Test slide

- Is there a chat box?
- Can you see my pointer?
- Can you hear this: Play

Learning acoustic units and words from unlabelled speech (with a bit of vision)

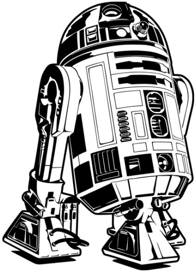
CLSP Seminar, Johns Hopkins University, Oct. 2020

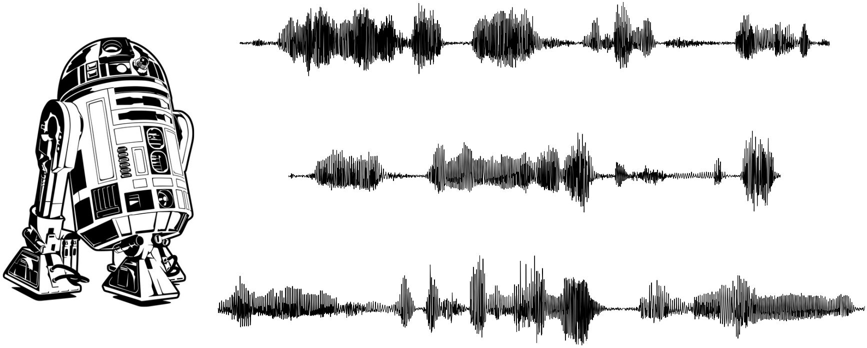
Herman Kamper

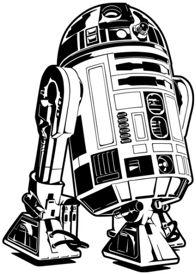
E&E Engineering, Stellenbosch University, South Africa

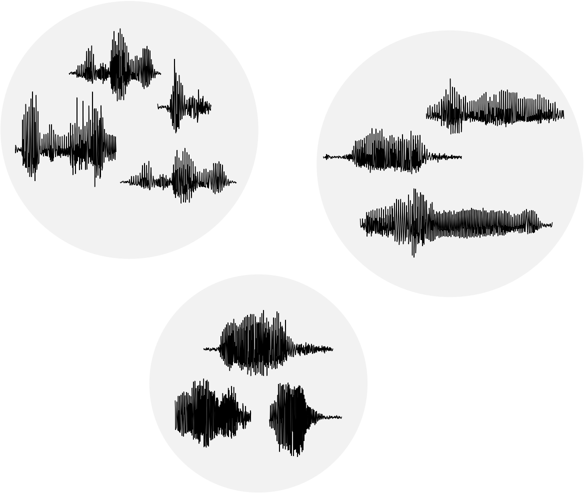
http://www.kamperh.com/

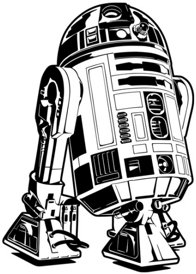
Photo: Leon Croukamp

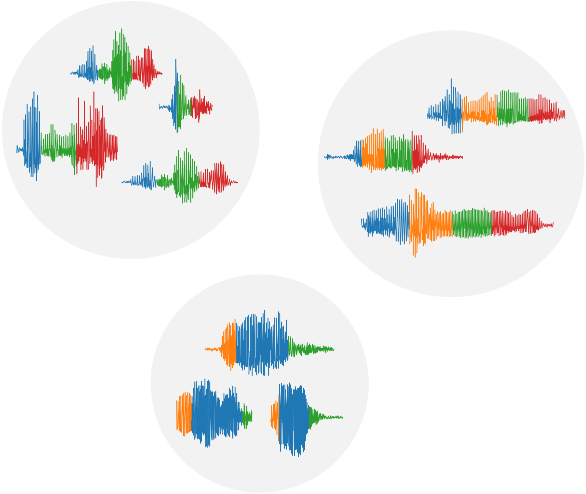












Why unsupervised speech processing?

Why unsupervised speech processing?

Bootstrap low-resource speech technology

Why unsupervised speech processing?

Bootstrap low-resource speech technology

Applications such as non-parallel voice conversion

Why unsupervised speech processing?

Bootstrap low-resource speech technology

Applications such as non-parallel voice conversion

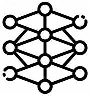
Cognitive models of language acquisition

Why unsupervised speech processing?

Bootstrap low-resource speech technology

Applications such as non-parallel voice conversion

Cognitive models of language acquisition



New insights and modelling approaches

Experience Grounds Language

Yonatan Bisk* **Ari Holtzman*** **Jesse Thomason***

Jacob Andreas Yoshua Bengio Joyce Chai Mirella Lapata
Angeliki Lazaridou Jonathan May Aleksandr Nisnevich Nicolas Pinto Joseph Turian

Experience Grounds Language

Yonatan Bisk* **Ari Holtzman*** **Jesse Thomason***

Jacob Andreas Yoshua Bengio Joyce Chai Mirella Lapata
Angeliki Lazaridou Jonathan May Aleksandr Nisnevich Nicolas Pinto Joseph Turian

You can't learn language ...

... from the radio (internet). **WS2 $\subset$ WS3**

> *A learner cannot be said to be in* WS3 *if it can perform its task without sensory perception such as visual, auditory, or tactile information.*

... from a television. **WS3 $\subset$ WS4**

> *A learner cannot be said to be in* WS4 *if the space of actions and consequences of its environment can be enumerated.*

... by yourself. **WS4 $\subset$ WS5**

> *A learner cannot be said to be in* WS5 *if its cooperators can be replaced with cleverly pre-programmed agents to achieve the same goals.*

Experience Grounds Language

Yonatan Bisk* **Ari Holtzman*** **Jesse Thomason***

Jacob Andreas Yoshua Bengio Joyce Chai Mirella Lapata
Angeliki Lazaridou Jonathan May Aleksandr Nisnevich Nicolas Pinto Joseph Turian

You can't learn language ...

... from the radio (internet). **WS2 ⊂ WS3**

> *A learner cannot be said to be in WS3 if it can perform its task without sensory perception such as visual, auditory, or tactile information.*

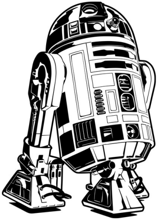
... from a television. **WS3 ⊂ WS4**

> *A learner cannot be said to be in WS4 if the space of actions and consequences of its environment can be enumerated.*

... by yourself. **WS4 ⊂ WS5**

> *A learner cannot be said to be in WS5 if its cooperators can be replaced with cleverly pre-programmed agents to achieve the same goals.*

But what can (and should) we learn at these different levels?



Levels of language learning (for word and phone acquisition)

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio?

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio?

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television?

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio?

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television?

3. What can we learn from interaction/feedback from our environment and other "agents"?

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio?
   — **Part 1**

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television?

3. What can we learn from interaction/feedback from our environment and other "agents"?

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio?
   — **Part 1**

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television? — **Part 2**

3. What can we learn from interaction/feedback from our environment and other "agents"?

1. Vector-quantised neural networks for unsupervised acoustic unit discovery

1. Vector-quantised neural networks for unsupervised acoustic unit discovery

Benjamin
van Niekerk

Leanne
Nortje

1. Vector-quantised neural networks for unsupervised acoustic unit discovery

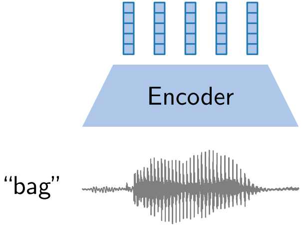
Benjamin
van Niekerk

Leanne
Nortje

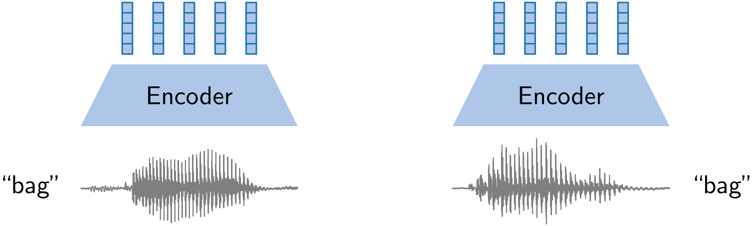
Van Niekerk et al., "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," *Interspeech*, 2020.

Phonetic representation learning

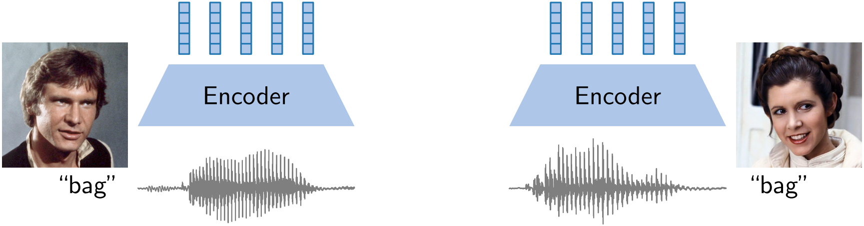
Phonetic representation learning



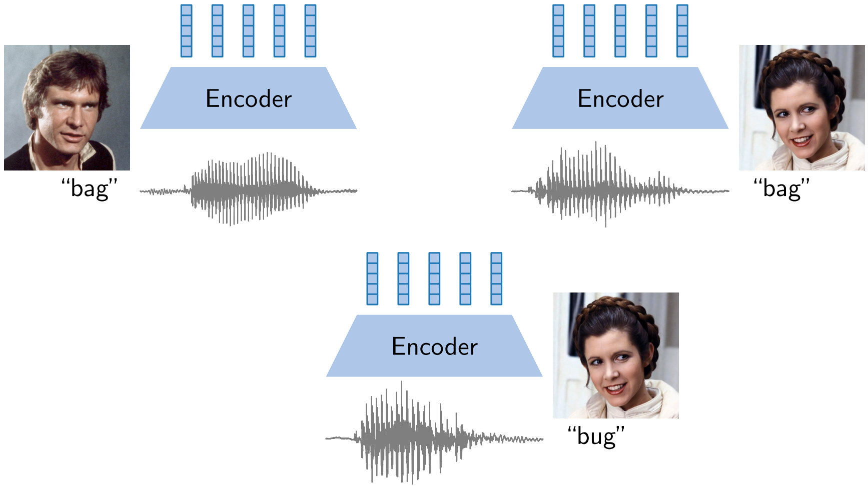
Phonetic representation learning



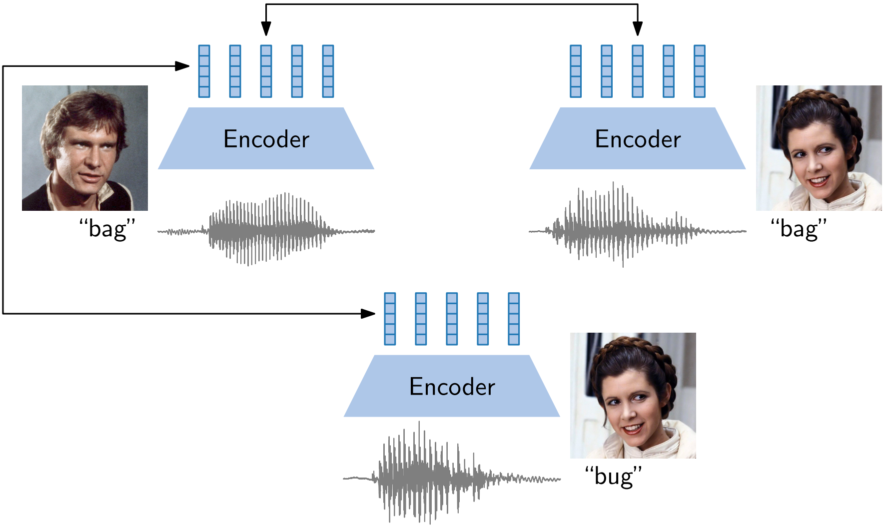
Phonetic representation learning



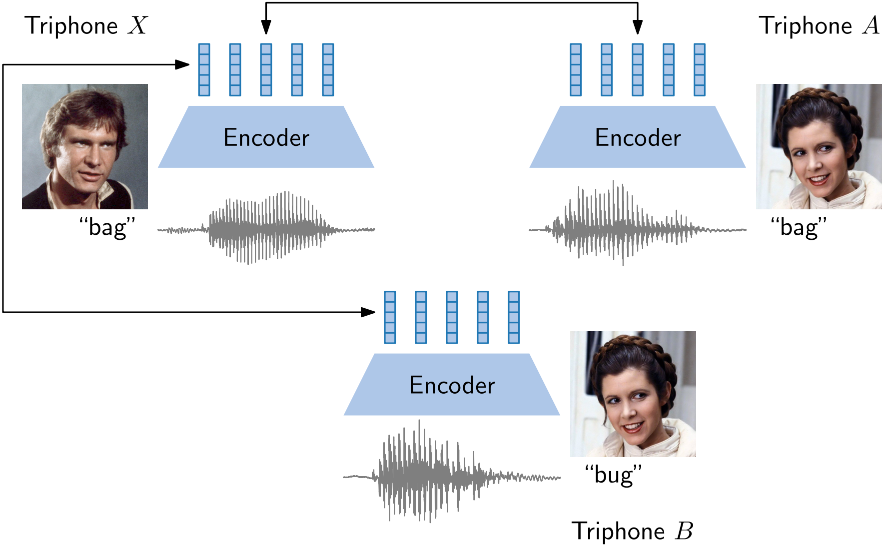
Phonetic representation learning



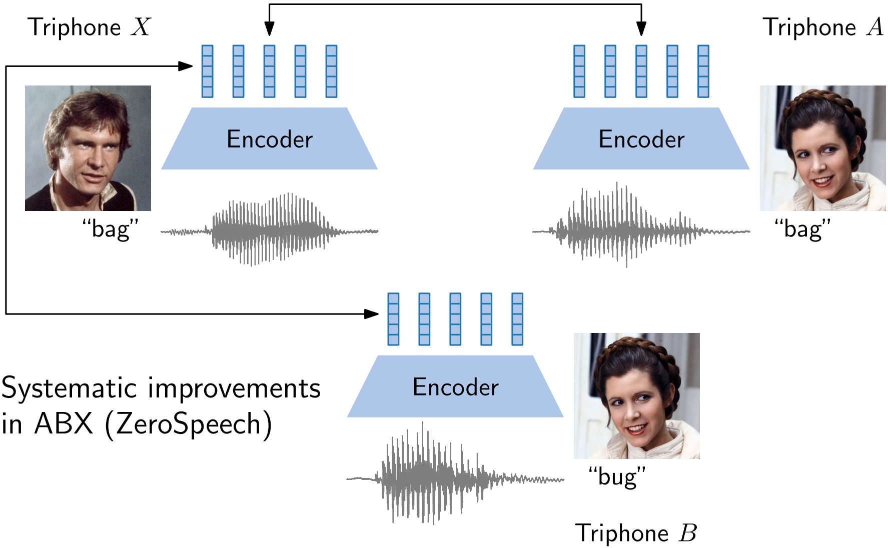
Phonetic representation learning



Phonetic representation learning

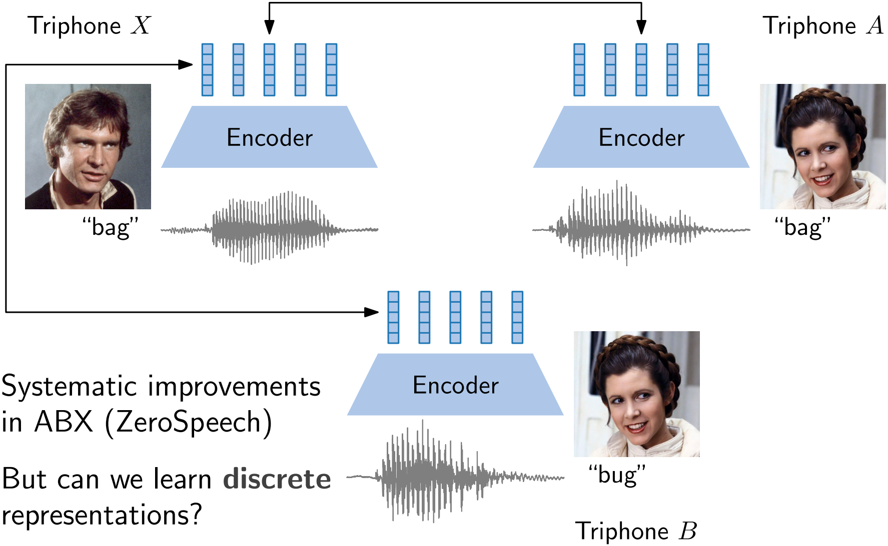


Phonetic representation learning



Systematic improvements in ABX (ZeroSpeech)

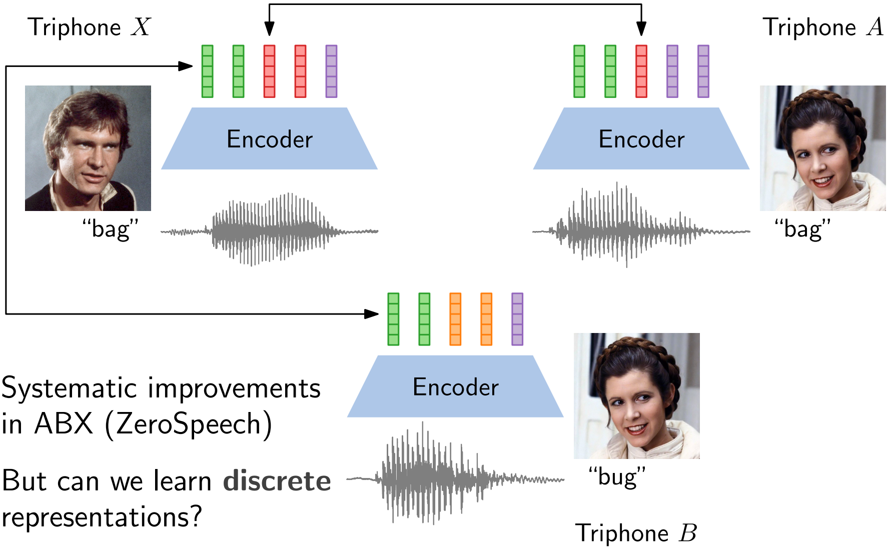
Phonetic representation learning



Systematic improvements in ABX (ZeroSpeech)

But can we learn **discrete** representations?

Phonetic representation learning

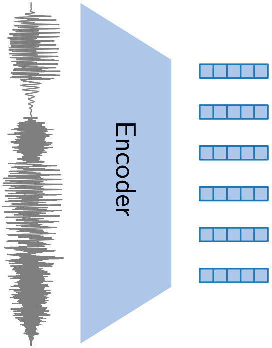


Systematic improvements in ABX (ZeroSpeech)

But can we learn **discrete** representations?

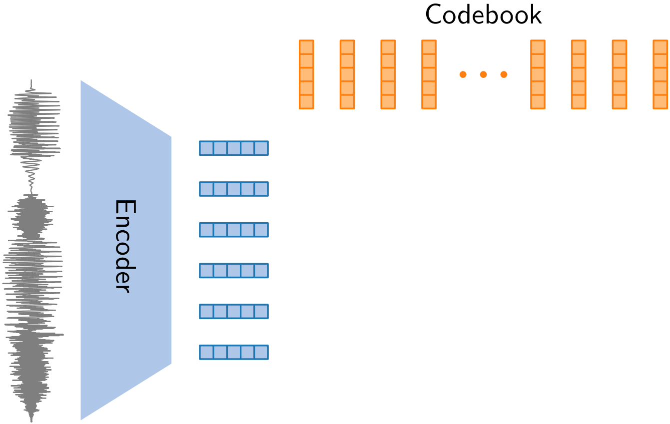
Vector quantisation in neural networks

Vector quantisation in neural networks



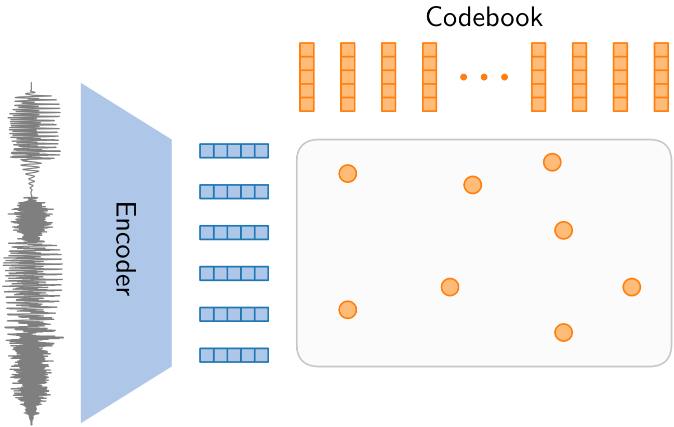
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



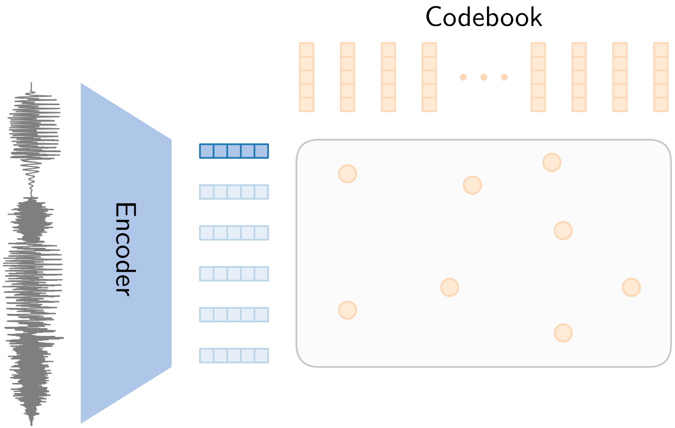
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



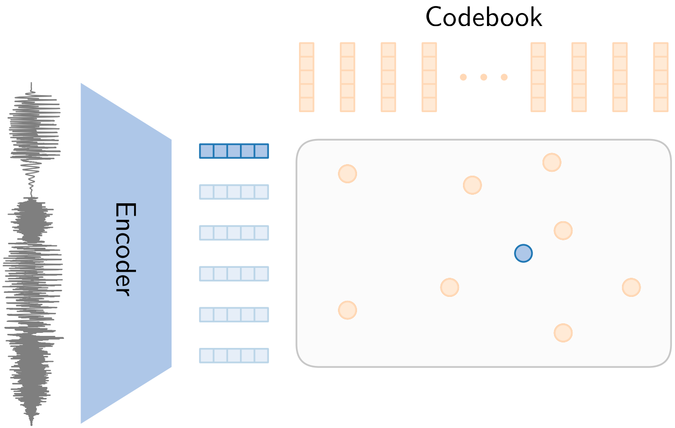
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



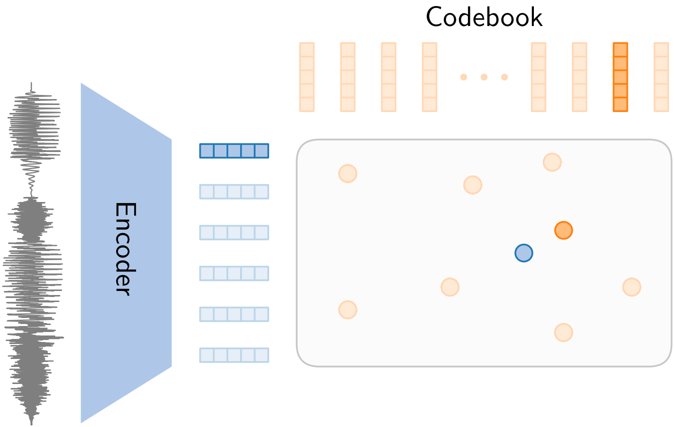
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



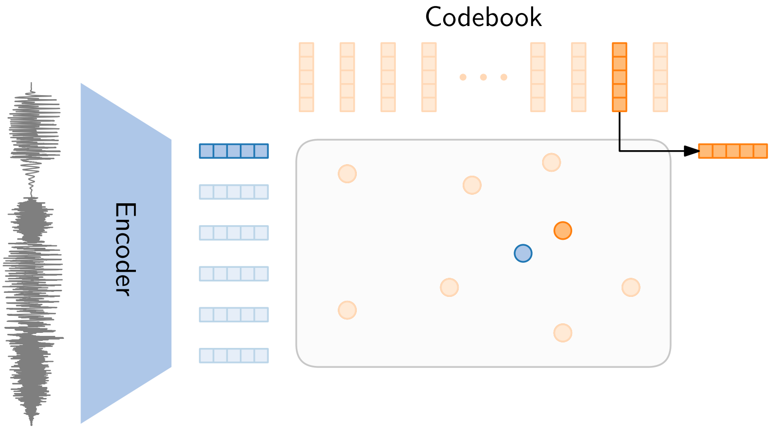
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



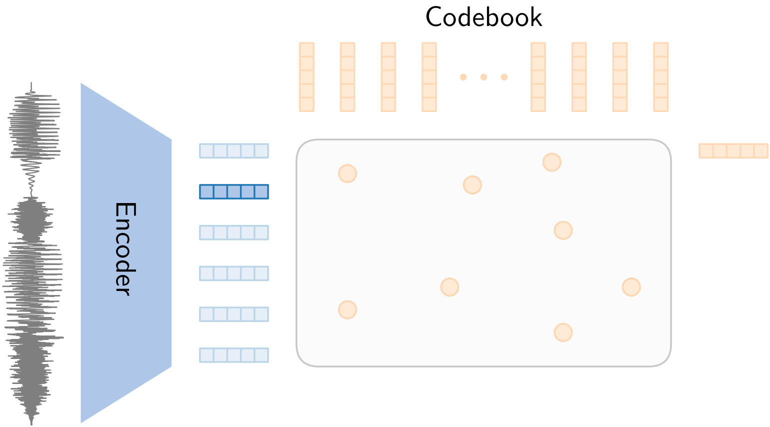
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



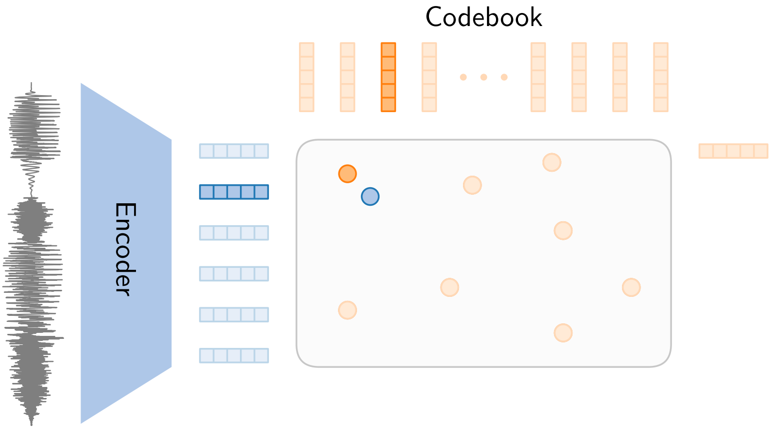
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



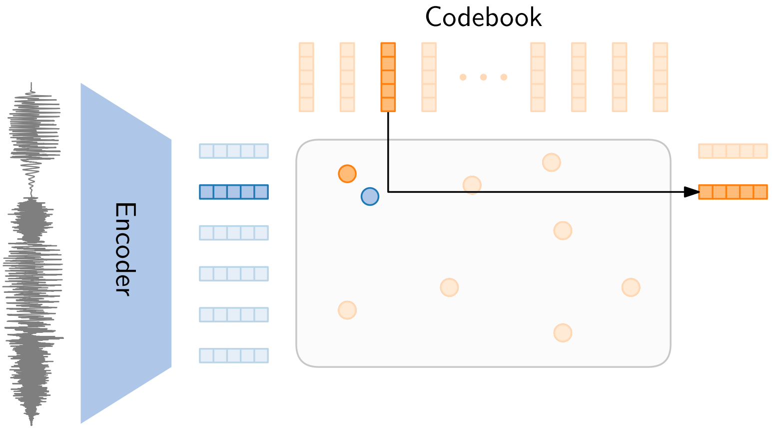
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



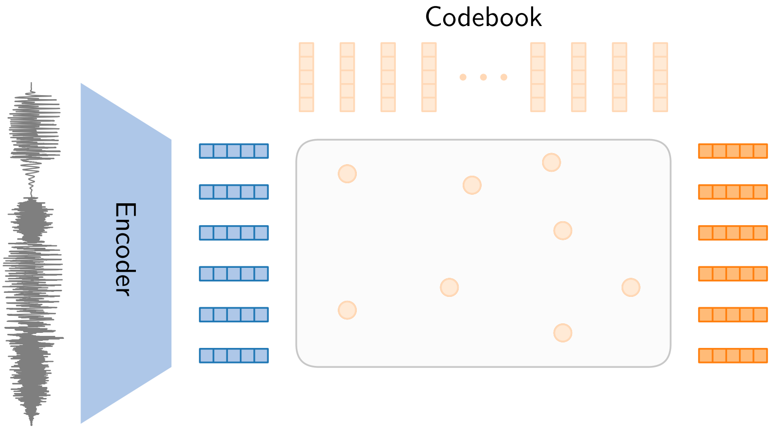
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



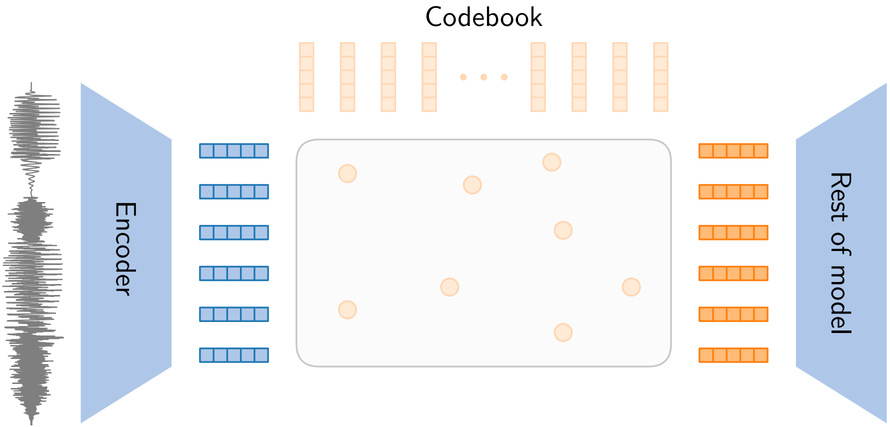
Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

Vector quantisation in neural networks



Van den Oord et al., "Neural discrete representation learning," *NeurIPS*, 2017.

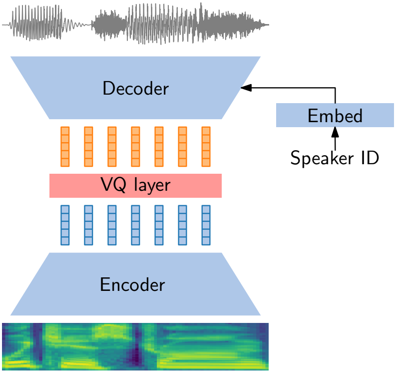
Our contribution

We propose and compare two models for unsupervised acoustic unit discovery:

Van Niekerk et al., "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," *Interspeech*, 2020.

Our contribution

We propose and compare two models for unsupervised acoustic unit discovery:



VQ-VAE: A vector-quantised variational autoencoder

Inspired by:
Chorowski, et al., "Unsupervised speech representation learning using wavenet autoencoders," *TASLP*, 2019.

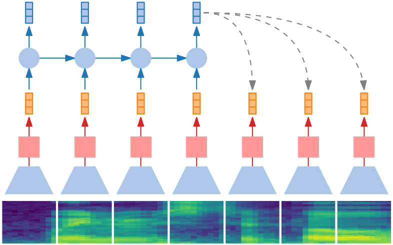
Van Niekerk et al., "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," *Interspeech*, 2020.

Our contribution

We propose and compare two models for unsupervised acoustic unit discovery:

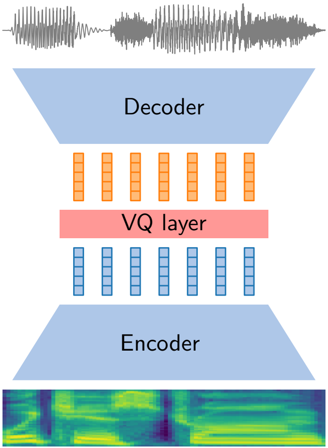
VQ-CPC: Combining vector quantisation with contrastive predictive coding

Inspired by:
Van den Oord, et al., "Representation learning with contrastive predictive coding," *arXiv*, 2018.

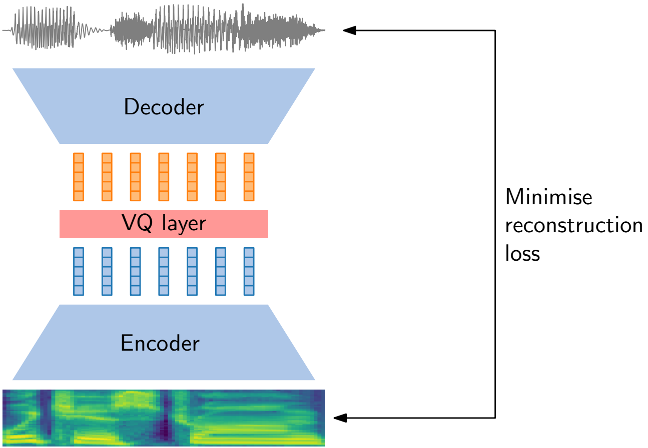


Van Niekerk et al., "Vector-quantized neural networks for acoustic unit discovery in the ZeroSpeech 2020 challenge," *Interspeech*, 2020.

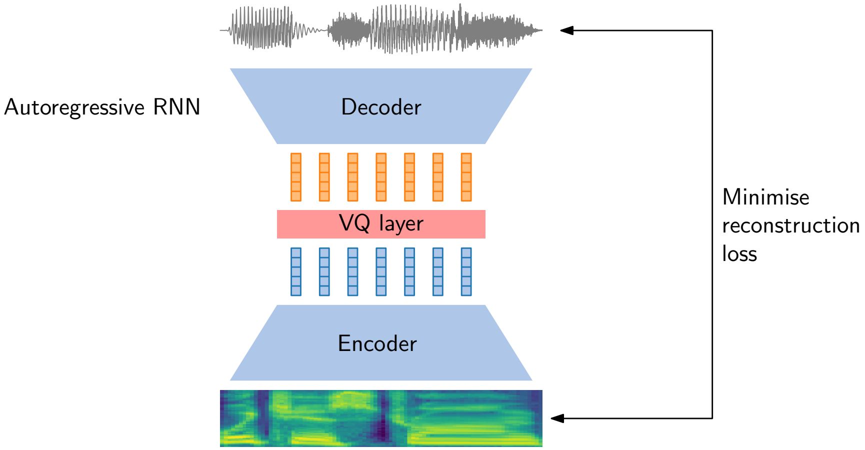
Vector-quantised variational autoencoder



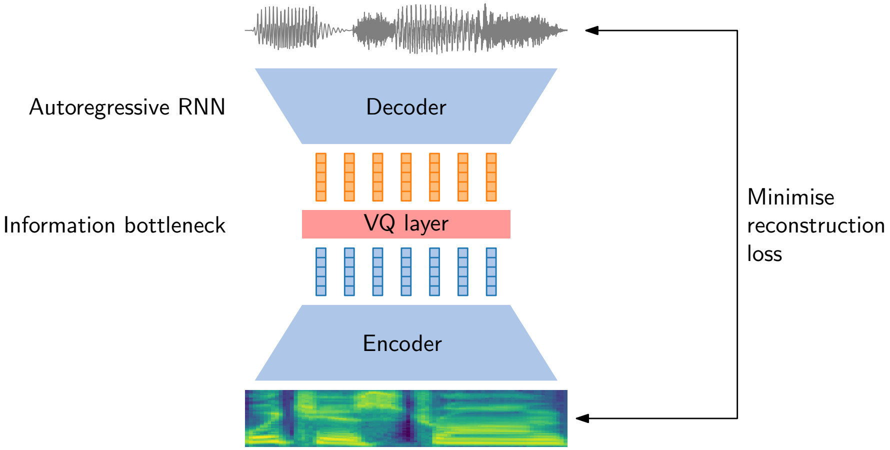
Vector-quantised variational autoencoder



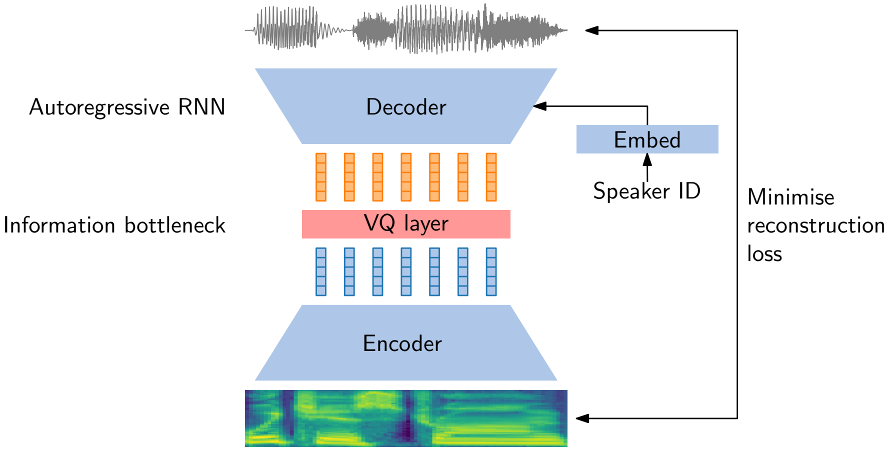
Vector-quantised variational autoencoder



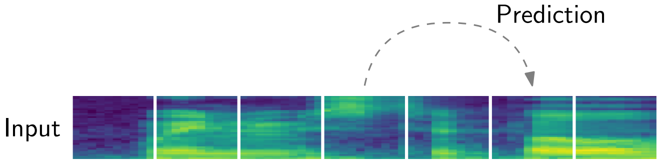
Vector-quantised variational autoencoder



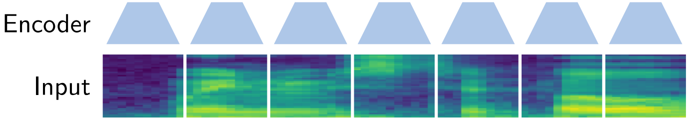
Vector-quantised variational autoencoder



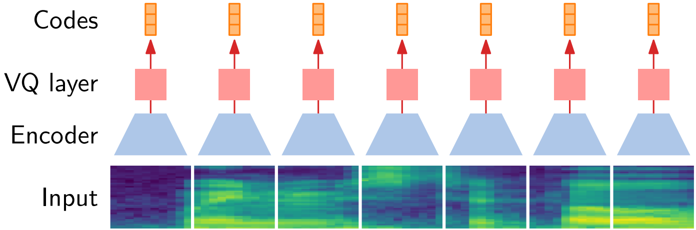
Vector-quantised contrastive predictive coding



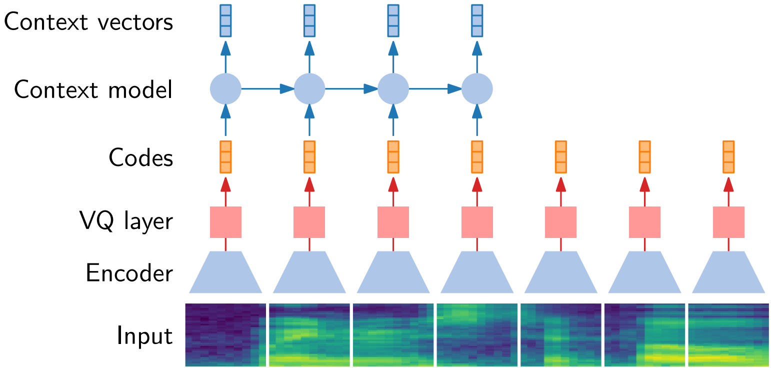
Vector-quantised contrastive predictive coding



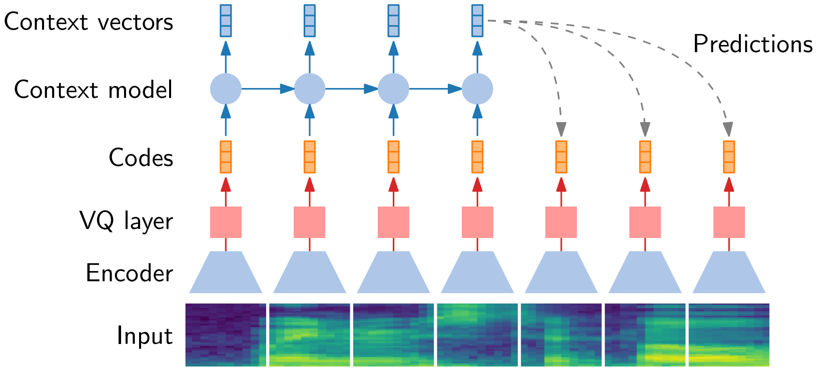
Vector-quantised contrastive predictive coding



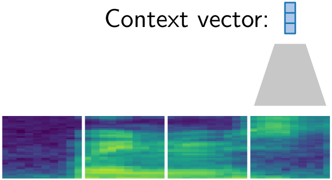
Vector-quantised contrastive predictive coding



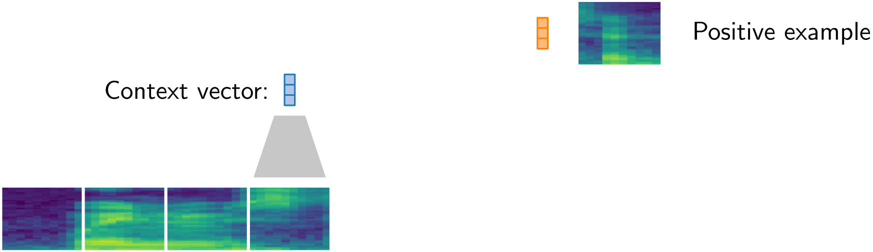
Vector-quantised contrastive predictive coding



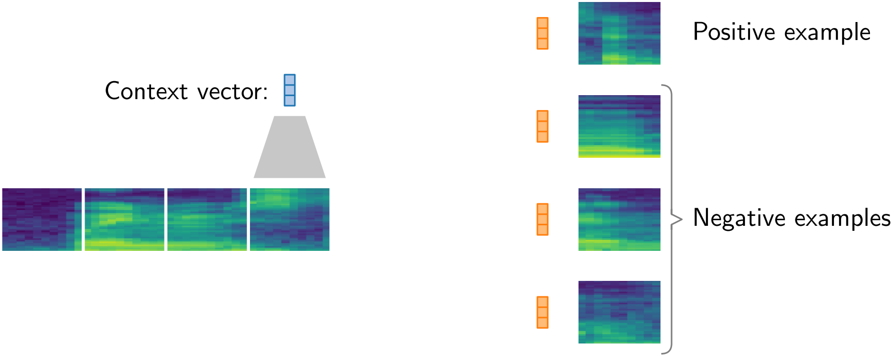
Vector-quantised contrastive predictive coding



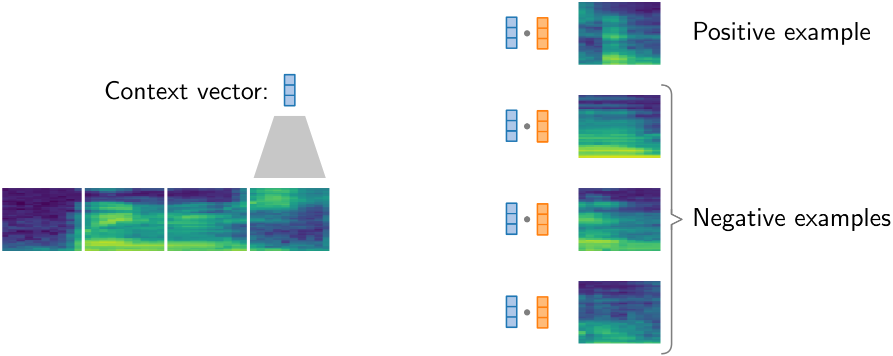
Vector-quantised contrastive predictive coding



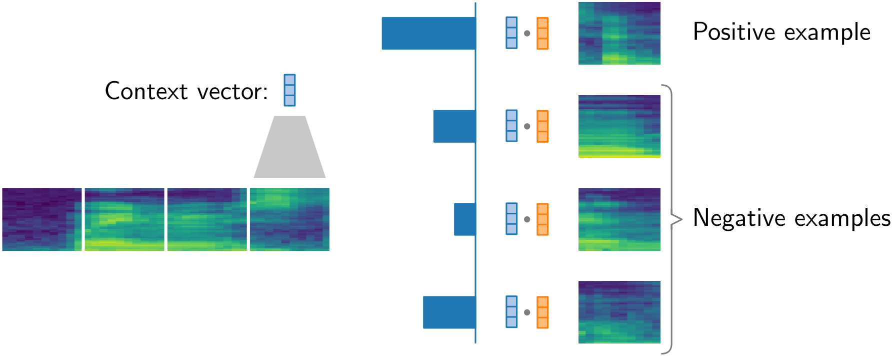
Vector-quantised contrastive predictive coding



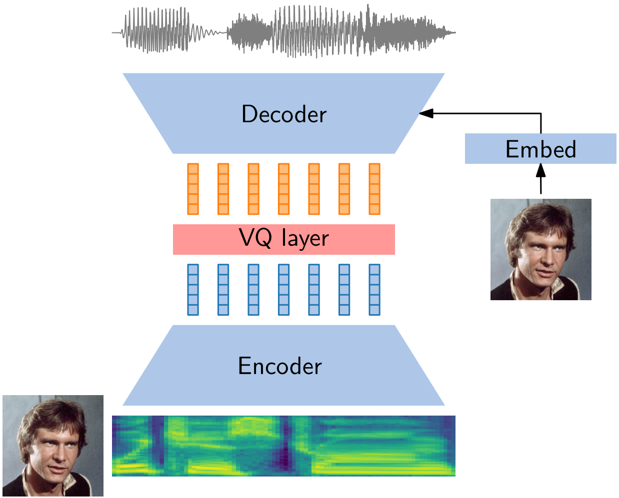
Vector-quantised contrastive predictive coding



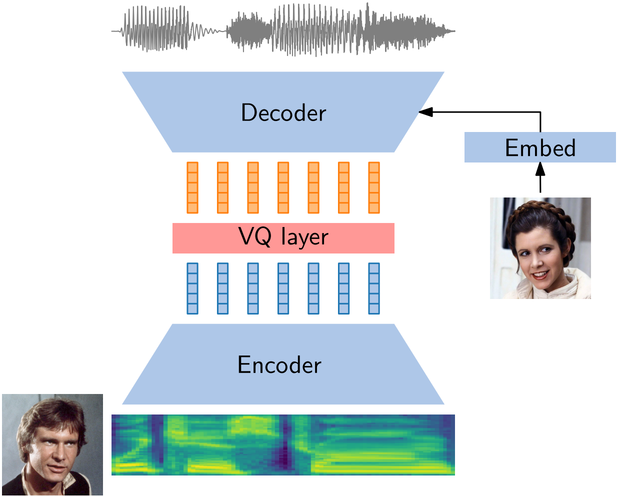
Vector-quantised contrastive predictive coding



Evaluation: Voice conversion



Evaluation: Voice conversion



Example conversions

Example 1:

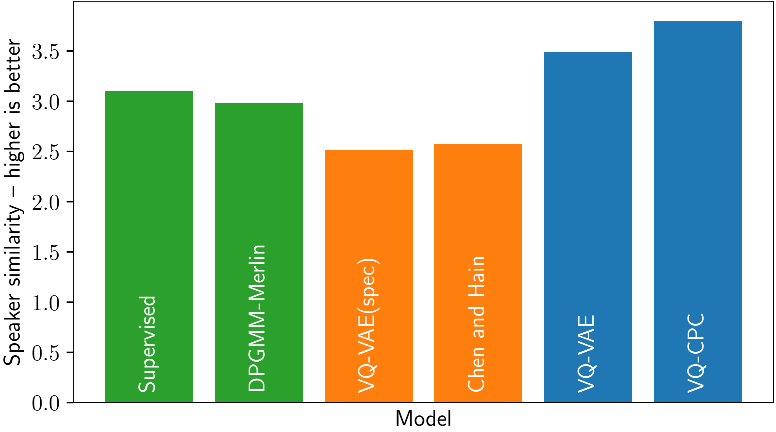
- Source: Play
- Converted: Play
- Target: Play

Example 2:

- Source: Play
- Converted: Play
- Target: Play

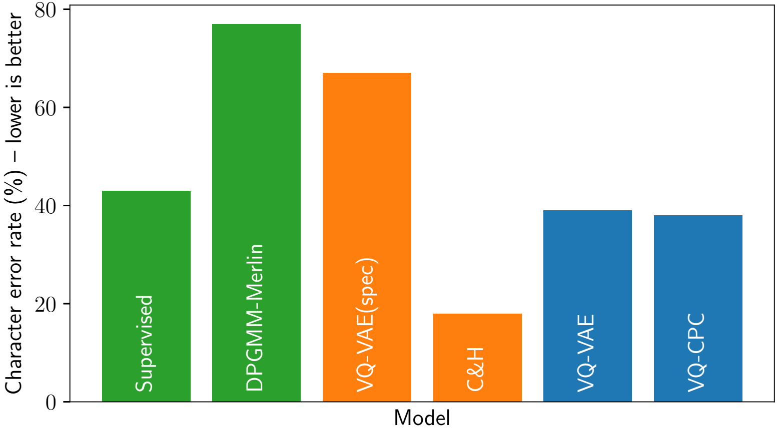
Evaluation: Speaker similarity

Evaluation: Speaker similarity

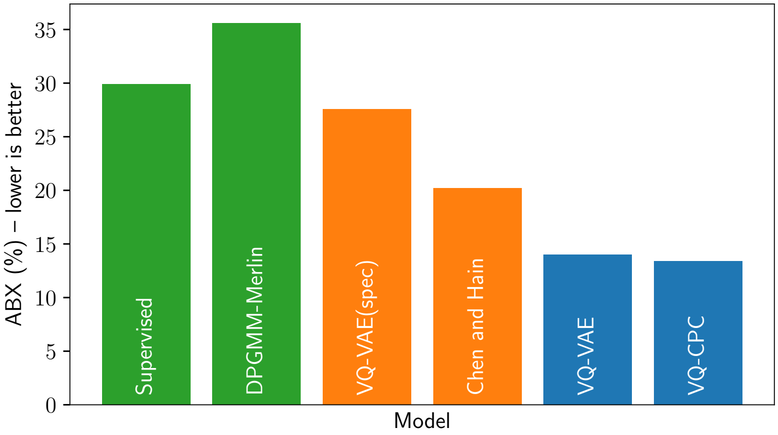


Chen and Hain, "Unsupervised acoustic unit representation learning for voice conversion using WaveNet auto-encoders," *Interspeech*, 2020.

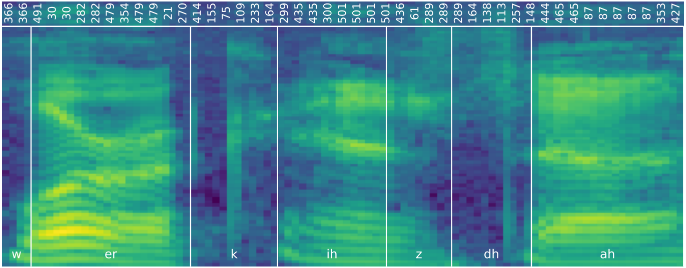
Evaluation: Intelligibility



Evaluation: ABX phone discrimination

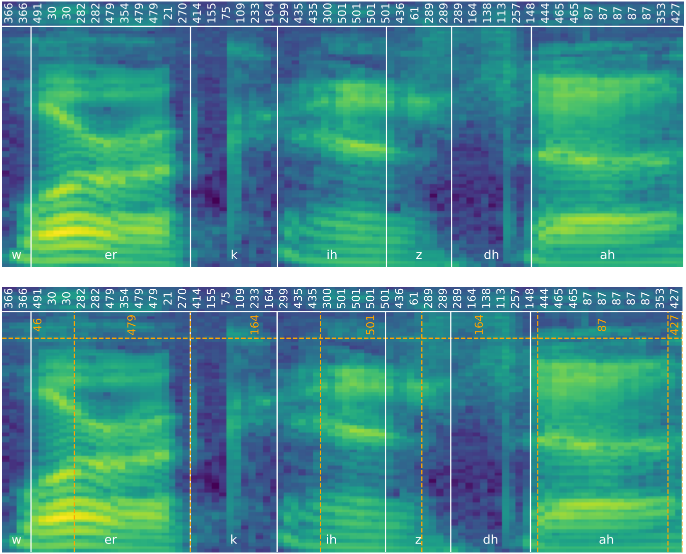


VQ-CPC codes



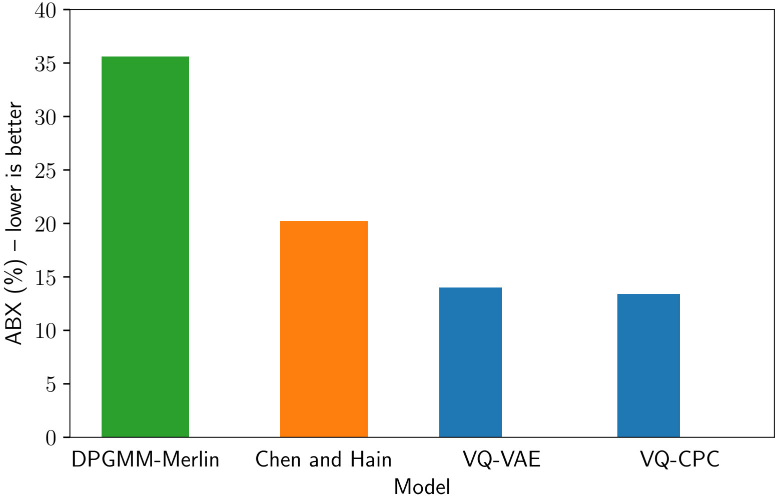
366 366 491 30 30 282 282 479 354 479 479 21 270 414 155 75 109 233 164 299 435 435 300 501 501 501 501 436 61 289 289 289 164 138 113 257 148 444 465 465 87 87 87 87 87 353 427
w
er
k
ih
z
dh
ah

VQ-CPC codes

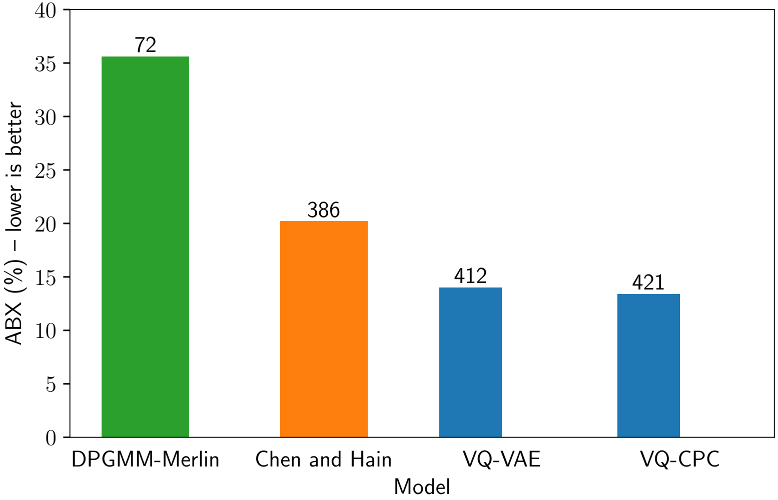


Inspired by:
Chorowski et al., "Unsupervised neural segmentation and clustering for unit discovery in sequential data," *PGR Workshop*, 2019.

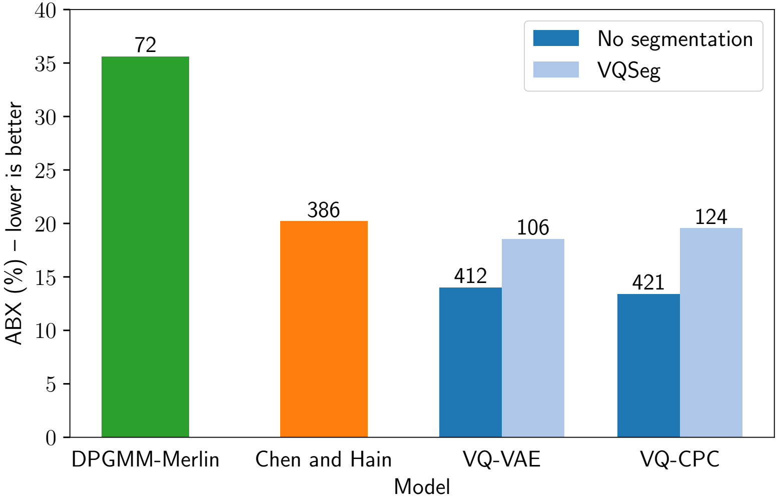
Evaluation: ABX phone discrimination



Evaluation: ABX phone discrimination



Evaluation: ABX phone discrimination



Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio? — **Part 1**
2. What can we learn from co-occurring (grounding) signals like vision, i.e. television?
3. What can we learn from interaction/feedback from our environment and other "agents"?

Levels of language learning (for word and phone acquisition)

1. What can we learn from unlabelled speech audio, i.e. radio? — **Part 1**

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television? — **Part 2**

3. What can we learn from interaction/feedback from our environment and other “agents”?

2. Multimodal few-shot learning from images and speech

2. Multimodal few-shot learning from images and speech

Ryan
Eloff

Herman
Engelbrecht

Leanne
Nortje

2. Multimodal few-shot learning from images and speech

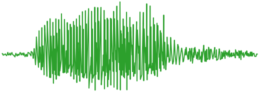
Ryan
Eloff

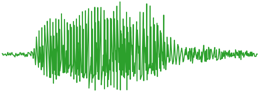
Herman
Engelbrecht

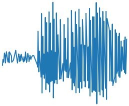
Leanne
Nortje

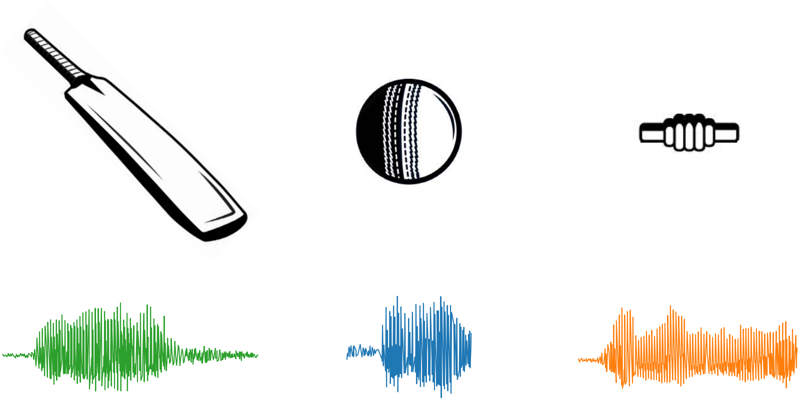
Nortje and Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images," *Interspeech*, 2020.


TOYOTA
HSR
104
Preferred Networks
Preferred Networks
Preferred Networks

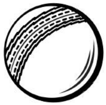


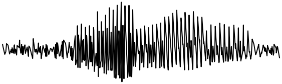


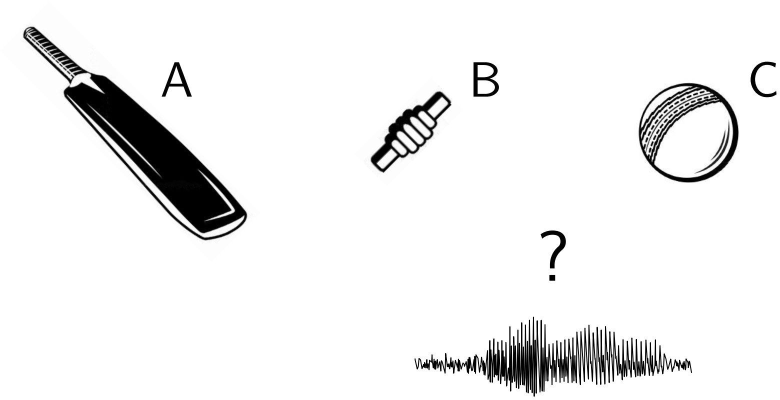






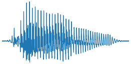


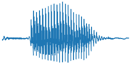


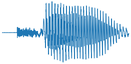

A
B
C
?

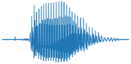
Unimodal one-shot learning and classification


– three


– one

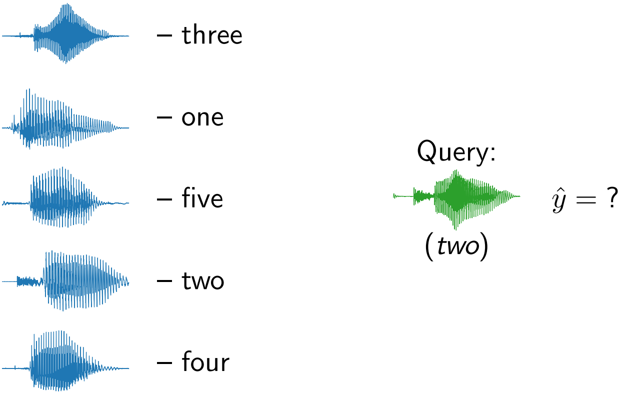

– five


– two


– four

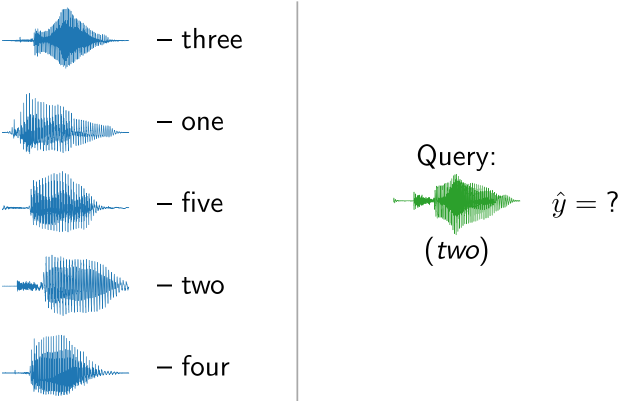
Fei-Fei et al., “One-shot learning of object categories,” *TPAMI*, 2006.
Lake et al., “One-shot learning of generative speech concepts,” *CogSci*, 2014.

Unimodal one-shot learning and classification



Fei-Fei et al., "One-shot learning of object categories," *TPAMI*, 2006.
Lake et al., "One-shot learning of generative speech concepts," *CogSci*, 2014.

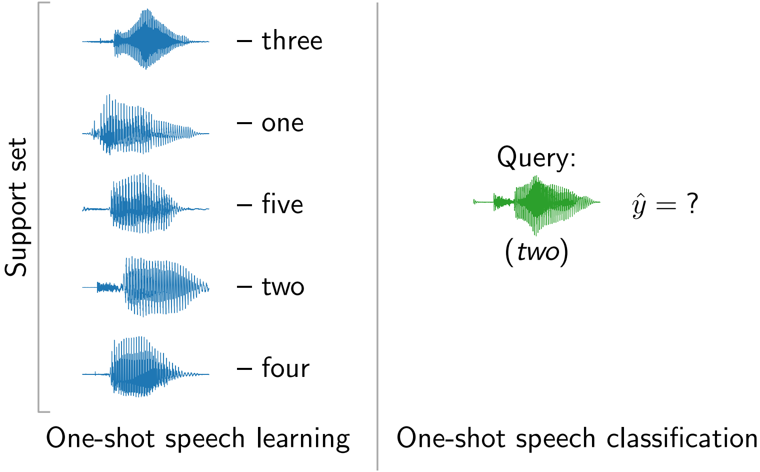
Unimodal one-shot learning and classification



One-shot speech learning | One-shot speech classification

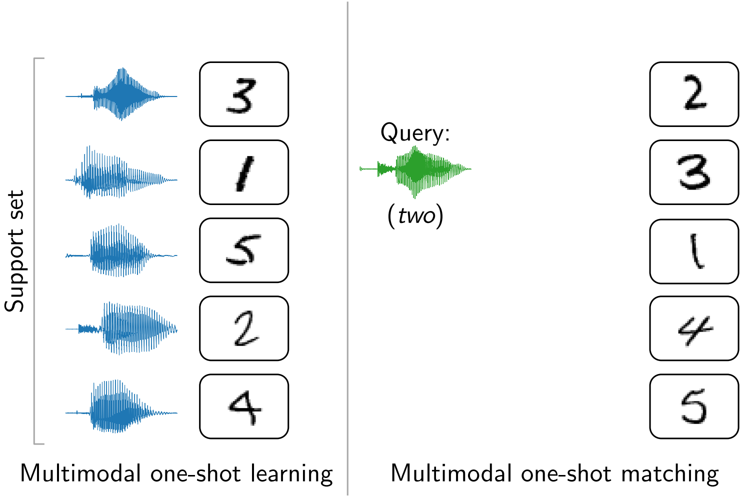
Fei-Fei et al., "One-shot learning of object categories," *TPAMI*, 2006.
Lake et al., "One-shot learning of generative speech concepts," *CogSci*, 2014.

Unimodal one-shot learning and classification



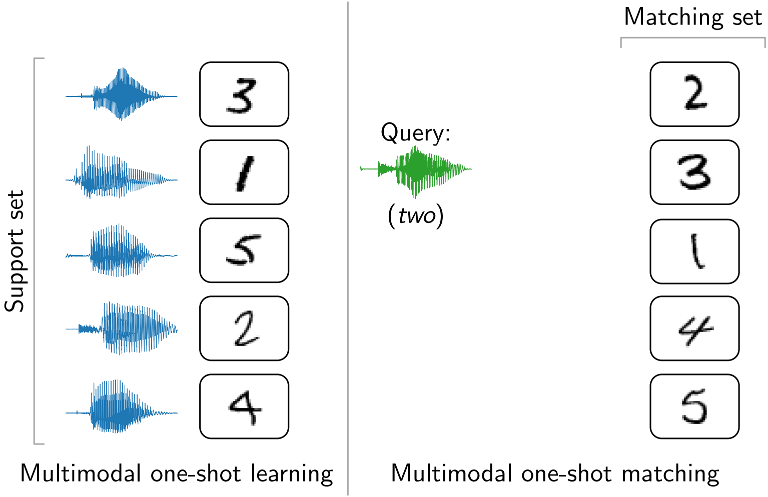
Fei-Fei et al., "One-shot learning of object categories," *TPAMI*, 2006.
Lake et al., "One-shot learning of generative speech concepts," *CogSci*, 2014.

Multimodal one-shot learning and matching



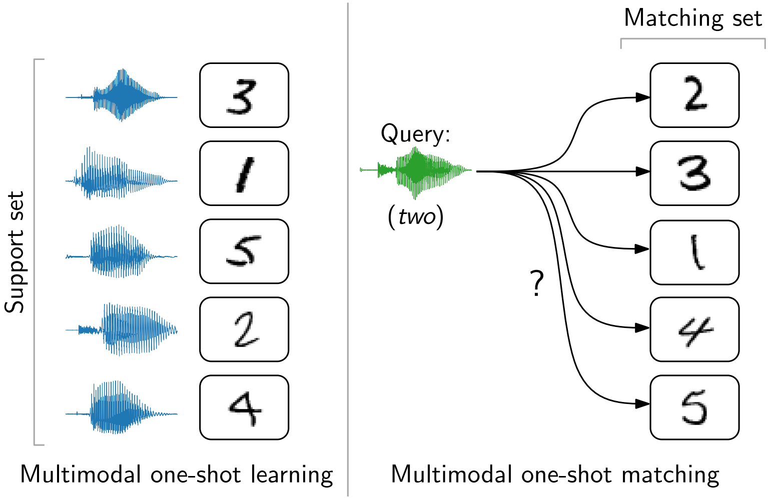
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Multimodal one-shot learning and matching



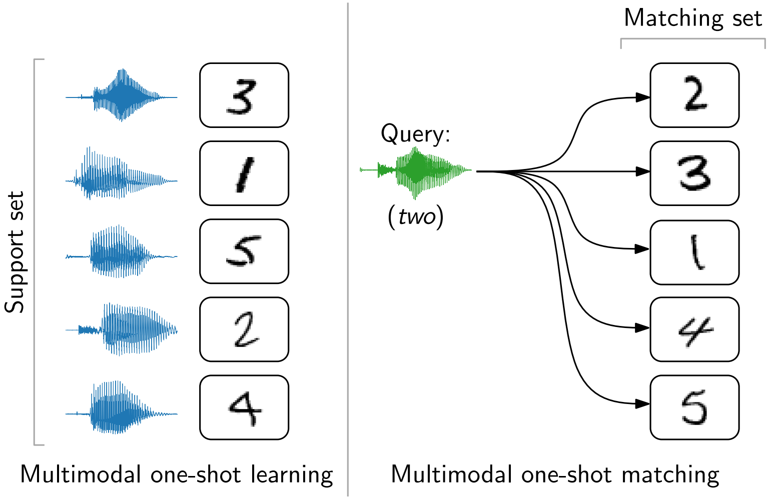
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Multimodal one-shot learning and matching



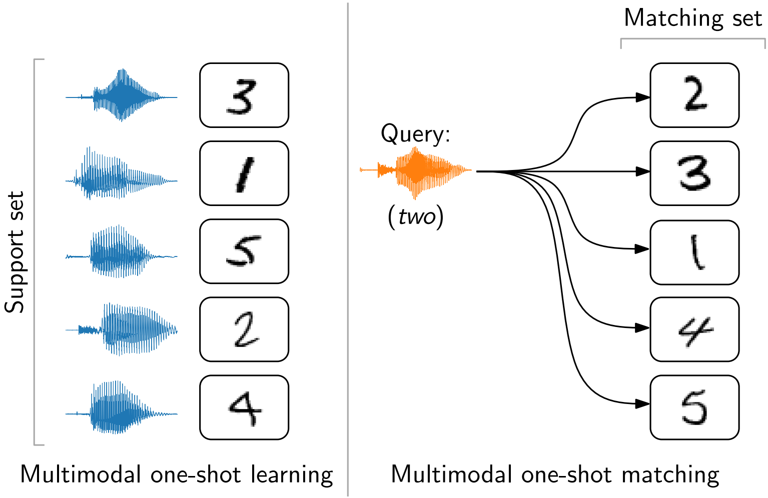
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



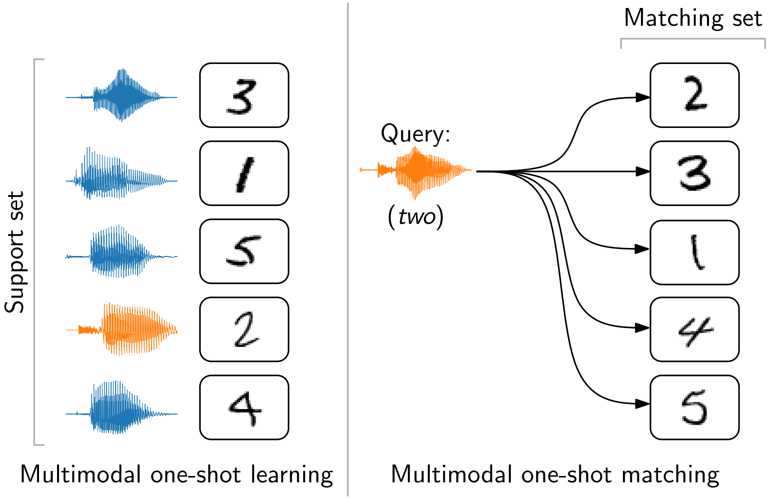
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



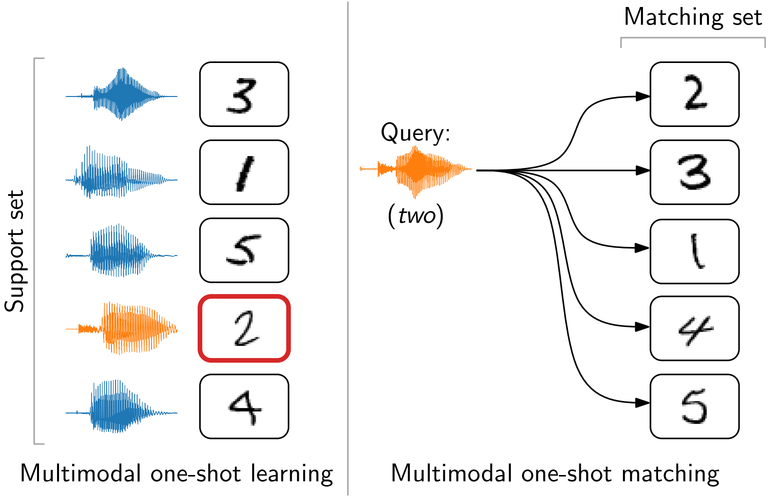
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



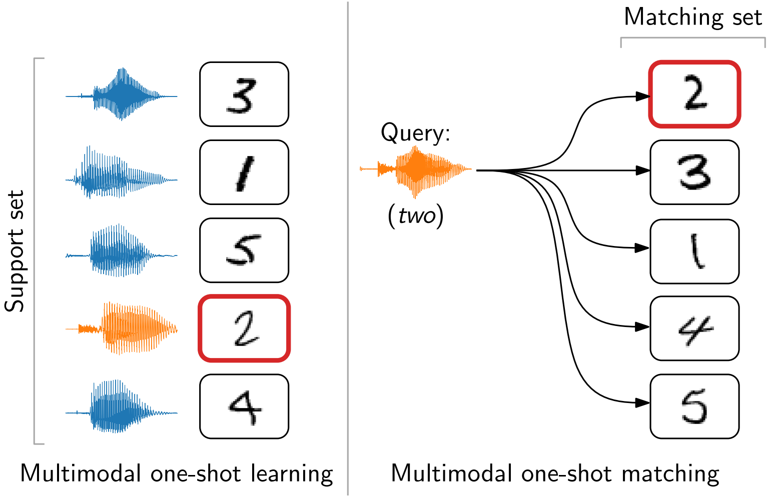
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



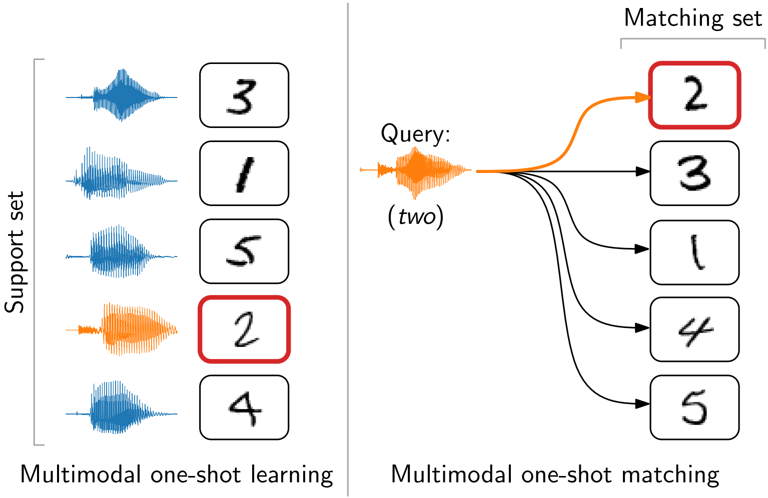
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



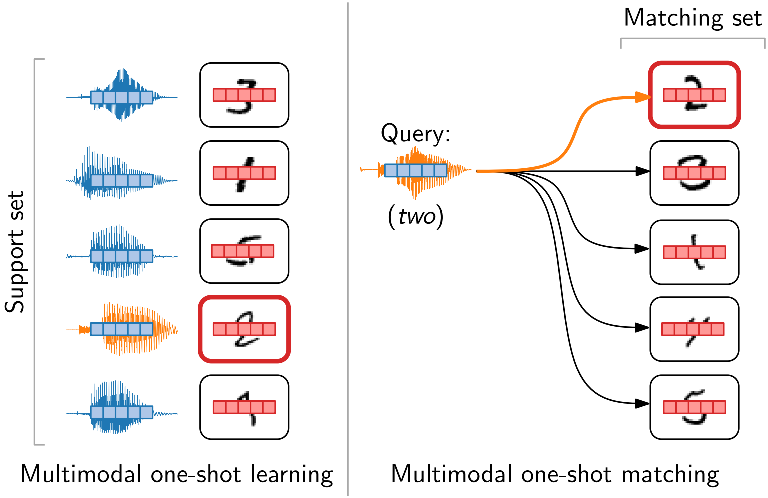
Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach



Eloff et al., "Multimodal one-shot learning of speech and images," *ICASSP*, 2019.

Two-step (indirect) multimodal one-shot approach

- Requires within-modality speech-to-speech and image-to-image distance metrics
- Baseline: DTW over speech, cosine over image pixels
- Or representations/distance metrics can be **learned**

Two-step (indirect) multimodal one-shot approach

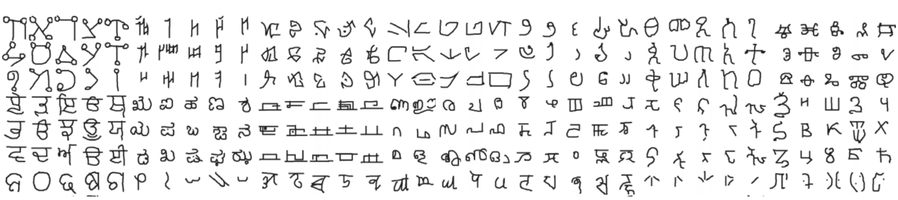
- Requires within-modality speech-to-speech and image-to-image distance metrics
- Baseline: DTW over speech, cosine over image pixels
- Or representations/distance metrics can be **learned**
- Compare two learning methodologies on TIDigits (speech) paired with MNIST (images)

Nortje and Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images," *Interspeech*, 2020.

1. Transfer learning from labelled background data

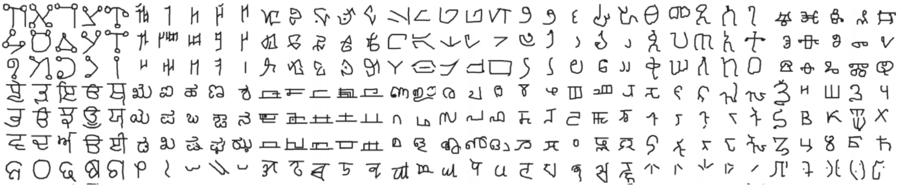
1. Transfer learning from labelled background data

Omniglot (no digits):

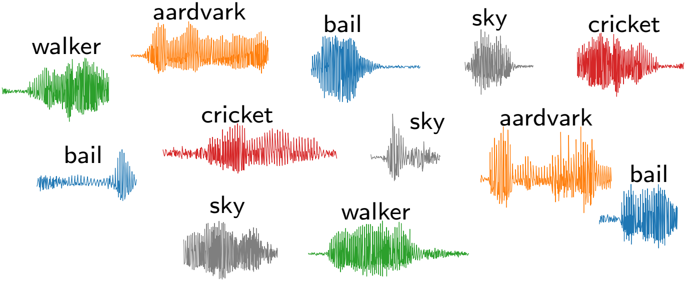


1. Transfer learning from labelled background data

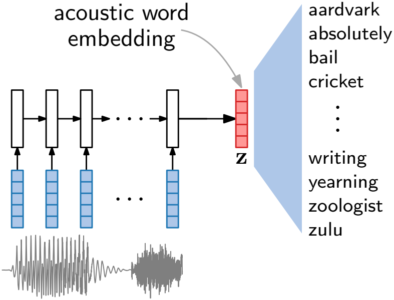
Omniglot (no digits):



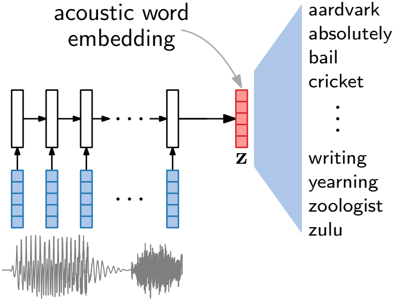
Isolated labelled words (no digits):

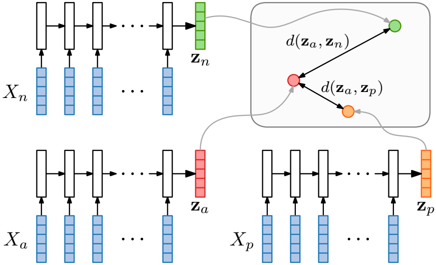


1. Supervised models for transfer learning

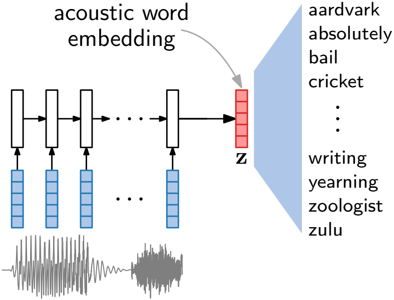


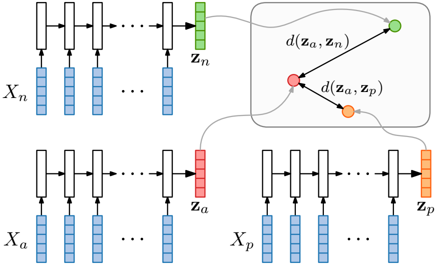
1. Supervised models for transfer learning





1. Supervised models for transfer learning

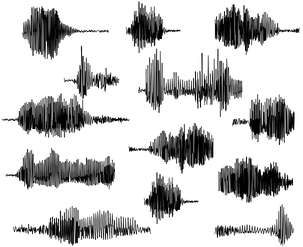




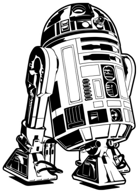
Settle and Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," *SLT*, 2016.

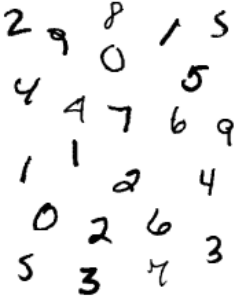
2. Unsupervised learning from unlabelled in-domain data

2. Unsupervised learning from unlabelled in-domain data



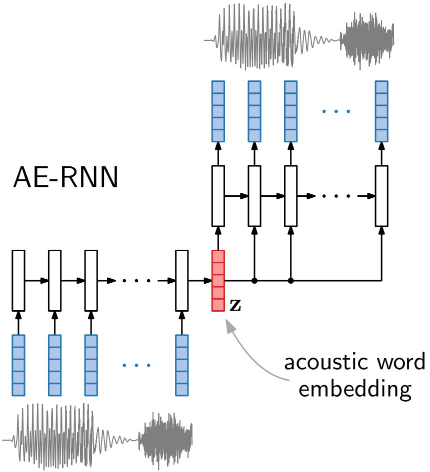
Unlabelled speech





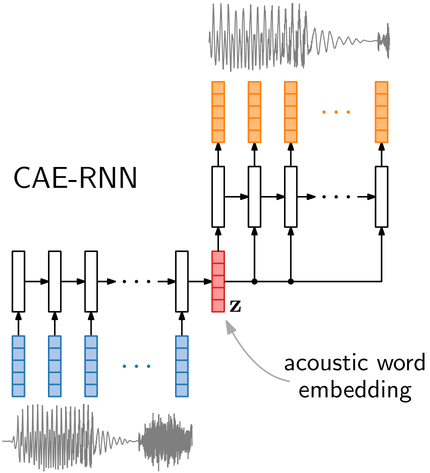
Unlabelled images

2. Unsupervised models



Chung et al., "Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks," *Interspeech*, 2016.

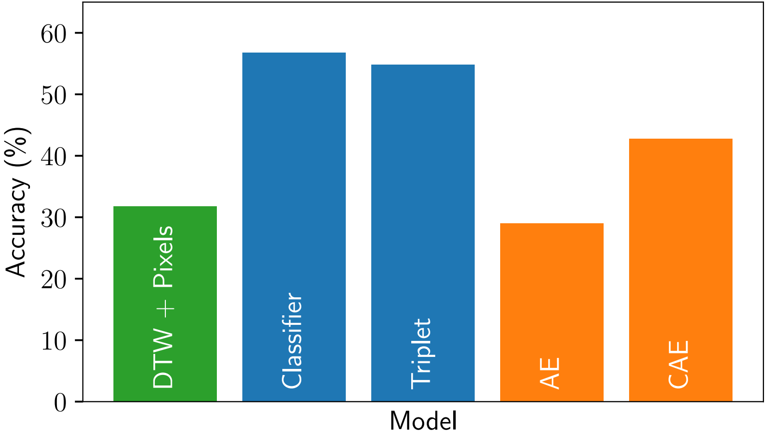
2. Unsupervised models



Kamper, "Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models," *ICASSP*, 2019.

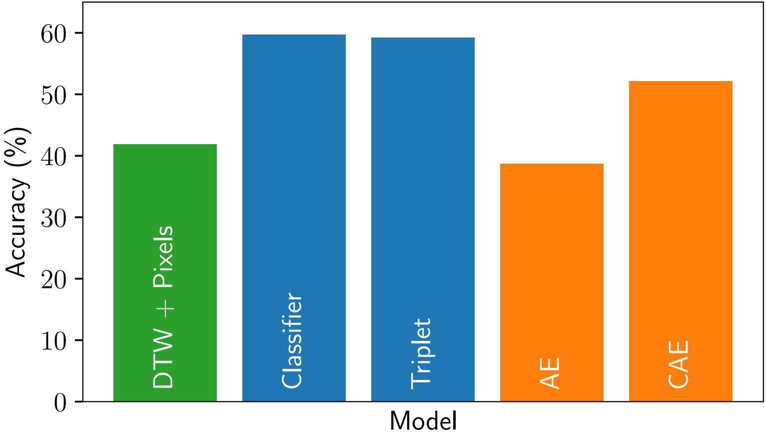
Evaluation: Multimodal one-shot matching

Evaluation: Multimodal one-shot matching



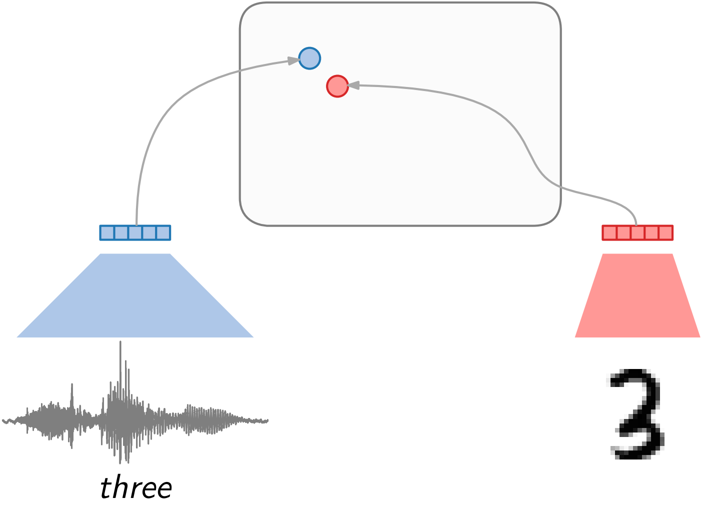
Nortje and Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images," *Interspeech*, 2020.

Evaluation: Multimodal five-shot matching

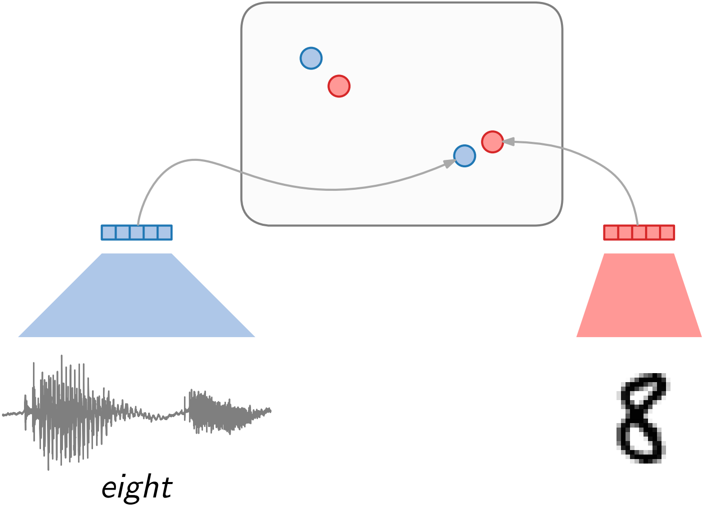


Nortje and Kamper, "Unsupervised vs. transfer learning for multimodal one-shot matching of speech and images," *Interspeech*, 2020.

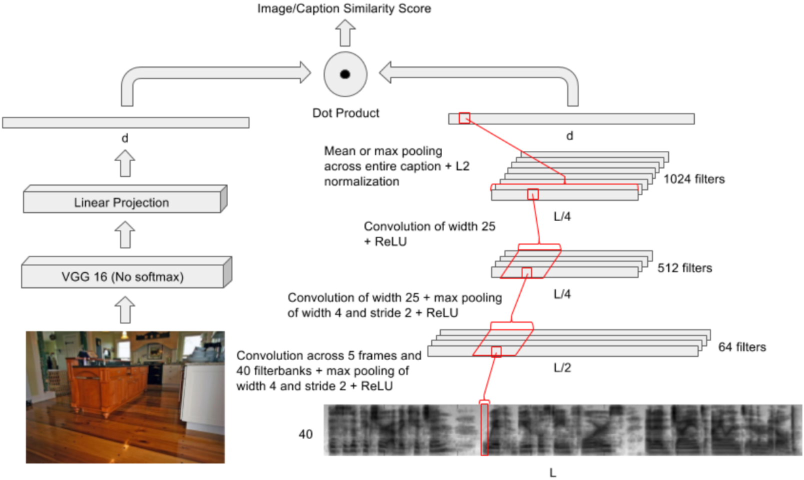
3. A direct approach?



3. A direct approach?

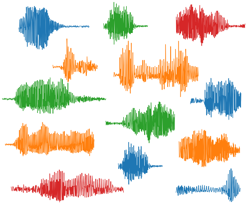


3. A direct approach?

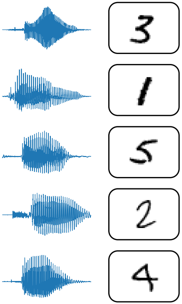


Harwath et al., "Unsupervised learning of spoken language with visual context," *NeurIPS*, 2016.

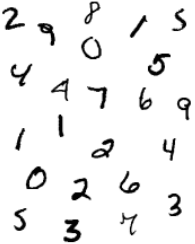
3. Pair mining for a direct model



Unlabelled speech

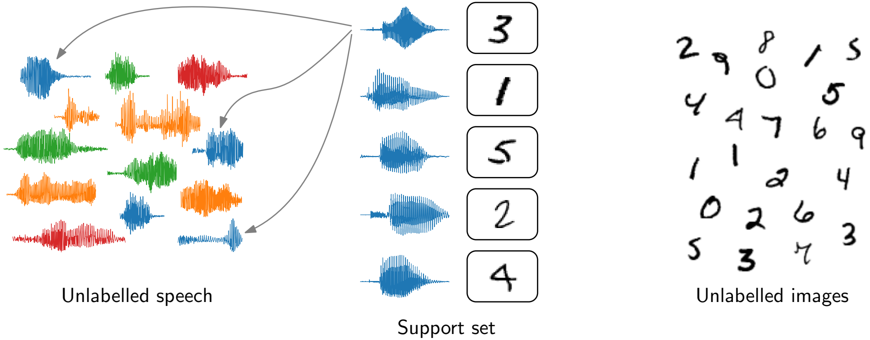


Support set

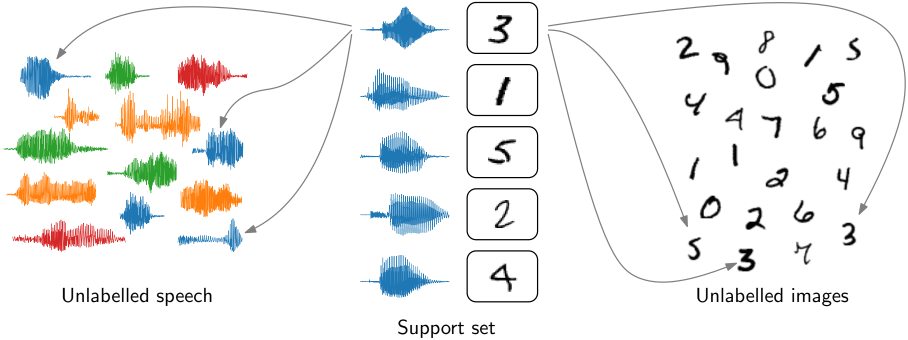


Unlabelled images

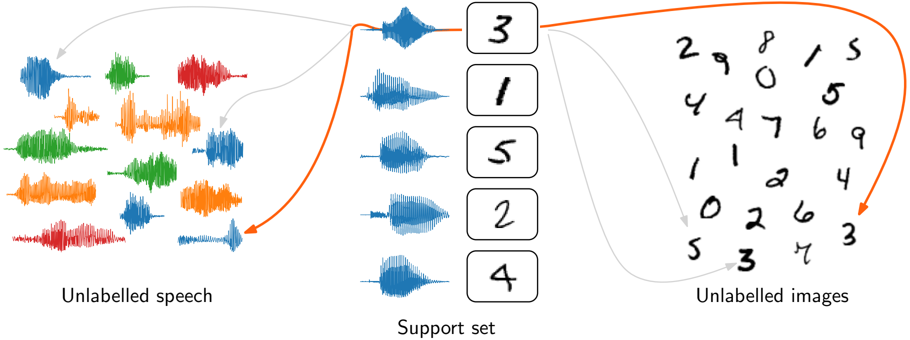
3. Pair mining for a direct model



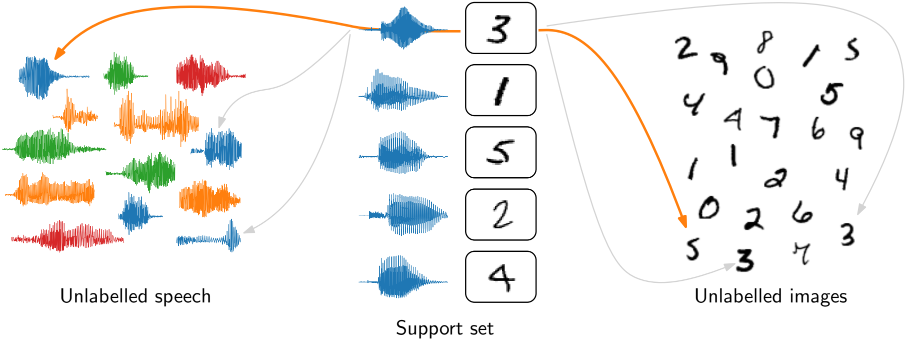
3. Pair mining for a direct model



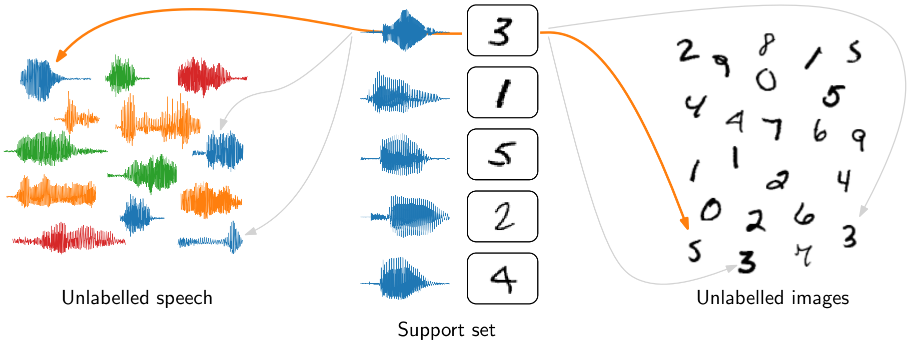
3. Pair mining for a direct model



3. Pair mining for a direct model

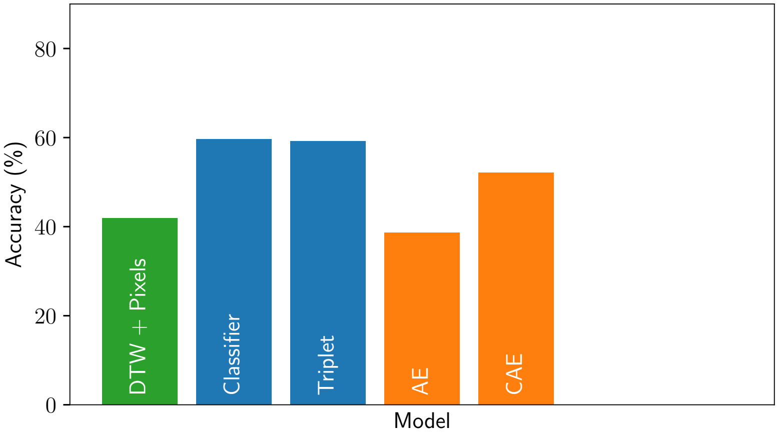


3. Pair mining for a direct model

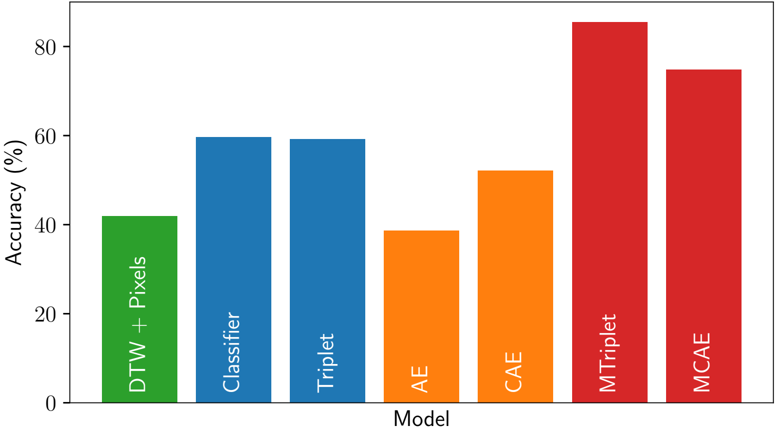


Involves combining (1) transfer learning and (2) unsupervised learning

Evaluation: Multimodal five-shot matching



Evaluation: Multimodal five-shot matching



Summary and conclusion

Summary and looking forward

1. What can we learn from unlabelled speech audio, i.e. radio? — **Part 1**

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television? — **Part 2**

3. What can we learn from interaction/feedback from our environment and other "agents"?

Summary and looking forward

0. What structures/knowledge should we start with/build in?

1. What can we learn from unlabelled speech audio, i.e. radio?
— **Part 1**

2. What can we learn from co-occurring (grounding) signals like vision, i.e. television? — **Part 2**

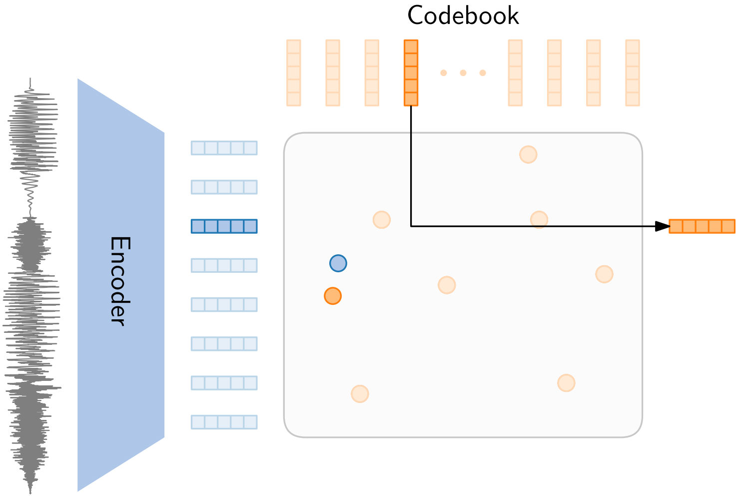
3. What can we learn from interaction/feedback from our environment and other "agents"?

https://github.com/bshall/ZeroSpeech/

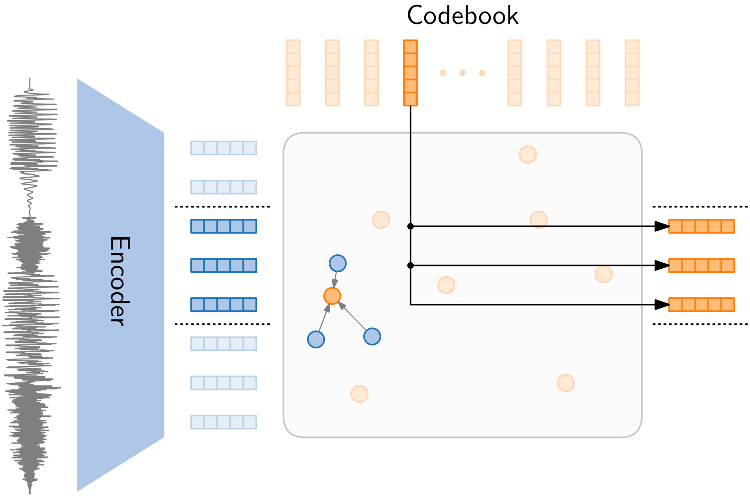
https://github.com/bshall/VectorQuantizedCPC/

https://github.com/LeanneNortje/multimodal_speech-image_matching/

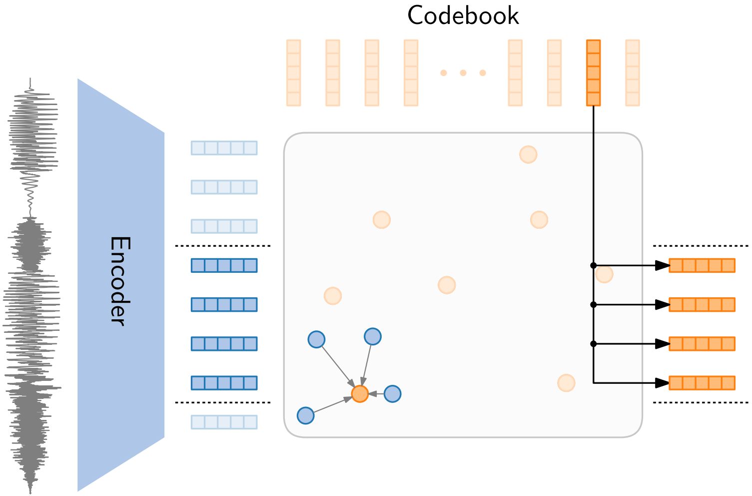
Segmentation on top of vector quantisation



Segmentation on top of vector quantisation



Segmentation on top of vector quantisation



Segmentation on top of vector quantisation