Unsupervised neural feature learning for speech using weak top-down constraints

Maties Machine Learning (MML), Oct. 2017

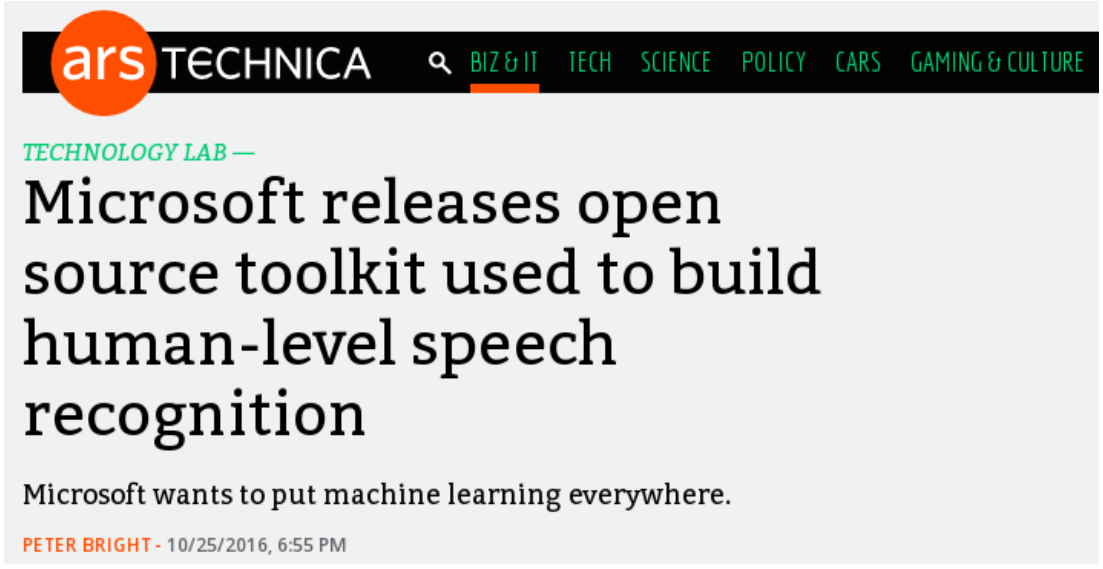
Herman Kamper

Stellenbosch University

http://www.kamperh.com/

Success in speech recognition

Success in speech recognition

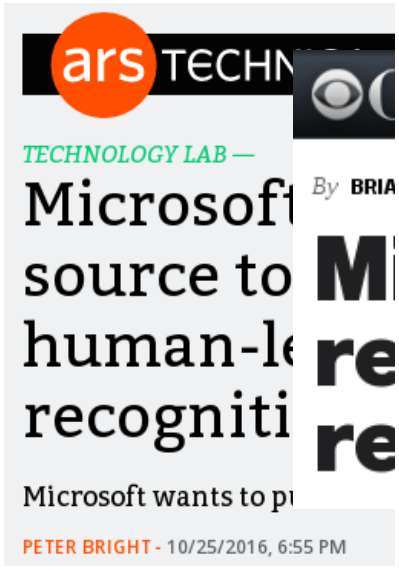


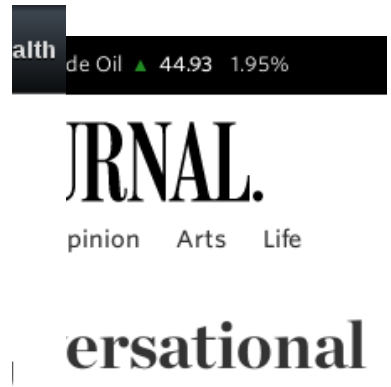
Success in speech recognition

Success in speech recognition

[Xiong et al., arXiv'16]; [Saon et al., arXiv'17]

Success in speech recognition





[Saon et al., arXiv'17]

Success in speech recognition

[Xiong et al., arXiv'16]; [Saon et al., arXiv'17]

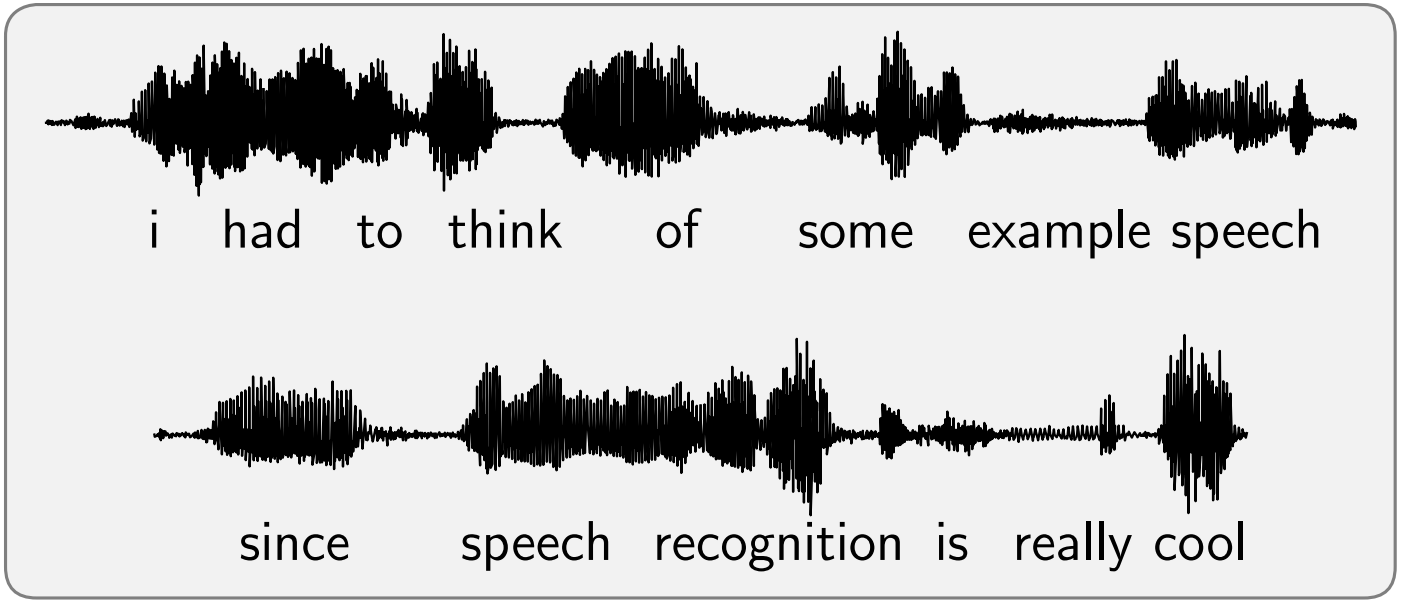
- Google Voice: English, Spanish, German, . . . , Zulu (∼50 languages)

Success in speech recognition

[Xiong et al., arXiv'16]; [Saon et al., arXiv'17]

- Google Voice: English, Spanish, German, . . . , Zulu (∼50 languages)
- An **addiction to labels**: 2000 hours transcribed speech audio; ∼350M/560M words text

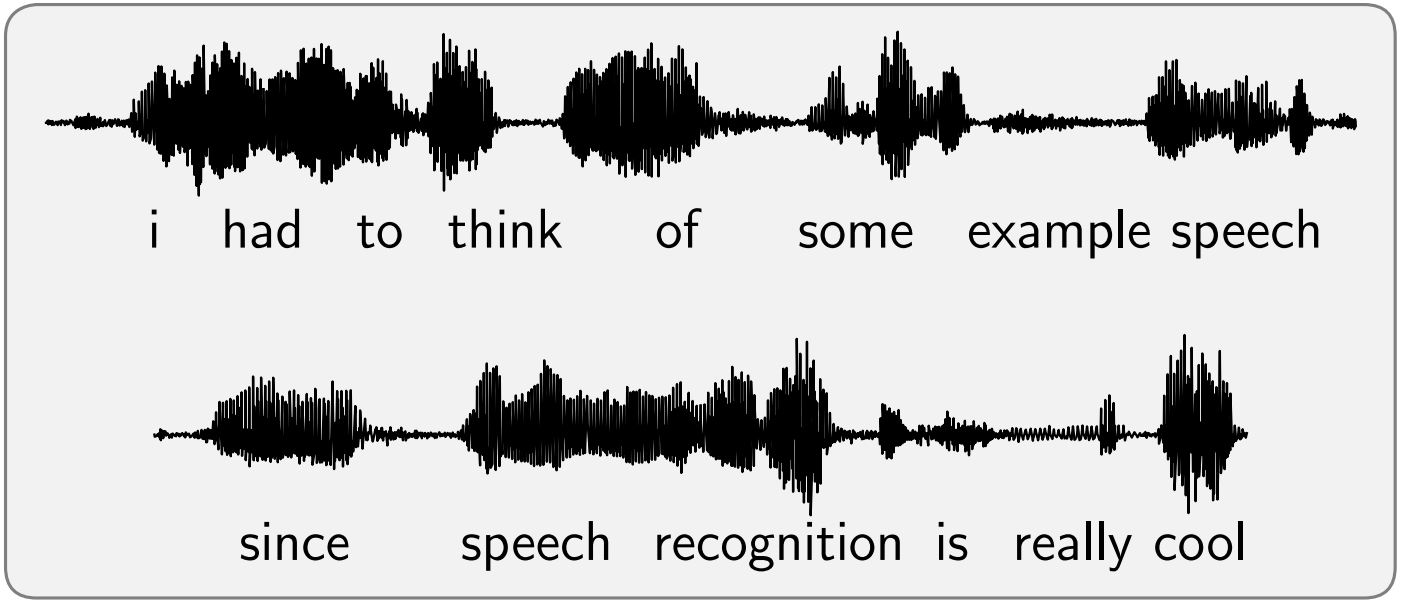
Success in speech recognition



[Xiong et al., arXiv'16]; [Saon et al., arXiv'17]

- Google Voice: English, Spanish, German, . . . , Zulu ($\sim$50 languages)
- An **addiction to labels**: 2000 hours transcribed speech audio; $\sim$350M/560M words text

Success in speech recognition



[Xiong et al., arXiv'16]; [Saon et al., arXiv'17]

- Google Voice: English, Spanish, German, . . . , Zulu ($\sim$50 languages)
- An **addiction to labels**: 2000 hours transcribed speech audio; $\sim$350M/560M words text
- But, there are around 7000 languages spoken in the world today

Why learn without labels?

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]

Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]

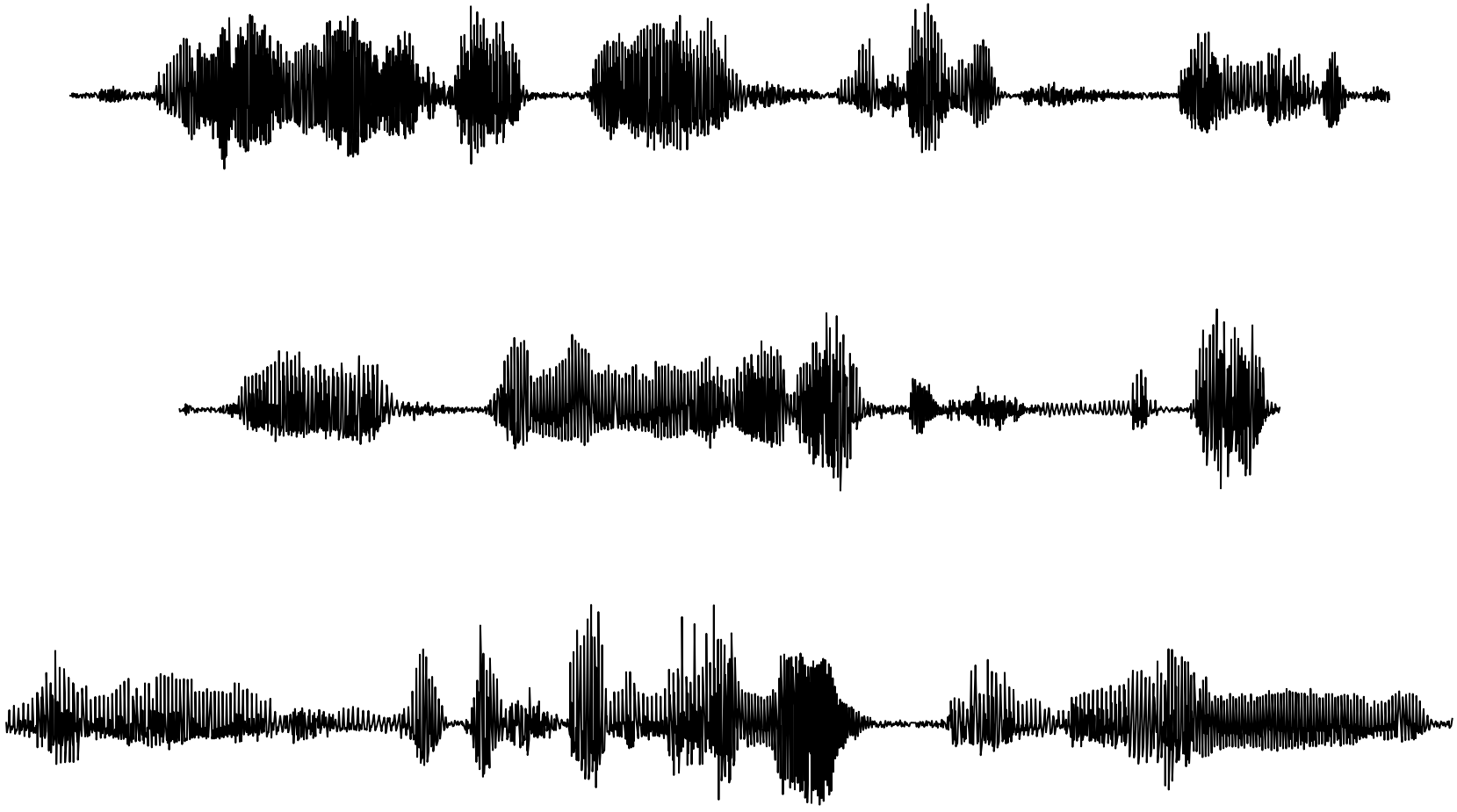
Why learn without labels?

- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]

Why learn without labels?

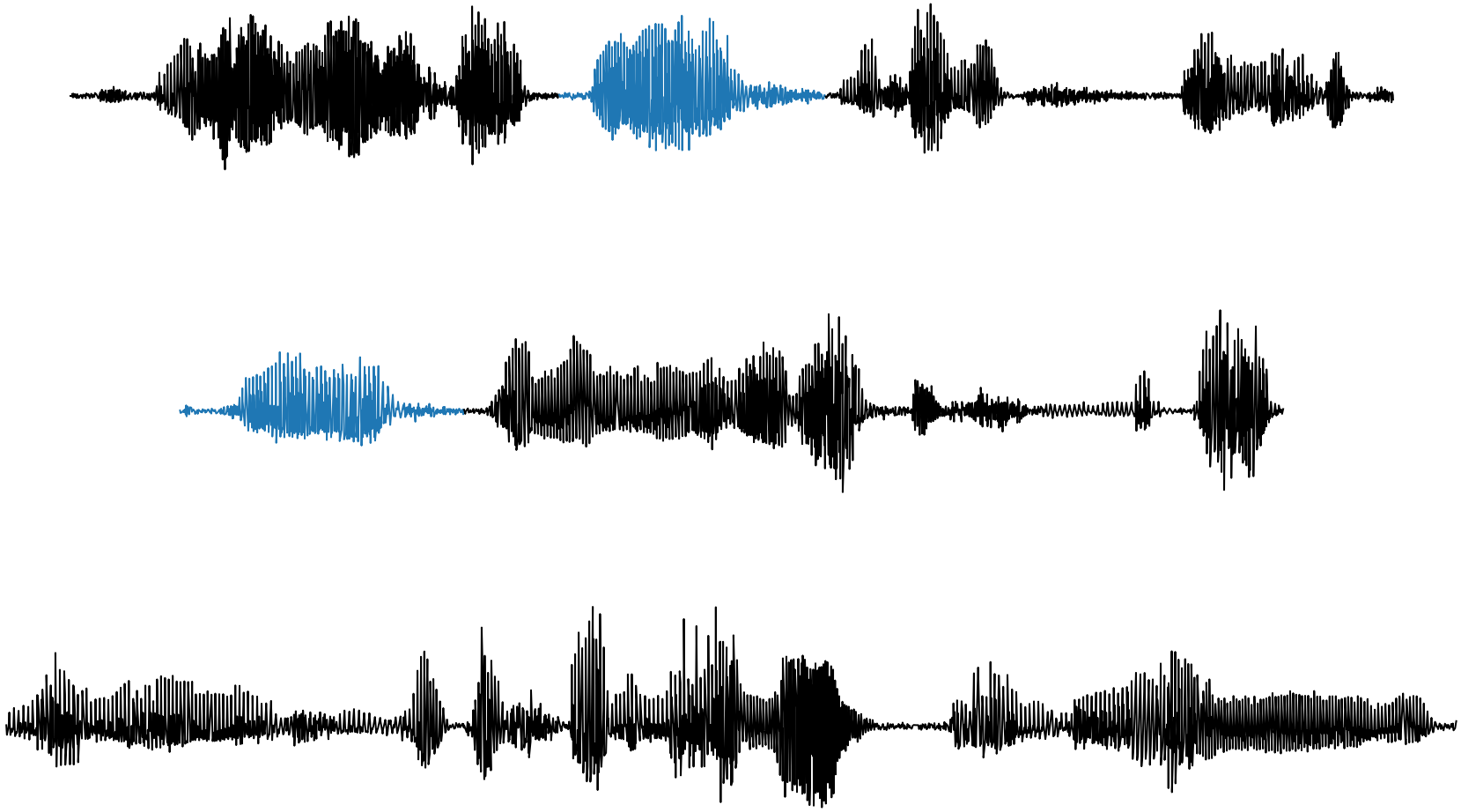
- Get insight into human **language acquisition** [Räsänen and Rasilo, '15]
- Language acquisition in **robots** [Roy, '99]; [Renkens and Van hamme, '15]
- Analysis of audio for unwritten languages [Besacier et al., '14]
- New **insights** and models for speech processing
  [Jansen et al., '13]

Unsupervised term discovery (UTD)



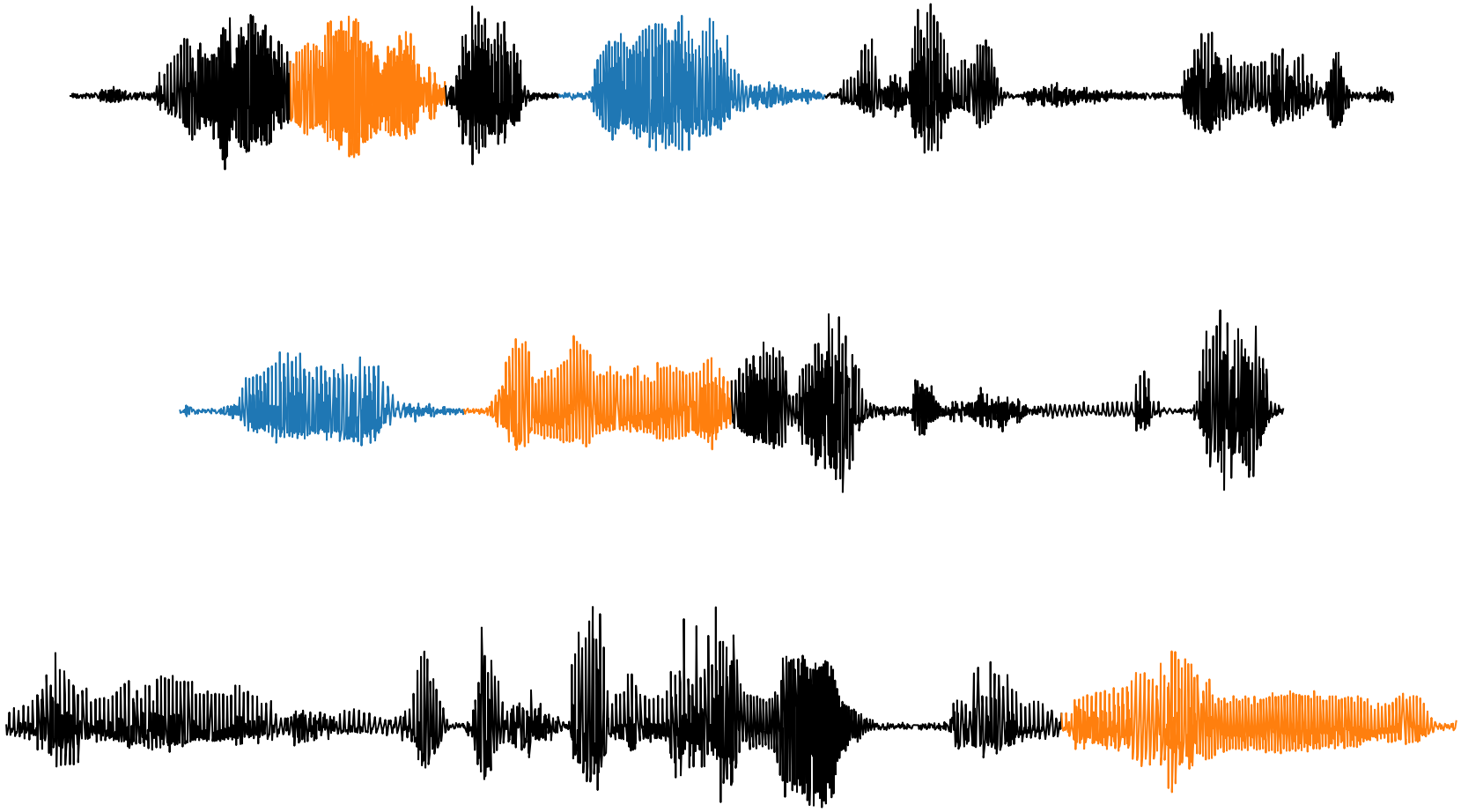
[Park and Glass, TASLP'08]

Unsupervised term discovery (UTD)



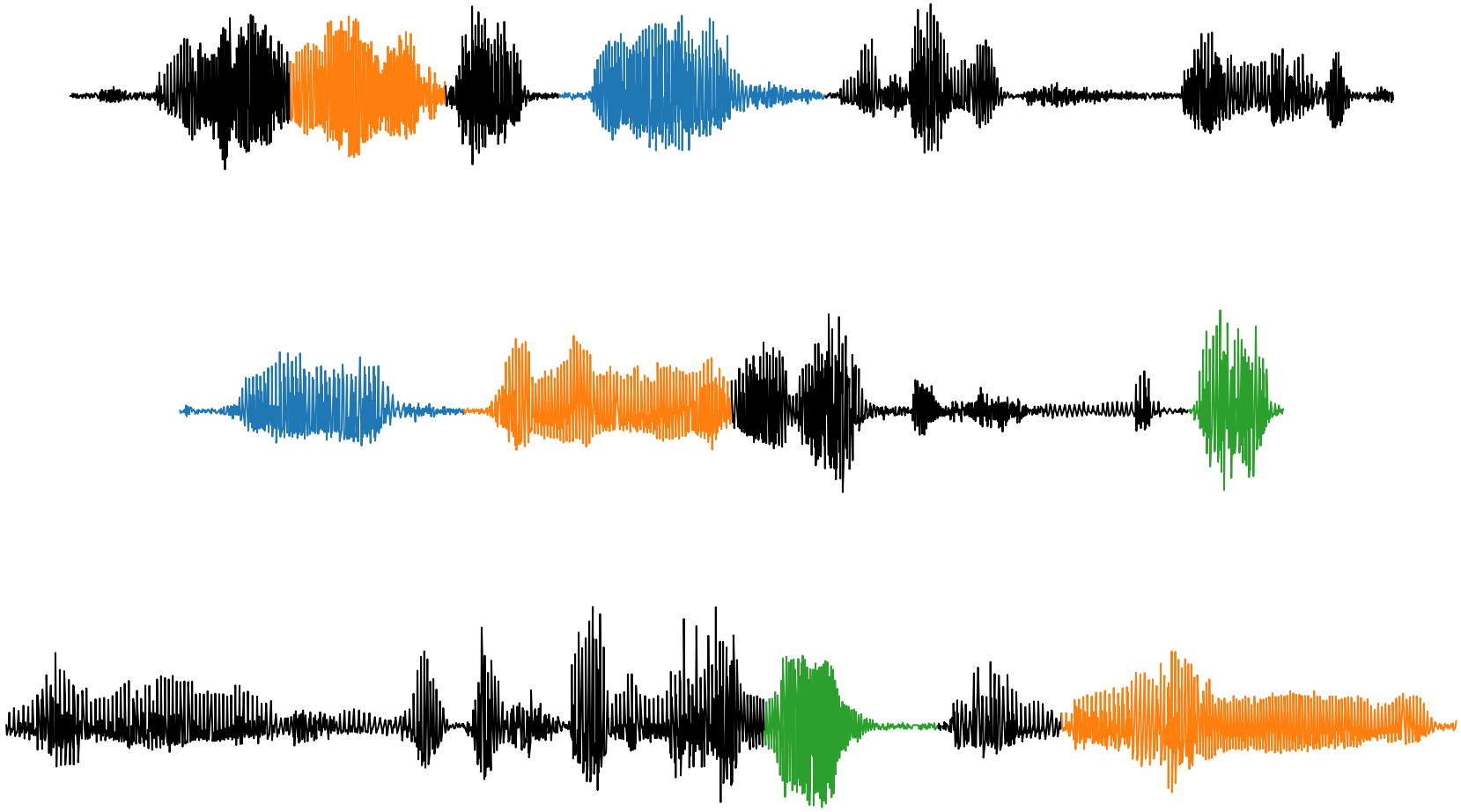
[Park and Glass, TASLP'08]

Unsupervised term discovery (UTD)



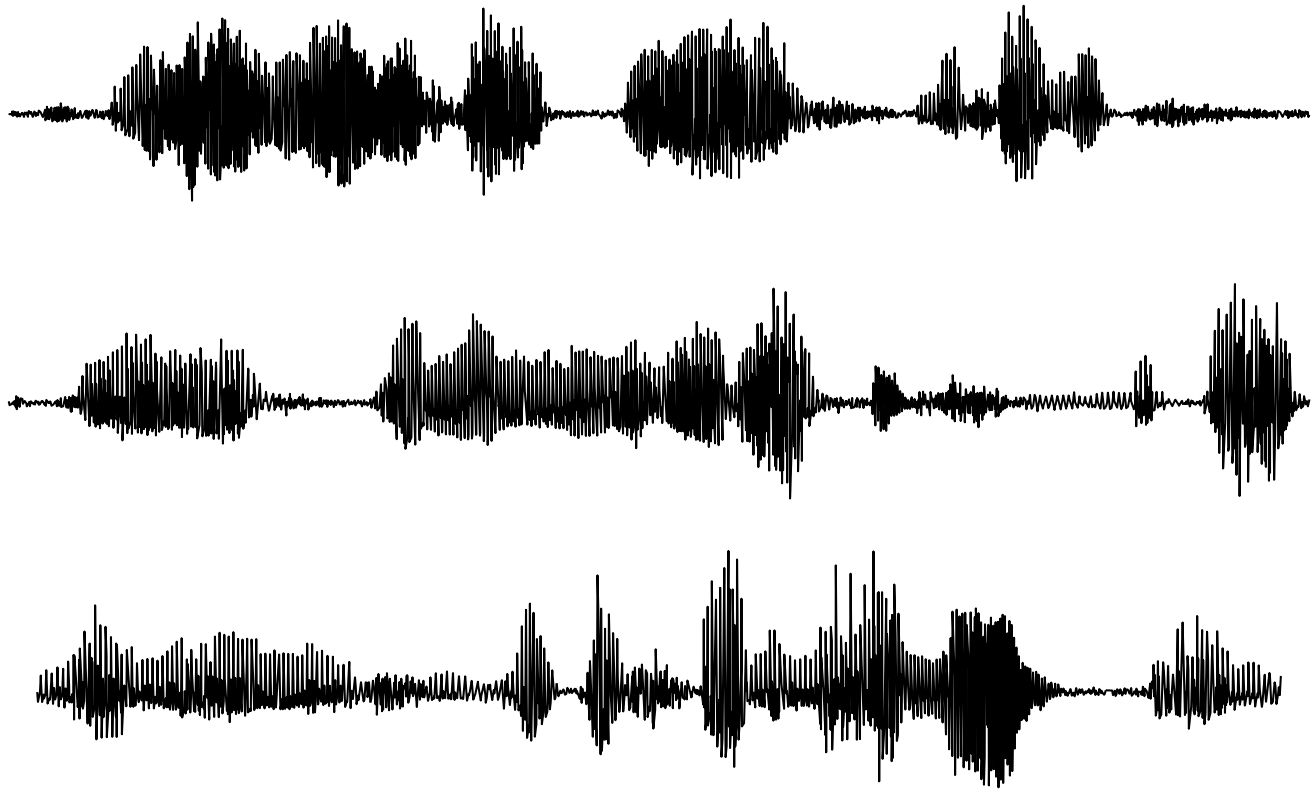
[Park and Glass, TASLP'08]

Unsupervised term discovery (UTD)



[Park and Glass, TASLP'08]

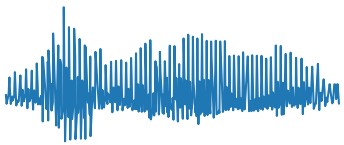
Example: Query-by-example search

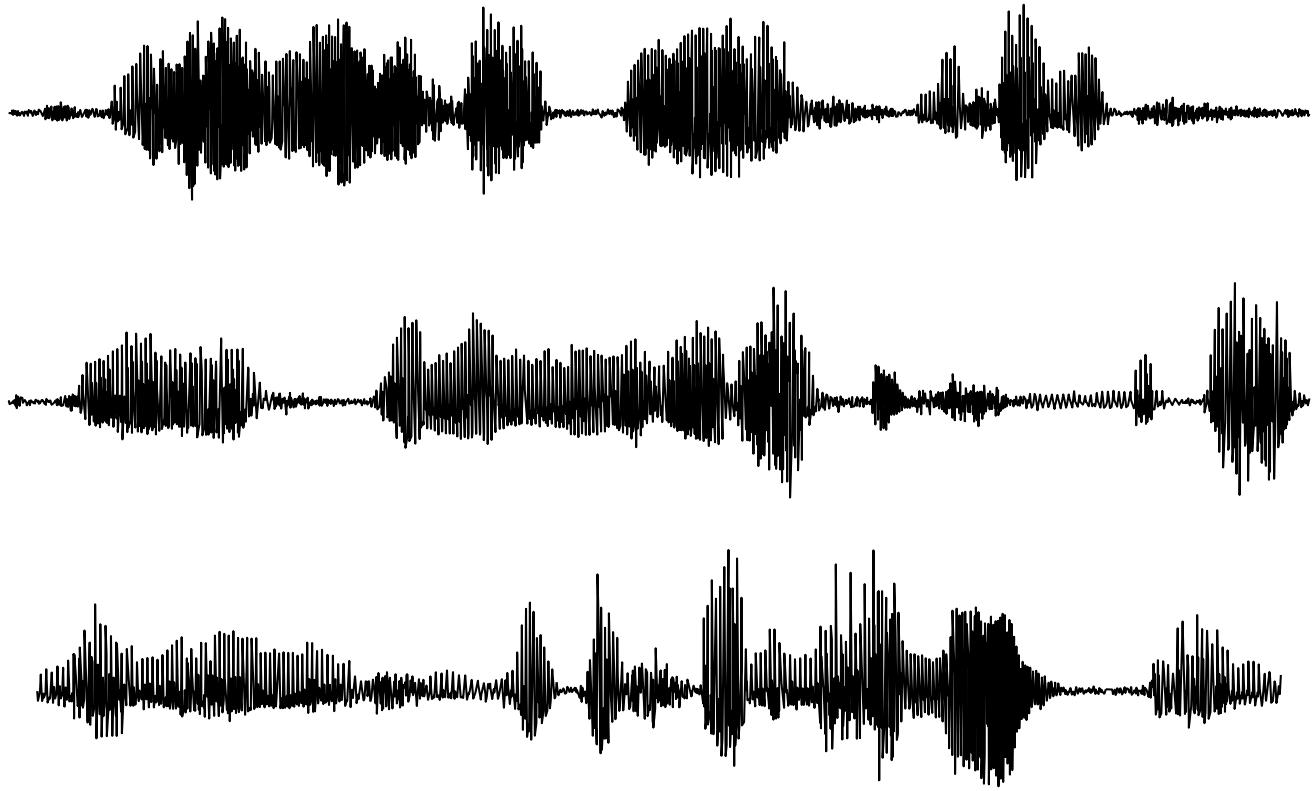


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Example: Query-by-example search

Spoken query:

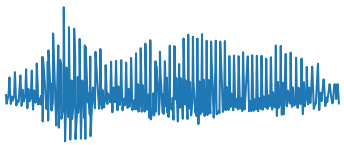


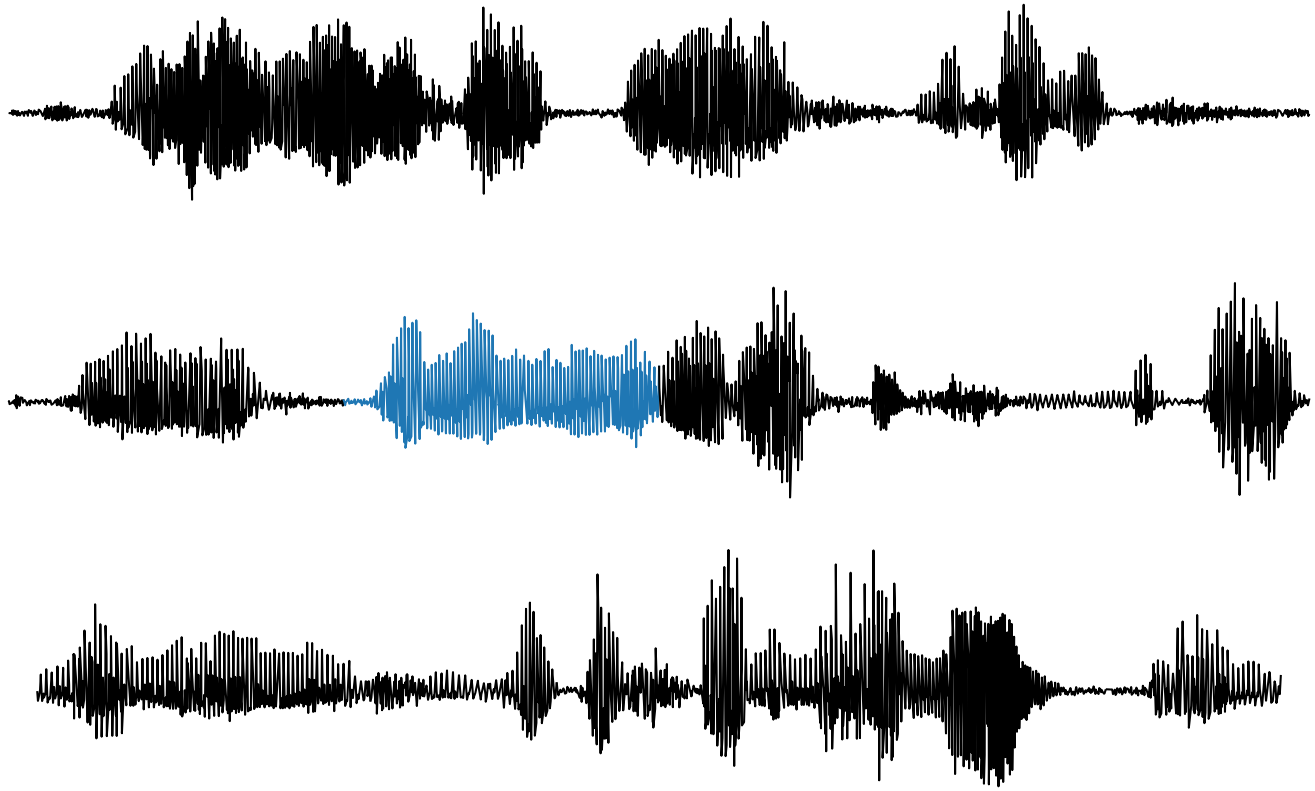


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Example: Query-by-example search

Spoken query:

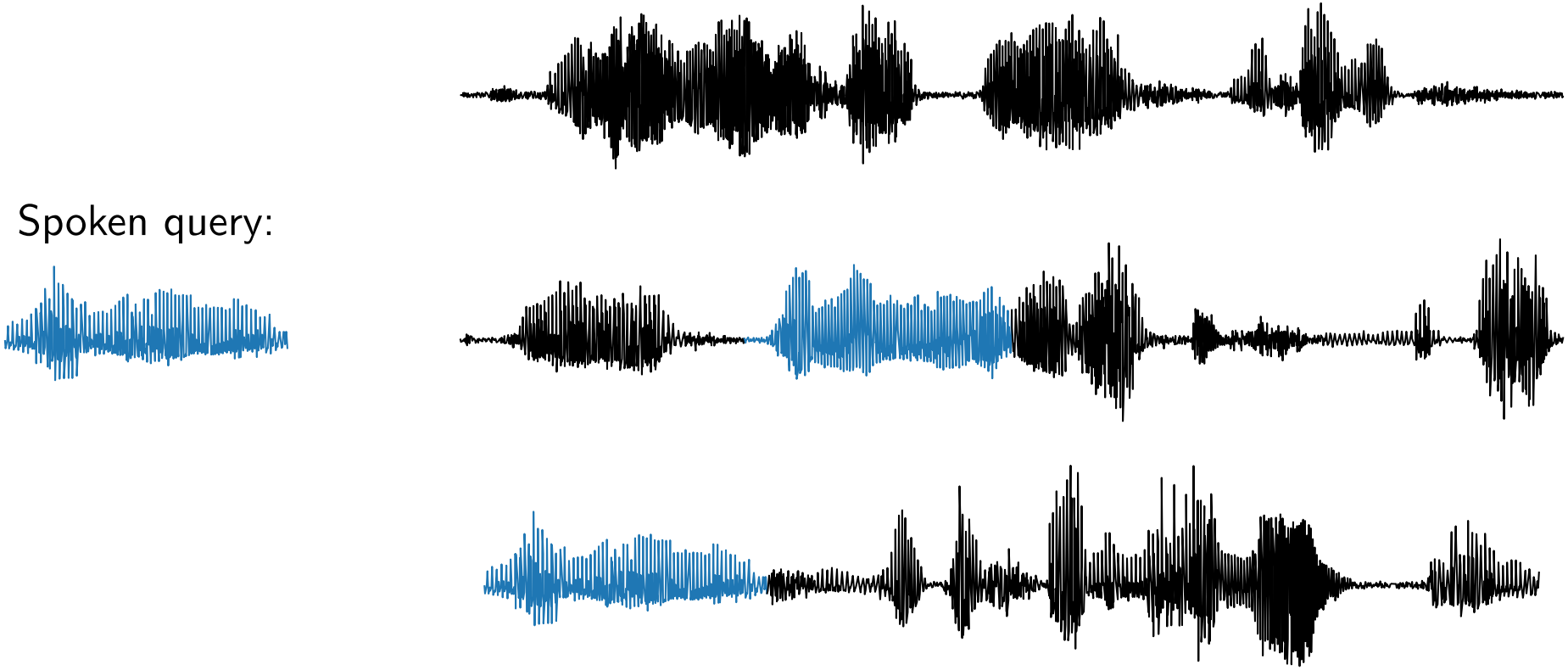




[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Example: Query-by-example search

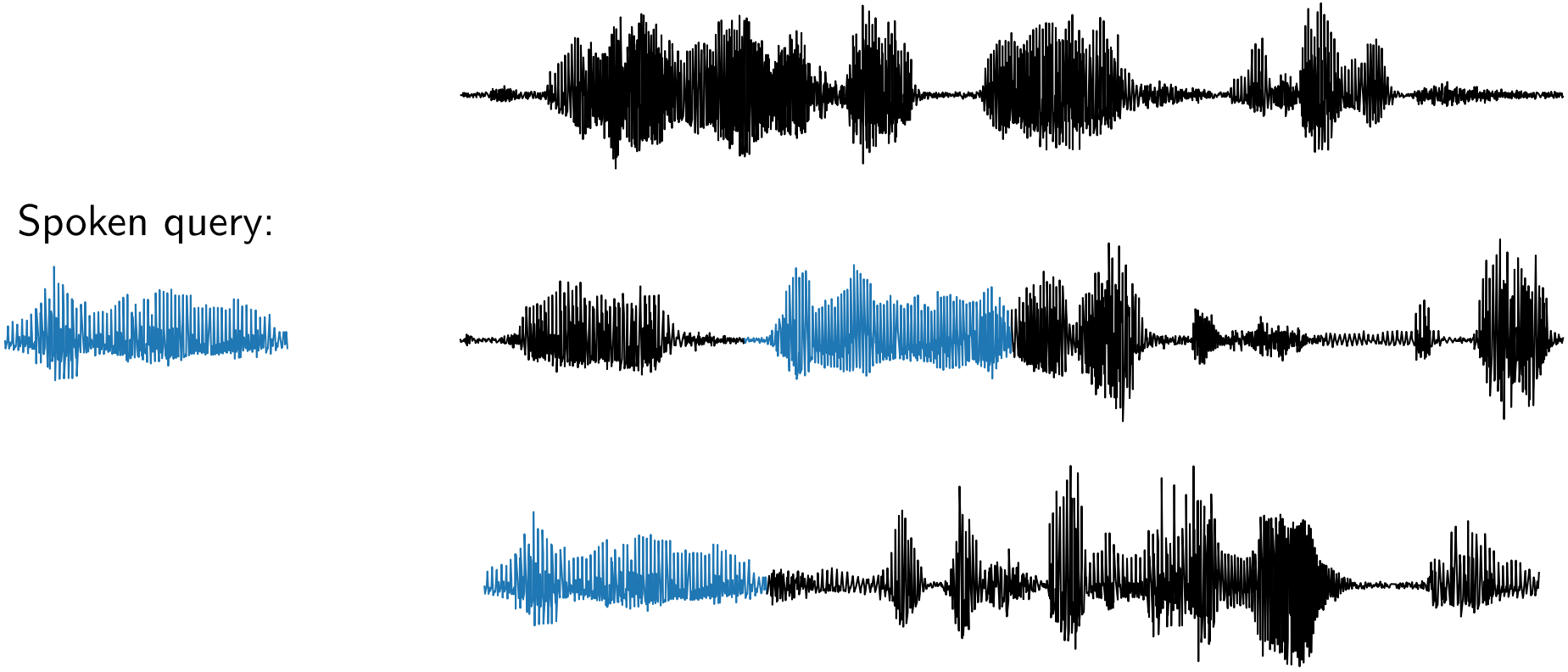
Spoken query:



[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Example: Query-by-example search

Spoken query:



Useful speech system, not requiring any transcribed speech

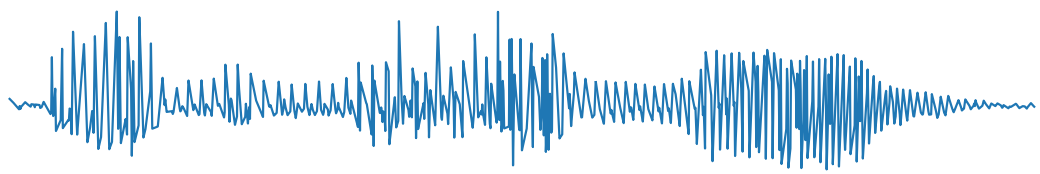
[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Unsupervised speech processing: Two problems

1. Unsupervised frame-level **representation learning**:

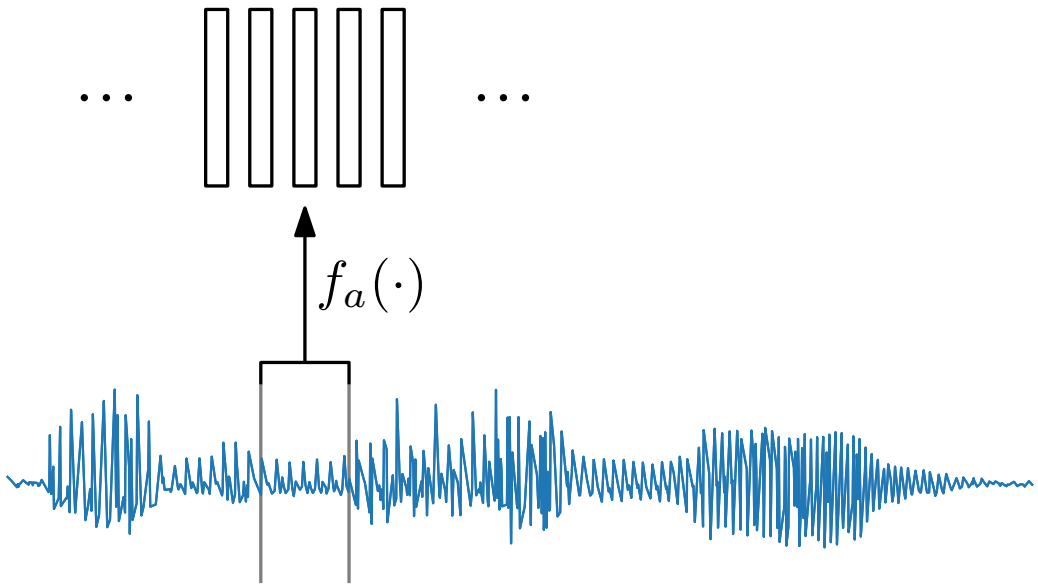
Unsupervised speech processing: Two problems

1. Unsupervised frame-level **representation learning**:



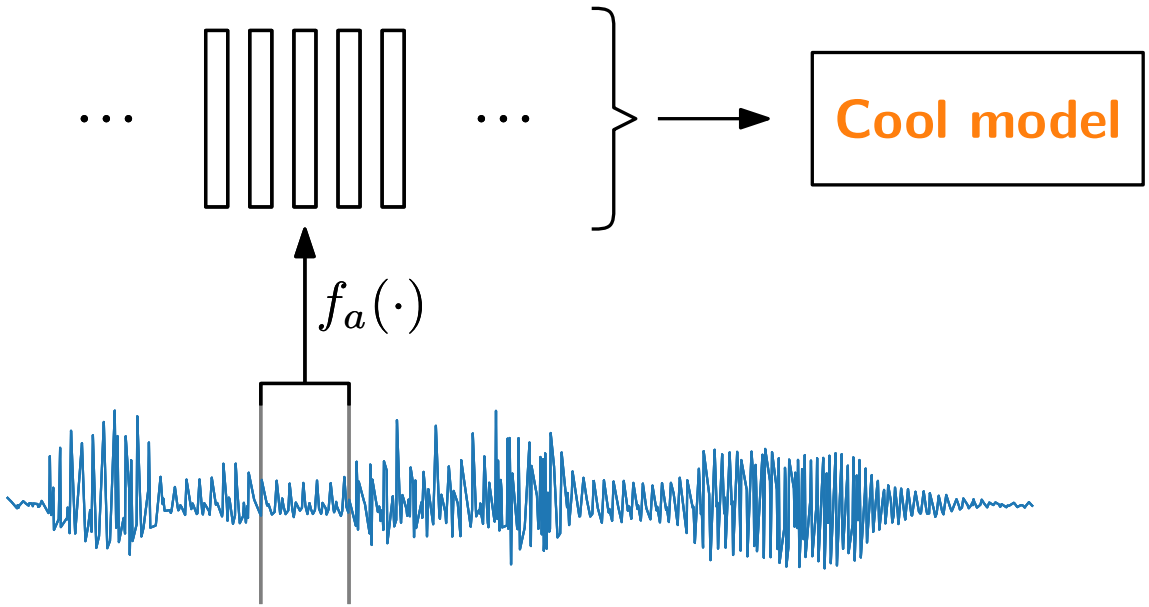
Unsupervised speech processing: Two problems

1. Unsupervised frame-level **representation learning**:



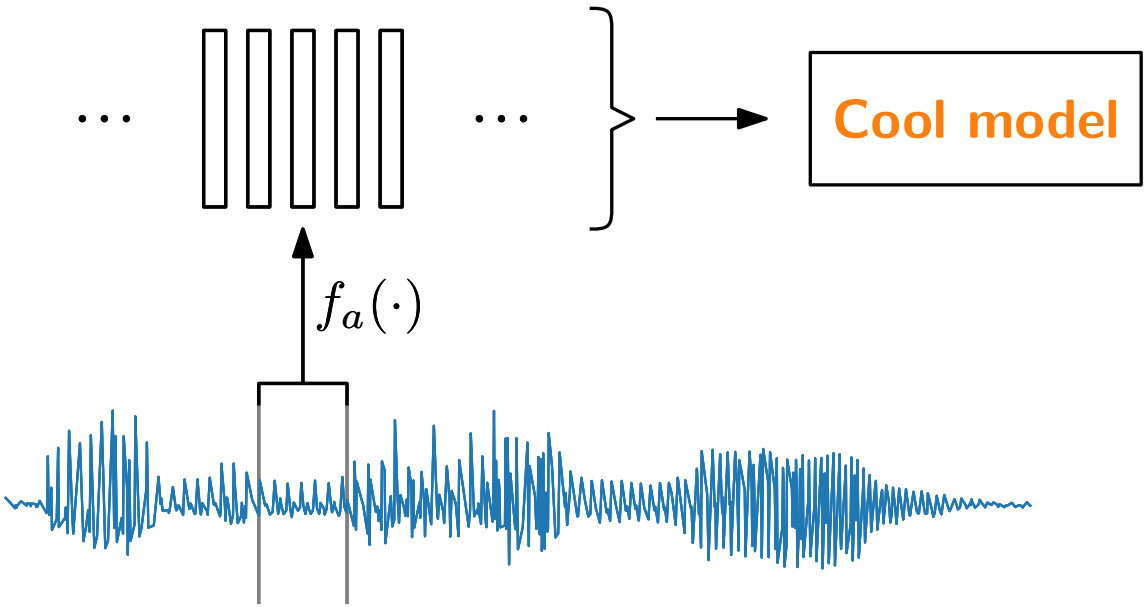
Unsupervised speech processing: Two problems

1. Unsupervised frame-level **representation learning**:



Unsupervised speech processing: Two problems

1. Unsupervised frame-level **representation learning**:



2. Unsupervised **segmentation** and **clustering**:
   How do we discover meaningful units in unlabelled speech?

Unsupervised frame-level representation learning:

The Correspondence Autoencoder

Unsupervised frame-level representation learning:

The Correspondence Autoencoder

Micha Elsner

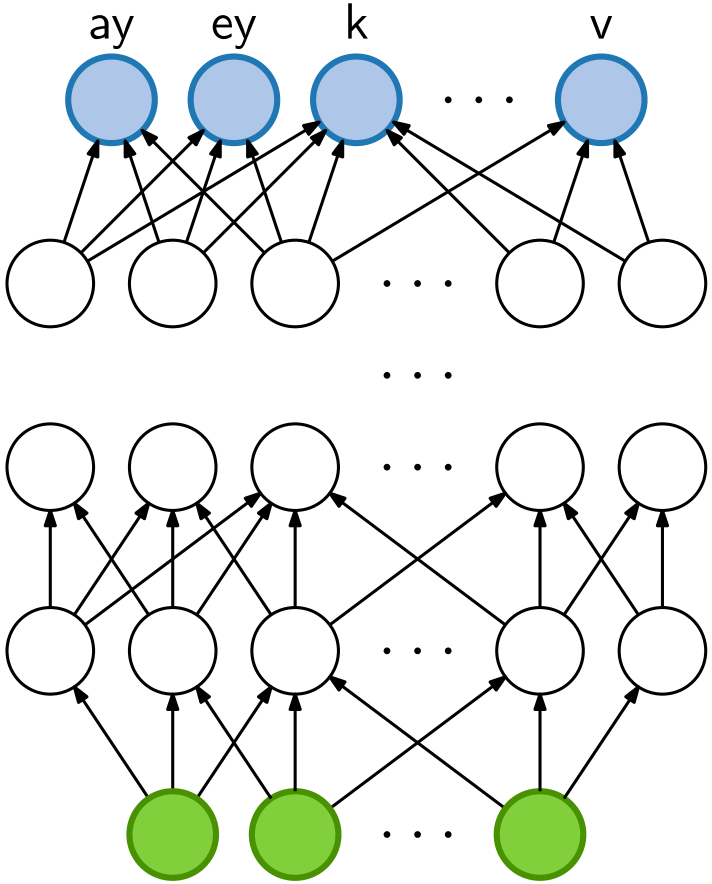
Daniel Renshaw

Aren Jansen

Sharon Goldwater

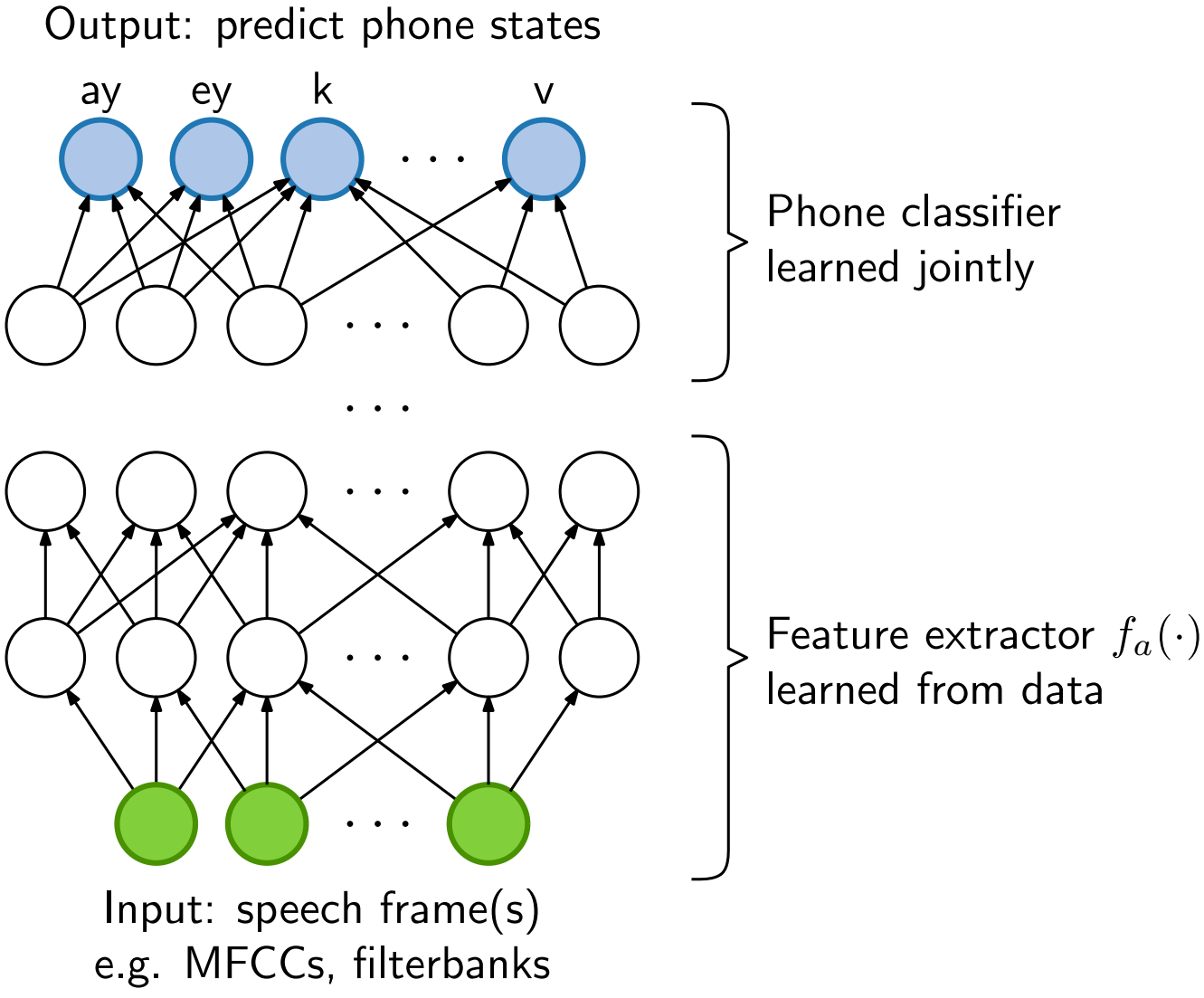
Supervised representation learning using DNNs

Output: predict phone states

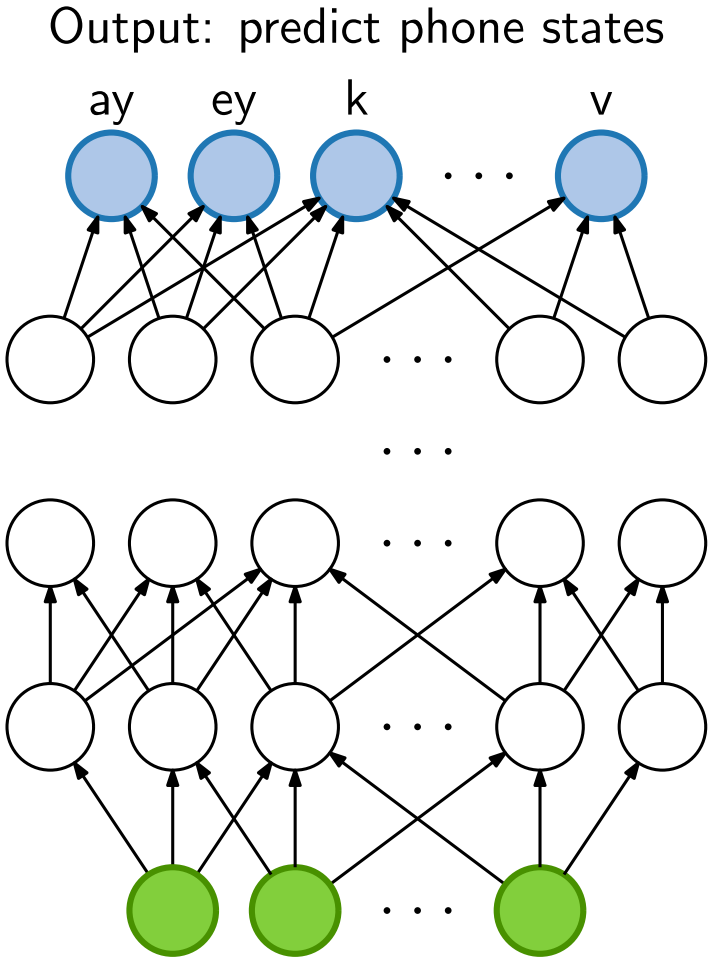


Input: speech frame(s)
e.g. MFCCs, filterbanks

Supervised representation learning using DNNs



Supervised representation learning using DNNs

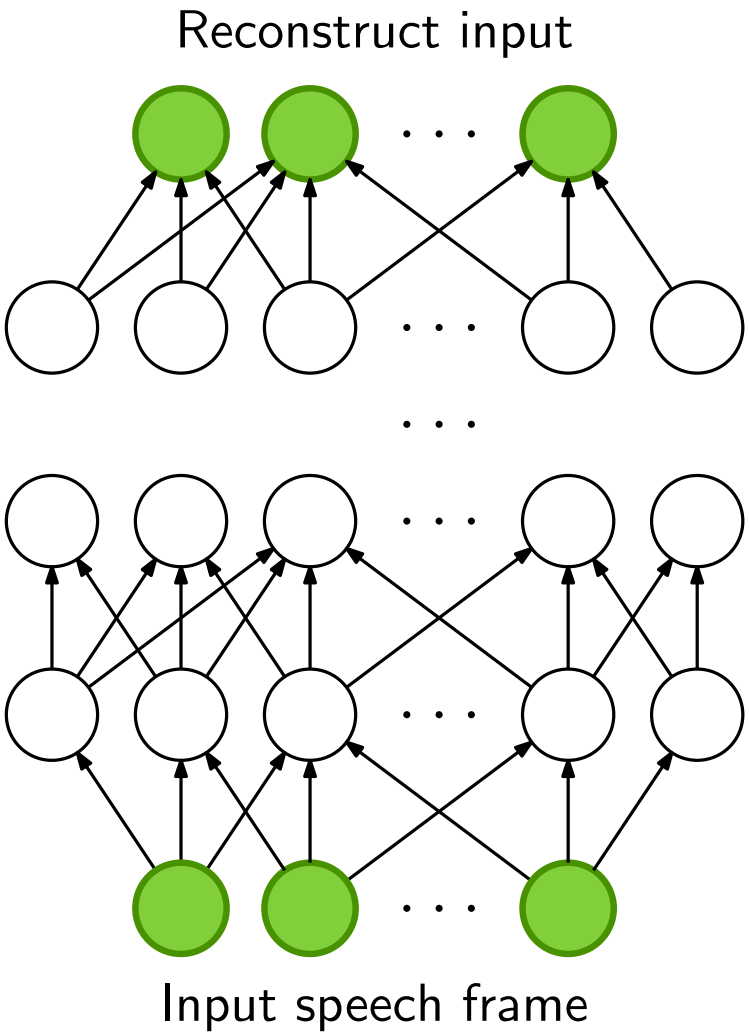


Phone classifier
learned jointly

Feature extractor $f_a(\cdot)$
learned from data

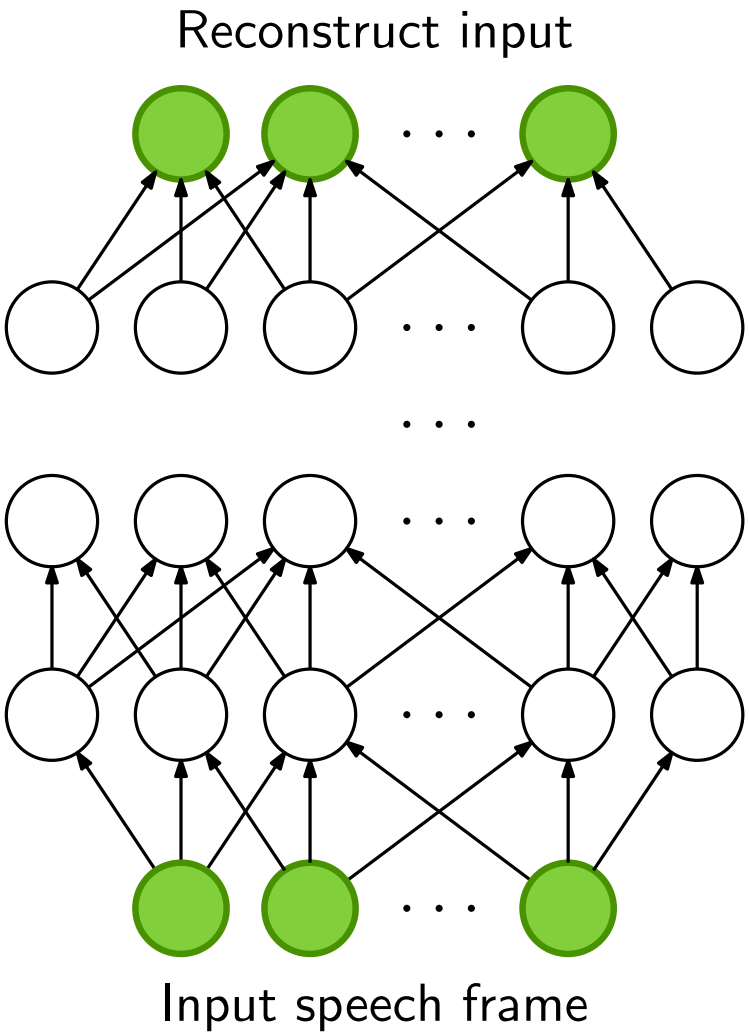
Unsupervised modelling:
No phone class targets to
train network on

Autoencoder (AE) neural network



[Badino et al., ICASSP'14]

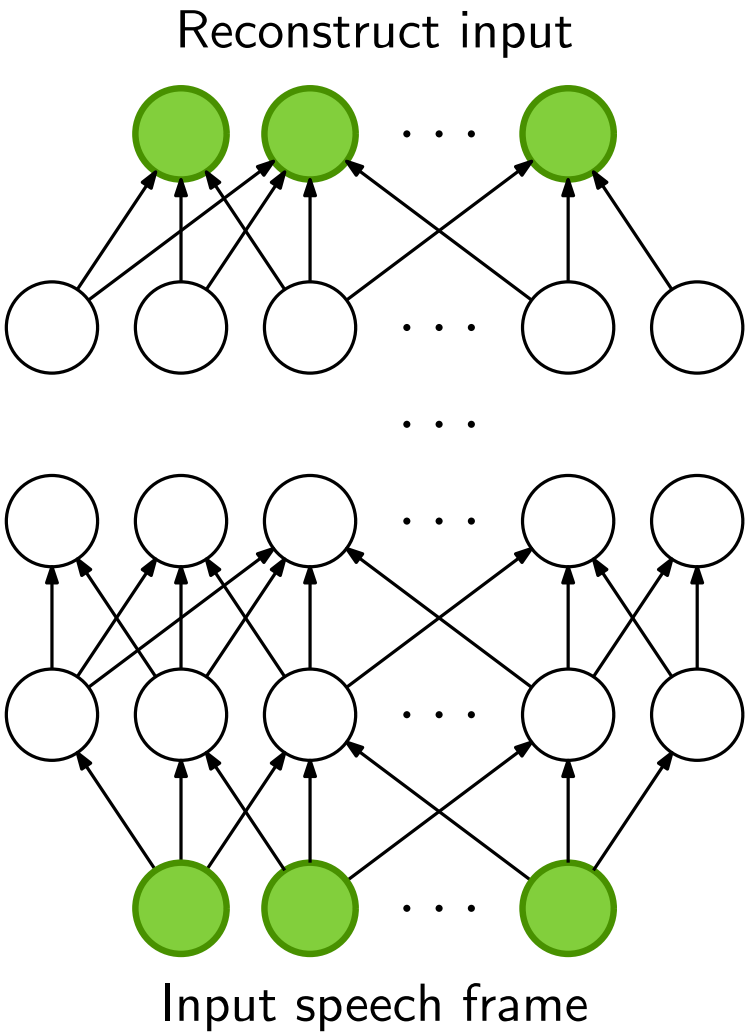
Autoencoder (AE) neural network



- Completely unsupervised
- But purely bottom-up
- Can we use top-down information?

[Badino et al., ICASSP'14]

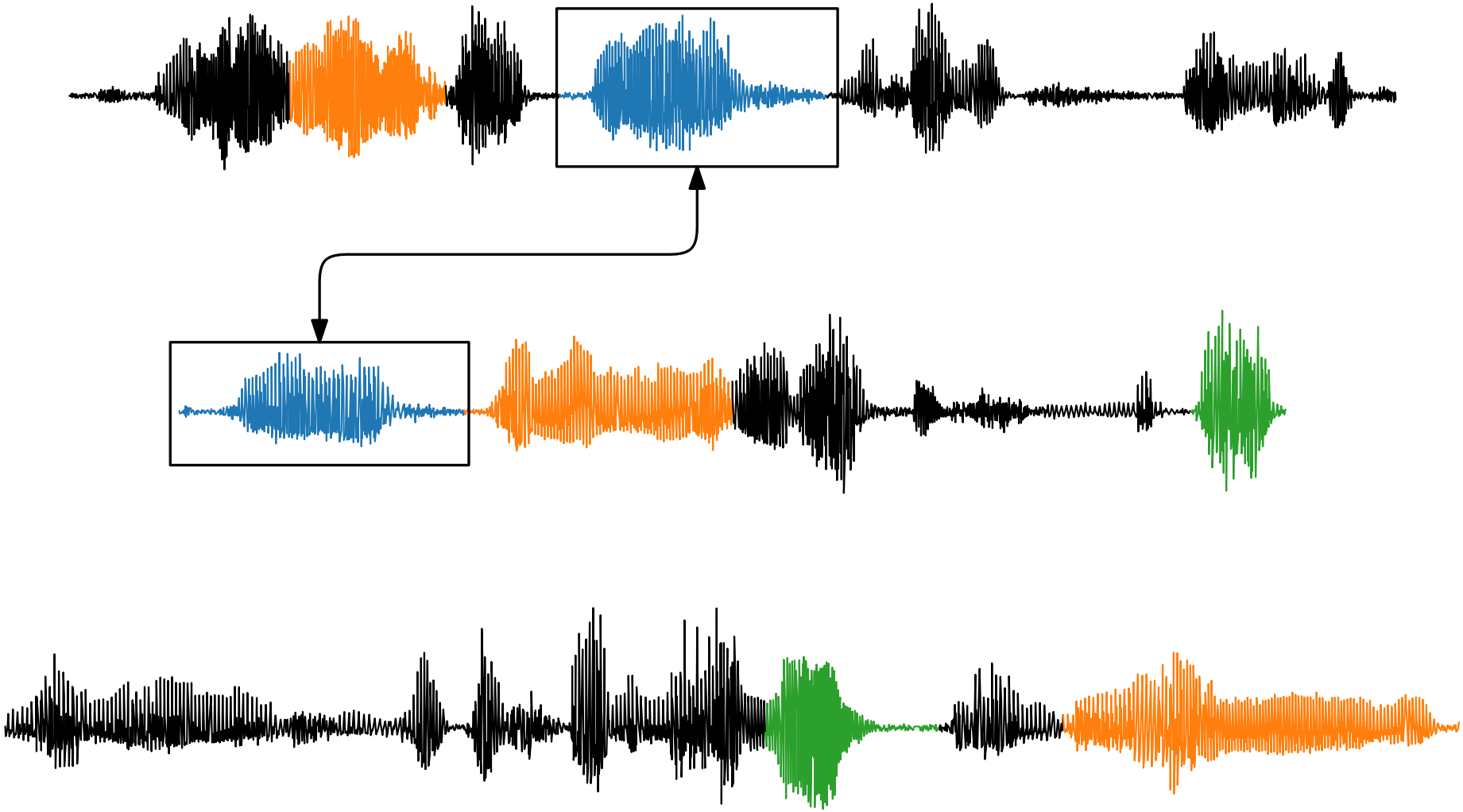
Autoencoder (AE) neural network



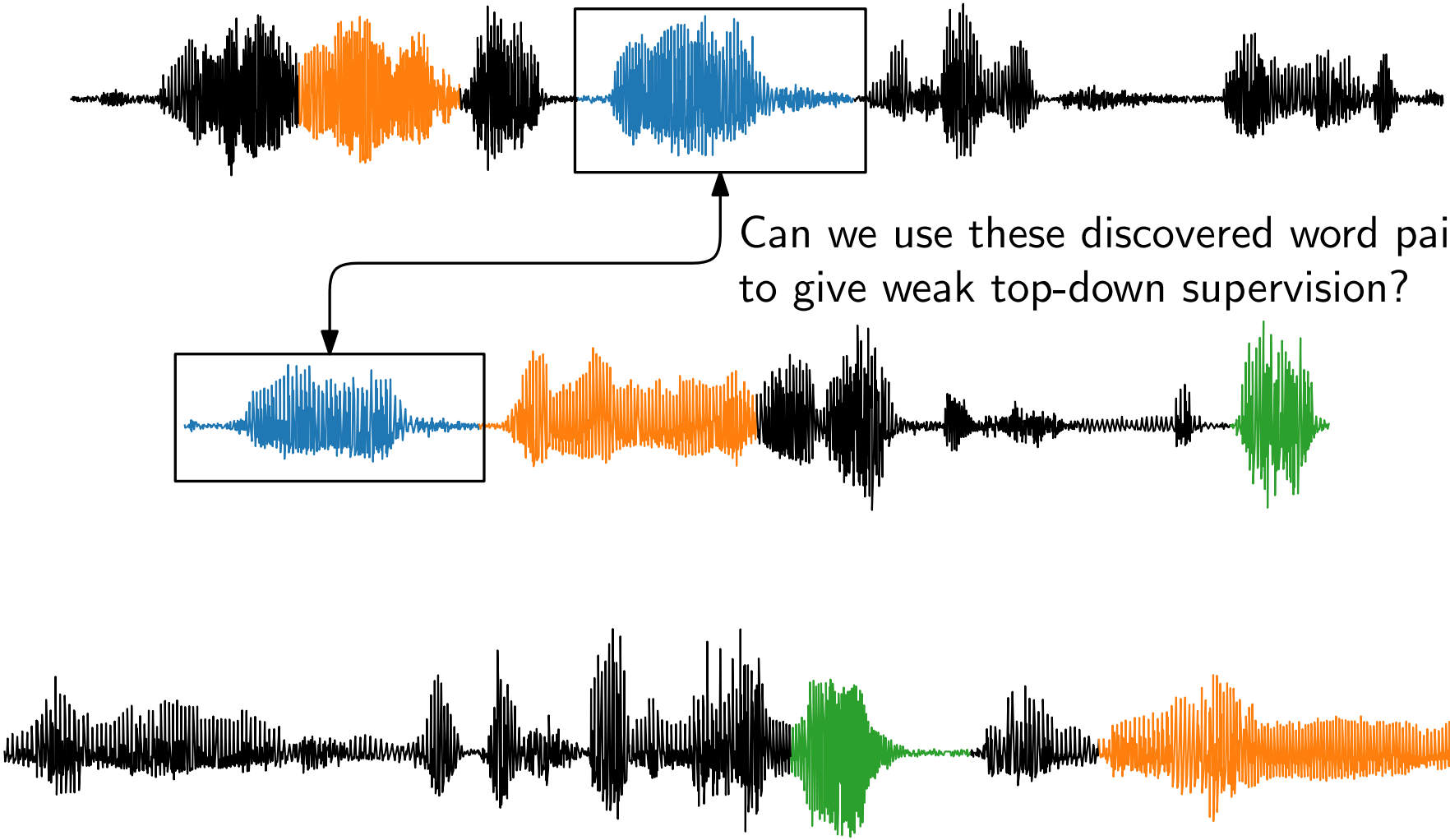
- Completely unsupervised
- But purely bottom-up
- Can we use top-down information?
- **Idea:** Unsupervised term discovery

[Badino et al., ICASSP'14]

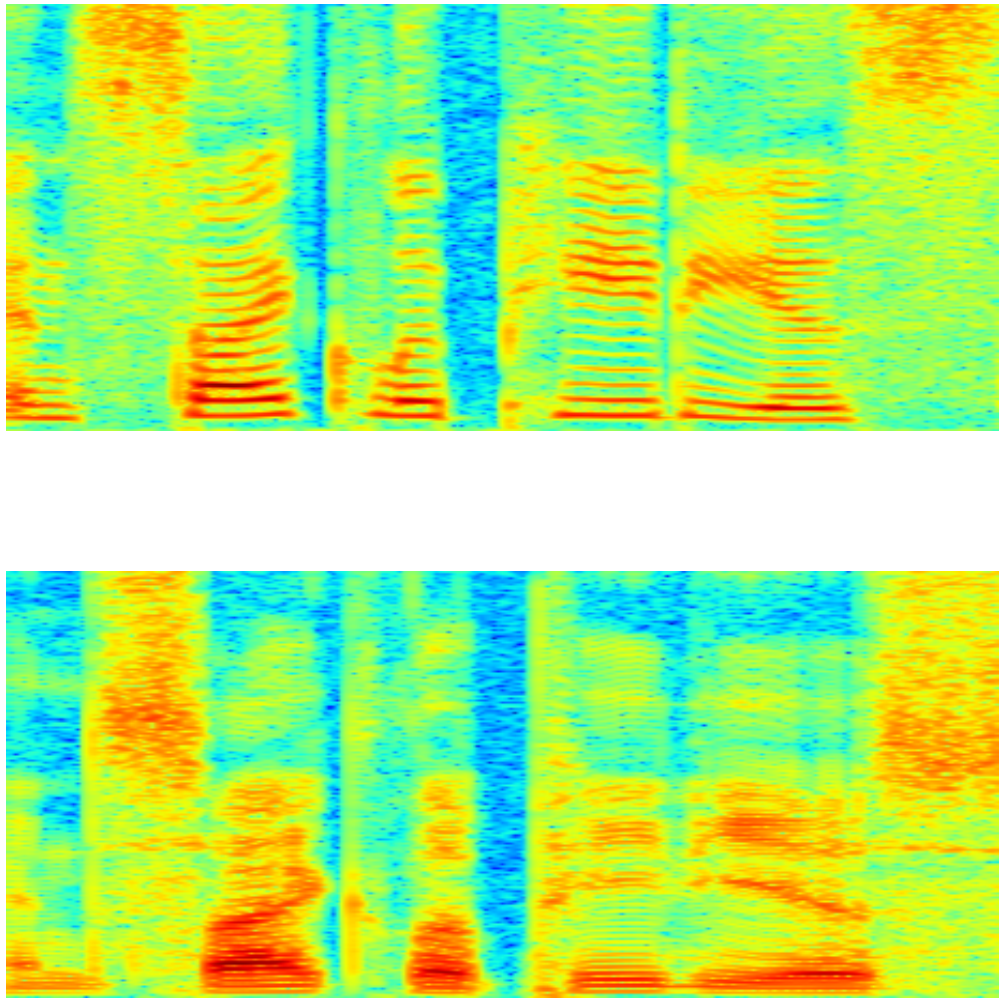
Unsupervised term discovery (UTD)



Unsupervised term discovery (UTD)

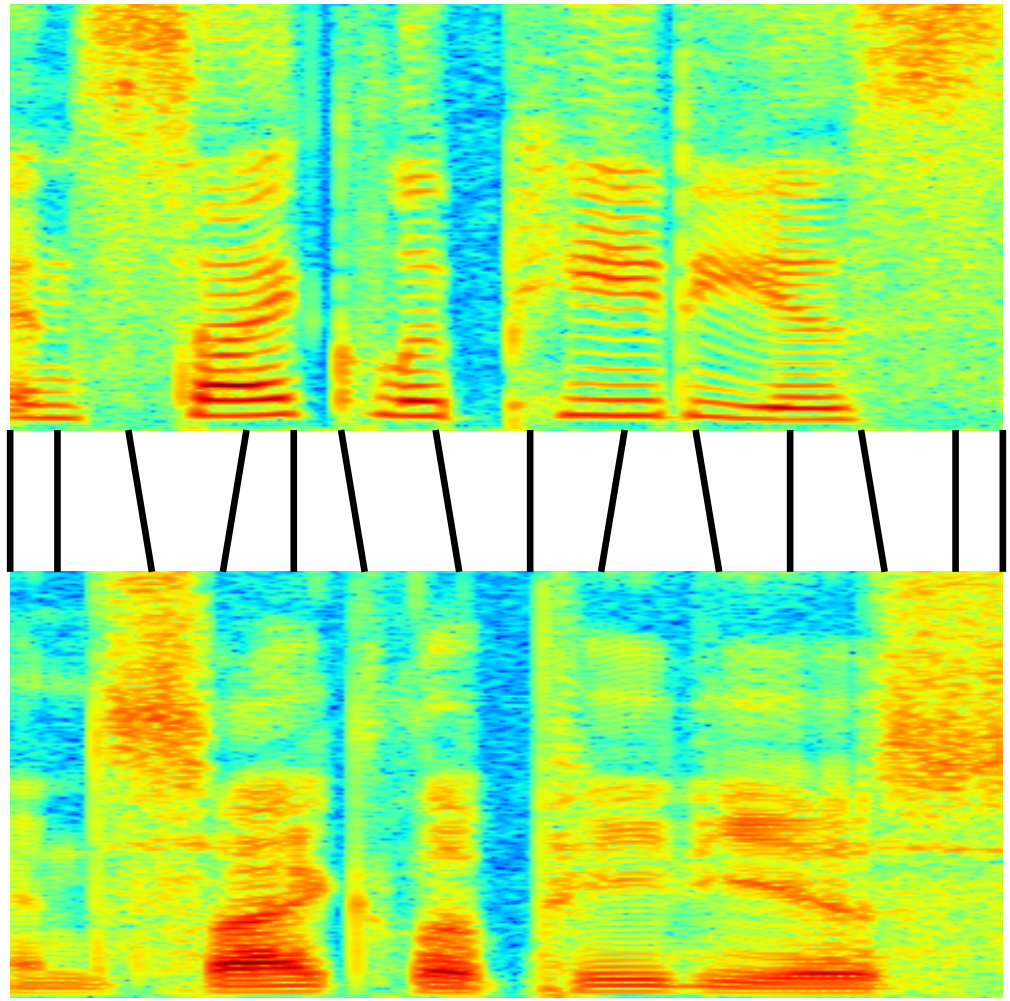


Weak top-down supervision: Align frames



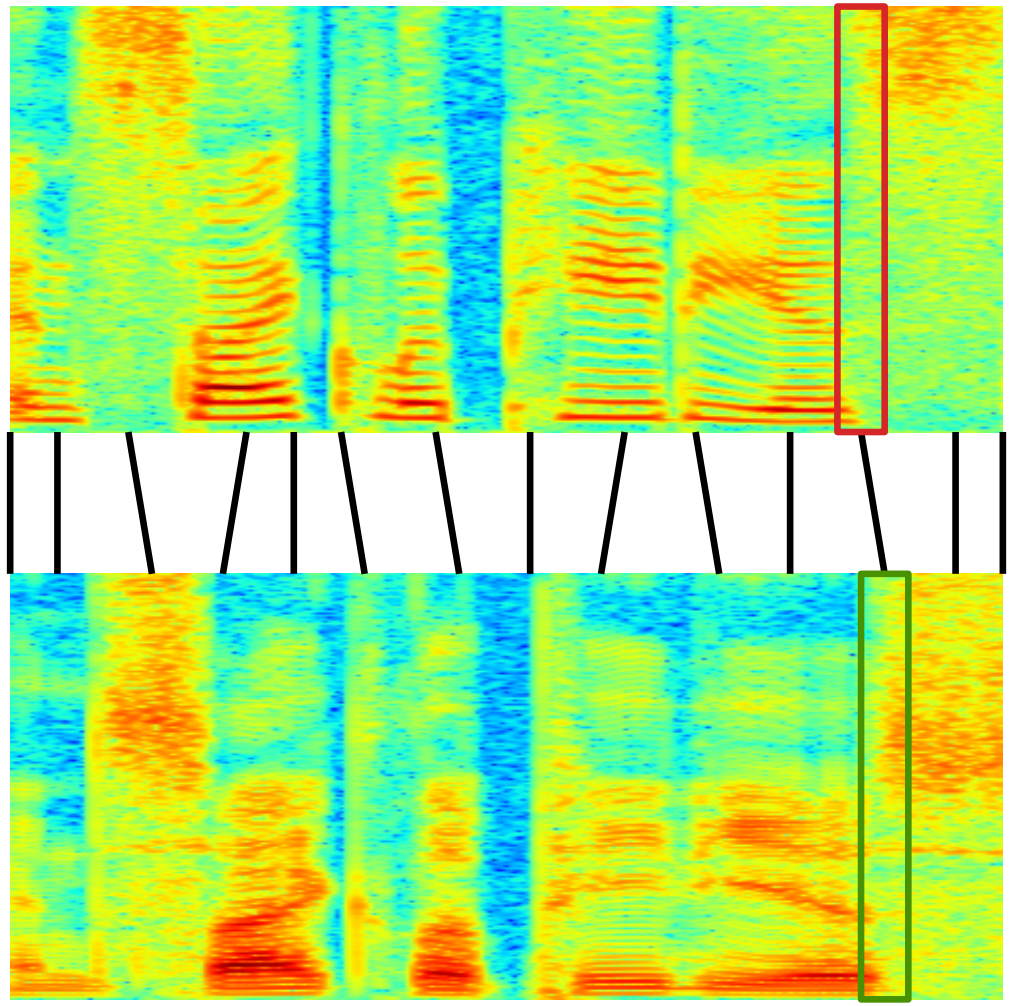
[Jansen et al., ICASSP'13]

Weak top-down supervision: Align frames



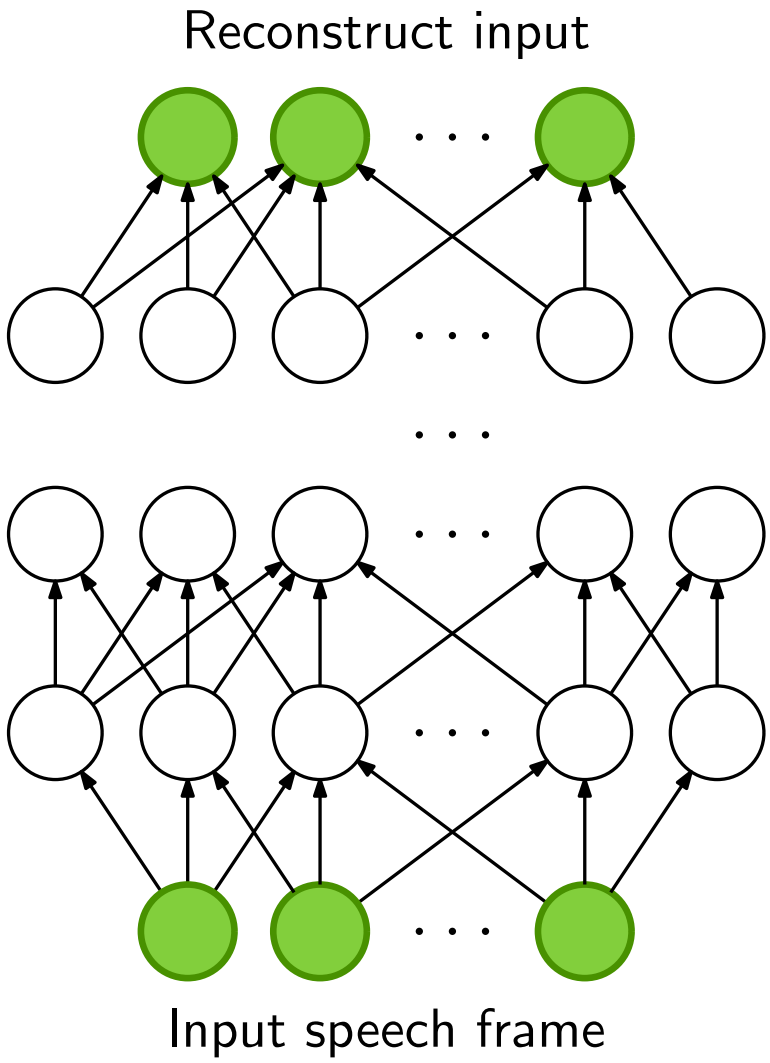
[Jansen et al., ICASSP'13]

Weak top-down supervision: Align frames

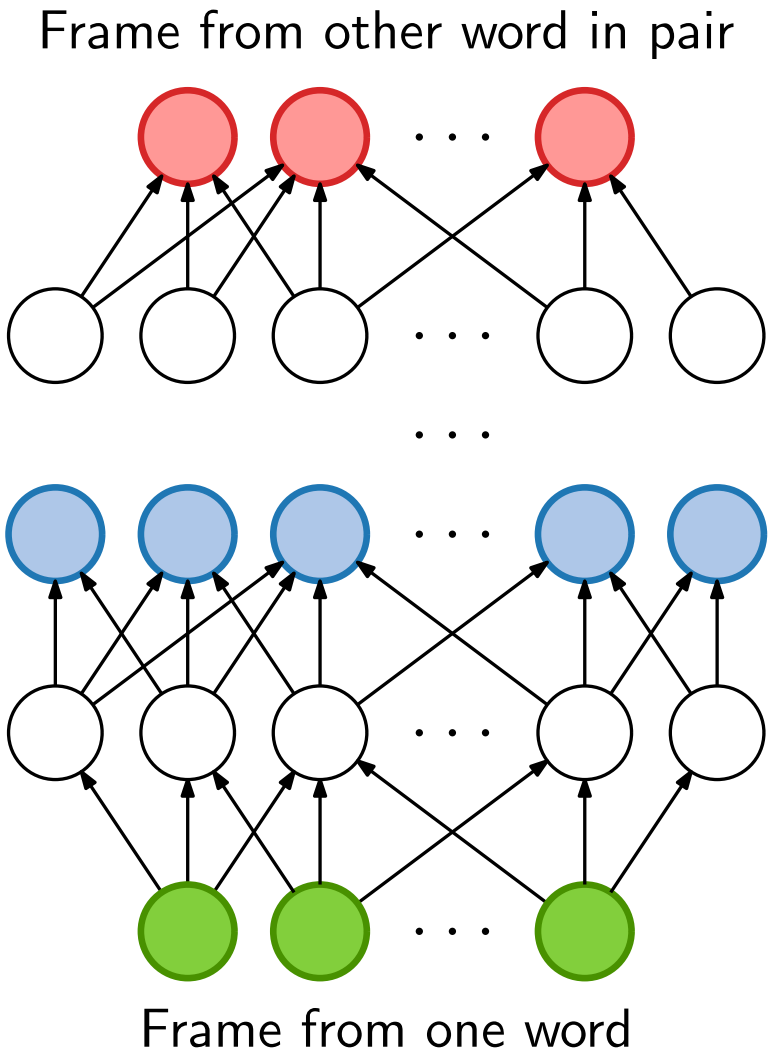


[Jansen et al., ICASSP'13]

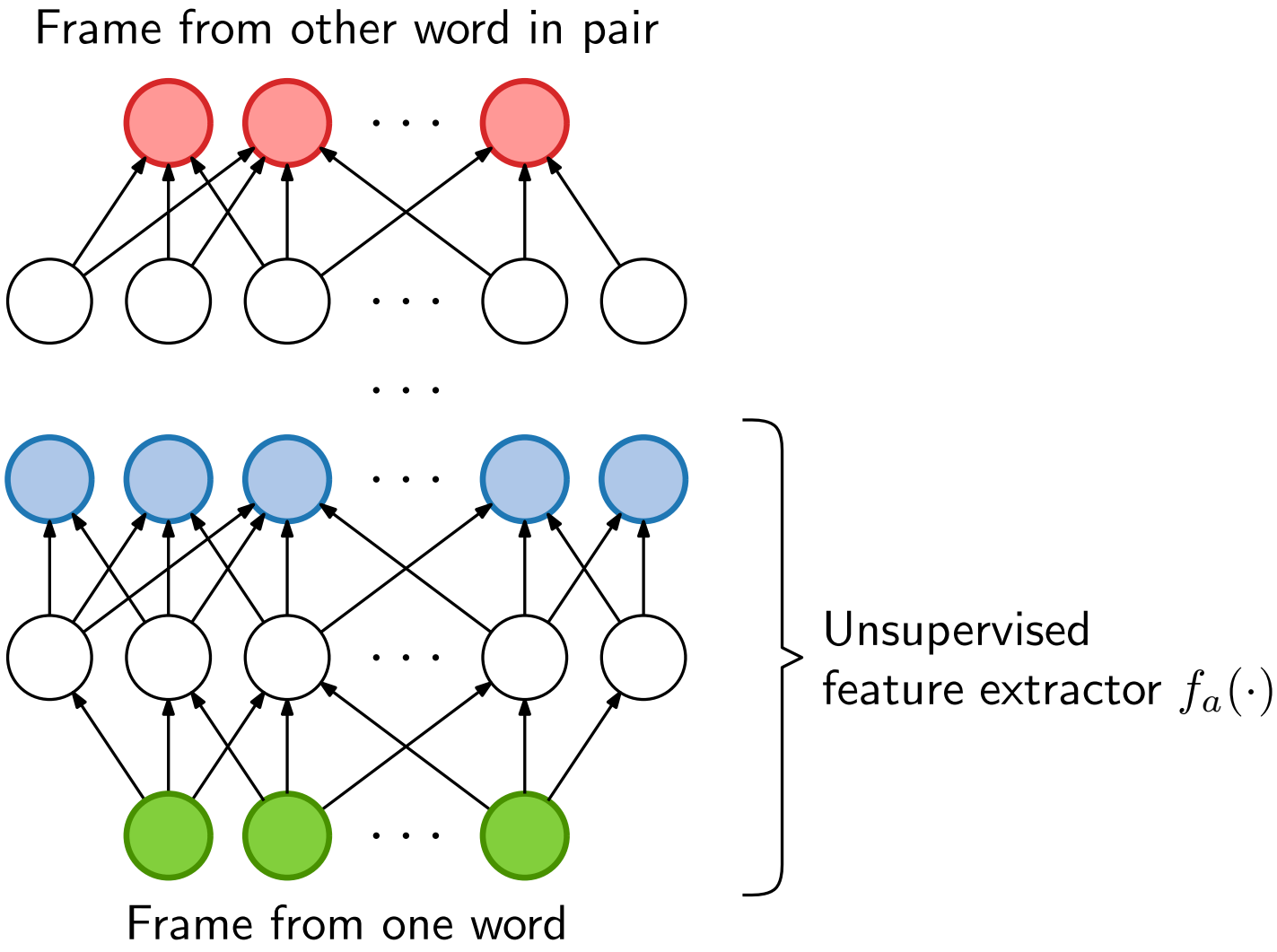
Autoencoder (AE)



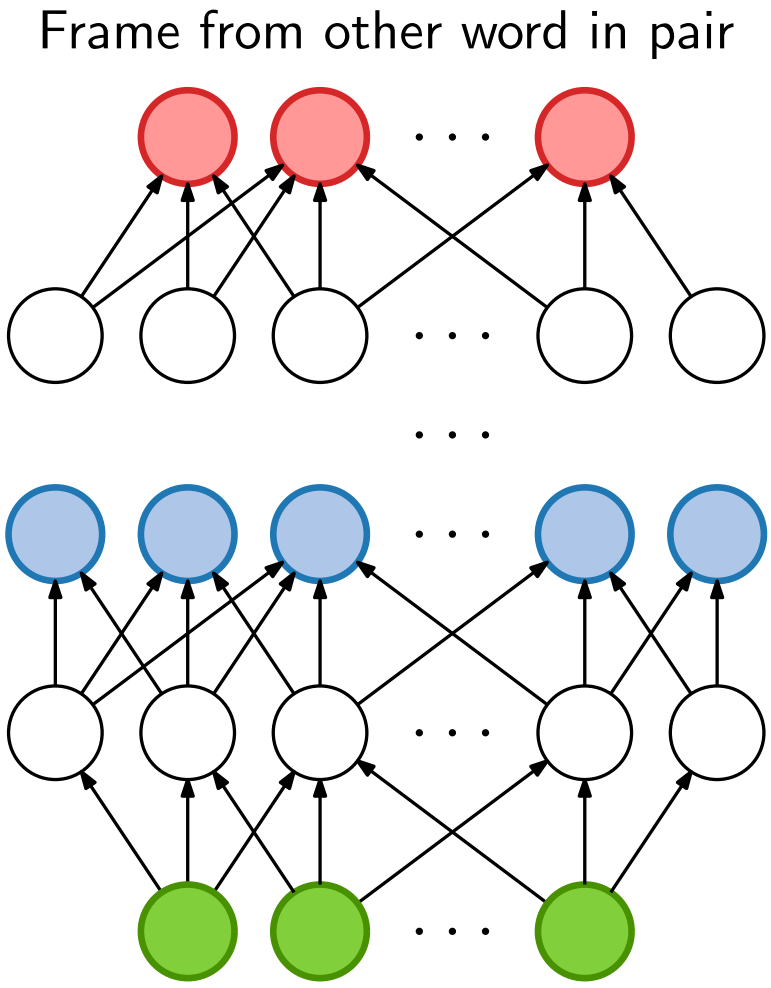
Correspondence autoencoder (cAE)



Correspondence autoencoder (cAE)



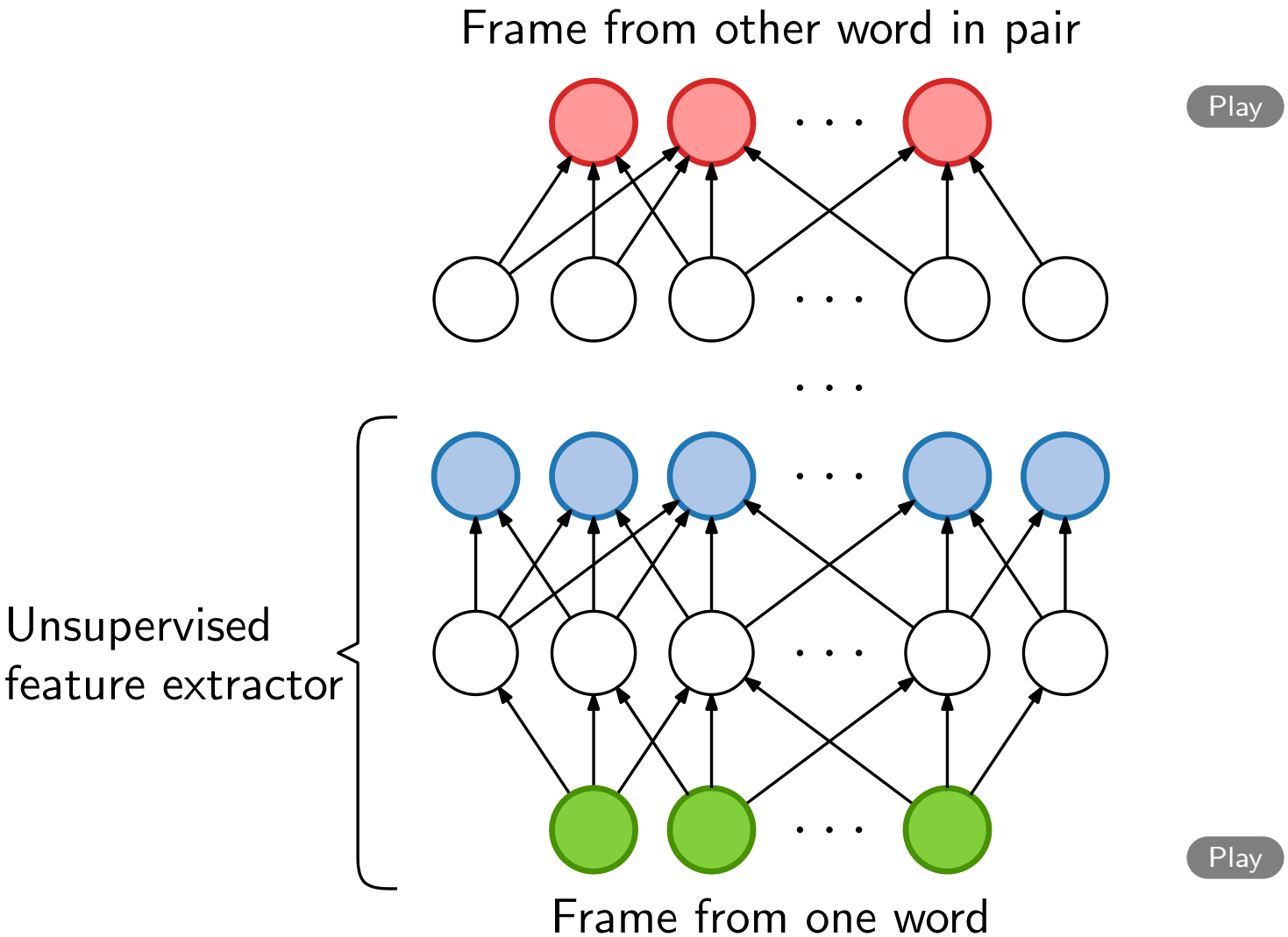
Correspondence autoencoder (cAE)



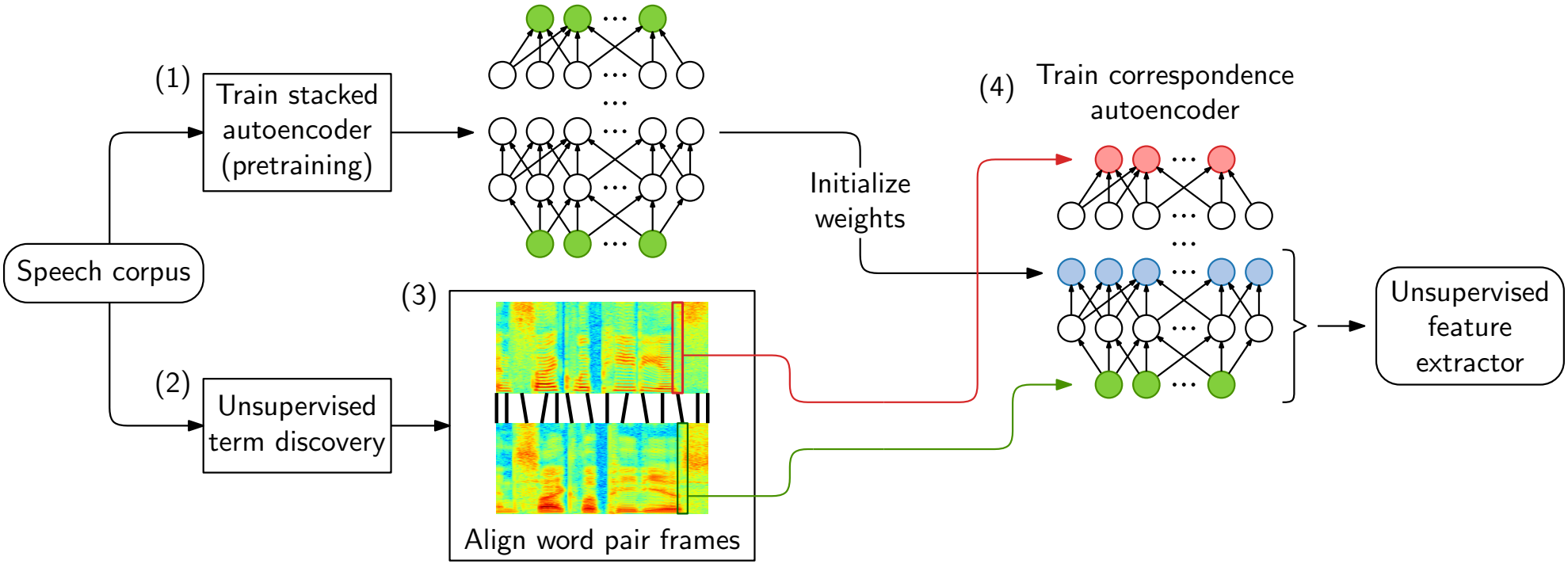
Combine **top-down** and **bottom-up** information

Unsupervised feature extractor $f_a(\cdot)$

Correspondence autoencoder (cAE)



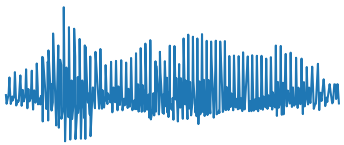
Correspondence autoencoder (cAE)

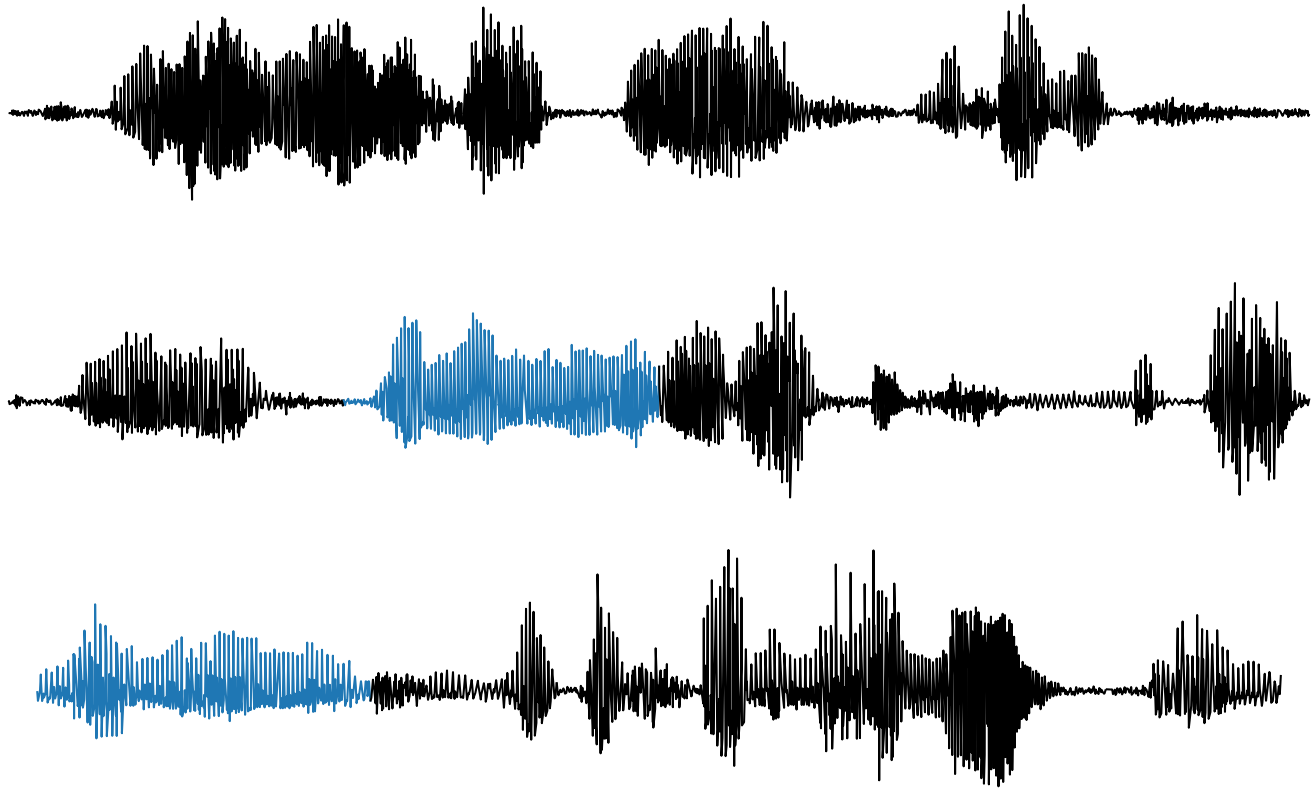


[Kamper et al., ICASSP'15]

Evaluation: Query-by-example search

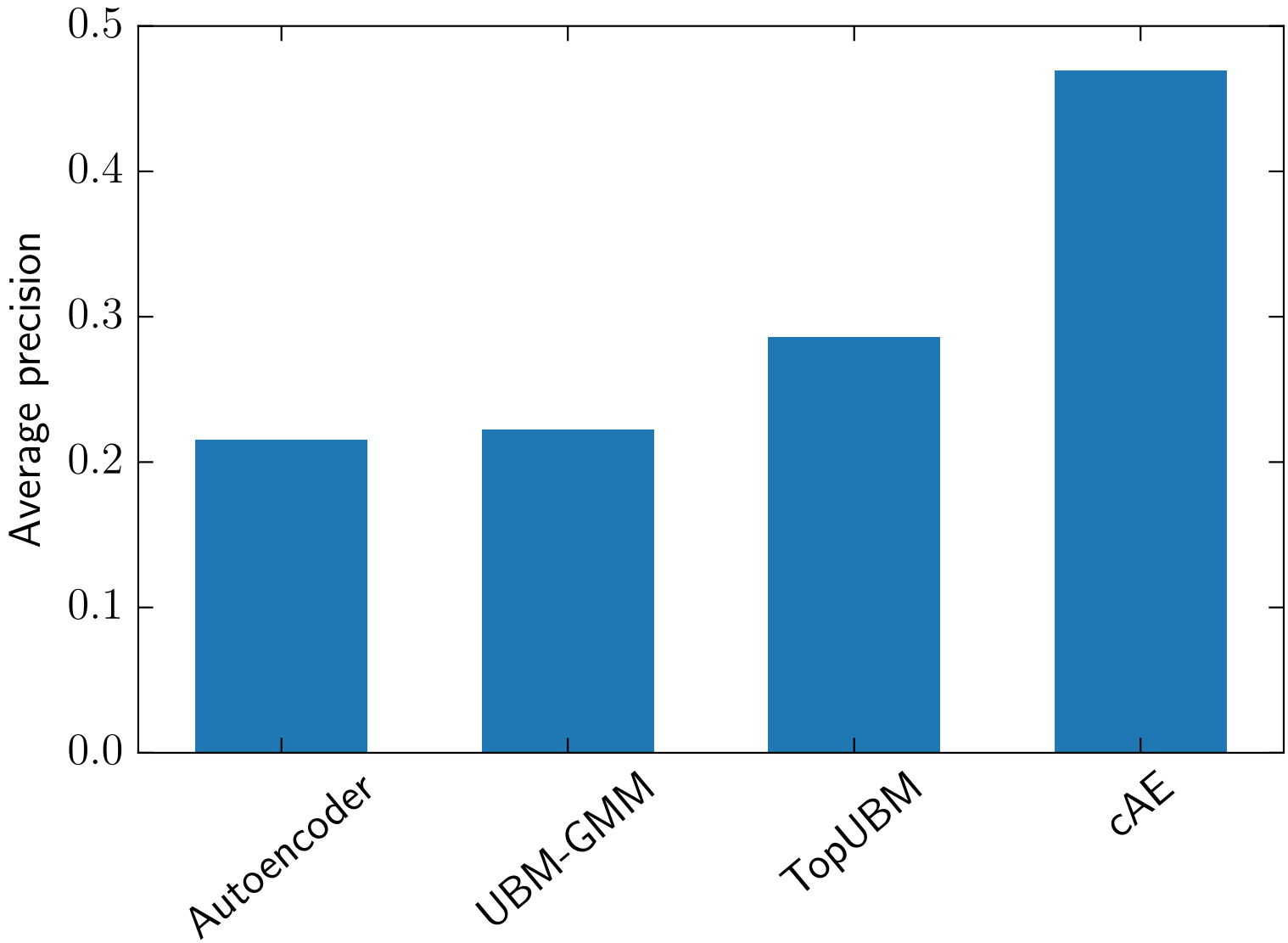
Spoken query:



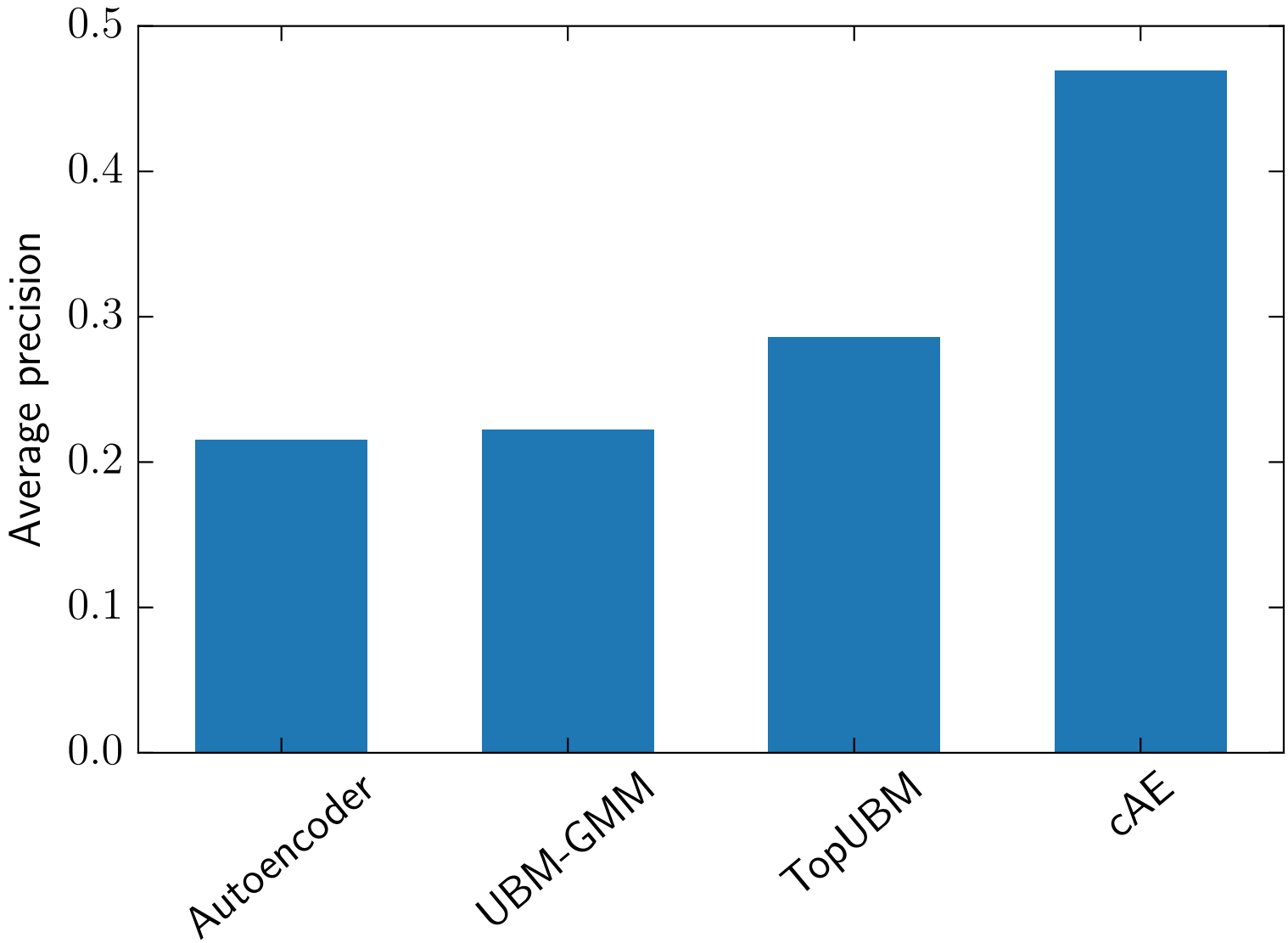


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

Evaluation: Isolated word query-by-example



Evaluation: Isolated word query-by-example



Extended: [Renshaw et al., IS'15] and [Yuan et al., IS'16]

Summary and conclusion

- Introduced correspondence autoencoder (cAE) for unsupervised frame-level representation learning
- Uses top-down information from unsupervised term discovery system
- Uses bottom-up initialization on large speech corpus
- Unsupervised neural network model that combines top-down and bottom-up information results in large intrinsic improvements
- Links with language acquisition research
- Future: More analysis; different domains; practical search systems

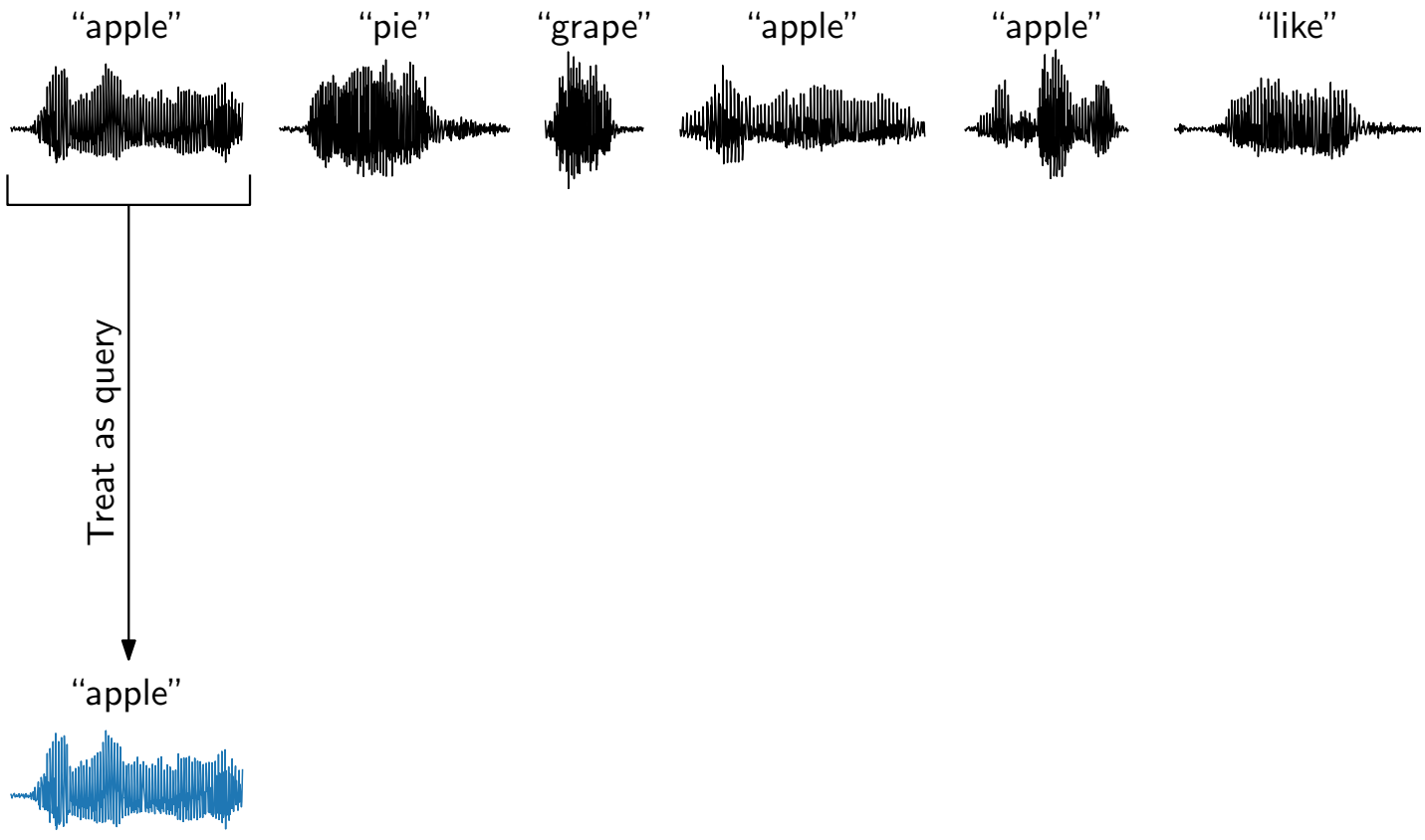
http://www.kamperh.com/

https://github.com/kamperh

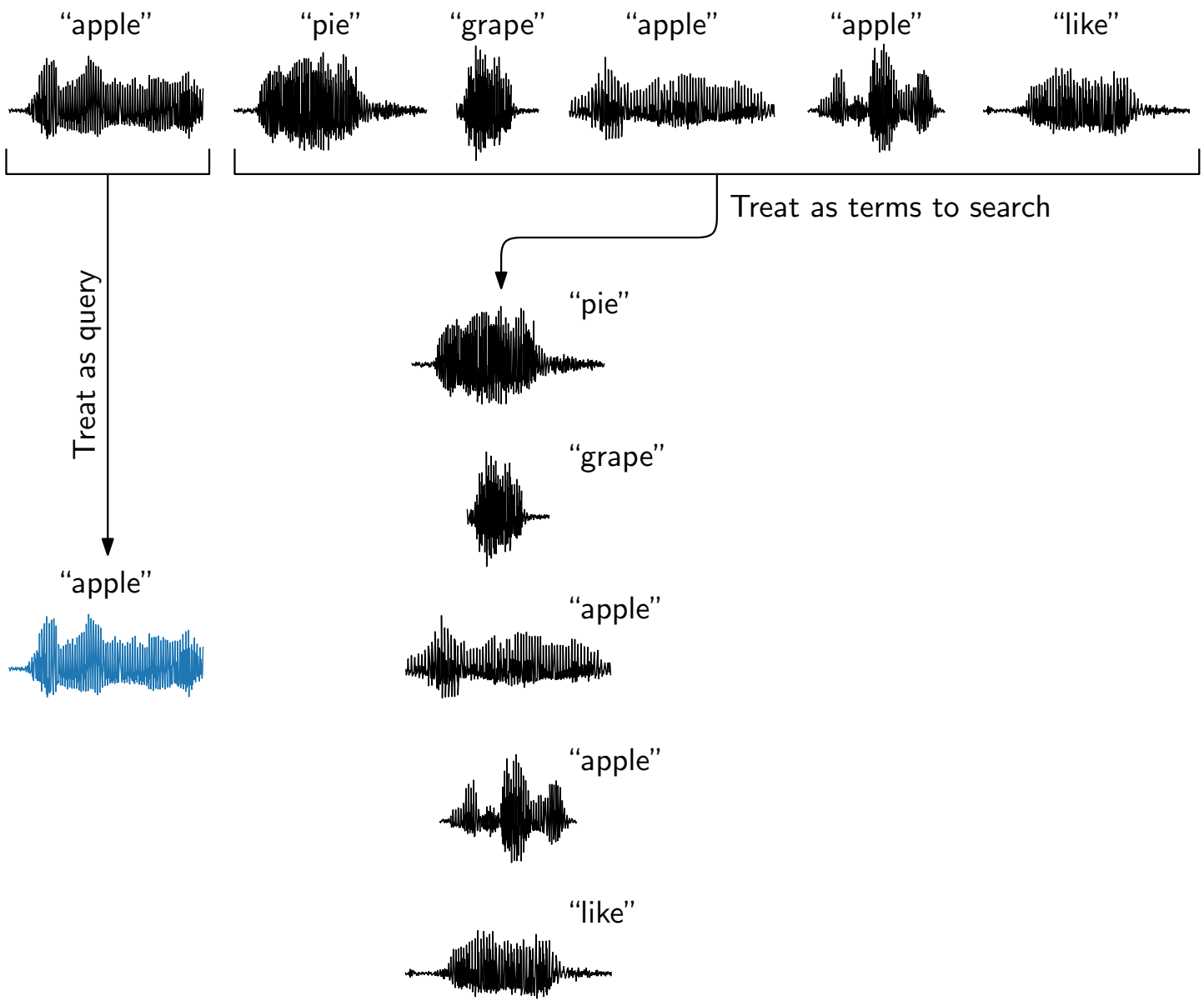
Evaluation of features: same-different task

Evaluation of features: same-different task

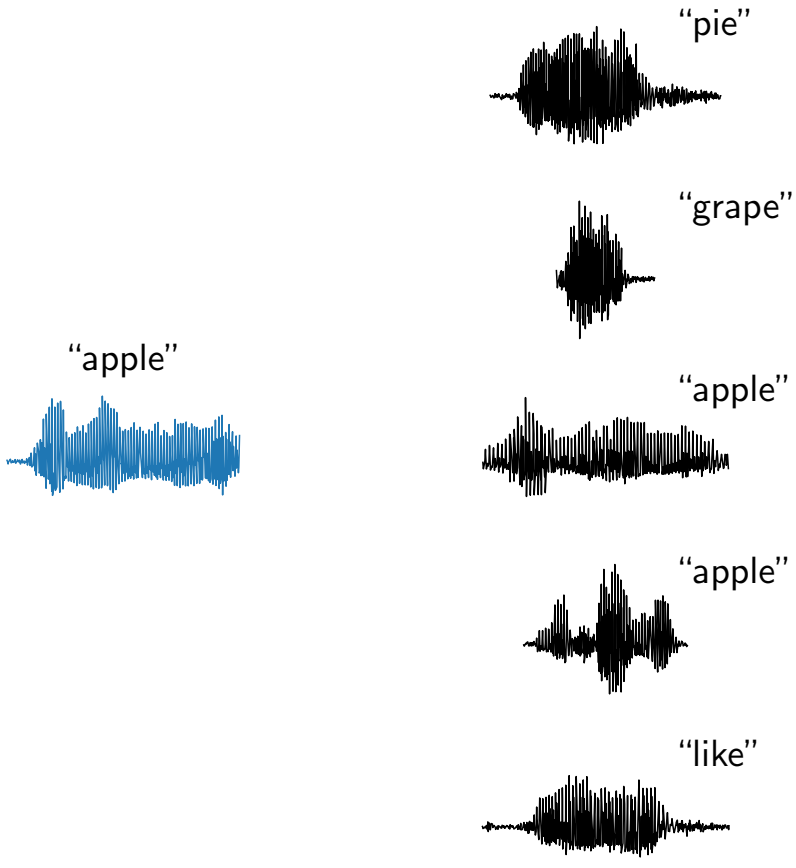
Evaluation of features: same-different task



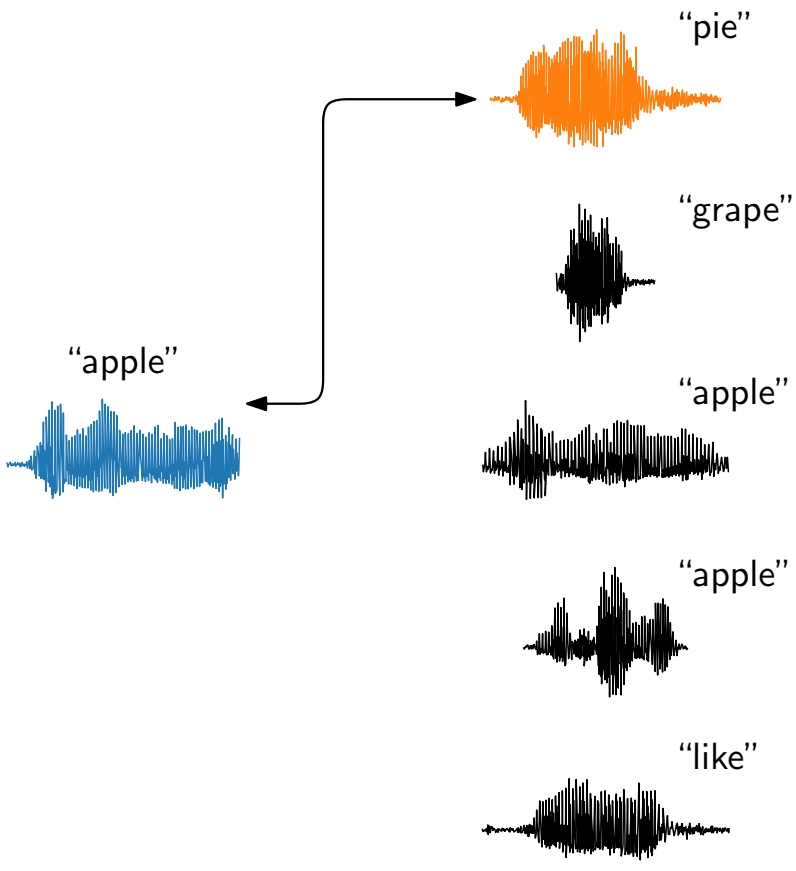
Evaluation of features: same-different task



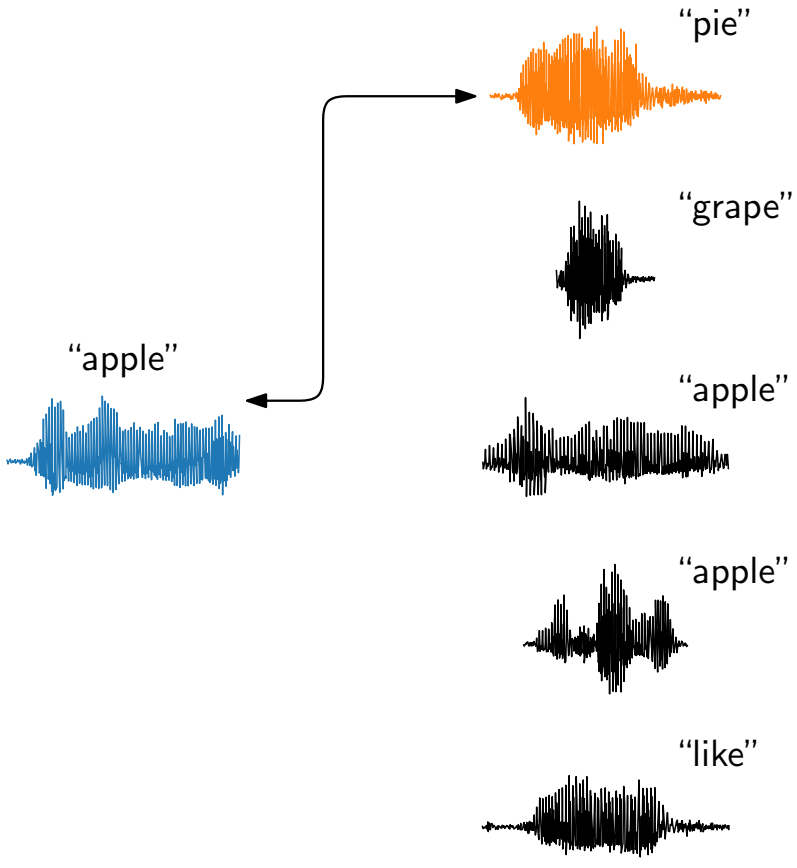
Evaluation of features: same-different task



Evaluation of features: same-different task



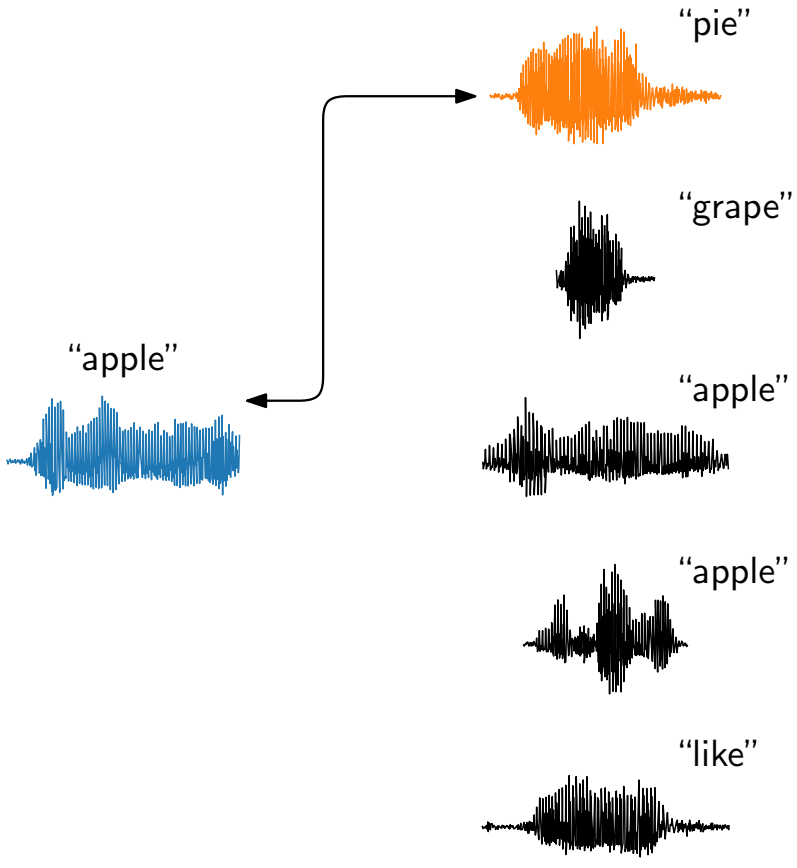
Evaluation of features: same-different task



DTW distance:

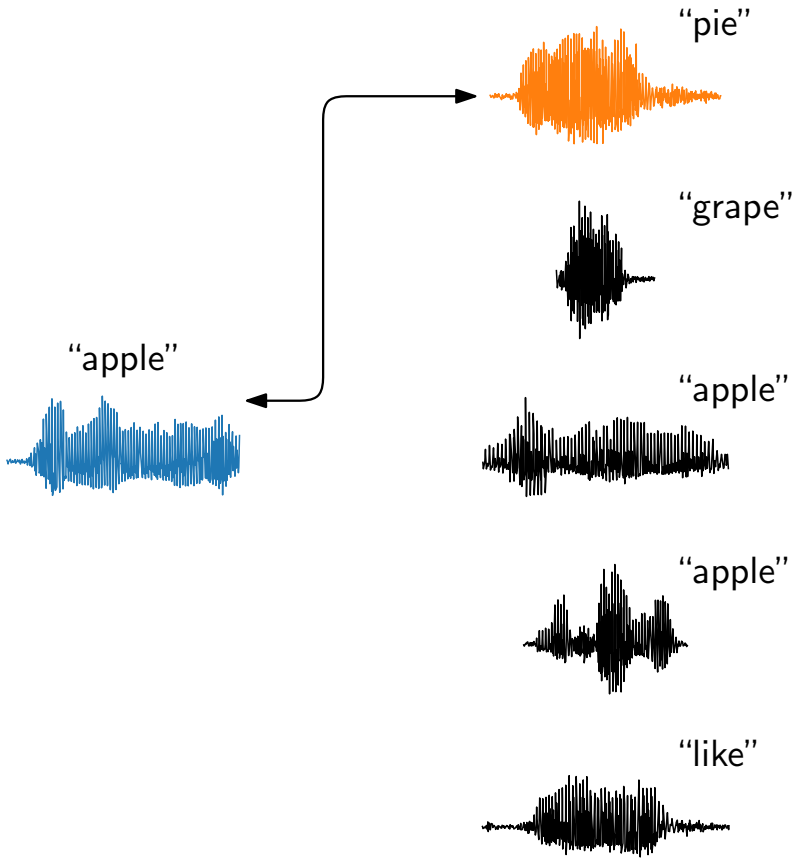
d_1

Evaluation of features: same-different task



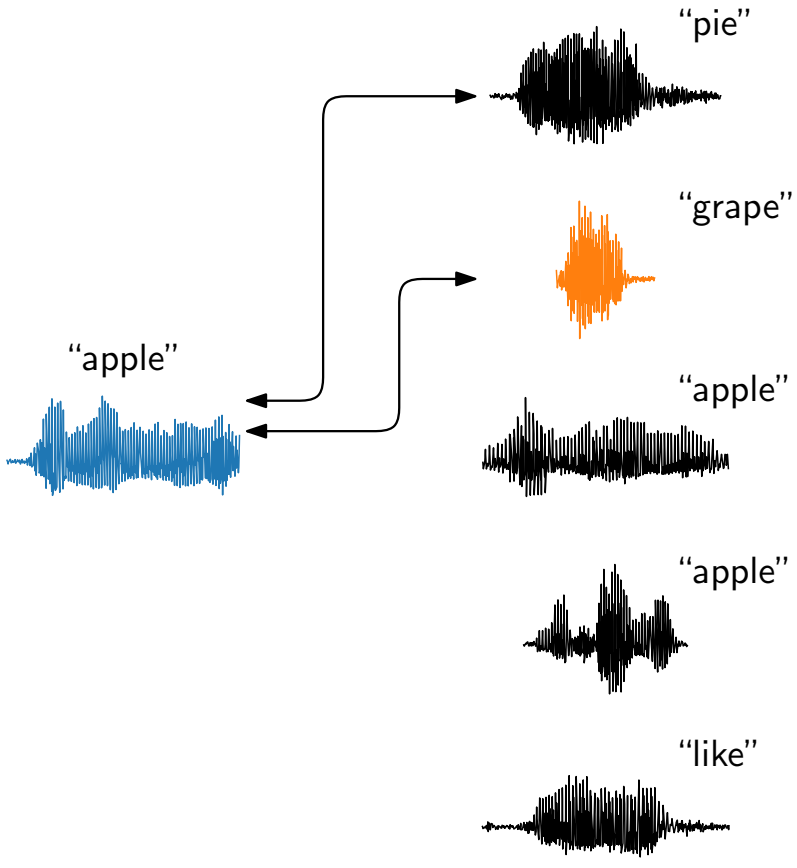
	DTW distance:	$d_i <$ threshold? predict:
"pie"	d_1	different

Evaluation of features: same-different task



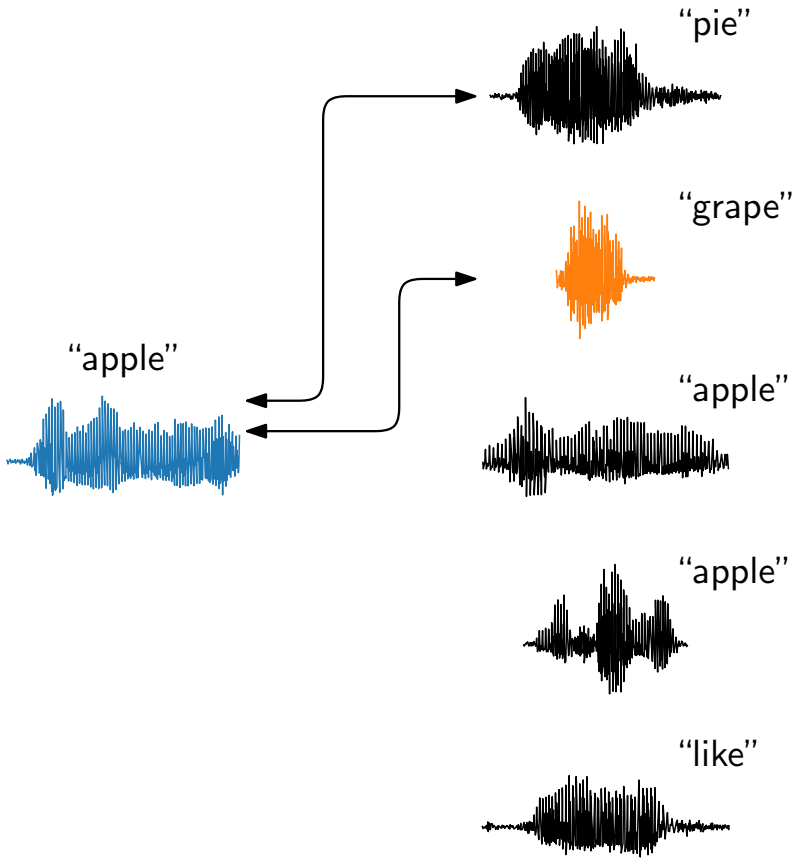
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓

Evaluation of features: same-different task



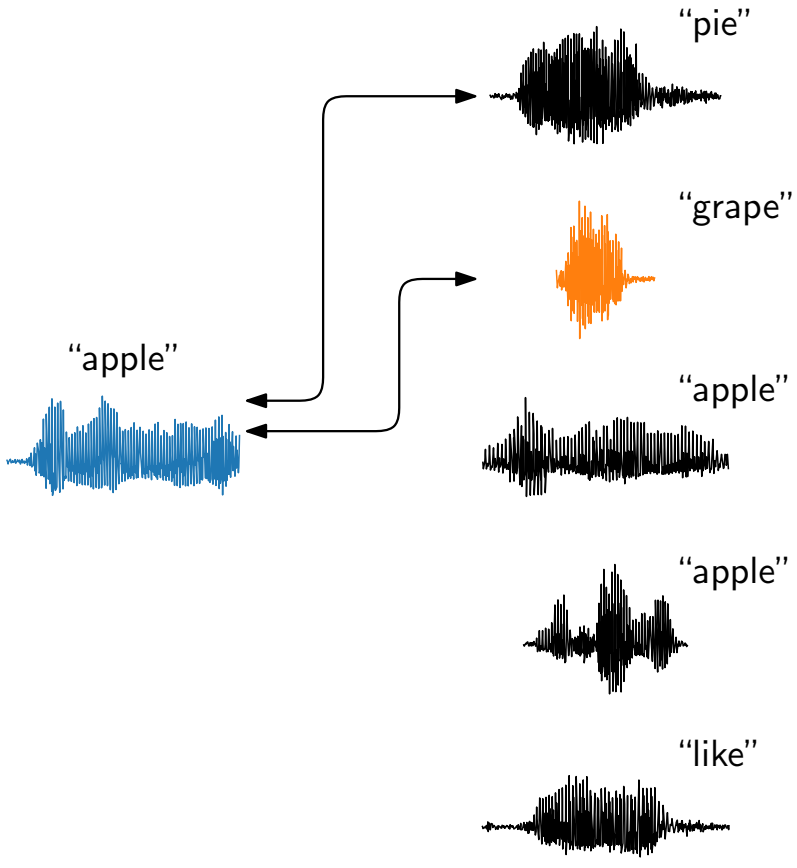
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓

Evaluation of features: same-different task



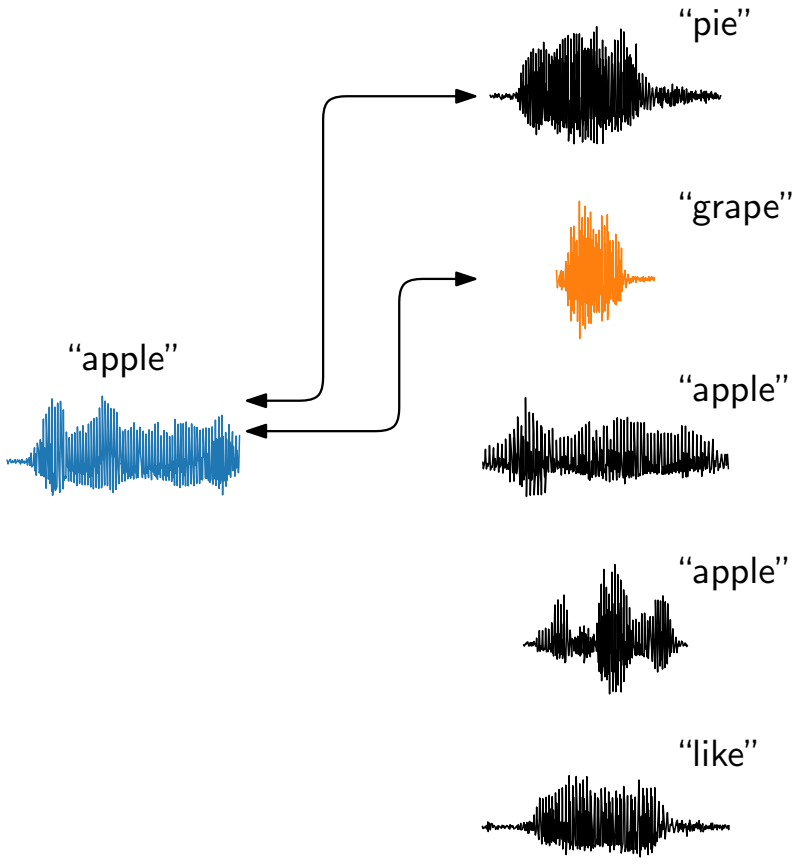
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	

Evaluation of features: same-different task



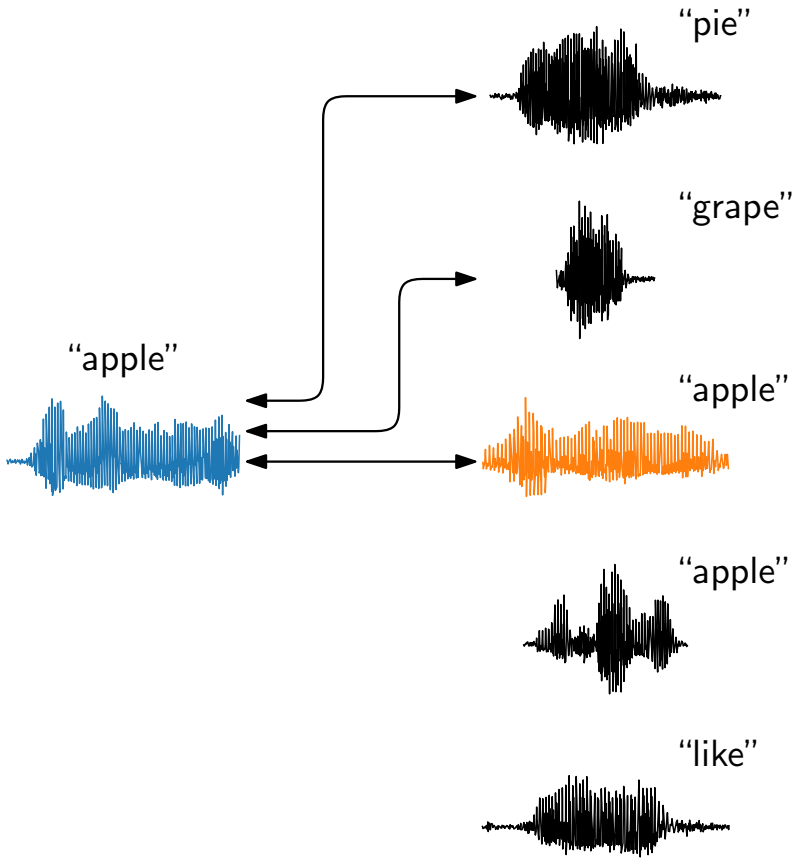
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same

Evaluation of features: same-different task



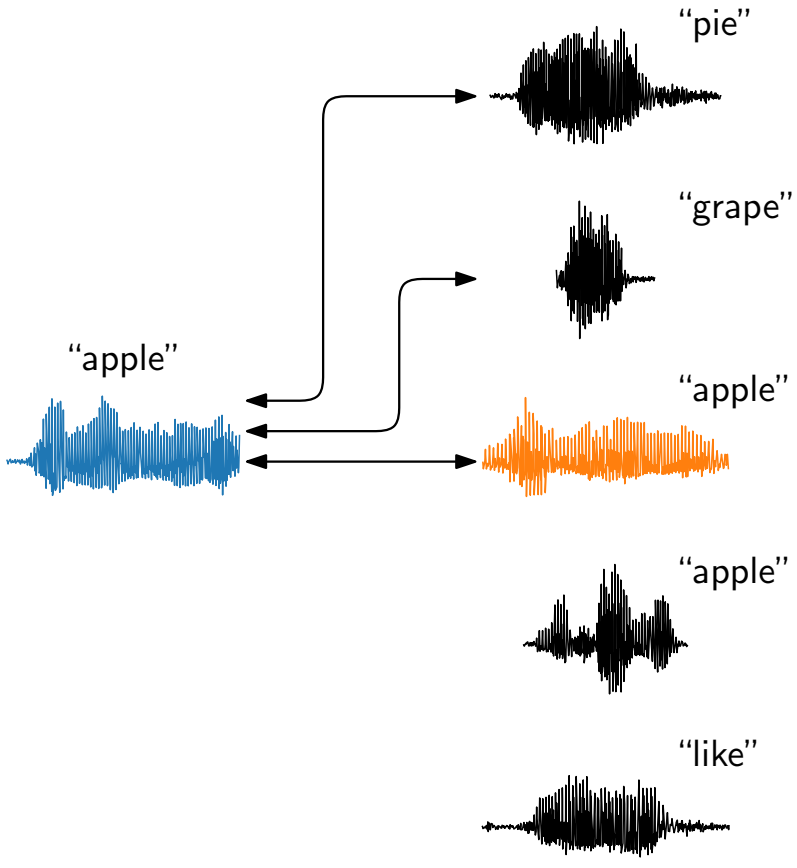
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗

Evaluation of features: same-different task



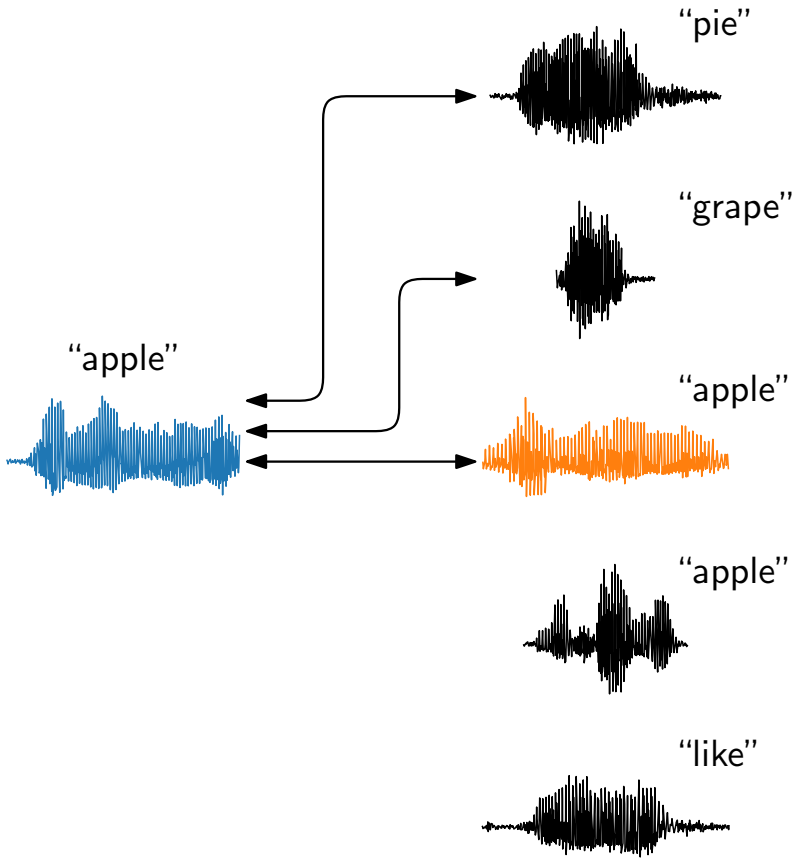
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗

Evaluation of features: same-different task



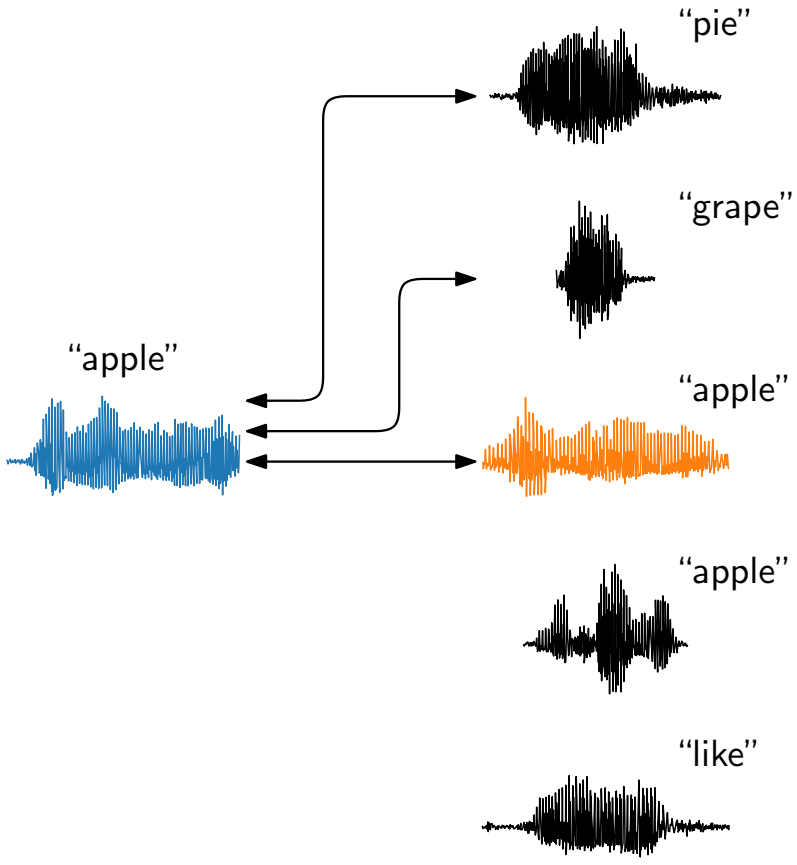
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗
d_3	

Evaluation of features: same-different task



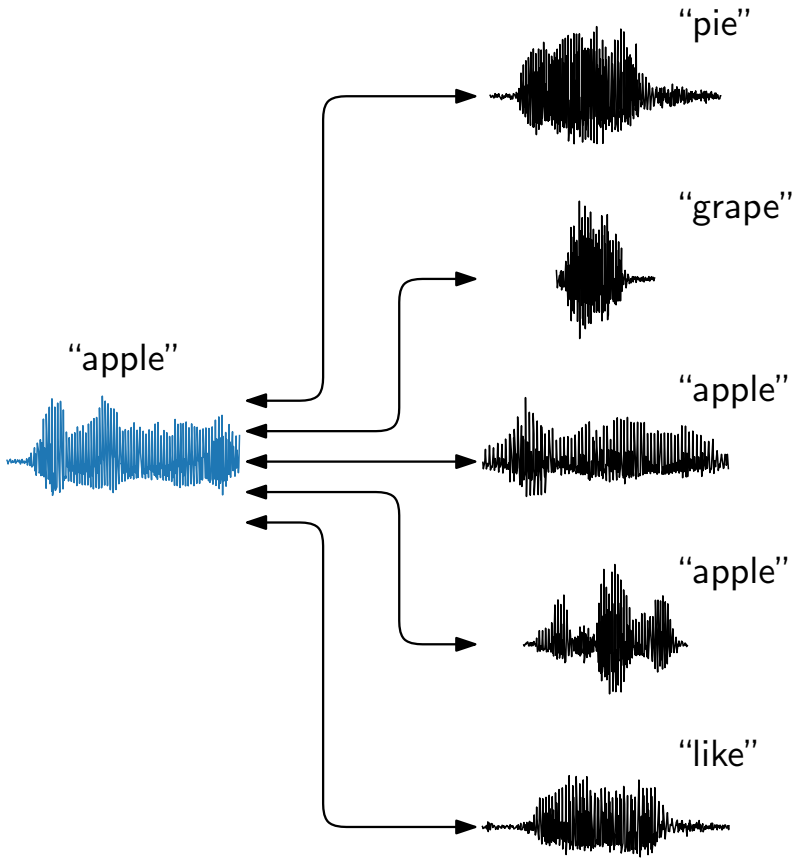
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗
d_3	same

Evaluation of features: same-different task



DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗
d_3	same ✓

Evaluation of features: same-different task



DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗
d_3	same ✓
d_4	different ✗
⋮	
d_N	different ✓

Maties Machine Learning (MML)

- Send "`subscribe mml`" in subject line to `sympa@sympa.sun.ac.za`
- Mailing list: `mml@sympa.sun.ac.za`
- Bring together machine learning researchers from across Stellenbosch University
- Format: Short (in)formal talks every second Friday over lunch
- Focus on machine learning **research**
- If you want to give a talk, or have any ideas, please let us know!
- `kamperh@sun.ac.za`, `wbrink@sun.ac.za`