

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

Maties Machine Learning (MML), Oct. 2017

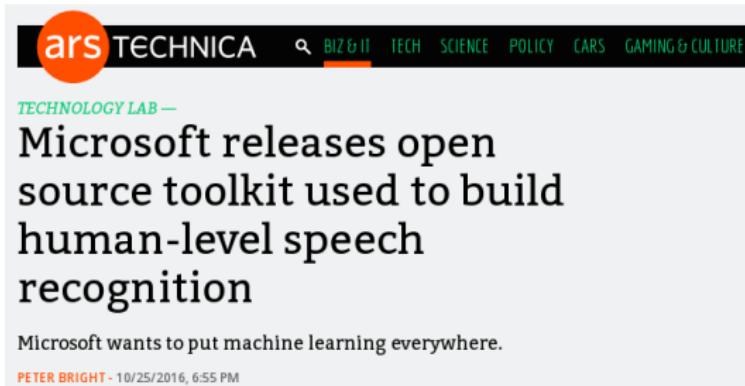
Herman Kamper

Stellenbosch University

<http://www.kamperh.com/>

# Success in speech recognition

# Success in speech recognition



The image shows a screenshot of an Ars Technica article. At the top, there's a navigation bar with the site's logo ('ars TECHNICA') and a search icon. Below the logo, a horizontal menu bar includes categories like 'BIZ & IT', 'TECH', 'SCIENCE', 'POLICY', 'CARS', and 'GAMING & CULTURE'. The main title of the article is 'Microsoft releases open source toolkit used to build human-level speech recognition'. A subtitle below it reads 'Microsoft wants to put machine learning everywhere.' The author's name, 'PETER BRIGHT', and the date, '10/25/2016, 6:55 PM', are also visible.

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

TECHNOLOGY LAB —

## Microsoft releases open source toolkit used to build human-level speech recognition

Microsoft wants to put machine learning everywhere.

PETER BRIGHT - 10/25/2016, 6:55 PM

# Success in speech recognition

The screenshot shows two adjacent news articles. On the left, Ars Technica's 'TECHNICA' section features a story titled 'Microsoft releases source toolkit to help build human-level speech recognition'. On the right, The Wall Street Journal's homepage features a prominent article titled 'Speech Recognition Gets Conversational'.

**Ars Technica Header:** TECHNICA | BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

Nasdaq ▲ 5166.17 2.37% U.S. 10 Yr ▼ -15/32 Yield 1.828% Crude Oil ▲ 44.93 1.95%

**WSJ Header:** THE WALL STREET JOURNAL. Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life DIGITS

**WSJ Article Headline:** Speech Recognition Gets Conversational

**WSJ Article Details:** By ROBERT MCMILLAN May 28, 2015 12:54 pm ET

**Bottom Left Text:** Microsoft wants to put machine learning into every app

**Bottom Left Author:** PETER BRIGHT - 10/25/2016, 6:55 PM

# Success in speech recognition

The screenshot shows a news article from CBS News. At the top, there is a navigation bar with the CBS News logo and links for Video, US, World, Politics, Entertainment, and Health. To the right of the navigation bar, there are some financial tickers showing oil prices. Below the navigation bar, the main headline reads "Microsoft says speech recognition technology reaches 'human parity'" in large, bold, black letters. The article is dated October 18, 2016, at 3:56 PM, and is attributed to Brian Mastroianni from CBS News. A sub-headline below the main one reads "Microsoft source to human-level recognition". On the left side of the article, there is a sidebar with the text "Microsoft wants to pi" and "PETER BRIGHT - 10/25/2016, 6:55 PM". The date "May 28, 2015 12:54 pm ET" is also visible near the bottom of the main article area.

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

# Success in speech recognition

**ars TECHN**  

TECHNOLOGY LAB —

By **BRIAN** 

## Microsoft source to human-language recognition

Microsoft wants to pi

**PETER BRIGHT** - 10/25/2016, 6:55 PM

## Google's Pixel Buds place 40 languages in your ear

Nick Lavars | October 5th, 2017



7 PICTURES

When paired with the Pixel smartphone, Google's Pixel Buds can let you carry out conversations in 40 languages



JRNAL.  
pinion Arts Life

ersational

Saon et al., arXiv'17]

[VIEW GALLERY - 7 IMAGES](#)

Software that can translate languages in real time would be huge news for travel, business and society as a whole. Names big and small have promised to make this technological leap in recent years, but it is by no means commonplace. Now Google has arrived on the scene with its first set of wireless earbuds that are claimed to translate 40 languages in real time, along with a few other handy features.

# Success in speech recognition

The screenshot shows a news article from CBS News. At the top, there is a navigation bar with the CBS News logo and links for Video, US, World, Politics, Entertainment, and Health. Below the navigation bar, the headline reads: "Microsoft says speech recognition technology reaches 'human parity'". The article is dated October 18, 2016, at 3:56 PM, by Brian Mastroianni. The text discusses Microsoft's claims and includes a quote from Peter Bright. The CBS News logo is visible in the bottom right corner.

ars TECH CBSNEWS Video US World Politics Entertainment Health de Oil ▲ 44.93 1.95%

TECHNOLOGY LAB —

By BRIAN MASTROIANNI / CBS NEWS / October 18, 2016, 3:56 PM

## Microsoft says speech recognition technology reaches "human parity"

Microsoft wants to pi

PETER BRIGHT - 10/25/2016, 6:55 PM

May 28, 2015 12:54 pm ET

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

- Google Voice: English, Spanish, German, . . . , Zulu ( $\sim$ 50 languages)

# Success in speech recognition

The screenshot shows a news article from CBS News. At the top, there is a navigation bar with links for Video, US, World, Politics, Entertainment, and Health. To the right of the navigation bar, there is some financial information: "de Oil ▲ 44.93 1.95%". Below the navigation bar, the CBS News logo is displayed. On the left side of the main content area, there is a sidebar with the text "TECHNOLOGY LAB — Microsoft source to human-like recognition". The main title of the article is "Microsoft says speech recognition technology reaches 'human parity'". Below the title, there is a sub-headline "Microsoft wants to pi". At the bottom of the article, there is a timestamp: "May 28, 2015 12:54 pm ET".

ars TECH CBSNEWS Video US World Politics Entertainment Health de Oil ▲ 44.93 1.95%

TECHNOLOGY LAB — Microsoft source to human-like recognition

## Microsoft says speech recognition technology reaches "human parity"

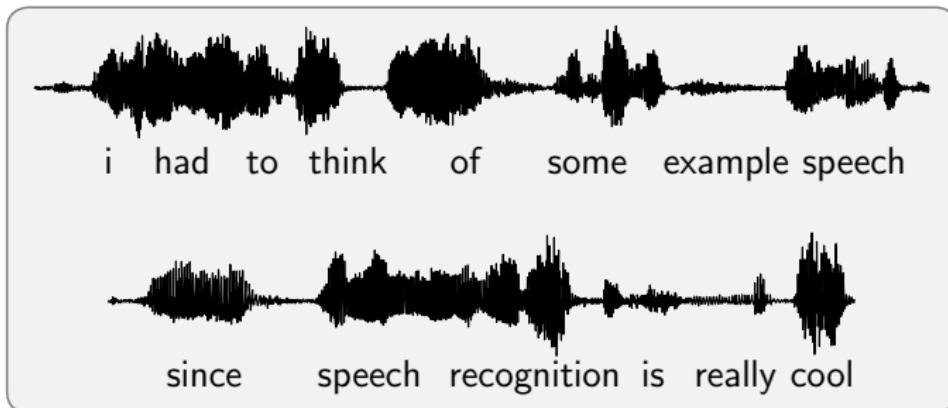
Microsoft wants to pi

PETER BRIGHT - 10/25/2016, 6:55 PM May 28, 2015 12:54 pm ET

[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

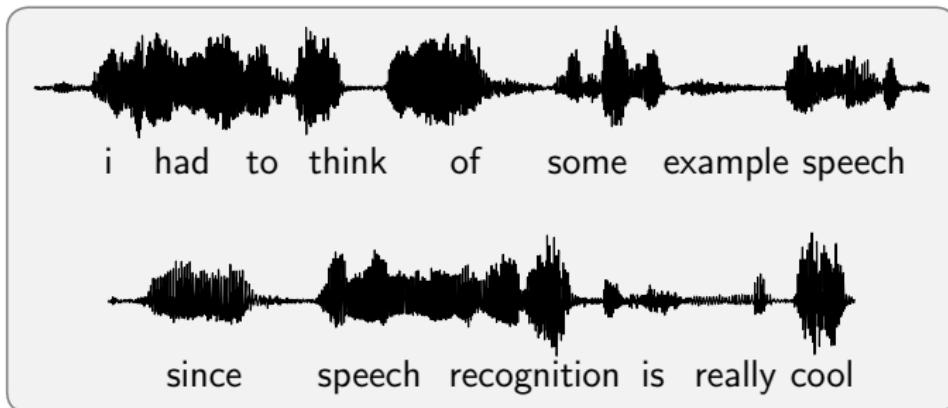
# Success in speech recognition



[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

# 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)



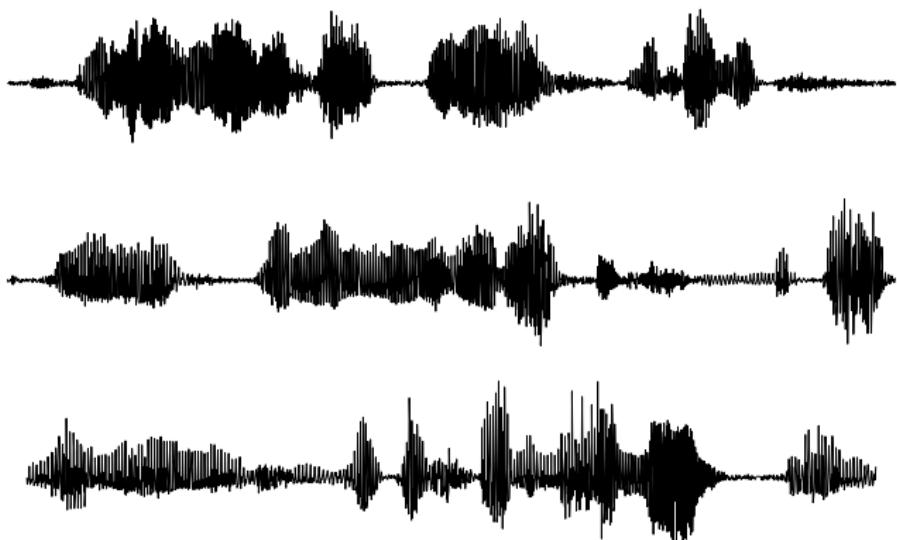
[Park and Glass, TASLP'08]

# Unsupervised term discovery (UTD)



[Park and Glass, TASLP'08]

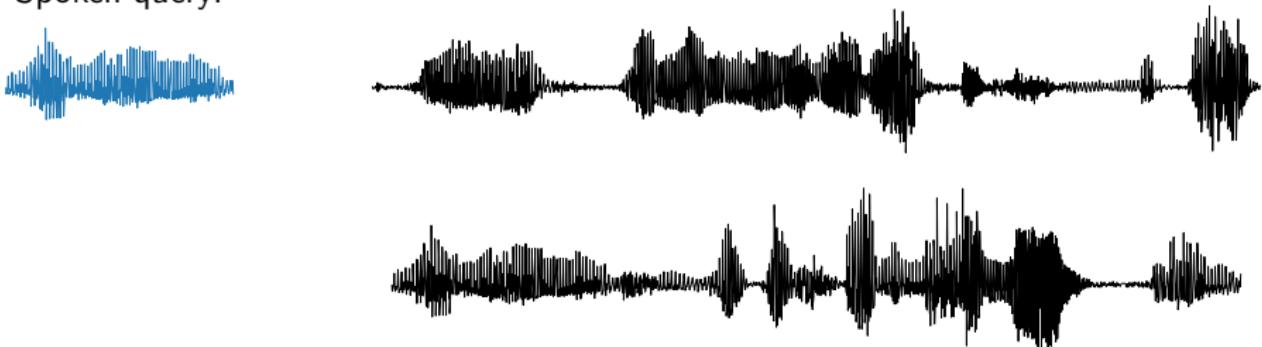
## Example: Query-by-example search



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

# Example: Query-by-example search

Spoken query:

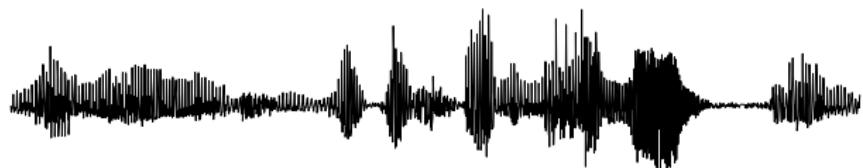
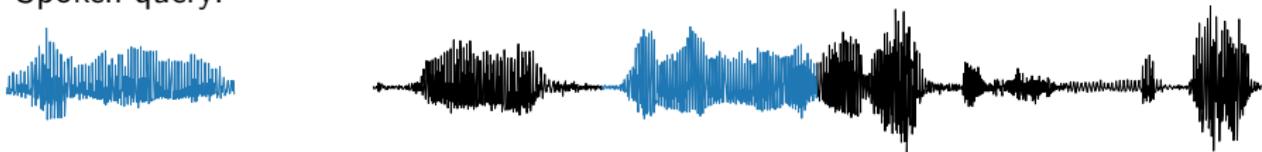


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

# Example: Query-by-example search



Spoken query:

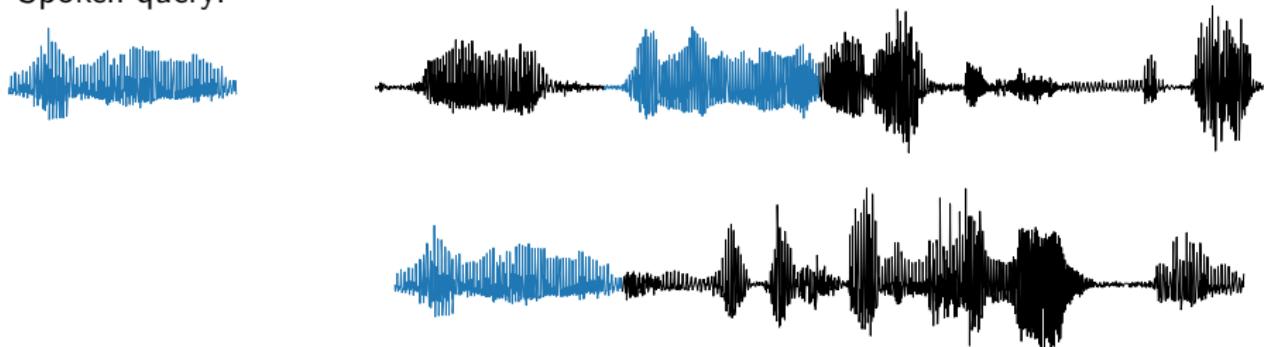


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

# Example: Query-by-example search



Spoken query:

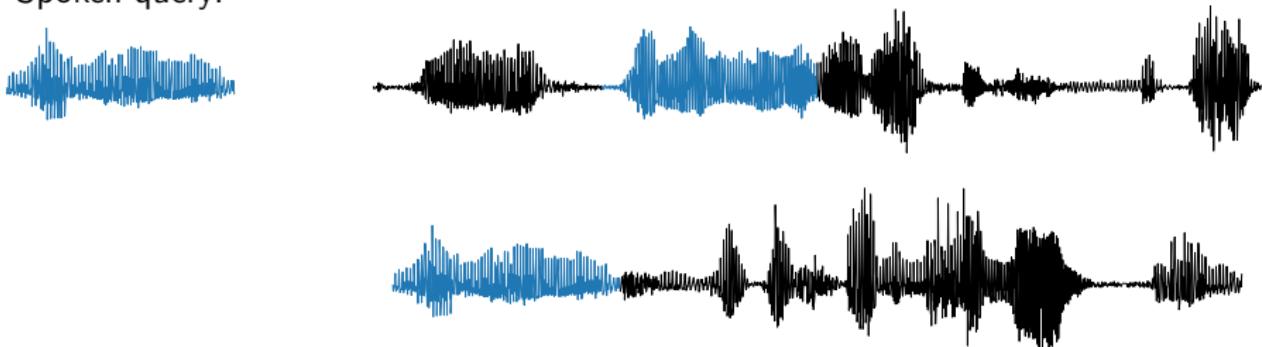


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

## 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**:

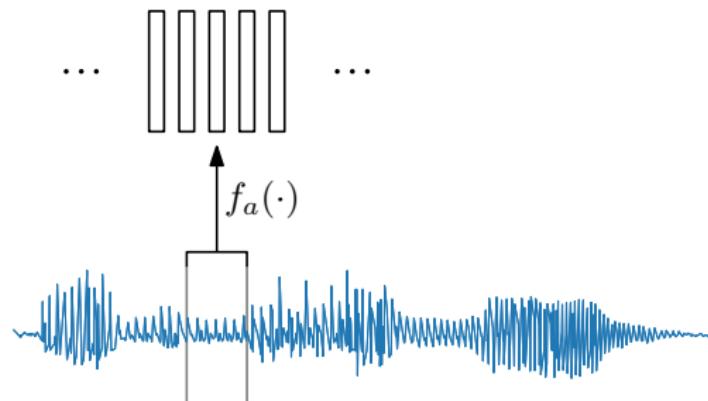
# Unsupervised speech processing: Two problems

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



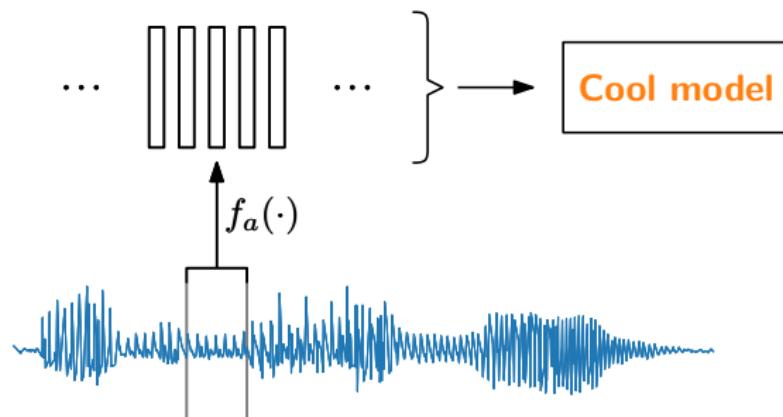
# Unsupervised speech processing: Two problems

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



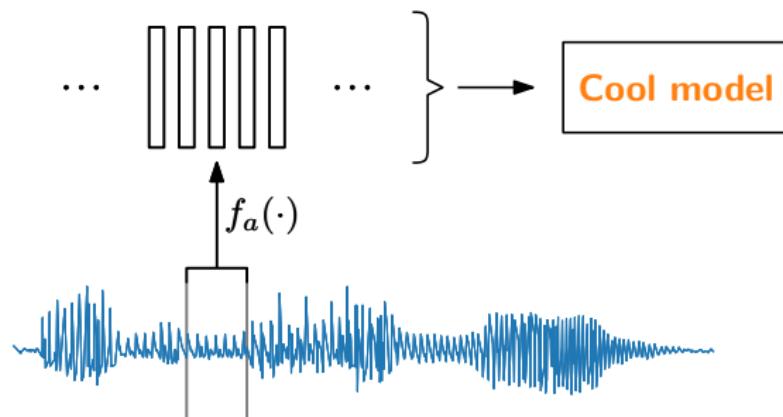
# Unsupervised speech processing: Two problems

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



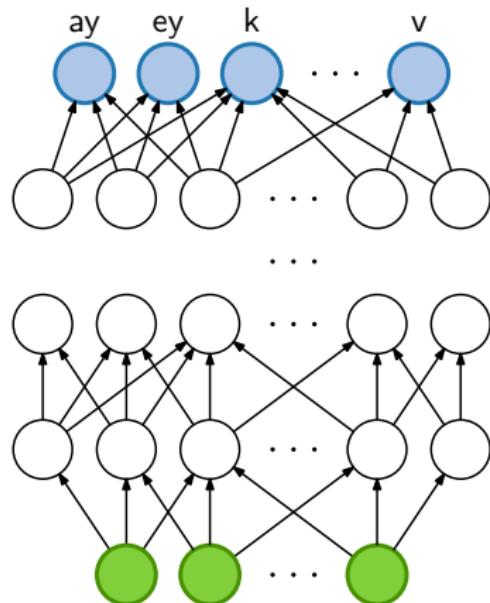
Aren Jansen



Sharon Goldwater

# Supervised representation learning using DNNs

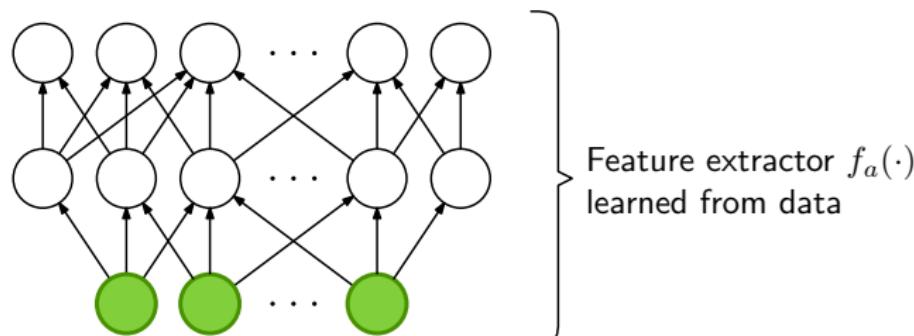
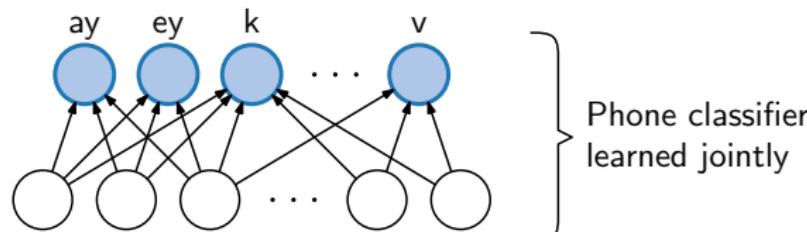
Output: predict phone states



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

# Supervised representation learning using DNNs

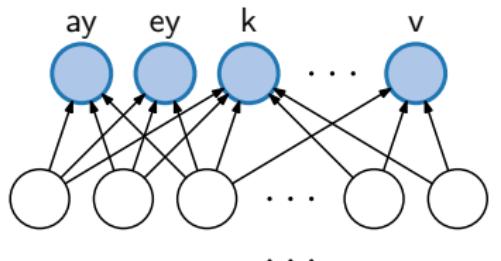
Output: predict phone states



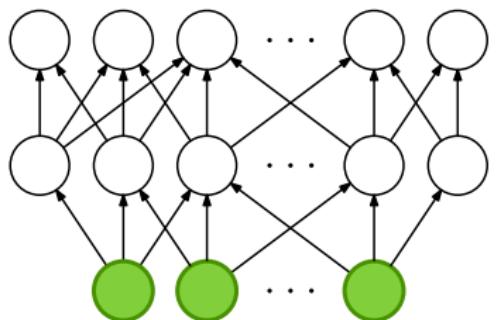
Input: speech frame(s)  
e.g. MFCCs, filterbanks

# Supervised representation learning using DNNs

Output: predict phone states



Phone classifier  
learned jointly

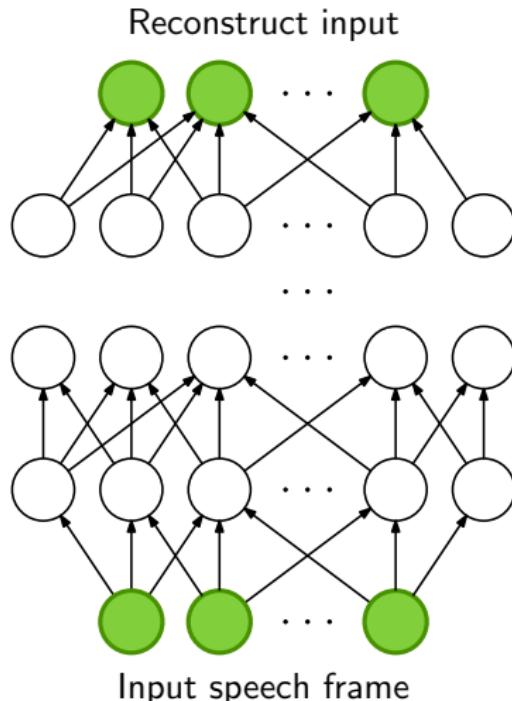


Input: speech frame(s)  
e.g. MFCCs, filterbanks

**Unsupervised modelling:**  
No phone class targets to  
train network on

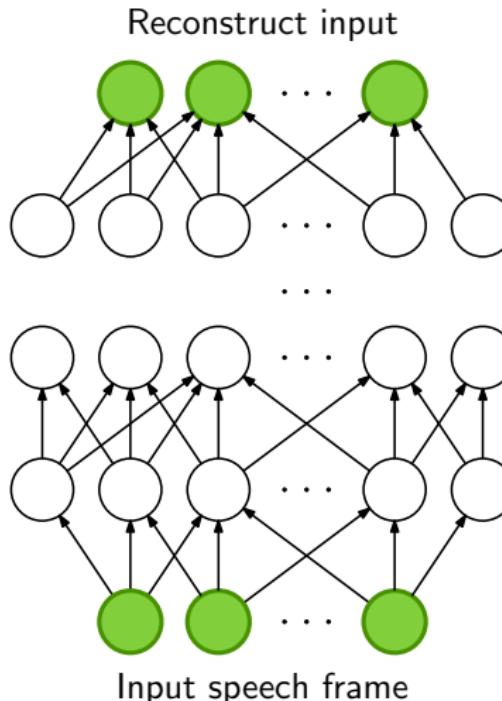
Feature extractor  $f_a(\cdot)$   
learned from data

# Autoencoder (AE) neural network



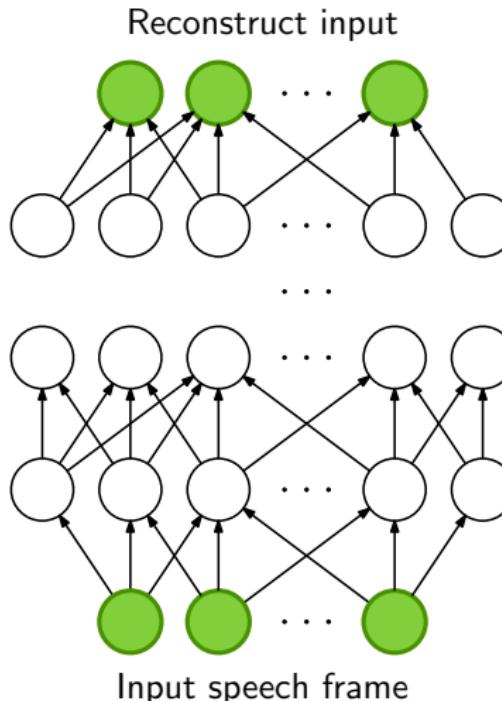
[Badino et al., ICASSP'14]

# Autoencoder (AE) neural network



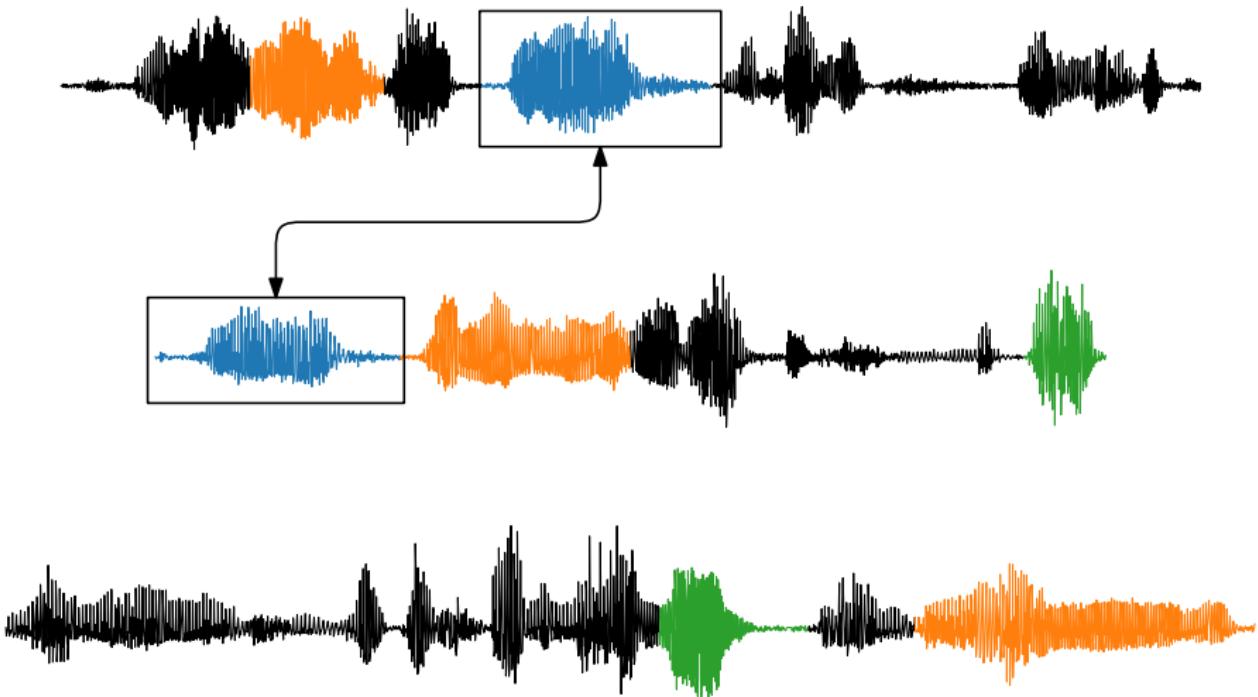
- Completely unsupervised
- But purely bottom-up
- Can we use top-down information?

# Autoencoder (AE) neural network

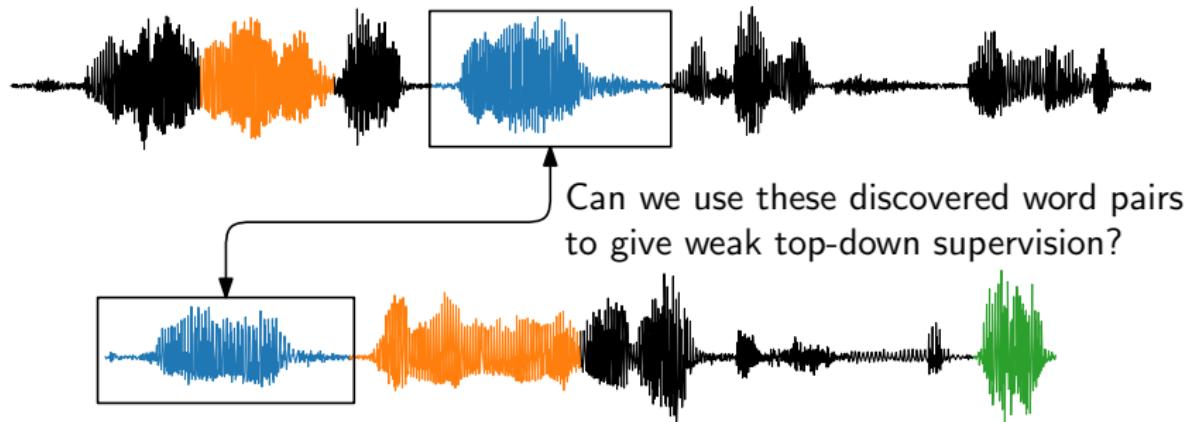


- Completely unsupervised
- But purely bottom-up
- Can we use top-down information?
- **Idea:** Unsupervised term discovery

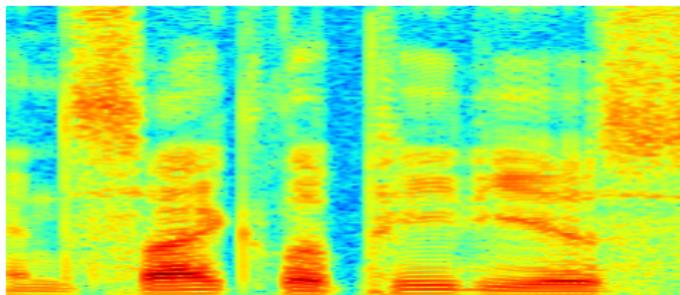
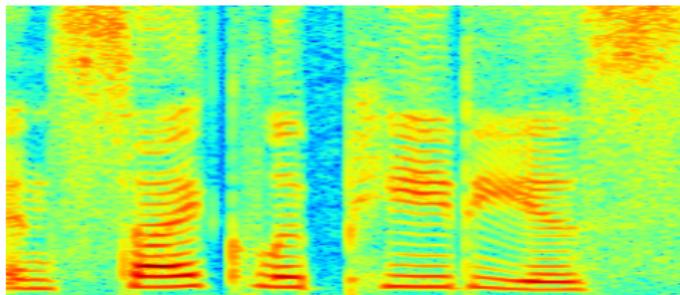
# Unsupervised term discovery (UTD)



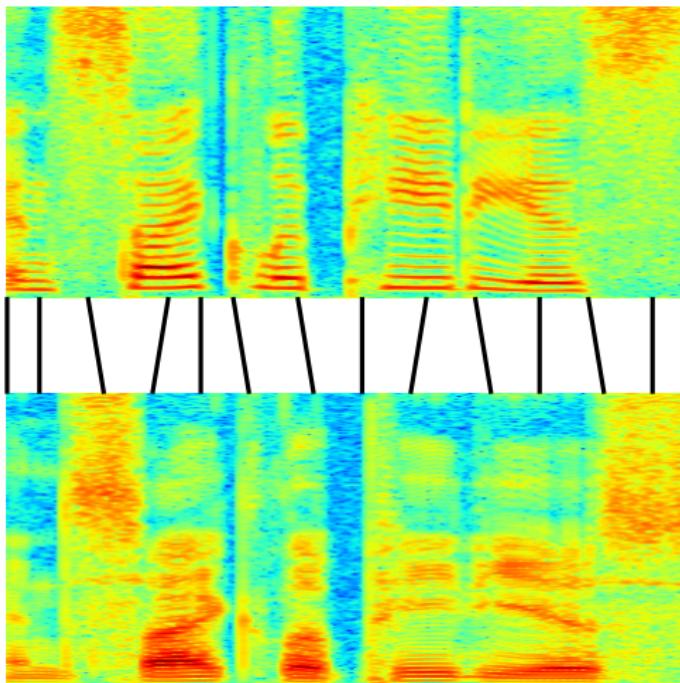
# Unsupervised term discovery (UTD)



# Weak top-down supervision: Align frames

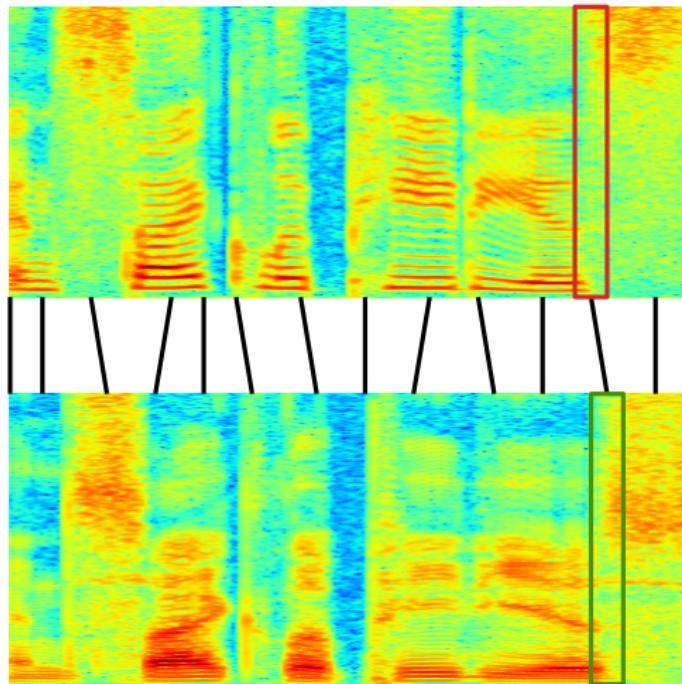


# Weak top-down supervision: Align frames

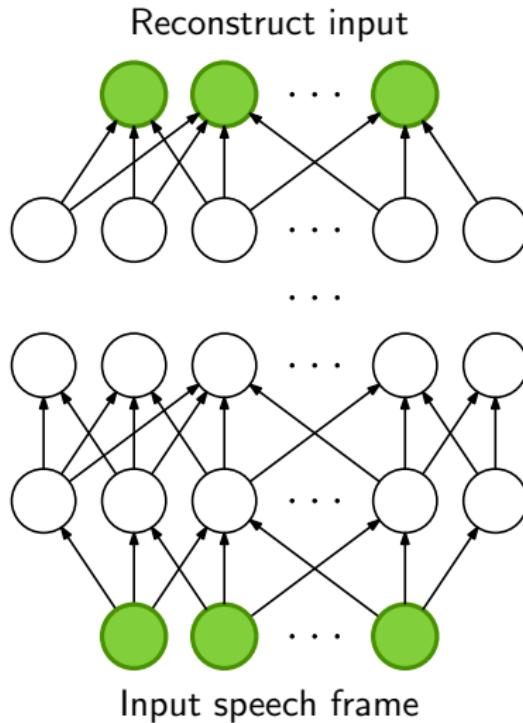


[Jansen et al., ICASSP'13]

# Weak top-down supervision: Align frames

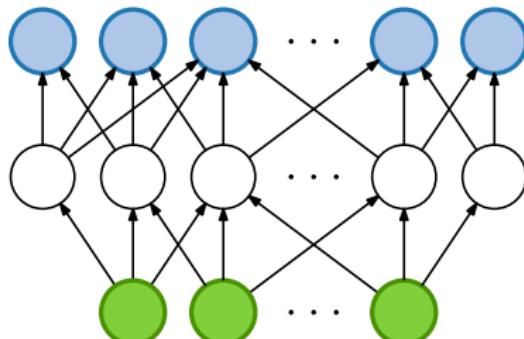
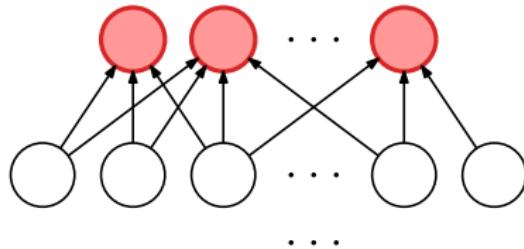


# Autoencoder (AE)



# Correspondence autoencoder (cAE)

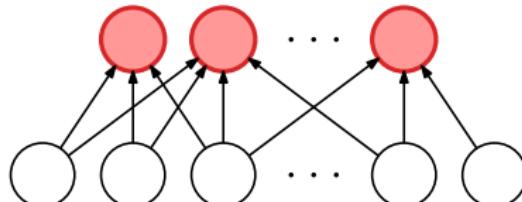
Frame from other word in pair



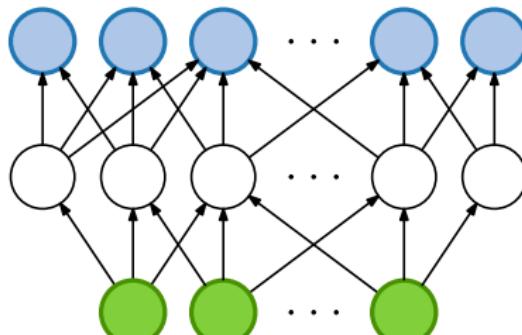
Frame from one word

# Correspondence autoencoder (cAE)

Frame from other word in pair



...



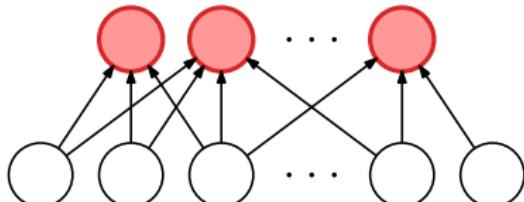
}

Unsupervised  
feature extractor  $f_a(\cdot)$

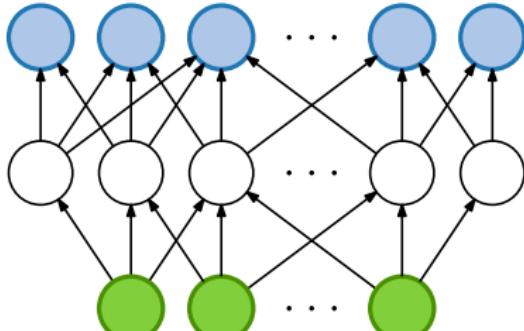
Frame from one word

# Correspondence autoencoder (cAE)

Frame from other word in pair



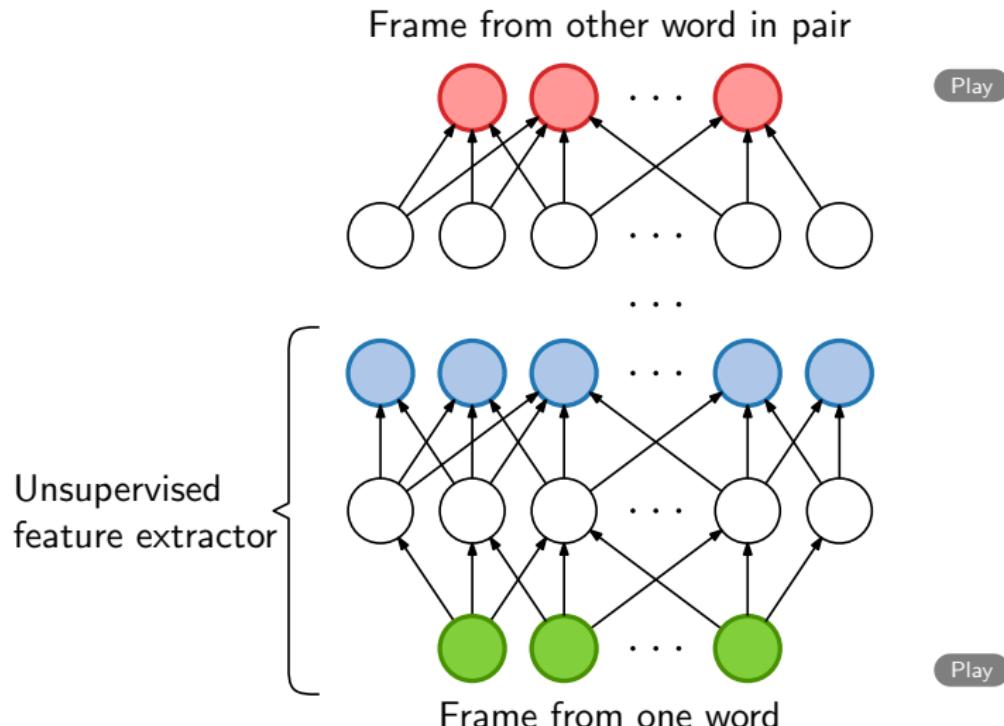
Combine **top-down** and  
**bottom-up** information



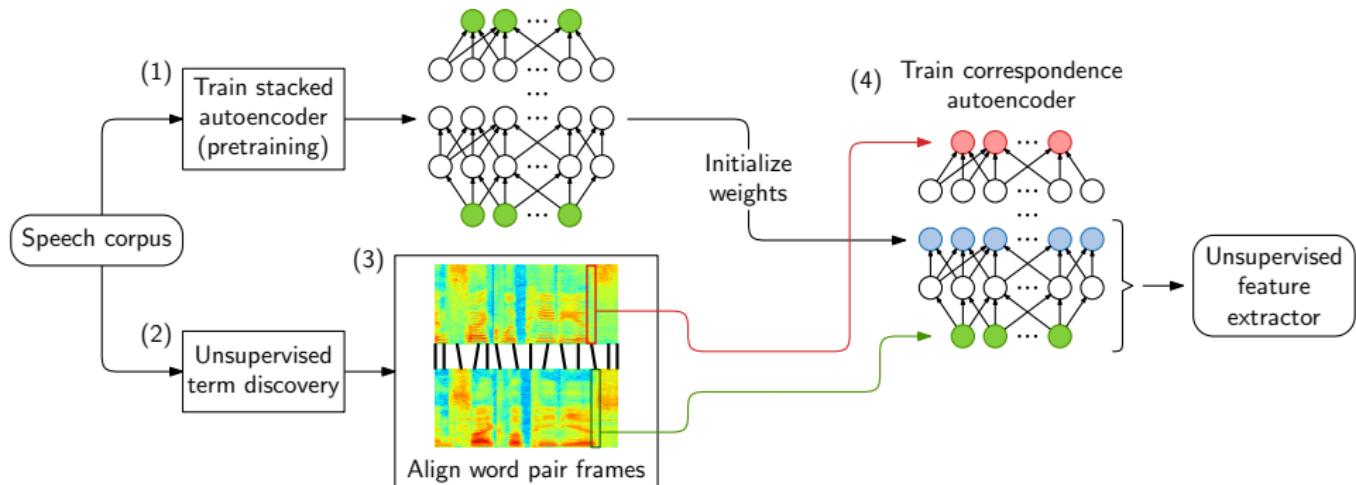
Frame from one word

} Unsupervised  
feature extractor  $f_a(\cdot)$

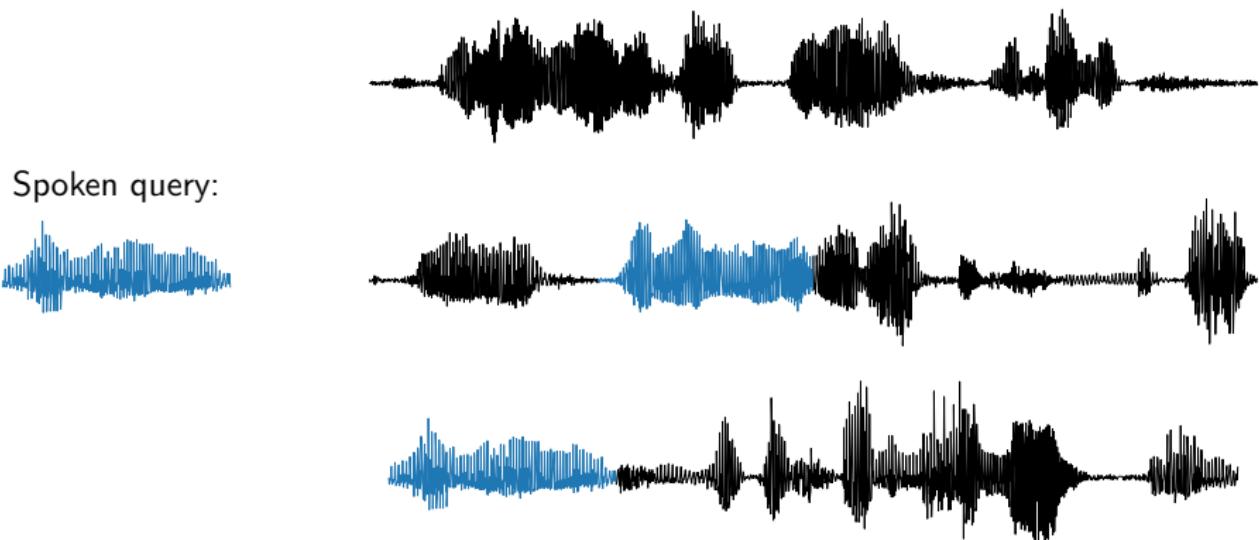
# Correspondence autoencoder (cAE)



# Correspondence autoencoder (cAE)

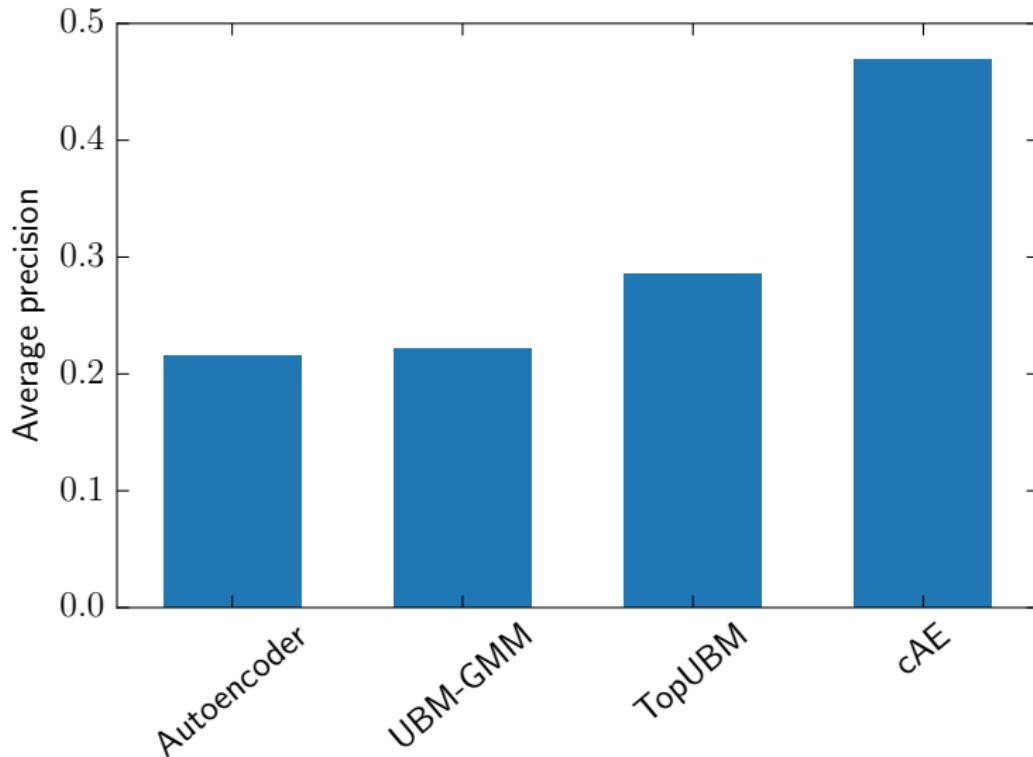


# Evaluation: Query-by-example search

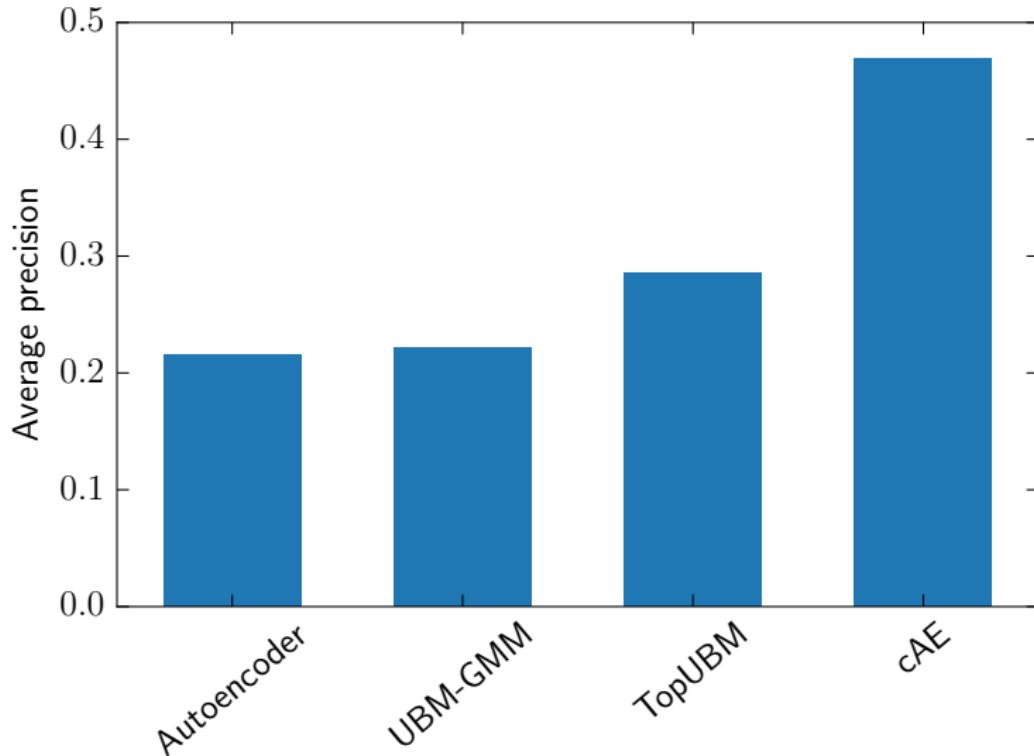


[Jansen and Van Durme, IS'12; Saeb et al., IS'17; Settle et al., IS'17]

# 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/>

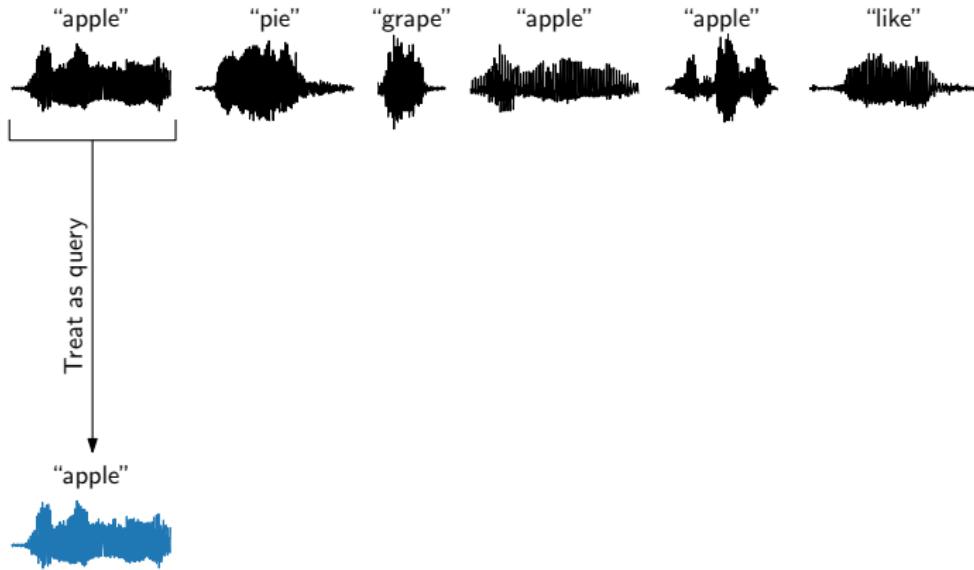
<https://github.com/kamperh>

## Evaluation of features: same-different task

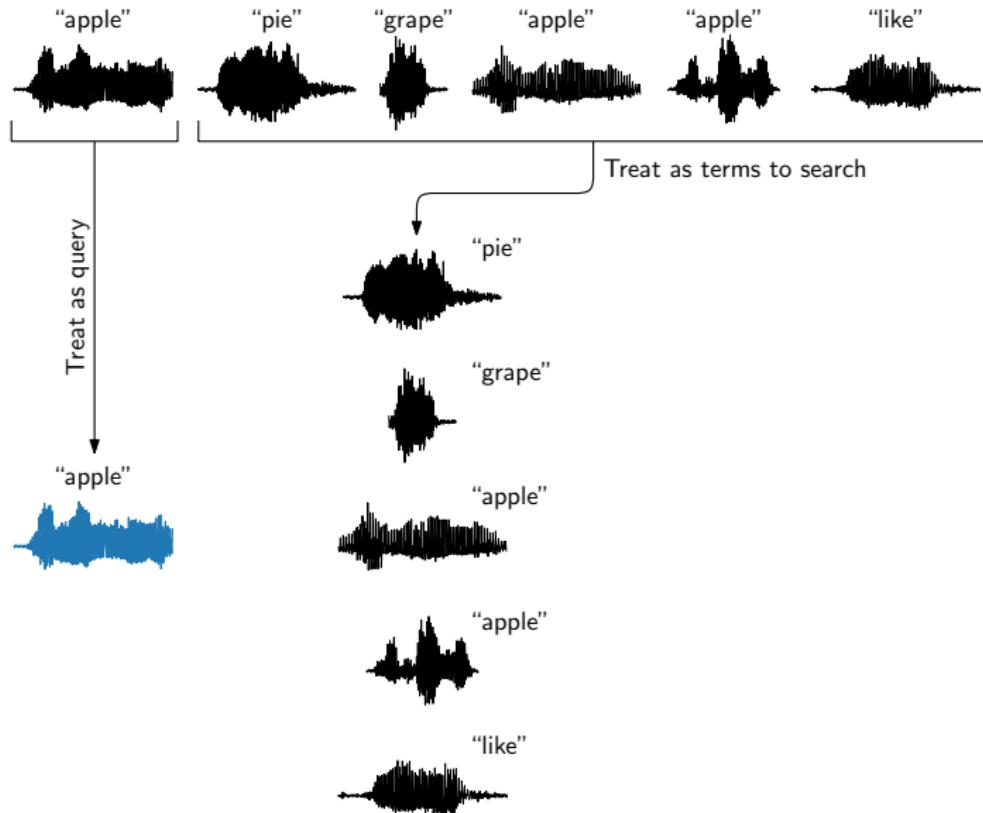
# Evaluation of features: same-different task



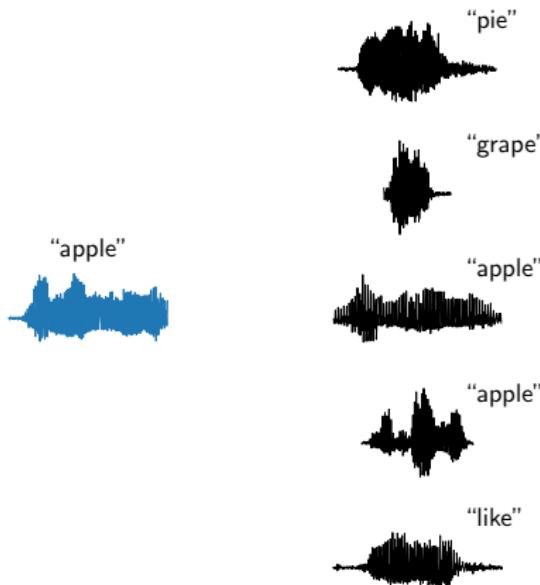
# Evaluation of features: same-different task



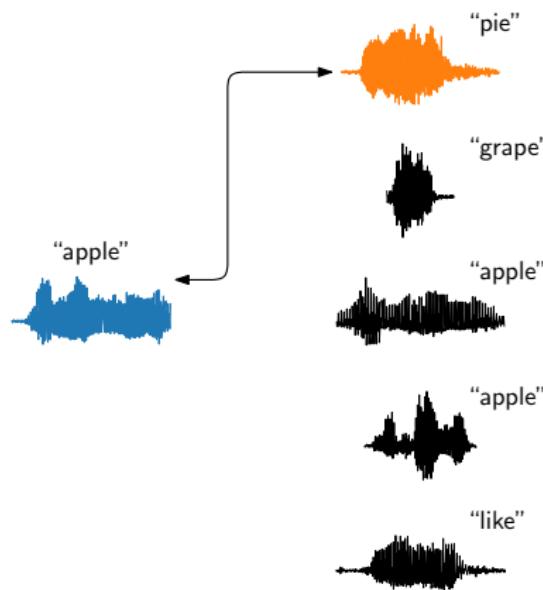
# Evaluation of features: same-different task



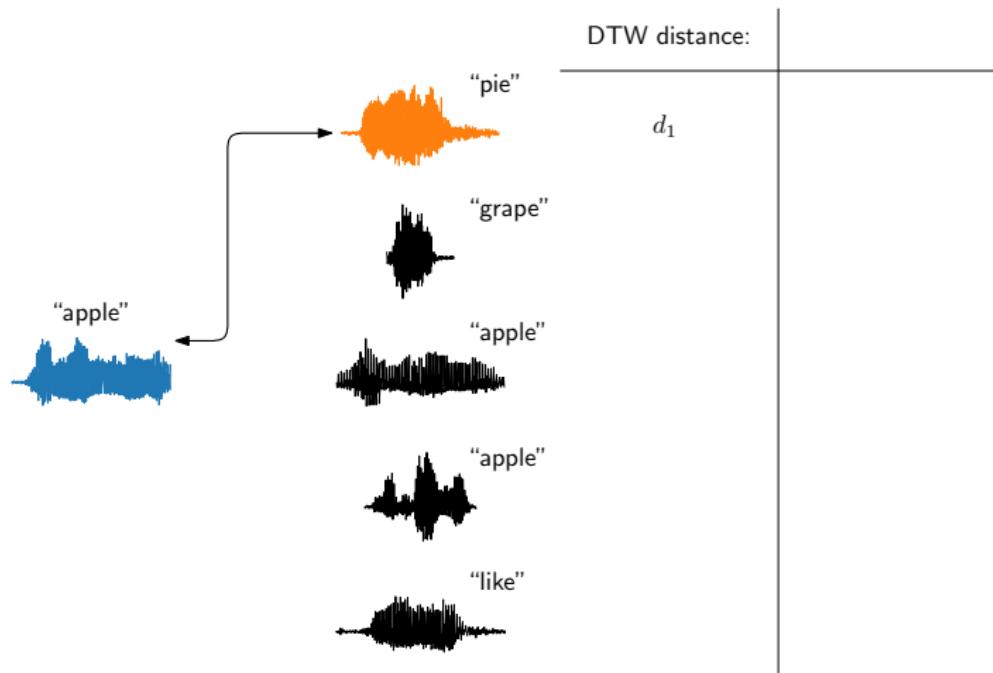
# Evaluation of features: same-different task



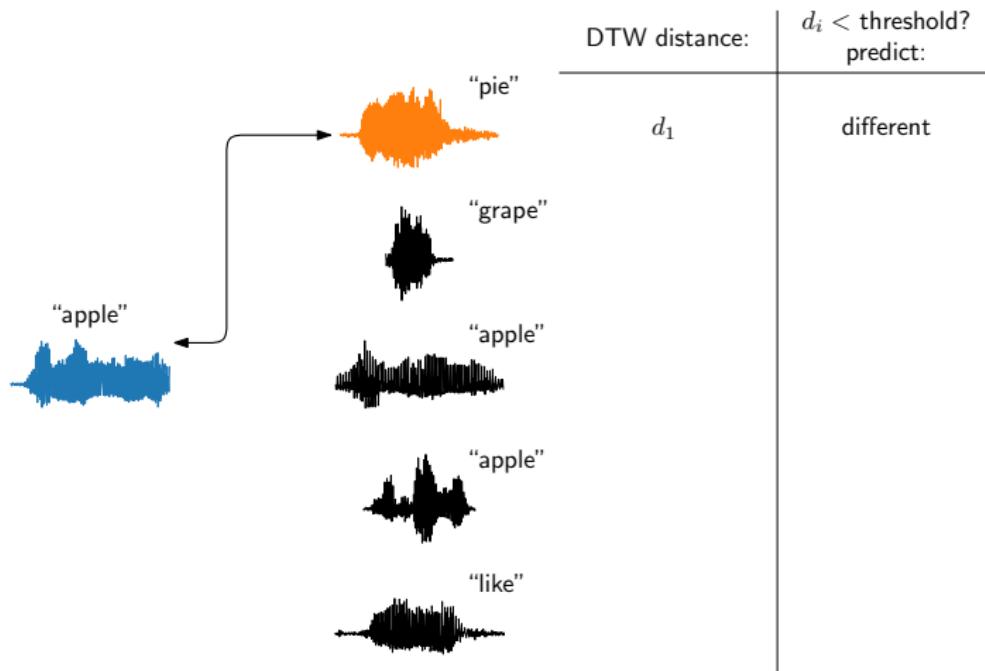
# Evaluation of features: same-different task



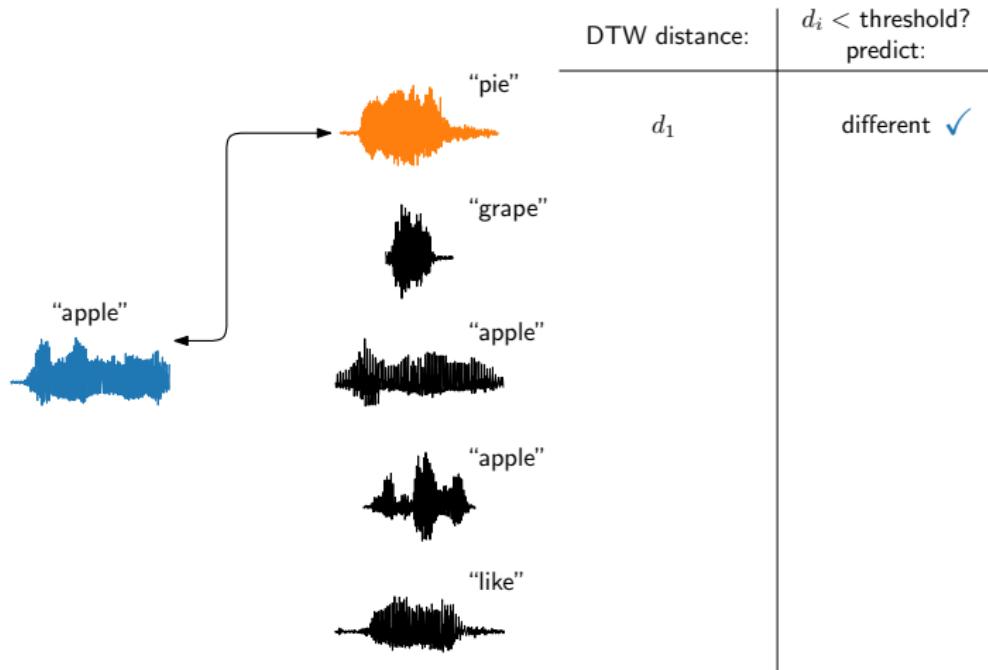
# Evaluation of features: same-different task



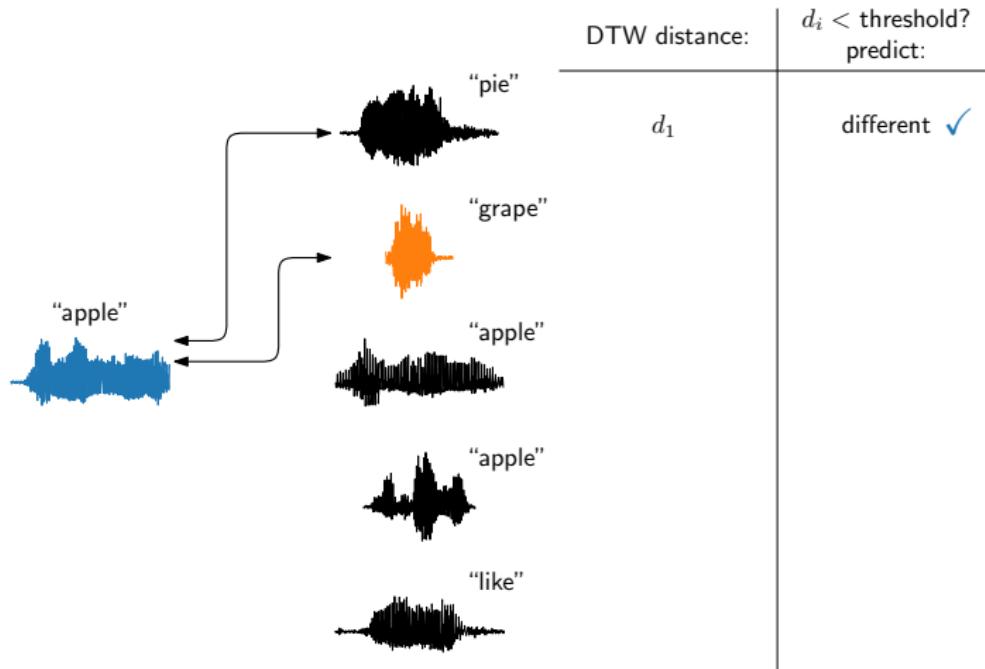
# Evaluation of features: same-different task



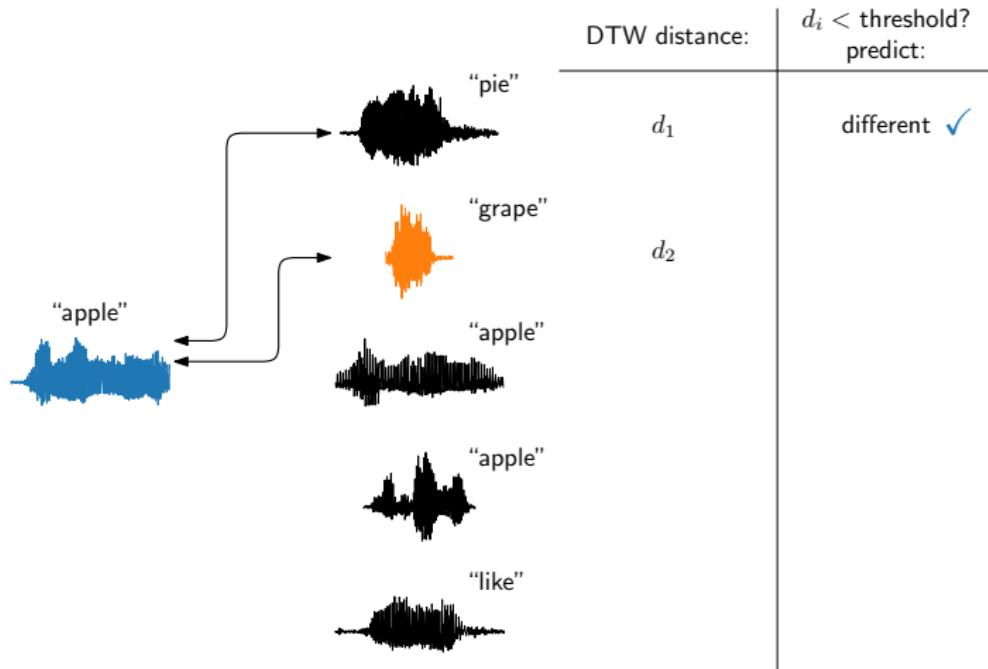
# Evaluation of features: same-different task



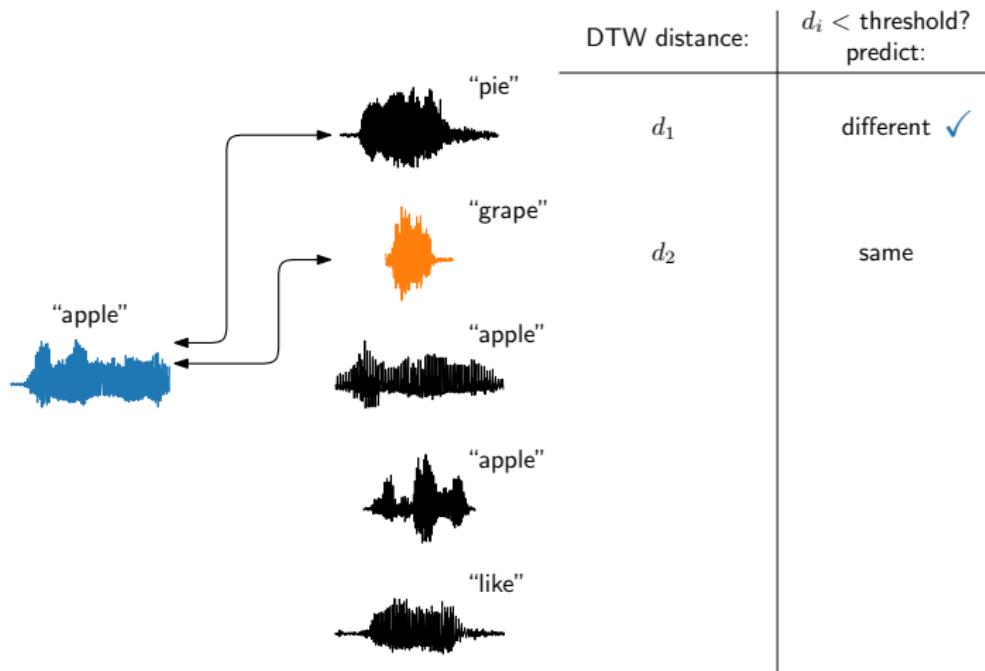
# Evaluation of features: same-different task



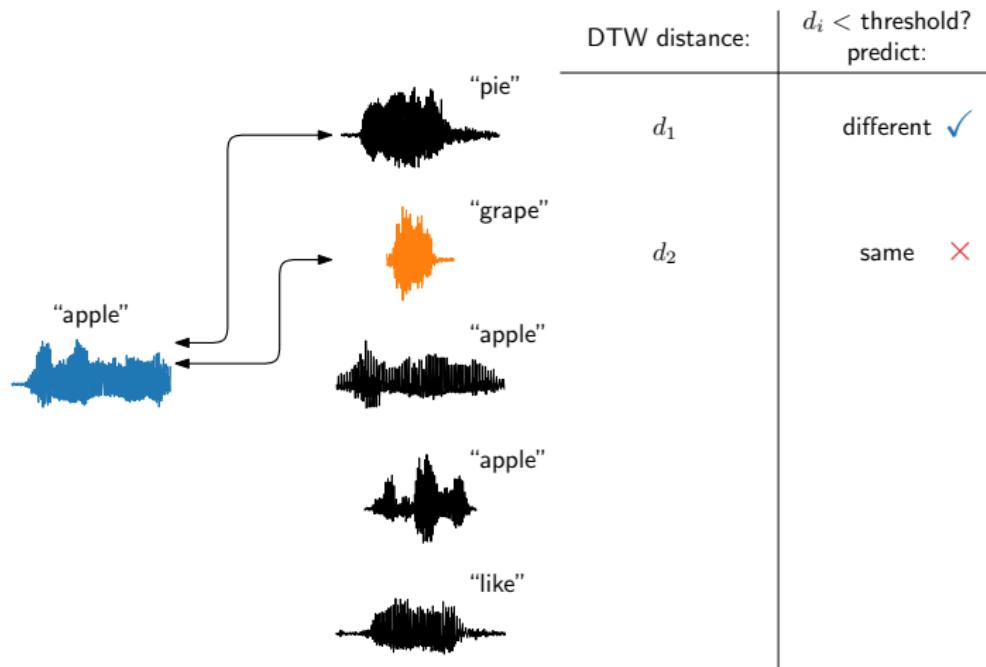
# Evaluation of features: same-different task



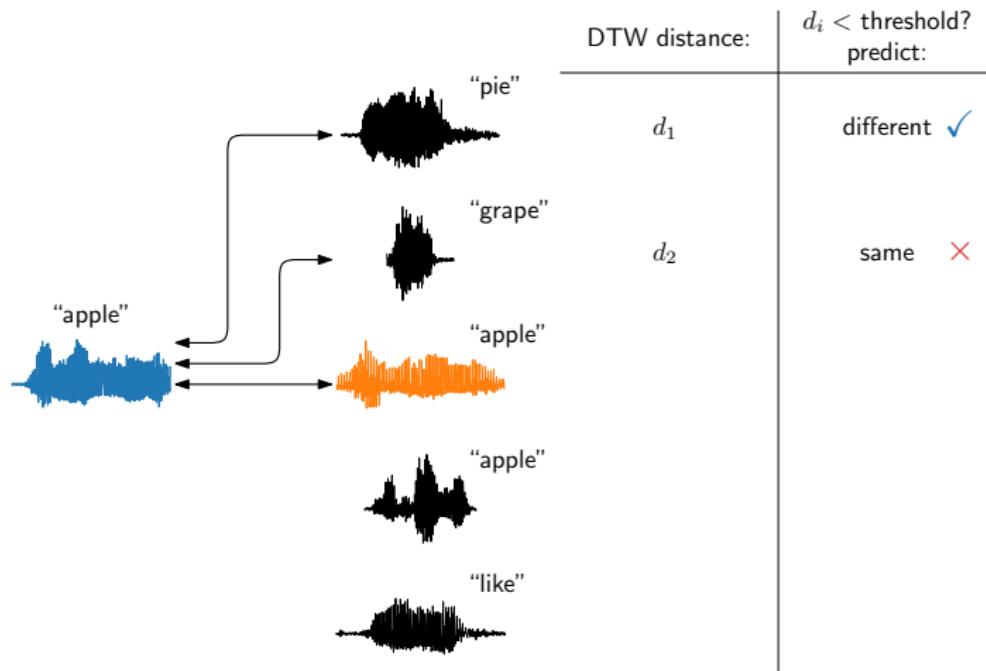
# Evaluation of features: same-different task



# Evaluation of features: same-different task



# Evaluation of features: same-different task



# Evaluation of features: same-different task



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

The diagram illustrates the DTW distance calculation between two audio features. A blue waveform labeled "apple" is compared against an orange waveform labeled "apple". Three arrows point from the blue waveform to the orange waveform, indicating the path of the dynamic programming algorithm. The resulting DTW distance is  $d_3$ . This distance is then compared against a threshold to determine if the two features are "same" or "different".

# Evaluation of features: same-different task



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

The diagram illustrates the DTW distance calculation between pairs of audio features. Arrows point from the first four waveforms to their corresponding DTW distances ( $d_1$ ,  $d_2$ ,  $d_3$ ). The first pair ("pie" vs "apple") has a large DTW distance ( $d_1$ ) and is classified as "different". The second pair ("grape" vs "apple") has a small DTW distance ( $d_2$ ) and is classified as "same" (incorrectly). The third pair ("apple" vs "apple") has a very small DTW distance ( $d_3$ ) and is correctly classified as "same".

# Evaluation of features: same-different task



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

The diagram illustrates the DTW distance calculation between pairs of audio features. Arrows point from the first four waveforms to their respective DTW distances:

- A blue arrow points from the first waveform ("apple") to  $d_1$ .
- A black arrow points from the second waveform ("pie") to  $d_1$ .
- A blue arrow points from the third waveform ("grape") to  $d_2$ .
- A black arrow points from the fourth waveform ("apple") to  $d_3$ .

The last two waveforms, "apple" and "like", are shown without arrows pointing to their DTW distances.

# Evaluation of features: same-different task



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

The table shows DTW distances between a reference waveform (blue) and other waveforms. The predict column indicates whether the system correctly identified the difference or similarity based on a threshold.

# 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`