Unsupervised word segmentation using dynamic programming on self-supervised speech representations

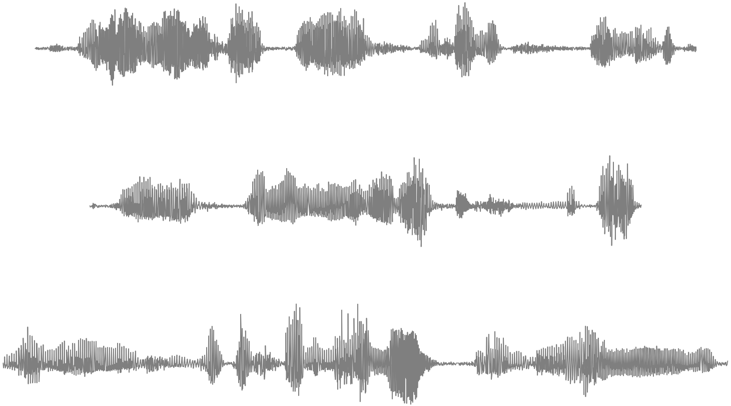
AAAI SAS Workshop 2022 – Invited Talk

Herman Kamper

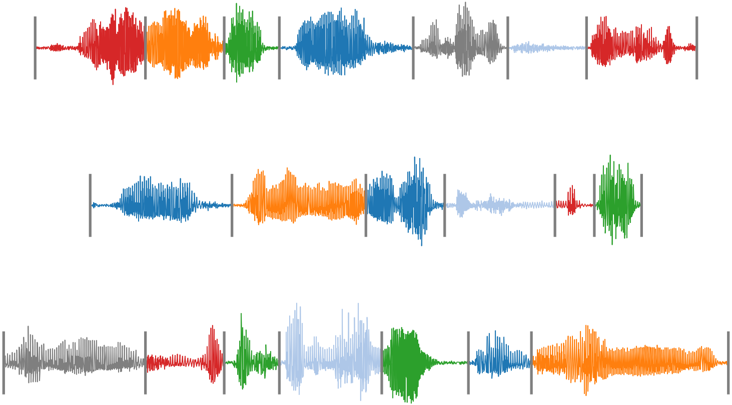
E&E Engineering, Stellenbosch University, South Africa

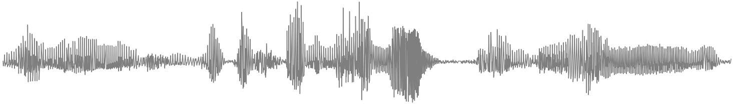
http://www.kamperh.com/

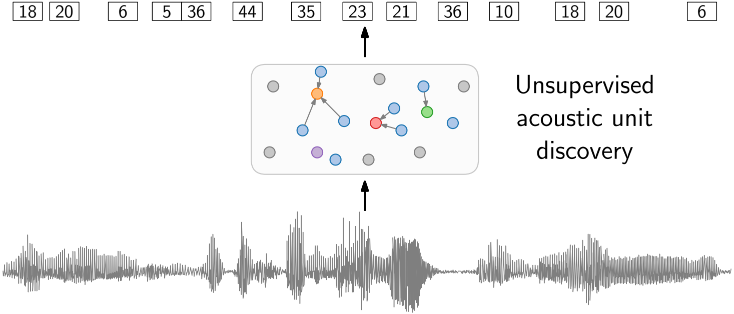
Unsupervised word segmentation



Unsupervised word segmentation

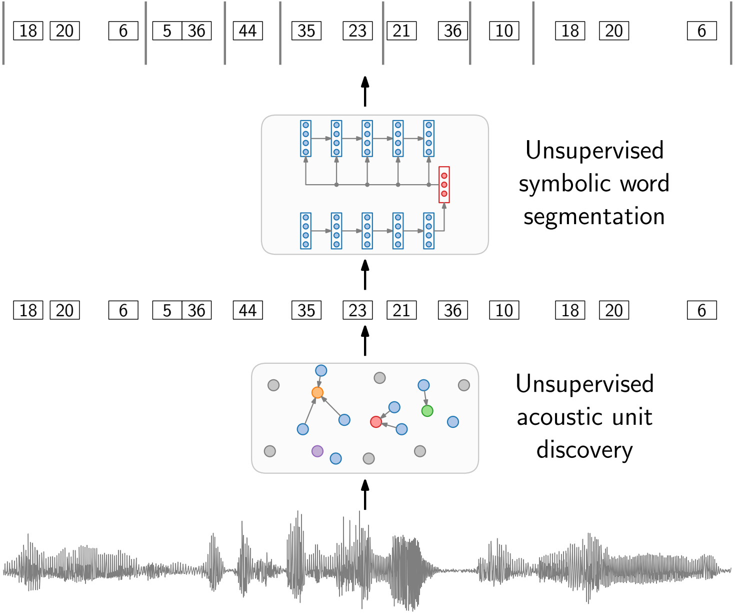






18 20 6 5 36 44 35 23 21 36 10 18 20 6

Unsupervised acoustic unit discovery

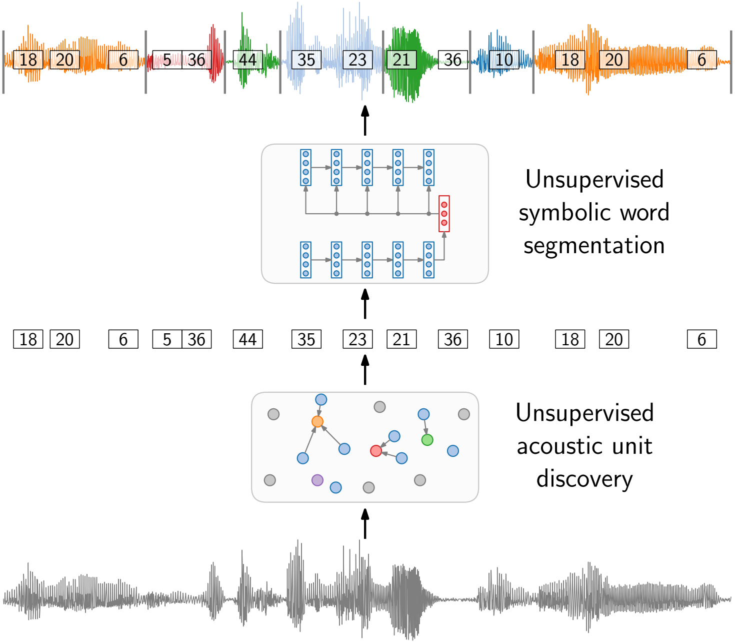


| 18 20 6 | 5 36 | 44 | 35 23 | 21 36 | 10 | 18 20 6 |

Unsupervised symbolic word segmentation

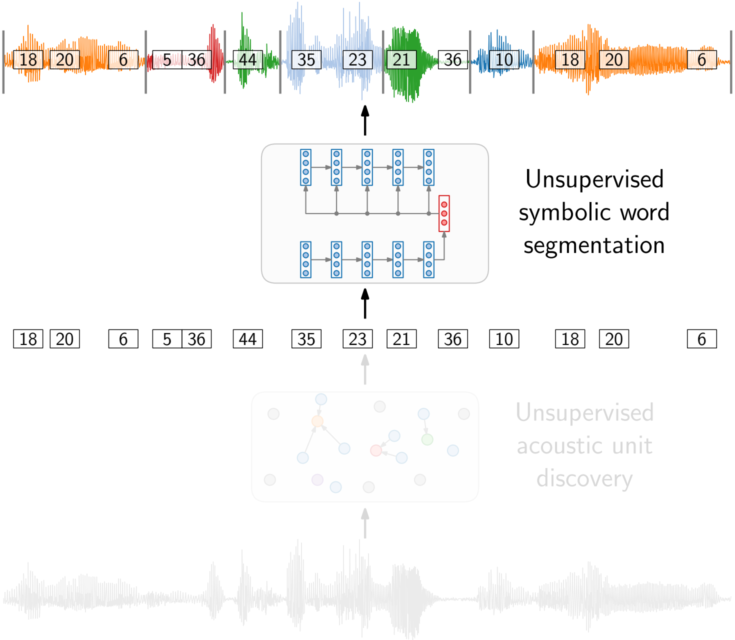
18 20 6 5 36 44 35 23 21 36 10 18 20 6

Unsupervised acoustic unit discovery



Unsupervised symbolic word segmentation

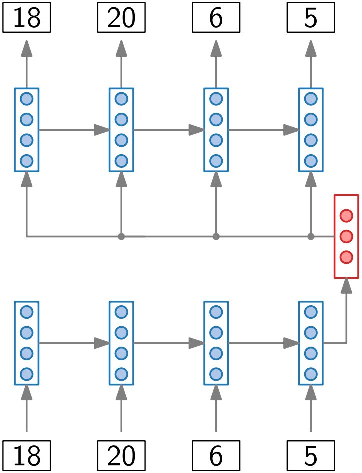
Unsupervised acoustic unit discovery



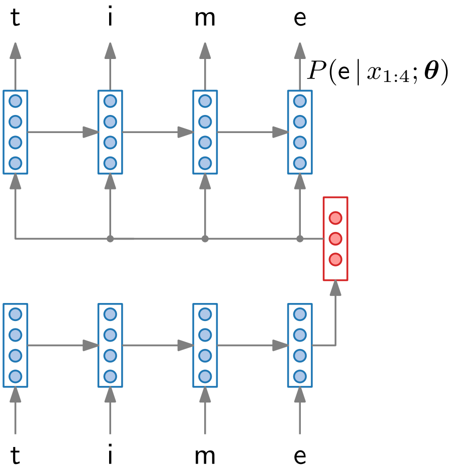
Unsupervised symbolic word segmentation

Unsupervised acoustic unit discovery

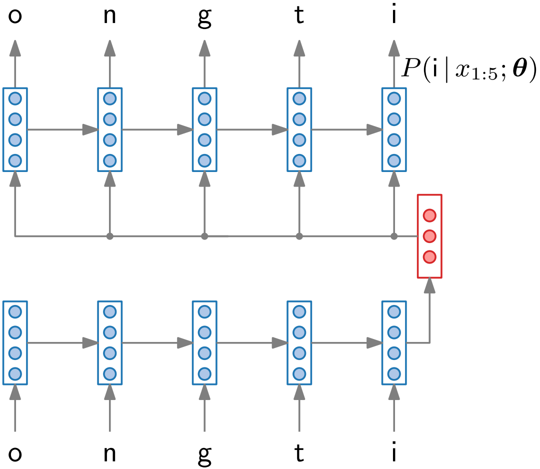
Autoencoding recurrent neural network (AE-RNN)



Autoencoding recurrent neural network (AE-RNN)



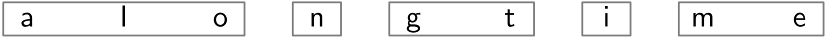
Autoencoding recurrent neural network (AE-RNN)



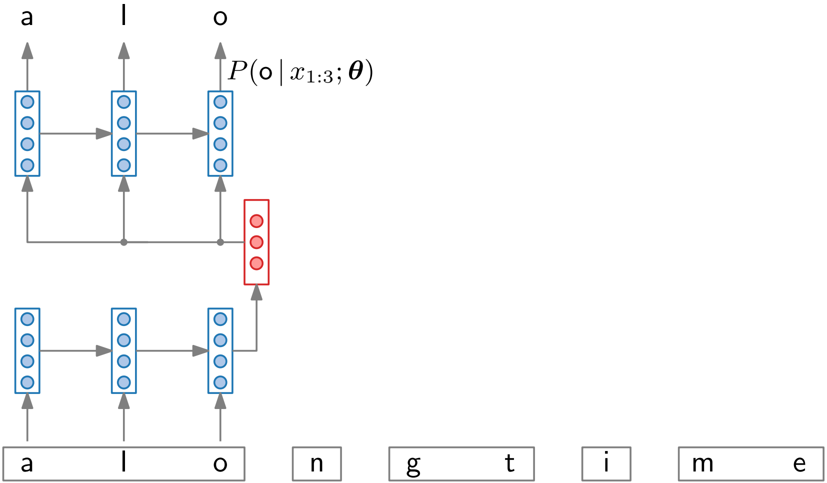
DPDP AE-RNN

a l o n g t i m e

DPDP AE-RNN



DPDP AE-RNN

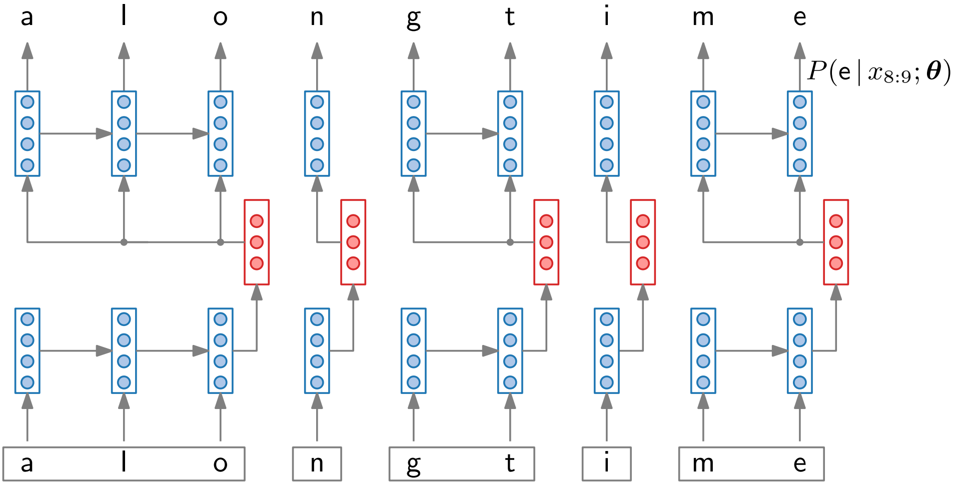


a l o

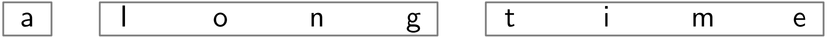
$P(\text{o} \mid x_{1:3}; \boldsymbol{\theta})$

a l o | n | g t | i | m e

DPDP AE-RNN



DPDP AE-RNN

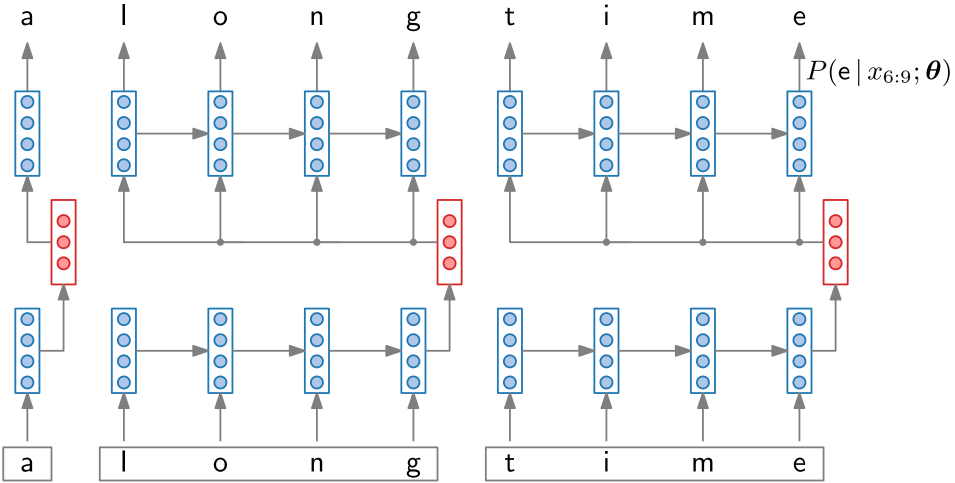


a

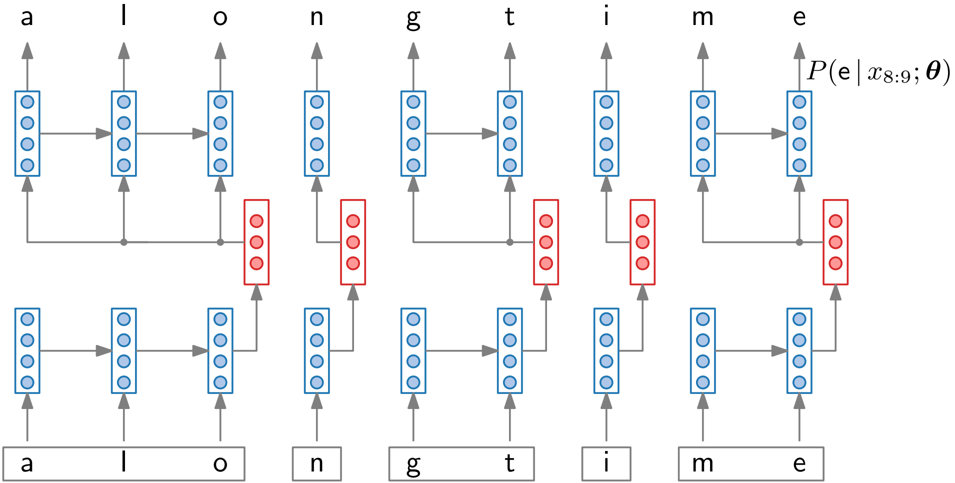
l o n g

t i m e

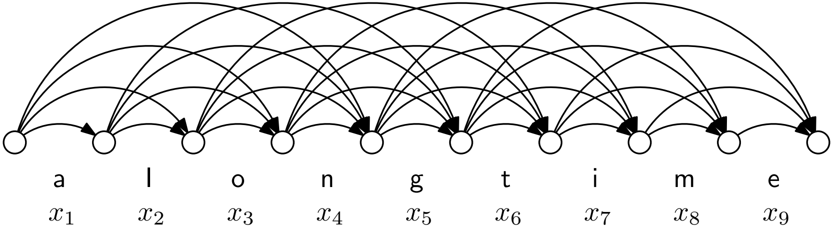
DPDP AE-RNN



DPDP AE-RNN



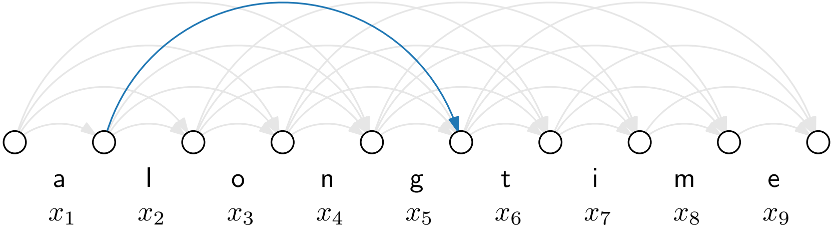
Duration-penalized dynamic programming (DPDP)



Assuming a maximum duration of 4 symbols

Duration-penalized dynamic programming (DPDP)

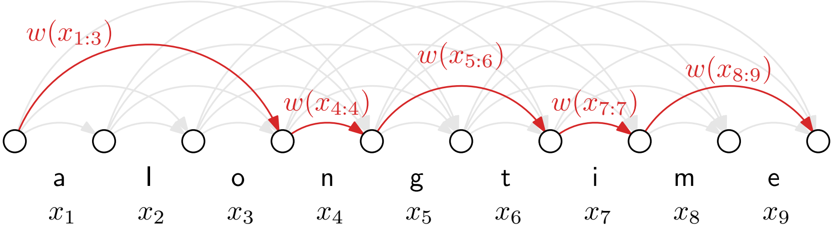
$$w_{\text{seg}}(x_{2:5}) = -\sum_{t=2}^{5} \log P(x_t | x_{2:5}; \boldsymbol{\theta})$$



Assuming a maximum duration of 4 symbols

Duration-penalized dynamic programming (DPDP)

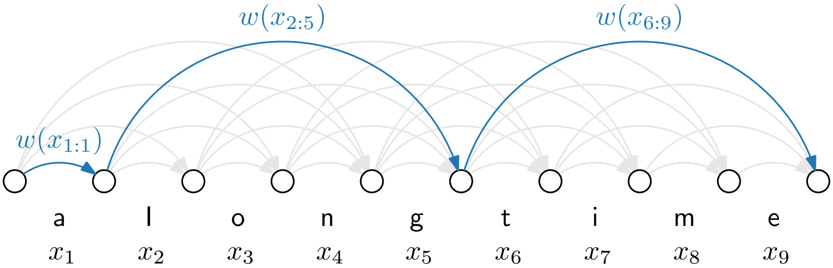
$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda\, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



Assuming a maximum duration of 4 symbols

Duration-penalized dynamic programming (DPDP)

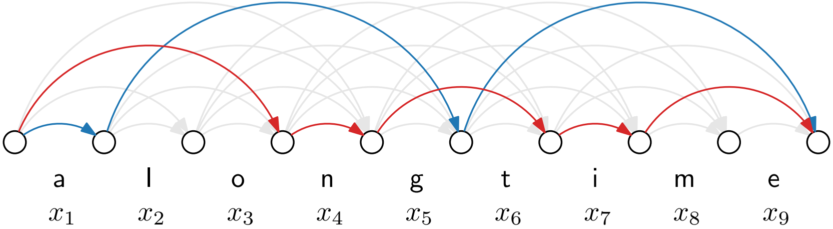
$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda\, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



Assuming a maximum duration of 4 symbols

Duration-penalized dynamic programming (DPDP)

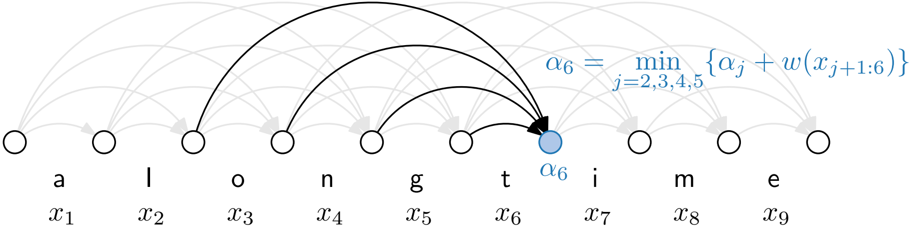
$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda\, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



Assuming a maximum duration of 4 symbols

Duration-penalized dynamic programming (DPDP)

$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda \, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



Assuming a maximum duration of 4 symbols

Where does the AE-RNN come from?

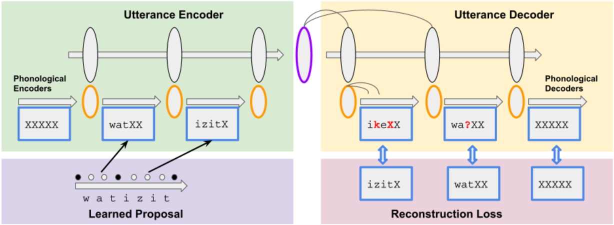
Where does the AE-RNN come from?

- Can refine iteratively: Start with random segmentation, train AE-RNN, DPDP segment, retrain AE-RNN, DPDP segment, etc.

Where does the AE-RNN come from?

- Can refine iteratively: Start with random segmentation, train AE-RNN, DPDP segment, retrain AE-RNN, DPDP segment, etc.

- Use probabilistic approach to sample from and explore the space of possible segmentations [Elsner and Shain, EMNLP'17]

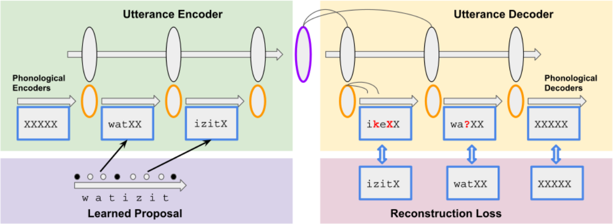


Where does the AE-RNN come from?

- Can refine iteratively: Start with random segmentation, train AE-RNN, DPDP segment, retrain AE-RNN, DPDP segment, etc.

- Use probabilistic approach to sample from and explore the space of possible segmentations [Elsner and Shain, EMNLP'17]

- Dumb idea:
  Train on full utterances

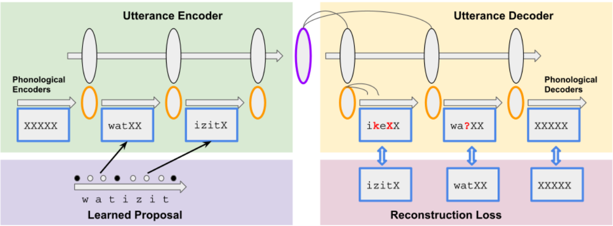


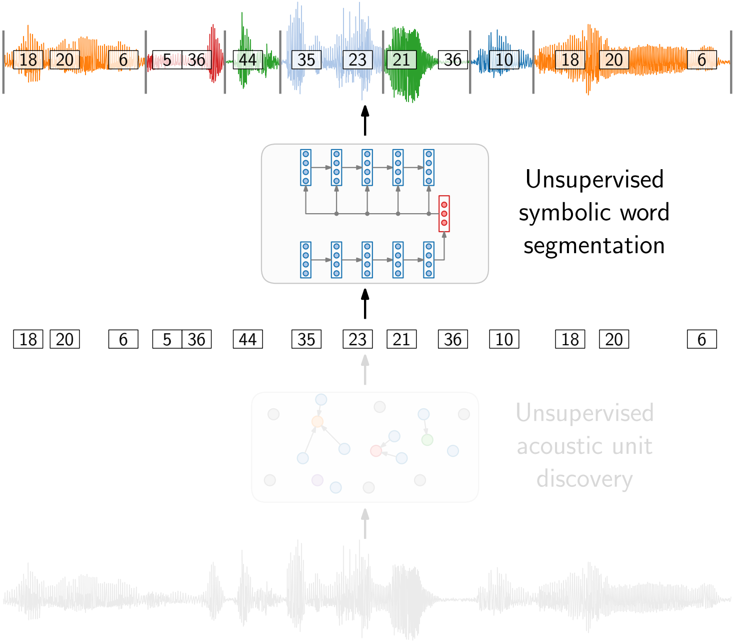
Where does the AE-RNN come from?

- Can refine iteratively: Start with random segmentation, train AE-RNN, DPDP segment, retrain AE-RNN, DPDP segment, etc.

- Use probabilistic approach to sample from and explore the space of possible segmentations [Elsner and Shain, EMNLP'17]

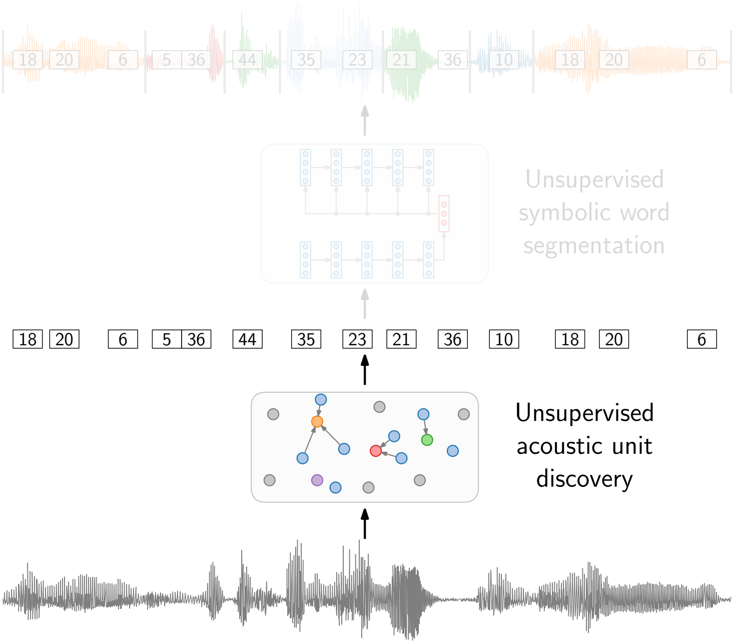
- Dumb idea:
  Train on full utterances
  (this works)





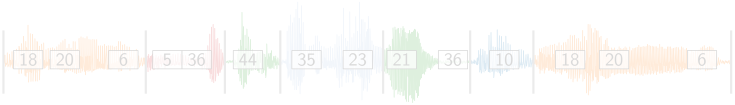
Unsupervised symbolic word segmentation

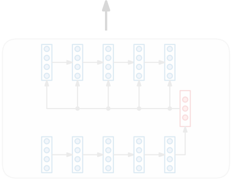
Unsupervised acoustic unit discovery



Unsupervised symbolic word segmentation

Unsupervised acoustic unit discovery

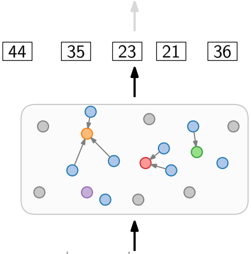




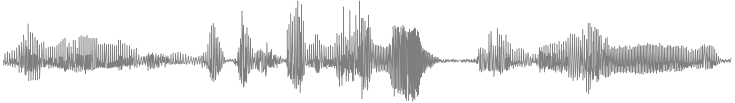
Unsupervised symbolic word segmentation

Benjamin van Niekerk

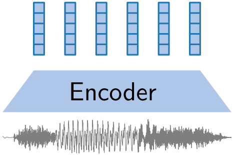
44 35 23 21 36 10 18 20 6



Unsupervised acoustic unit discovery

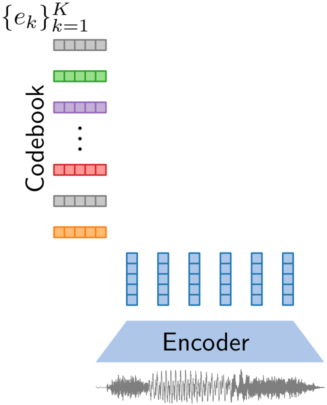


DPDP on self-supervised features with vector quantization



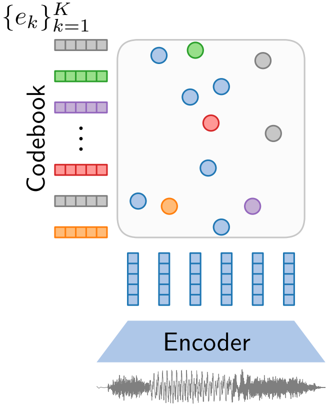
H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

DPDP on self-supervised features with vector quantization



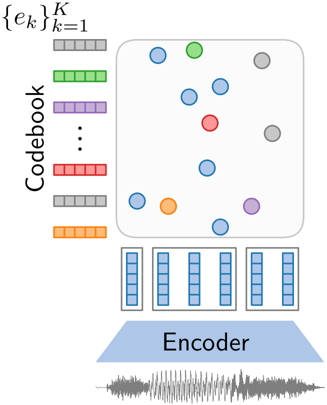
H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

DPDP on self-supervised features with vector quantization



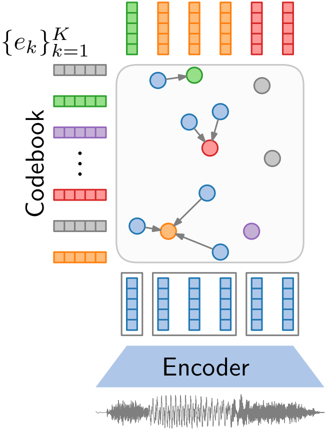
H. Kamper and B. van Niekerk, “Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks,” in *Proc. Interspeech*, 2021.

DPDP on self-supervised features with vector quantization



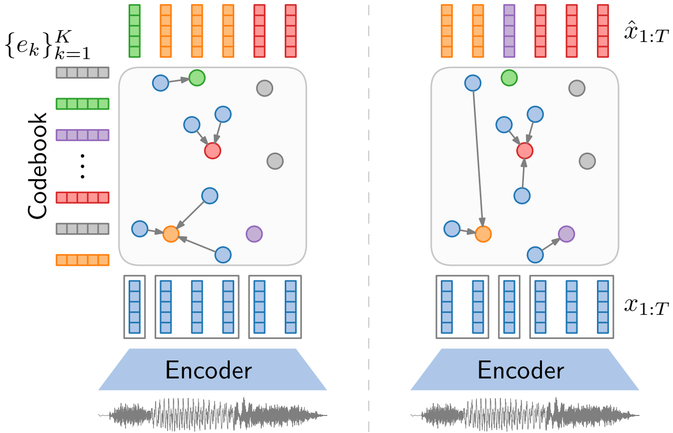
H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

DPDP on self-supervised features with vector quantization



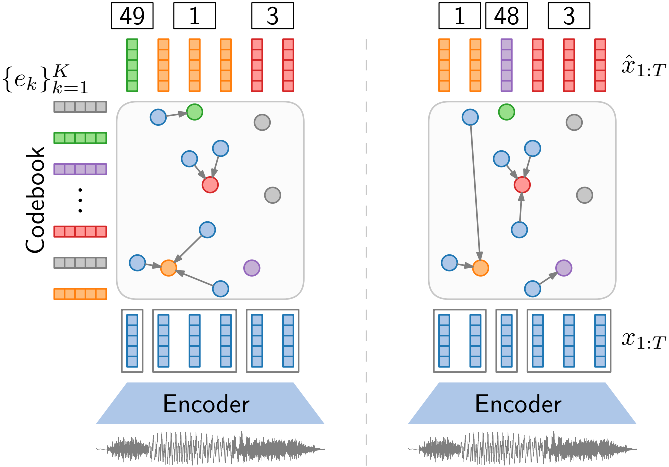
H. Kamper and B. van Niekerk, “Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks,” in *Proc. Interspeech*, 2021.

DPDP on self-supervised features with vector quantization



H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

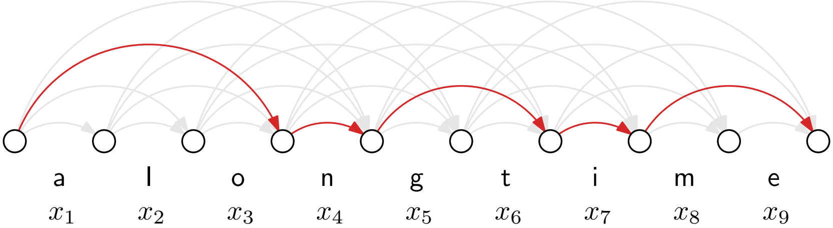
DPDP on self-supervised features with vector quantization



H. Kamper and B. van Niekerk, "Towards unsupervised phone and word segmentation using self-supervised vector-quantized neural networks," in *Proc. Interspeech*, 2021.

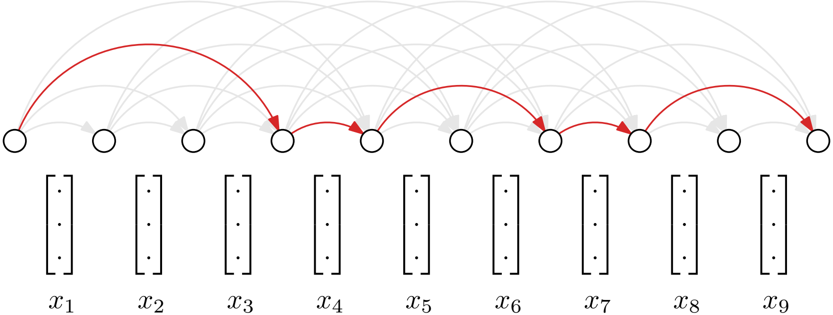
DPDP on self-supervised features with vector quantization

$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda\, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



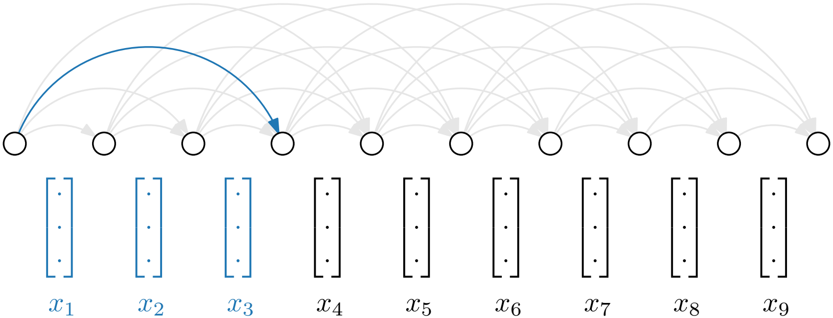
DPDP on self-supervised features with vector quantization

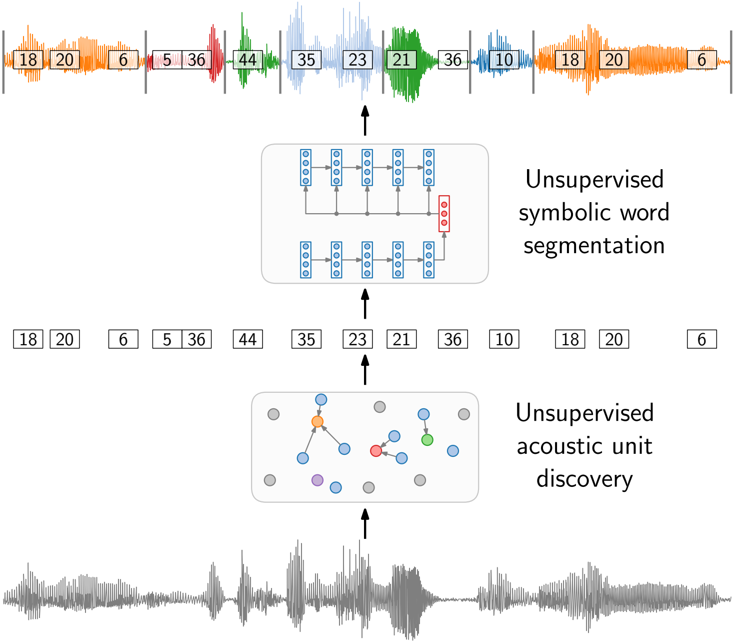
$$w(x_{a:b}) = w_{\text{seg}}(x_{a:b}) + \lambda\, w_{\text{dur}}(\text{dur}(x_{a:b}))$$



DPDP on self-supervised features with vector quantization

$$w_{\text{seg}}(x_{1:3}) = \min_{k=1}^{K} \sum_{t=1}^{3} ||x_t - e_k||^2$$



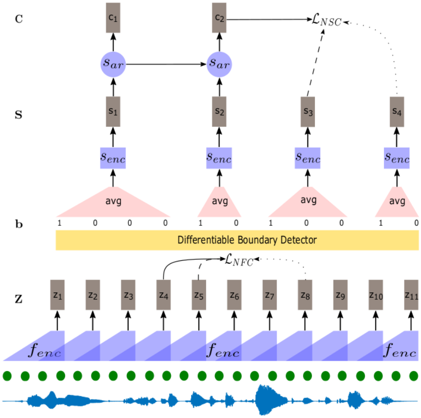


Unsupervised symbolic word segmentation

Unsupervised acoustic unit discovery

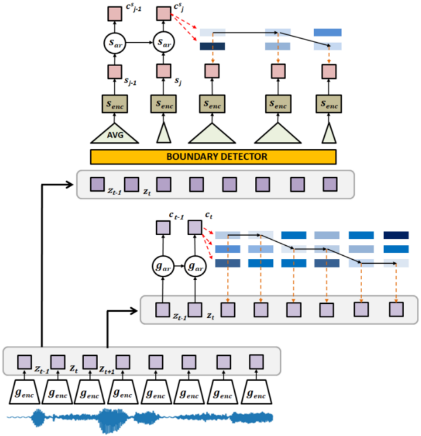
Segmental CPC (SCPC):

[Bhati et al., Interspeech'21]



Multi-level aligned CPC (mACPC):

[Cuervo et al., arXiv'21]



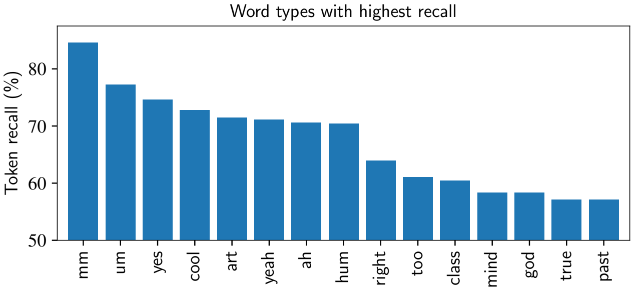
S. Bhati et al., "Segmental contrastive predictive coding for unsupervised word segmentation," in *Proc. Interspeech*, 2021.
S. Cuervo et al., "Contrastive prediction strategies for unsupervised segmentation and categorization of phonemes and words," *arXiv preprint arXiv:2110.15909*, 2021.

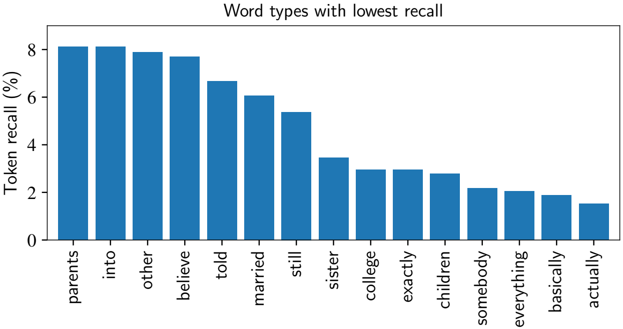
Evaluation on English

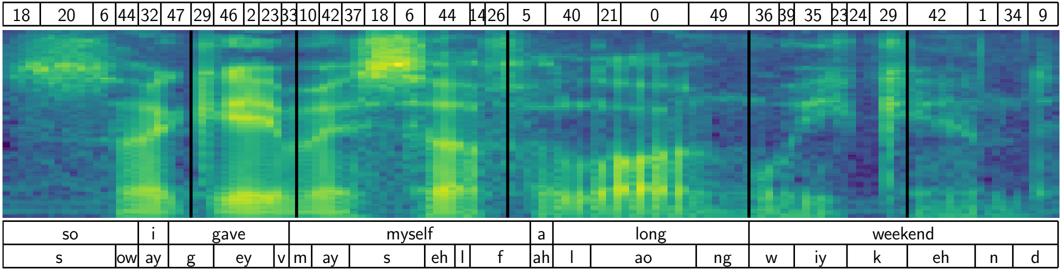
	Word boundary				Token
Model	Prec.	Rec.	F_1	R-val.	F_1
ES-KMeans [Kamper et al., ASRU'17]	30.3	16.6	21.4	39.1	19.2
BES-GMM [Kamper et al., CSL'17]	31.5	12.4	17.8	37.2	18.6
SCPC [Bhati et al., Interspeech'21]	36.9	29.9	33.0	45.6	-
mACPC [Cuervo et al., arXiv'21]	**42.1**	30.3	35.1	**47.4**	-
DPDP AE-RNN on DPDP CPC+K-means	35.3	**37.7**	**36.4**	44.3	**25.0**

ZeroSpeech 2017/2020 evaluation

	Word boundary			Token
Model	Prec.	Rec.	F_1	F_1
French:				
ES-KMeans [Kamper et al., ASRU'17]	37.0	52.2	43.3	6.3
Probabilistic DTW [Räsänen and Blandon, IS'20]	31.6	**86.4**	46.3	5.1
Self-expressing autoencoder [Bhati et al., IS'20]	34.0	83.9	48.4	8.3
DPDP AE-RNN on DPDP CPC+K-means	**49.8**	57.9	**53.5**	**12.2**
Mandarin:				
ES-KMeans [Kamper et al., ASRU'17]	42.6	75.6	54.5	8.1
Probabilistic DTW [Räsänen and Blandon, IS'20]	34.2	87.4	49.2	4.4
Self-expressing autoencoder [Bhati et al., IS'20]	36.5	**91.9**	52.2	12.1
DPDP AE-RNN on DPDP CPC+K-means	**66.2**	70.7	**68.3**	**26.3**


Word types with highest recall
Token recall (%)
50
60
70
80
mm
um
yes
cool
art
yeah
ah
hum
right
too
class
mind
god
true
past


Word types with lowest recall
Token recall (%)
0
2
4
6
8
parents
into
other
believe
told
married
still
sister
college
exactly
children
somebody
everything
basically
actually


18 20 6 44 32 47 29 46 2 23 33 10 42 37 18 6 44 14 26 5 40 21 0 49 36 39 35 23 24 29 42 1 34 9
so i gave myself a long weekend
s ow ay g ey v m ay s eh l f ah l ao ng w iy k eh n d

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive

- DPDP improvements: Duration modelling and lexicon discovery

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive
- DPDP improvements: Duration modelling and lexicon discovery
- Absolute results are still low: Word token $F_1 \approx 25\%$
- Are we starting to reach limits of word segmentation from unlabelled speech?

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive

- DPDP improvements: Duration modelling and lexicon discovery

- Absolute results are still low: Word token $F_1 \approx 25\%$

- Are we starting to reach limits of word segmentation from unlabelled speech?

- Start to integrate external top-down and grounding signals, e.g. other modalities
  [Bisk et al., EMNLP'20; Moulin-Frier and Oudeyer, AAAI'21]

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive
- DPDP improvements: Duration modelling and lexicon discovery
- Absolute results are still low: Word token $F_1 \approx 25\%$
- Are we starting to reach limits of word segmentation from unlabelled speech?
- Start to integrate external top-down and grounding signals, e.g. other modalities
  [Bisk et al., EMNLP'20; Moulin-Frier and Oudeyer, AAAI'21]

Conclusion & Where do we go from here?

- Bottom-up acoustic unit discovery with symbolic word segmentation is competitive
- DPDP improvements: Duration modelling and lexicon discovery
- Absolute results are still low: Word token $F_1 \approx 25\%$
- Are we starting to reach limits of word segmentation from unlabelled speech?
- Start to integrate external top-down and grounding signals, e.g. other modalities
  [Bisk et al., EMNLP'20; Moulin-Frier and Oudeyer, AAAI'21]

H. Kamper, "Word segmentation on discovered phone units with dynamic programming and self-supervised scoring," *arXiv preprint arXiv:2202.11929*, 2022

`https://github.com/kamperh/dpdp_aernn`

`https://github.com/kamperh/vqwordseg`