Unsupervised Word Segmentation from Speech with Attention

Pierre Godard[1], **Marcely Zanon Boito[2]**, Lucas Ondel[3], Alexandre Berard[2,4], François Yvon[1], Aline Villavicencio[5] and Laurent Besacier[2]

[1]LIMSI, CNRS, Univ. Paris-Saclay, **France**
[2]LIG, Univ. Grenoble Alpes, **France**
[3]BUT, Brno, **Czech Republic**
[4]CRISTAL, Univ. Lille, **France**
[5]CSEE, Univ. of Essex, **UK**

OUTLINE

01. MOTIVATION

02. APPROACH

03. RESULTS

MOTIVATION

Computational Language Documentation

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

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

Computational Language Documentation (CLD)

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

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

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

Endangered Languages Corpora

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

SPEECH

Translations
to a well-documented language[1]

[1]Adda et al. **Breaking the Unwritten Language Barrier: The BULB Project.** SLTU 2016.

THE TASK: Unsupervised Word Segmentation from Speech using Attention

We focus on **UNSUPERVISED WORD SEGMENTATION**.

- From speech

- The system must output timestamps delimiting stretches of speech corresponding to real words in the language

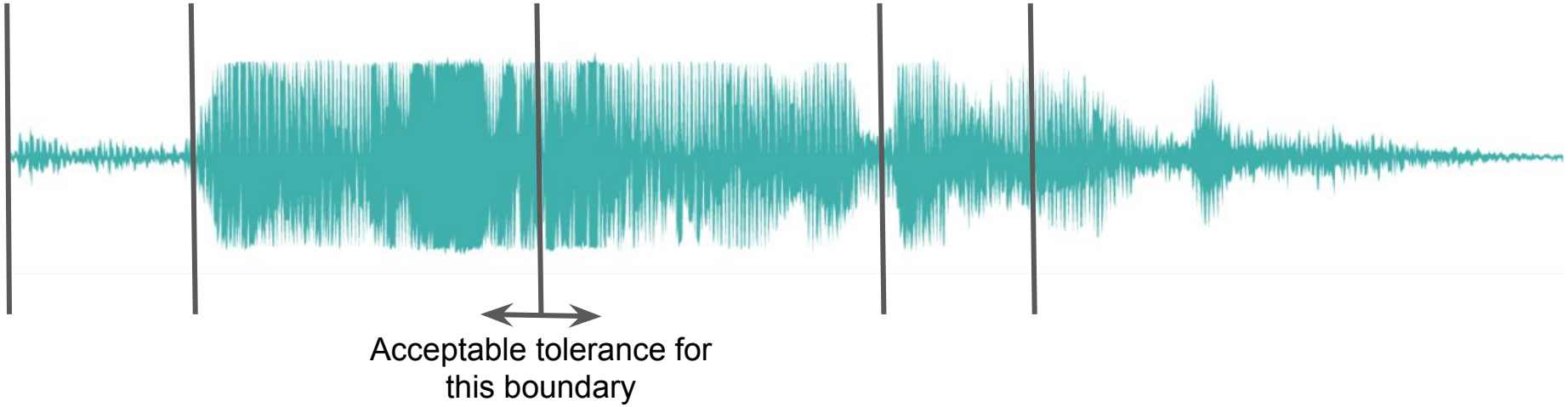
THE TASK: Unsupervised Word Segmentation from Speech using Attention

We focus on **UNSUPERVISED WORD SEGMENTATION**.

- From speech,
- The system must output timestamps delimiting stretches of speech corresponding to real words in the language;
- **Slightly more favorable setup:** the speech utterances are multilingually grounded (text translation in another language is available)

THE TASK: Unsupervised Word Segmentation from Speech using Attention

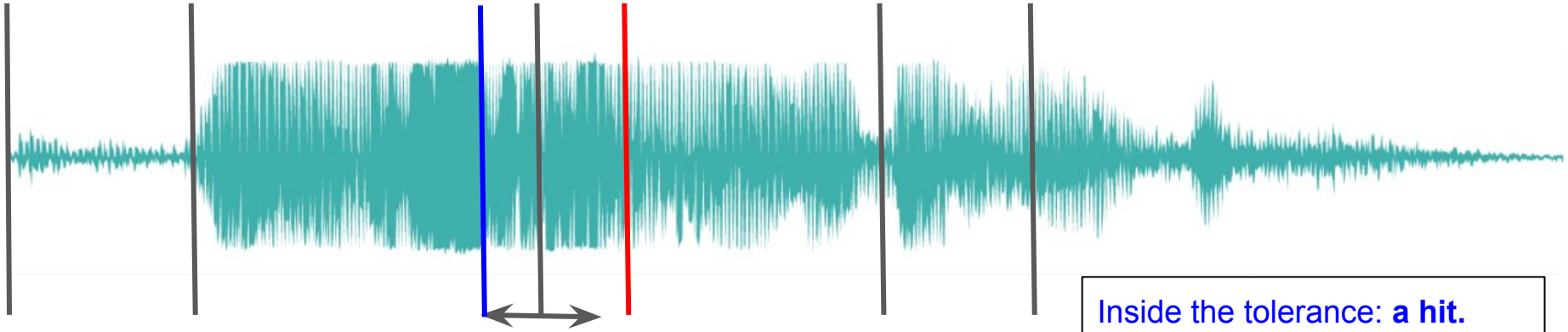
We focus on **UNSUPERVISED WORD SEGMENTATION**.



The tolerance window is defined on the **Zero Resource Challenge 2017 Track 2.**

THE TASK: Unsupervised Word Segmentation from Speech using Attention

We focus on **UNSUPERVISED WORD SEGMENTATION**.



The tolerance window is defined on the **Zero Resource Challenge 2017 Track 2.**

CONTRIBUTION:

- First attempt of performing attentional (neural) word segmentation on speech
  - Previously proposed: a model working from symbolic level[1] (not speech)

- Low-resource setup, using only 5k sentences of the Mboshi-French parallel corpus[2]

[1]MZ Boito et al. **Unwritten Languages Demand Attention too! Word Discovery Using Encoder-Decoder Models.** ASRU 2017.
[2]P Godard et al. **A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments.** LREC 2017.

OUR APPROACH

Unsupervised Word Segmentation from Speech **using Attention**

BACKGROUND:

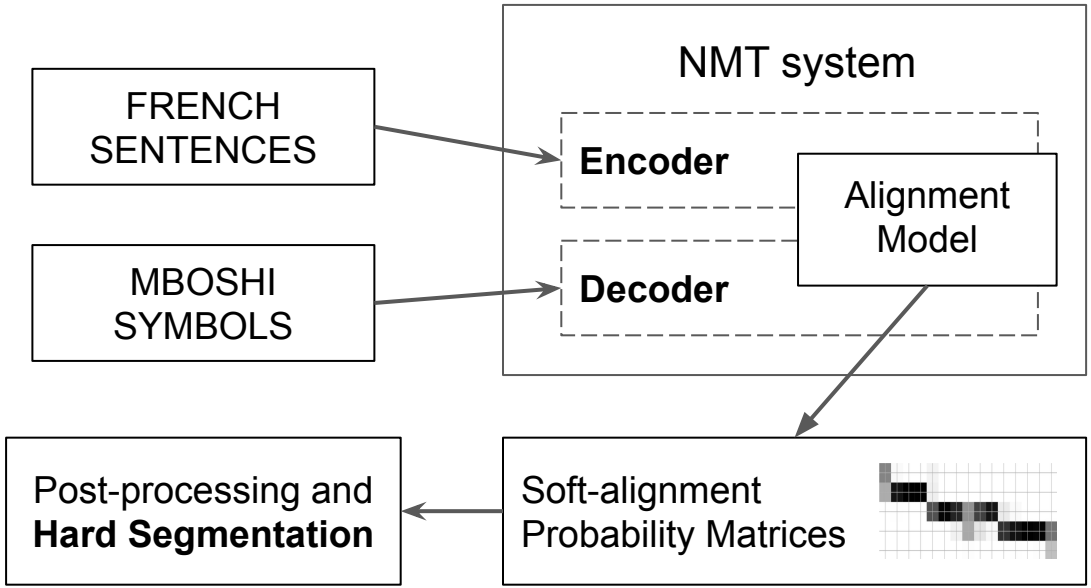
Attention-based encoder decoder models for Neural Machine Translation (NMT) are known to **jointly align and translate** a source into a target language[1]

→ We use soft-alignment probability matrices learned during training to segment[2]

[1] D. Bahdanau et al. **Neural Machine Translation by Jointly Learning to Align and Translate.** ICLR 2015.
[2] MZ Boito et al. **Unwritten Languages Demand Attention too! Word Discovery Using Encoder-Decoder Models.** ASRU 2017.

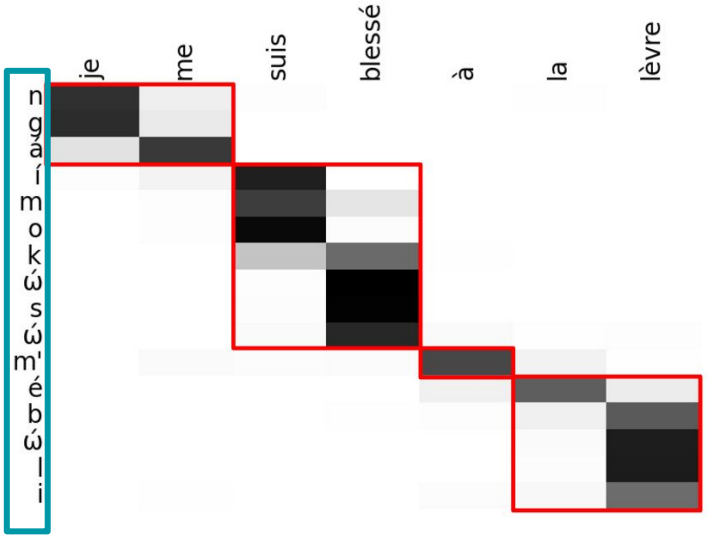
Unsupervised Word Segmentation from Speech **using Attention**

➔ NMT systems[1] are trained with only 5k sentences



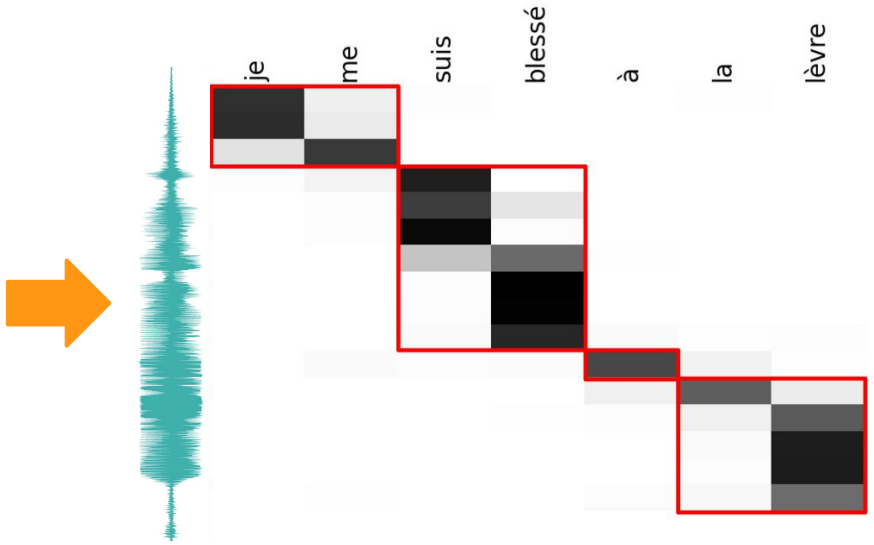
[1]NMT implementation available at github.com/eske/seq2seq

Unsupervised Word Segmentation from Speech **using Attention**



Unsupervised Word Segmentation from Speech **using Attention**

We would like to do the same, but from speech!



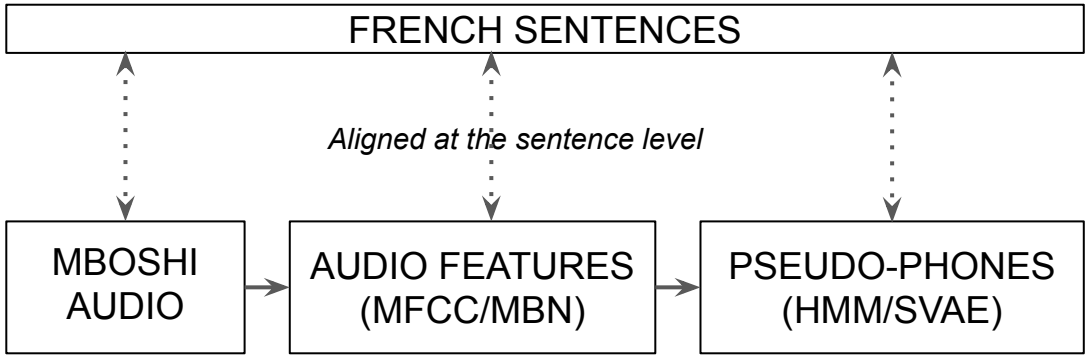
Unsupervised Word Segmentation **from Speech** using Attention

Problem: Infeasible training the model directly from speech with only 5k sentences

Solution: to extract pseudo-phones before training the network

We investigate Acoustic Unit Discovery (AUD) using two different audio feature extraction methods

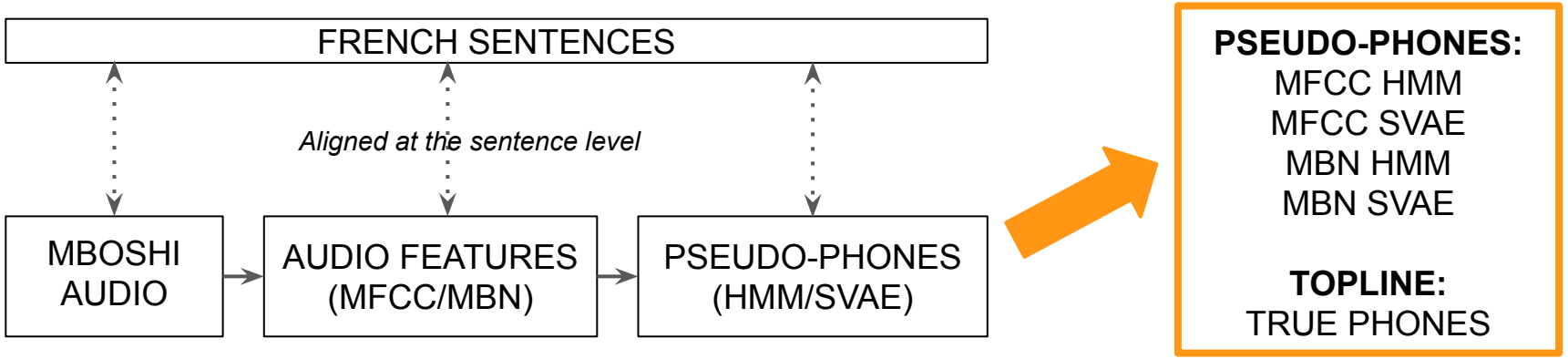
Acoustic Unit Discovery (AUD)



Two AUD models based on Bayesian Non-parametric HMM[1]

[1] Implementation available a github.com/iondel/amdtk
Variational Inference for Acoustic Unit Discovery; L Ondel, L Burget, J Černocký; SLTU 2016.

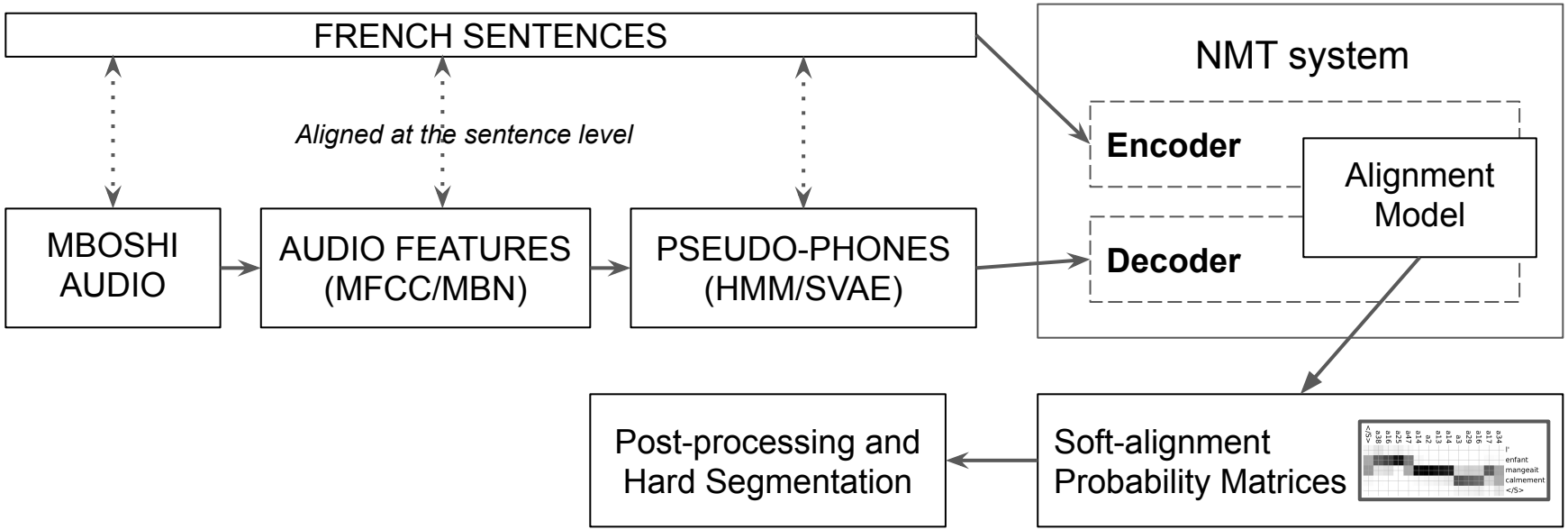
Acoustic Unit Discovery (AUD)



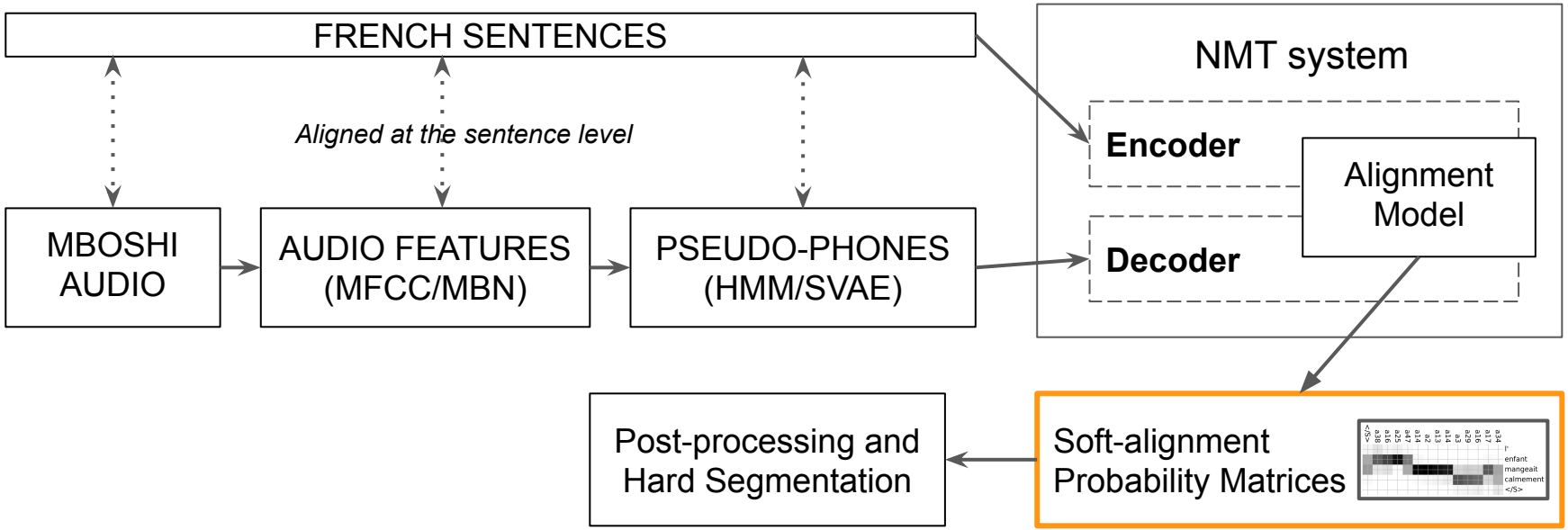
Two AUD models based on Bayesian Non-parametric HMM[1]

[1] Implementation available a github.com/iondel/amdtk
Variational Inference for Acoustic Unit Discovery; L Ondel, L Burget, J Černocký; SLTU 2016.

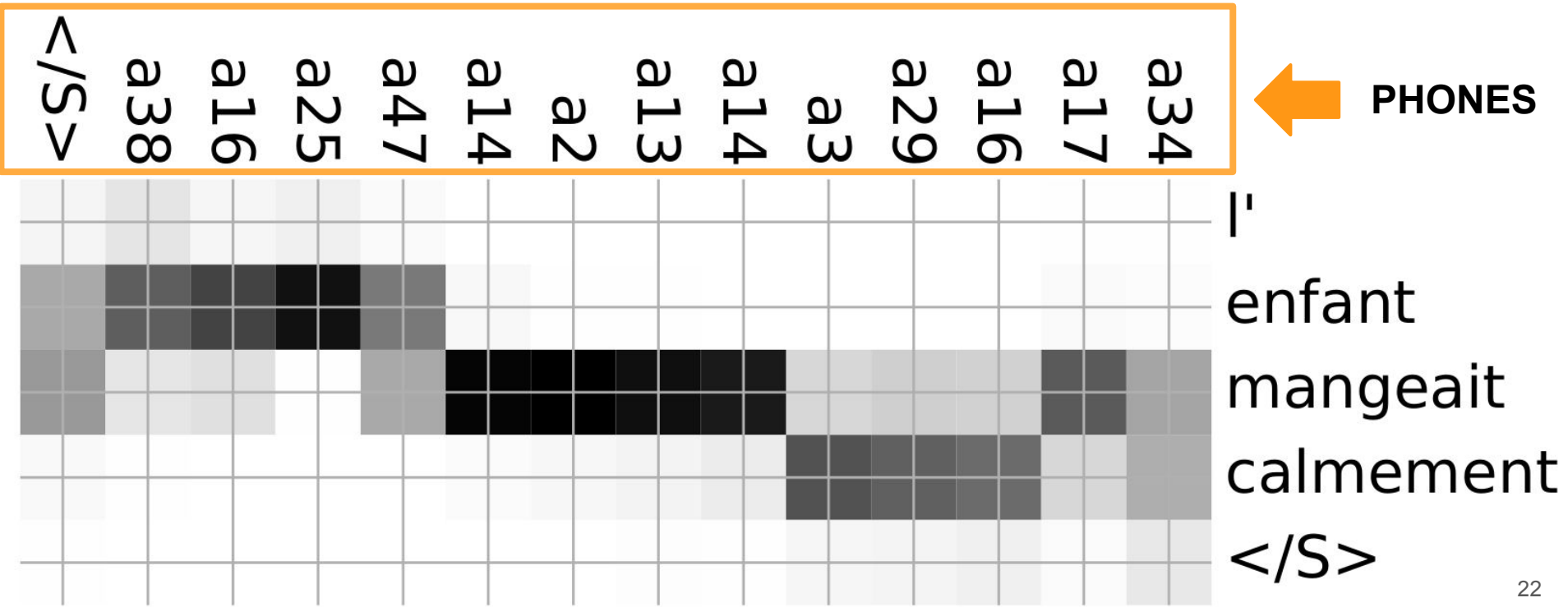
Unsupervised Word Segmentation from Speech using Attention



Unsupervised Word Segmentation from Speech using Attention



Unsupervised Word Segmentation from Speech using Attention



RESULTS

Baselines Comparison

- Dpseg (Dirichlet process-based bigram LM[1]) **(monolingual)**
- Proportional Segmentation **(bilingual)**
- Neural Word Segmentation[2] **(bilingual)**

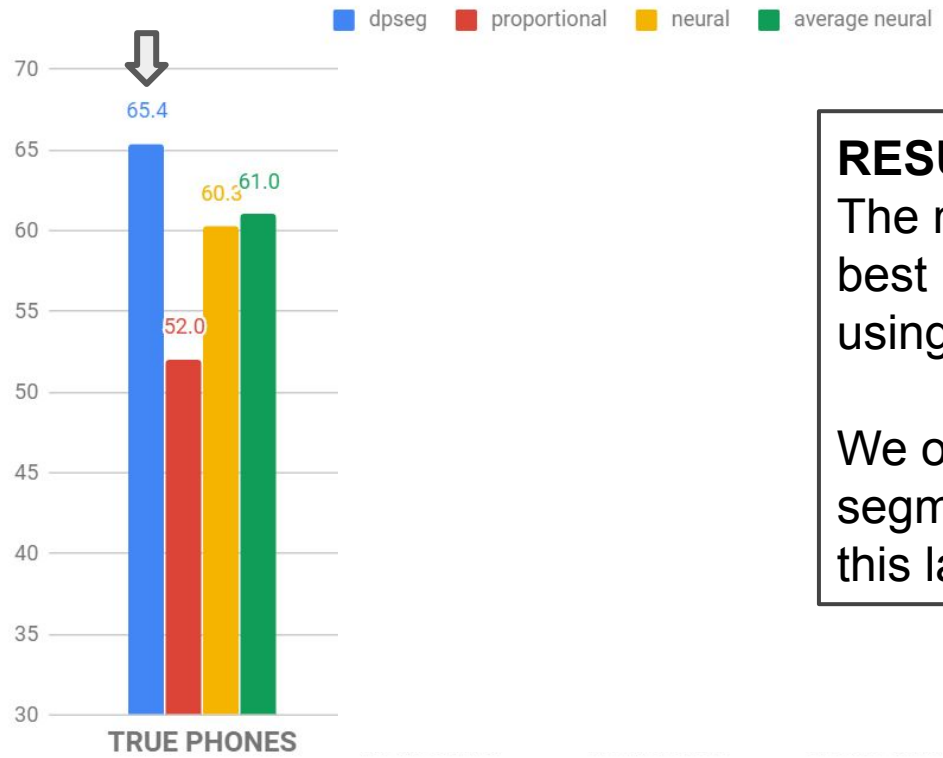
Results are averaged over 5 runs with different splits.

- **Average** Neural Word Segmentation **(bilingual)**

Results are obtained through averaging 5 different soft-alignment matrices for each sentence.

[1]Available at https://homepages.inf.ed.ac.uk/sgwater/, parameters choice described in P. Godard et al **"Preliminary Experiments on Unsupervised Word Discovery in Mboshi"**, Interspeech 2016.
Sharon J Goldwater. ***Nonparametric Bayesian models of lexical acquisition***. PhD thesis, Brown University. 2006;
S. Goldwater. **A Bayesian framework for word segmentation: Exploring the effects of context.** Cognition. 2009.
[2]**Unwritten languages demand attention too! Word Discovery using Encoder-Decoder Models**. M. Zanon Boito, A. Berard, A. Villavicencio, L. Besacier, ASRU 2017

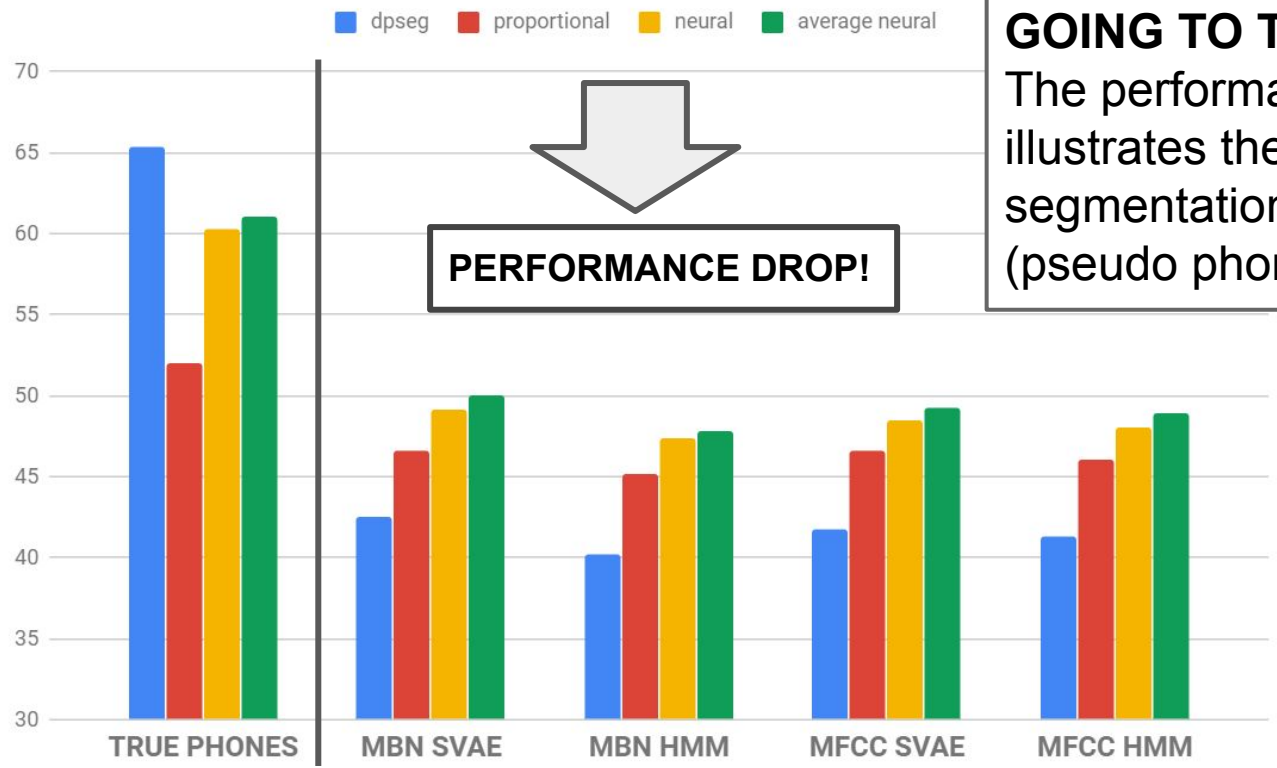
Results: Word Boundary Scores



RESULTS FOR TOPLINE:
The monolingual dpseg achieves the best F-score for word segmentation using the true phones.

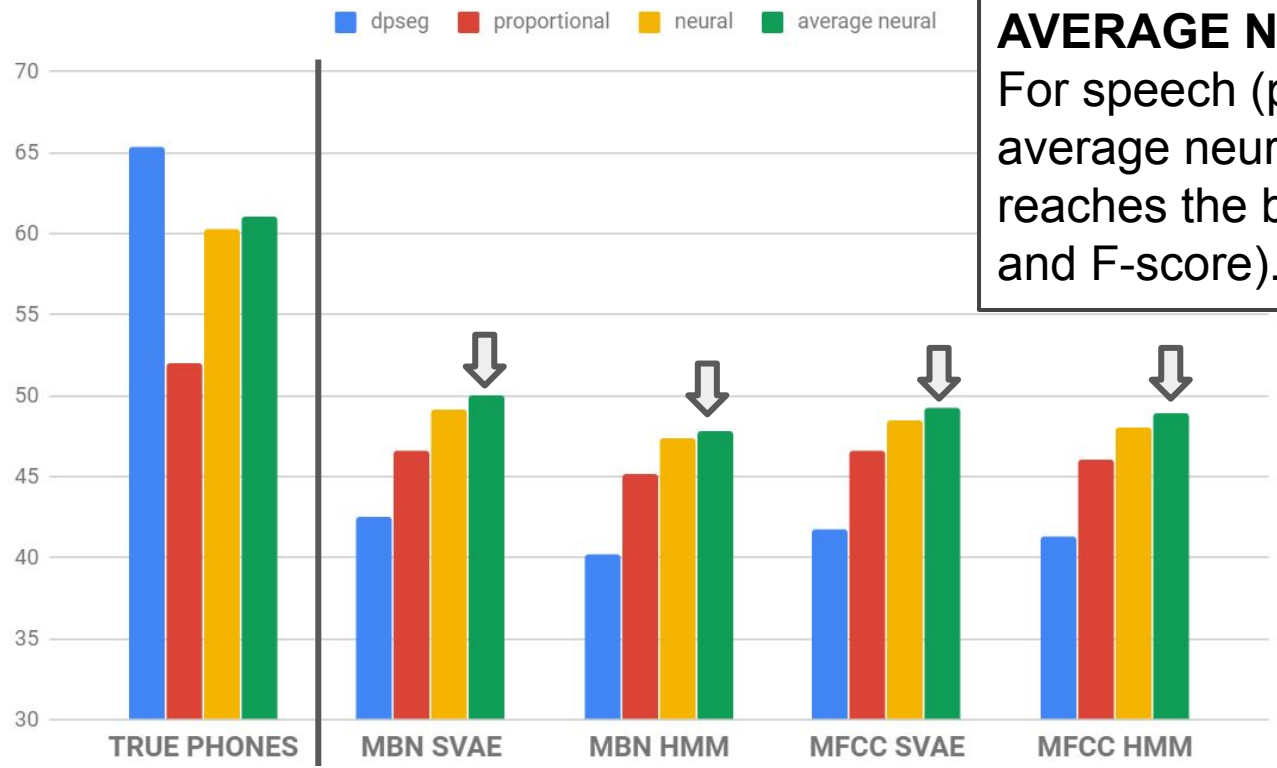
We observe that the proportional segmentation is particularly strong for this language pair.

Results: Word Boundary Scores



GOING TO THE NOISIER SETUP:
The performance drop (all models) illustrates the challenge of the word segmentation task from speech input (pseudo phones).

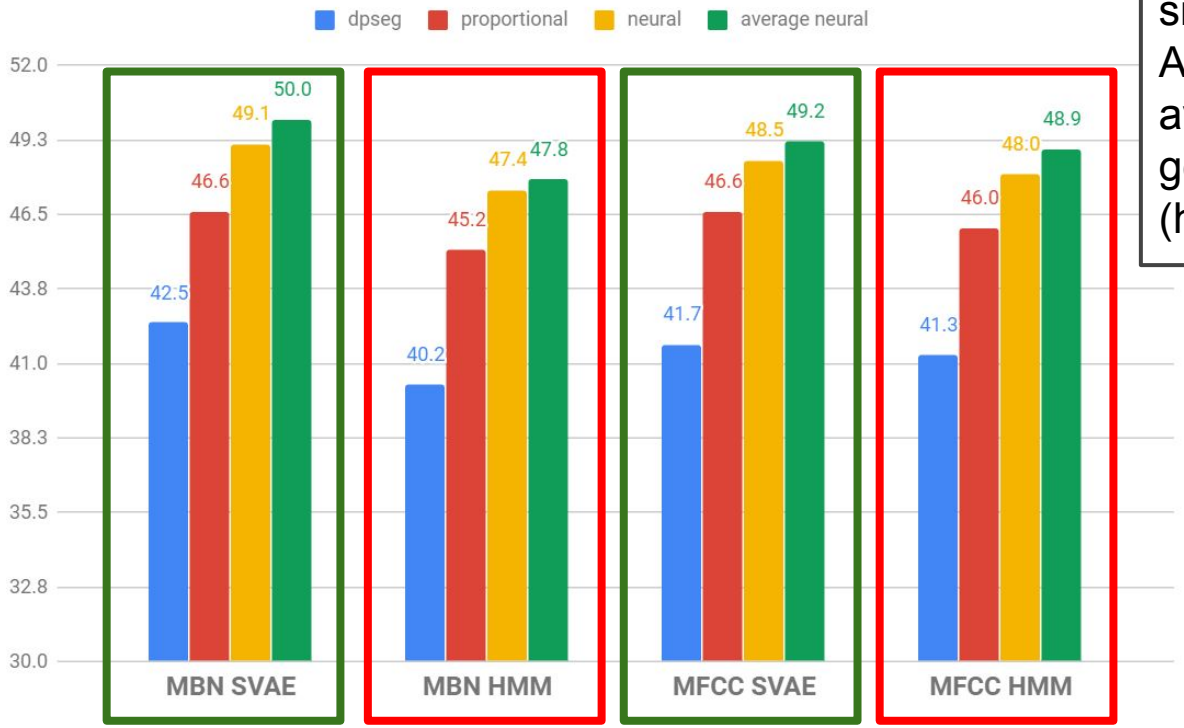
Results: Word Boundary Scores



AVERAGE NEURAL:
For speech (pseudo-phones), our average neural system consistently reaches the best results (precision and F-score).

Average Neural is approximately 8 points better than dpseg.

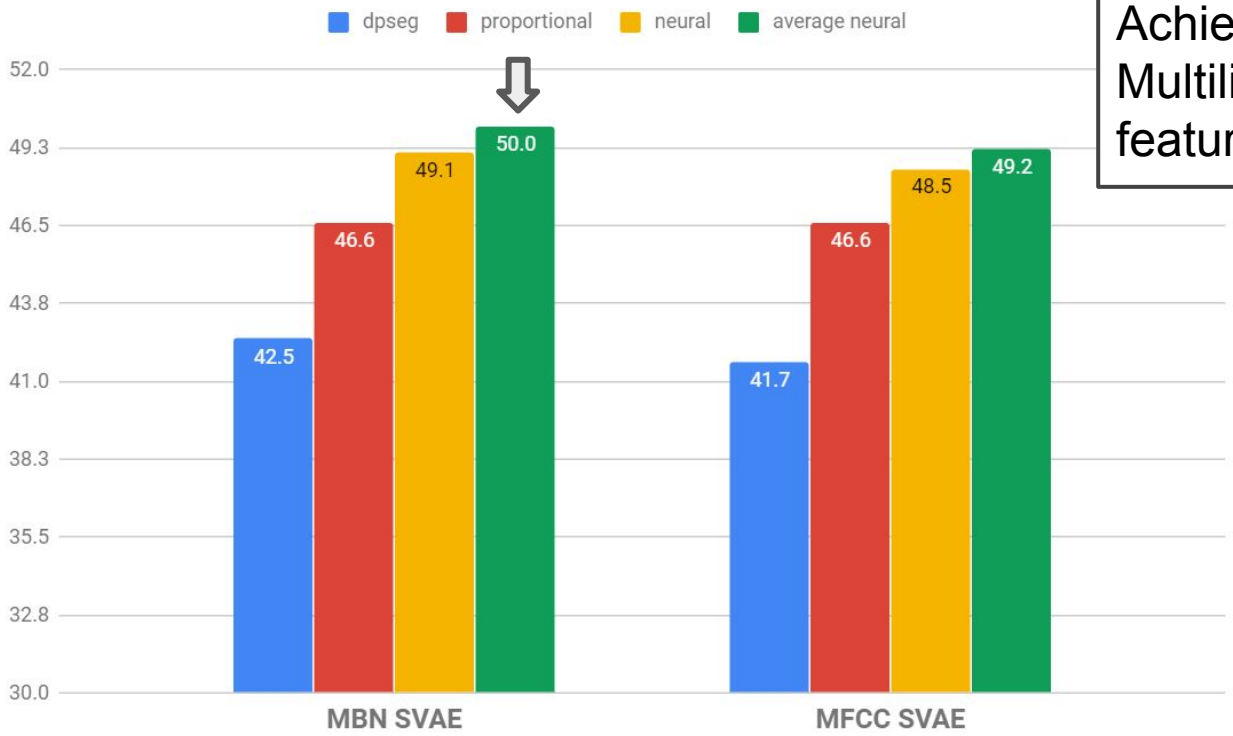
Results: Word Boundary Scores



HMM RESULTS ARE WORSE:
This can be explained by the simplicity of the model.
Another explanation is the higher average number of pseudo-phones generated by sentence on HMMs (harder to segment).

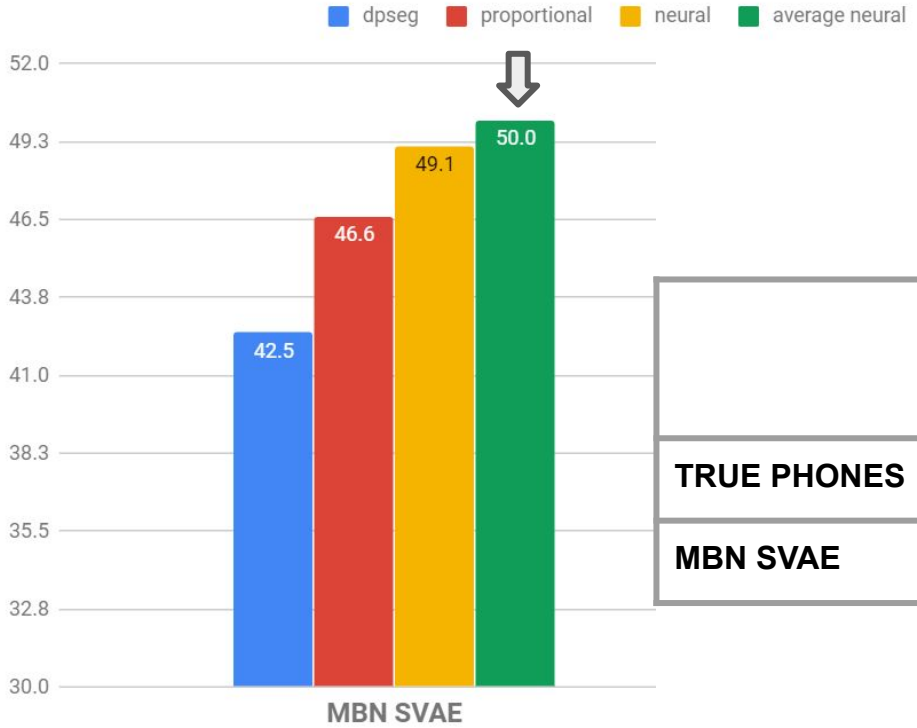
It's a 3 points difference between HMMs and SVAEs.

Results: Word Boundary Scores



BEST RESULTS:
Achieved by using the Multilingual BottleNeck (MBN) features with the SVAE model.

Results: Word Boundary Scores



BEST RESULTS:
This indicates the SVAE model extract more consistent pseudo-phones units.

	Phones per Sentence			Tokens per Sentence		
	avg	max	min	avg	max	min
TRUE PHONES	21.8	60	4	6.0	21	1
MBN SVAE	23.4	71	7	5.4	21	1

Example: Soft-alignment Probability Matrices

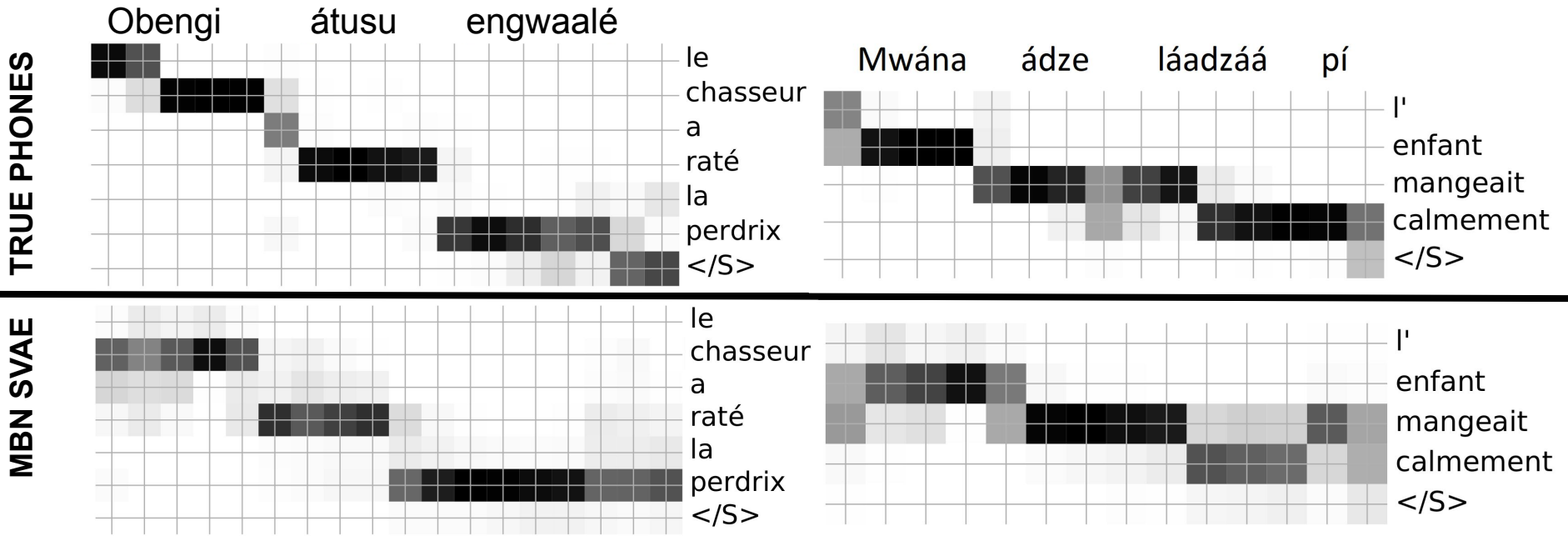


Figure: Successful segmentation, examples of soft-alignment probability matrices. True Phones on top and MBN SVAE setup in the bottom.

Example: Soft-alignment Probability Matrices

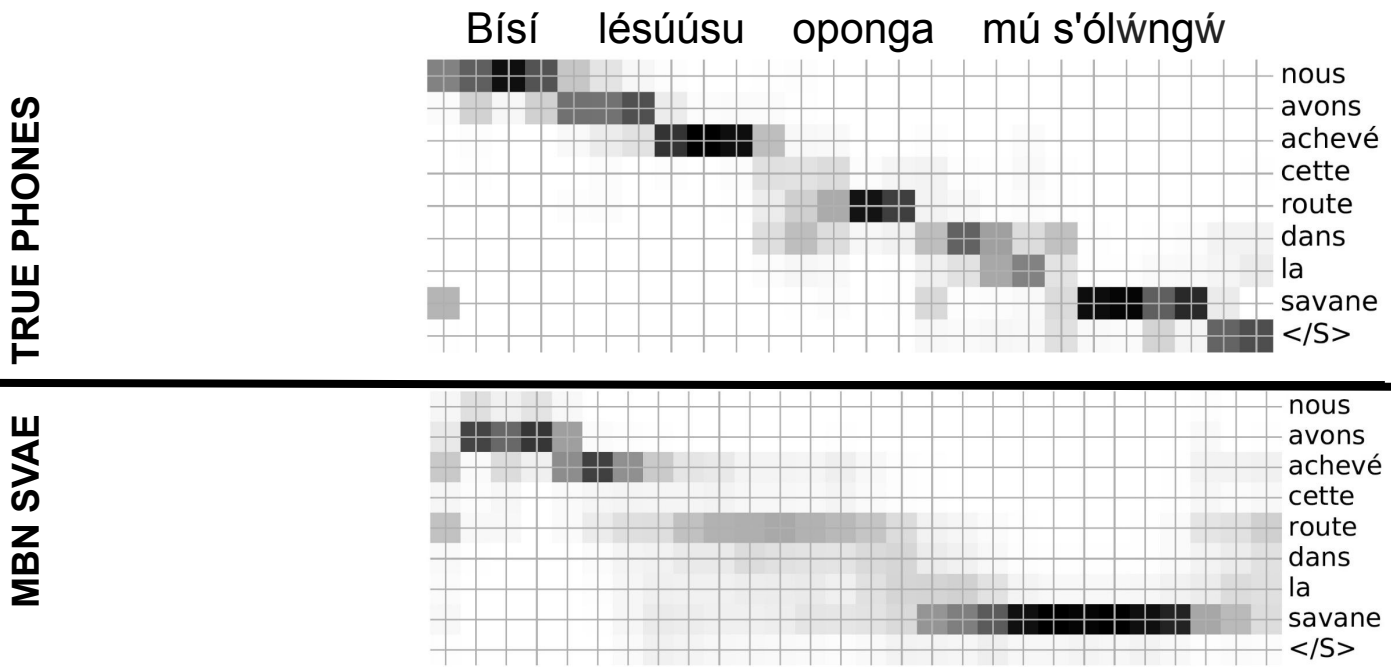


Figure: Failed segmentation, example of soft-alignment probability matrices. True Phones on top and MBN SVAE setup in the bottom.

CONCLUSION

Conclusion

- Promising results for word segmentation from speech, outperforming two baselines in noisy (pseudo-phones) setup
- A deeper analysis of the world clusters obtained is needed to better understand how AUD affects the word discovery task
- Word type results need improvement:
  30.7% true phones, 14.1% best pseudo-phones setup

Thank you!

Questions?

Contact: marcely.zanon-boito@univ-grenoble-alpes.fr **and** pierre.godard@limsi.fr

Unsupervised Word Segmentation from Speech with Attention

Pierre Godard[1], **Marcely Zanon Boito[2]**, Lucas Ondel[3], Alexandre Berard[2,4], François Yvon[4], Aline Villavicencio[5] and Laurent Besacier[2]

[1]LIMSI, Univ. Paris-Saclay, **France**
[2]LIG, Univ. Grenoble Alpes, **France**
[3]BUT, Brno, **Czech Republic**
[4]CRISTAL, Univ. Lille, **France**
[5]CSEE, Univ. of Essex, **UK**

Results: Word Boundary Scores

AUD feat.	AUD model	dpseg			proportional			neural			average neural		
		P	R	F	P	R	F	P	R	F	P	R	F
MFCC	HMM	27.9	80.2	41.3	42.6	49.9	46.0	51.6	44.9	48.0	55.5	43.7	48.9
MFCC	SVAE	29.8	69.1	41.7	42.2	51.9	46.6	52.7	45.0	48.5	55.7	44.1	49.2
MBN	HMM	27.8	72.6	40.2	42.5	48.1	45.2	50.8	44.5	47.4	54.1	42.9	47.8
MBN	SVAE	30.0	72.9	42.5	42.5	51.6	46.6	57.2	43.0	49.1	60.6	42.5	50.0
TRUE PHONES		53.8	**83.5**	**65.4**	44.5	62.6	52.0	60.5	59.9	60.3	**62.8**	59.3	61.0

Table: The monolingual dpseg achieves the best recall and F-score for word segmentation using the true phones (topline). We observe that the proportional segmentation is particularly strong for this language pair.

Results: Boundary Scores

AUD feat.	AUD model	dpseg			proportional			neural			average neural		
		P	R	F	P	R	F	P	R	F	P	R	F
MFCC	HMM	27.9	80.2	41.3	42.6	49.9	46.0	51.6	44.9	48.0	55.5	43.7	48.9
MFCC	SVAE	29.8	69.1	41.7	42.2	51.9	46.6	52.7	45.0	48.5	55.7	44.1	49.2
MBN	HMM	27.8	72.6	40.2	42.5	48.1	45.2	50.8	44.5	47.4	54.1	42.9	47.8
MBN	SVAE	30.0	72.9	42.5	42.5	51.6	46.6	57.2	43.0	49.1	60.6	42.5	50.0
TRUE PHONES		53.8	83.5	65.4	44.5	62.6	52.0	60.5	59.9	60.3	62.8	59.3	61.0

Table: The performance drop of all models illustrates the challenge of the word segmentation task from speech input (pseudo phones).

Results: Boundary Scores

AUD feat.	AUD model	dpseg			proportional			neural			average neural		
		P	R	F	P	R	F	P	R	F	P	R	F
MFCC	HMM	27.9	**80.2**	41.3	42.6	49.9	46.0	51.6	44.9	48.0	**55.5**	43.7	**48.9**
MFCC	SVAE	29.8	**69.1**	41.7	42.2	51.9	46.6	52.7	45.0	48.5	**55.7**	44.1	**49.2**
MBN	HMM	27.8	**72.6**	40.2	42.5	48.1	45.2	50.8	44.5	47.4	**54.1**	42.9	**47.8**
MBN	SVAE	30.0	**72.9**	42.5	42.5	51.6	46.6	57.2	43.0	49.1	**60.6**	42.5	**50.0**
TRUE PHONES		53.8	83.5	65.4	44.5	62.6	52.0	60.5	59.9	60.3	62.8	59.3	61.0

Table: For speech (pseudo-phones), our average neural system consistently reaches the best results (precision and F-score).

Results: Boundary Scores

AUD feat.	AUD model	dpseg			proportional			neural			average neural		
		P	R	F	P	R	F	P	R	F	P	R	F
MFCC	**HMM**	27.9	80.2	41.3	42.6	49.9	46.0	51.6	44.9	48.0	55.5	43.7	**48.9**
MFCC	**SVAE**	29.8	69.1	41.7	42.2	51.9	46.6	52.7	45.0	48.5	55.7	44.1	**49.2**
MBN	**HMM**	27.8	72.6	40.2	42.5	48.1	45.2	50.8	44.5	47.4	54.1	42.9	**47.8**
MBN	**SVAE**	30.0	72.9	42.5	42.5	51.6	46.6	57.2	43.0	49.1	60.6	42.5	**50.0**
TRUE PHONES		53.8	83.5	65.4	44.5	62.6	52.0	60.5	59.9	60.3	62.8	59.3	61.0

Table: HMM models achieved worse results than SVAE models. This can be explained by the simplicity of the model. Another explanation would be the average number of pseudo-phones generated by sentence, which is higher on HMM models (harder to segment).

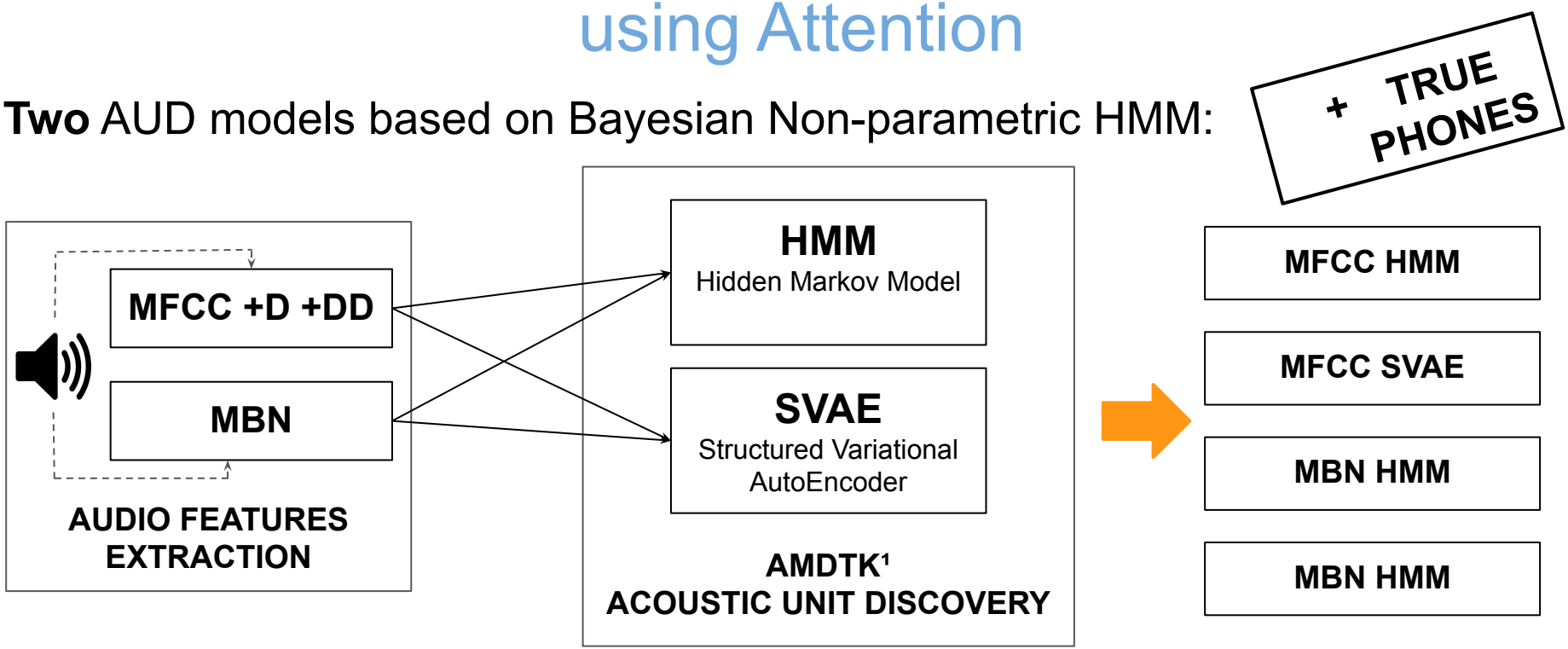
Results: Boundary Scores

AUD feat.	AUD model	dpseg			proportional			neural			average neural		
		P	R	F	P	R	F	P	R	F	P	R	F
MFCC	**HMM**	27.9	80.2	41.3	42.6	49.9	46.0	51.6	44.9	48.0	55.5	43.7	**48.9**
MFCC	**SVAE**	29.8	69.1	41.7	42.2	51.9	46.6	52.7	45.0	48.5	55.7	44.1	**49.2**
MBN	**HMM**	27.8	72.6	40.2	42.5	48.1	45.2	50.8	44.5	47.4	54.1	42.9	**47.8**
MBN	**SVAE**	30.0	72.9	42.5	42.5	51.6	46.6	57.2	43.0	49.1	60.6	42.5	**50.0**
TRUE PHONES		53.8	83.5	65.4	44.5	62.6	52.0	60.5	59.9	60.3	62.8	59.3	61.0

Table: Best results were achieved by using the Multilingual BottleNeck (MBN) features with the SVAE model. This indicates this model extract more consistent pseudo-phones units.

Unsupervised Word Segmentation **from Speech** using Attention

Two AUD models based on Bayesian Non-parametric HMM:



(These + translation are the NMT system input!)

[1] Implementation available a github.com/iondel/amdtk

[2] **Variational Inference for Acoustic Unit Discovery**; L Ondel, L Burget, J Černocký; 2016