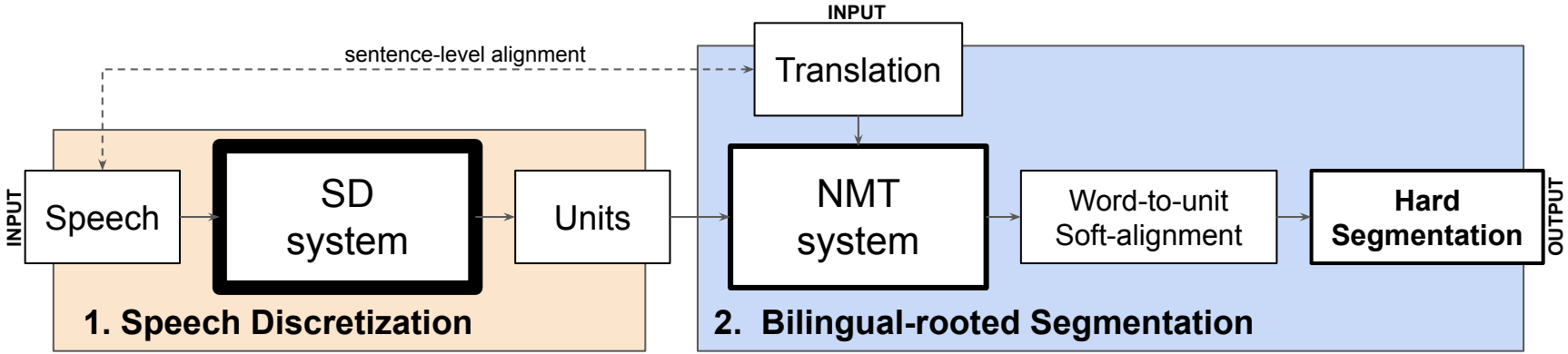
Reference

VQ-VAE output



Reference

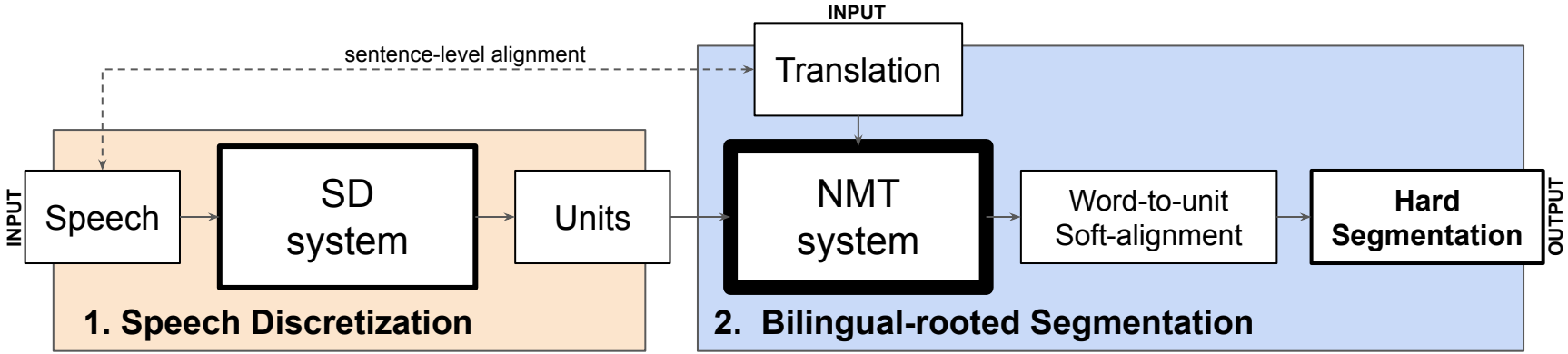
Bilingual UWS from Speech in Low-resource Settings



→ 6 setups for SD:

- Bayesian Models: HMM, SHMM, H-SHMM
- VQ Neural Models: VQ-VAE, VQ-WAV2VEC (V=16), VQ-WAV2VEC (V=36)

Bilingual UWS from Speech in Low-resource Settings



→ Best NMT model: RNN from Bahdanau et al. [12]

Bilingual UWS from Speech: Results

- Results for Mboshi
- 5 models, 6 setups
  1. HMM
  2. SHMM
  3. H-SHMM
  4. VQ-VAE
  5. VQ-W2V V=16
  6. VQ-W2V V=36

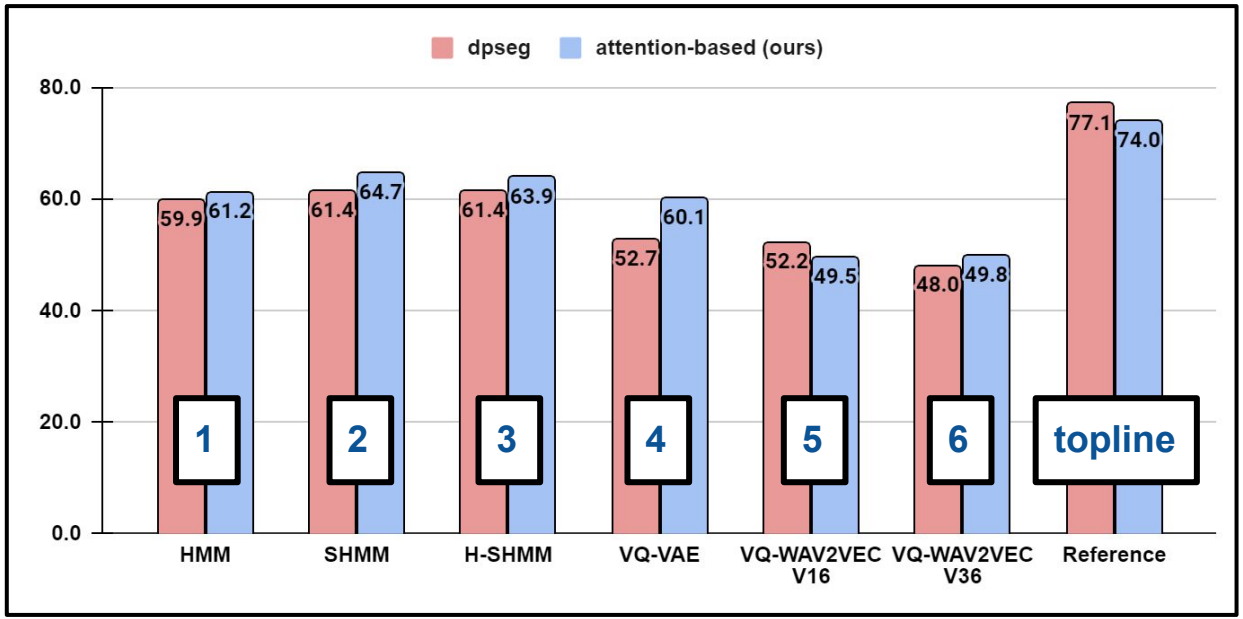


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Bilingual UWS from Speech: Results

- We notice a drop in performance, but **we still successfully generate segmentation (R5)**
- **We** are competitive against **dpseg**. Why?
  - The bilingual information might be helping us for this **noisier setup**

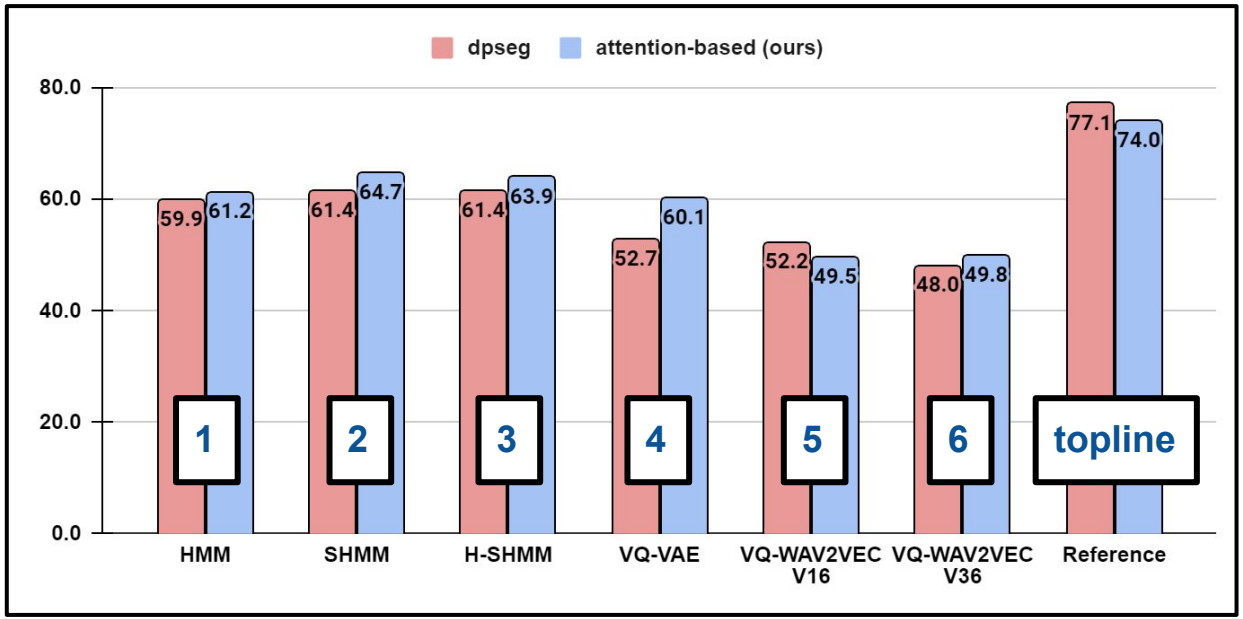


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Bilingual UWS from Speech: Results

- **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 [31]

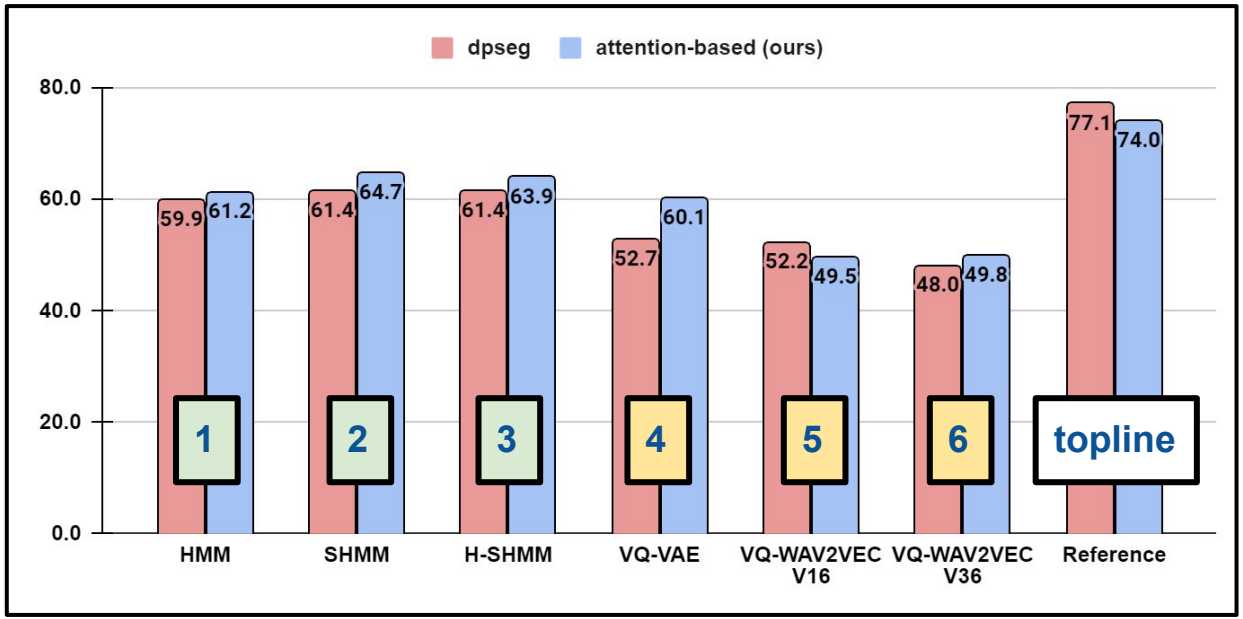


Figure: Boundary UWS F-score results for the different SD models, using the MB-FR dataset. The result is the average over 5 runs.

Conclusion

Bilingual UWS from Speech in Low-resource Settings

➔ We proposed a pipeline for CLD able to:

- Process speech in low-resource settings
- Incorporate bilingual information, generating bilingual links

SPEECH

Translations
to a well-documented language

Bilingual UWS from Speech in Low-resource Settings

- We proposed a pipeline for CLD able to:
  - Process speech in low-resource settings
  - Incorporate bilingual information, generating bilingual links

- In this process we:
  - Investigated different speech discretization approaches for UWS [32]
    - **Bayesian models produce a better representation, due to their Acoustic Unit Discovery modeling**

Bilingual UWS from Speech in Low-resource Settings

- We proposed a pipeline for CLD able to:
  - Process speech in low-resource settings
  - Incorporate bilingual information, generating bilingual links

- In this process we:
  - Investigated different speech discretization approaches for UWS [32]
  - Compared different attention-based NMT models in low-resource [19]
    - **Found the following ranking: RNN > 2D-CNN > Transformer**
    - **Proposed a task-agnostic metric (ANE) for assessing quality in soft-alignment probability matrices**

Bilingual UWS from Speech in Low-resource Settings

- We proposed a pipeline for CLD able to:
  - Process speech in low-resource settings
  - Incorporate bilingual information, generating bilingual links

- In this process we:
  - Investigated different speech discretization approaches for UWS [32]
  - Compared different attention-based NMT models in low-resource [19]
  - Achieved competitive results in a realistic scenario (only 5k sentences) against a strong monolingual baseline (dpseg).
    - **While not shown here, this trend was also verified in 4 other languages: Finnish, Hungarian, Romanian and Russian.**

Future Work

- Evaluation of the output of the system with linguists
  - Focus on the aligned annotation that the model produces

Future Work

- Evaluation of the output of the system with linguists
  - Focus on the aligned annotation that the model produces

- Leveraging information inside the attention layer during training
  - Biasing the alignment discovered, similar to Garg et al. [36] and Godard et al. [37]

Future Work

- Evaluation of the output of the system with linguists
  - Focus on the aligned annotation that the model produces
- Leveraging information inside the attention layer during training
  - Biasing the alignment discovered, similar to Garg et al. [36] and Godard et al. [37]
- **Investigation of the attention mechanism in end-to-end speech translation models**
  - If attention remains exploitable, we could perform UWS from speech

SELMA Consortium Project[1]

- **Investigation of the attention mechanism in end-to-end speech translation models**

SELMA stands for **S**tre**a**m **L**earning for **M**ultilingual Knowledge Tr**a**nsfer

- Platform for journalists to browse multilingual data from colleagues
- The goal is to develop speech technologies in 30 different languages, many of them low-resource
  - Speech Recognition;
  - **Speech-to-Text Translation;**
  - Speech-to-Speech Translation;
  - Speech and Textual Named Entity Recognition.

[1] SELMA is a European Horizon 2020 Research and Innovation Action (grant agreement No 957017). https://selma-project.eu/

Thank you.

Bibliography

[1] Austin, Sallabank. **The Cambridge handbook of endangered languages.** Cambridge University Press, 2011.

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

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

[4] Kawakami et al. **Learning to discover, ground and use words with segmental neural language models.** ACL 2019.

[5] Berg-Kirkpatrick et al. **Painless unsupervised learning with features.** NAACL 2010.

[6] Liang et al. **Online EM for unsupervised models.** NAACL 2009.

[7] Lee et al. **Unsupervised lexicon discovery from acoustic input.** ACL 2015.

[8] Kamper et al. **Unsupervised word segmentation and lexicon discovery using acoustic word embeddings.** ACM 2016.

[9] Lyzinski et al. **An evaluation of graph clustering methods for unsupervised term discovery.** Interspeech 2015.

[10] Duong et al. **An attentional model for speech translation without transcription.** NAACL 2016.

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

[12] Bahdanau et al. **Neural Machine Translation by Jointly Learning to Align and Translate.** ICLR 2015.

[13] Boito et al. **Investigating Language Impact in Bilingual Approaches for Computational Language Documentation.** SLTU-CCURL WORKSHOP: LREC 2020.

[14] Godard et al. **Preliminary Experiments on Unsupervised Word Discovery in Mboshi.** Interspeech 2016.

[15] Vaswani et al. **Attention is all you need.** NeurIPS 2017.

[16] Elbayad et al. **Pervasive attention: 2D convolutional neural networks for sequence-to-sequence prediction.** CoNLL 2018.

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

[18] Kocabiyikoglu et al. **Augmenting librispeech with french translations: A multimodal corpus for direct speech translation evaluation.** LREC 2018.

[19] Boito et al. **Empirical Evaluation of Sequence-to-Sequence Models for Word Discovery in Low-resource Settings.** Interspeech 2019.

[20] Michel et al. **Are Sixteen Heads Really Better than One?** NeurIPS 2019.

Bibliography

[21] Voita et al. **Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned.** ACL 2019.

[22] Watanabe et al. **Hybrid CTC/attention architecture for end-to-end speech recognition.** IEEE Journal of Selected Topics in Signal Processing, 2017.

[23] Chorowski et al. **Attention-based models for speech recognition.** NeurIPS 2015.

[24] Wang et al. **Tacotron: Towards end-to-end speech synthesis.** Interspeech 2017.

[25] Shen et al. **Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions.** ICASSP 2018.

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

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

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

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

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

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

[32] Boito et al. **Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings.** ArXiv 2021.

[33] Boito et al. **Investigating Alignment Interpretability for low-resource NMT.** Machine Translation Journal: Special Issue on Machine Translation for Low-resource Languages. Springer Netherlands 2021.

[34] Boito et al. **A small Griko-Italian speech translation corpus.** SLTU 2018.

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

[36] Garg et al. **Jointly Learning to Align and Translate with Transformer Models.** EMNLP 2019.

[37] Godard et al. **Controlling Utterance Length in NMT-based Word Segmentation with Attention.** IWSLT 2019.

Models and Resources for Attention-based Unsupervised Word Segmentation

An Application to Computational Language Documentation

Marcely Zanon Boito

September 28, 2021