Deep learning for (more than) speech recognition

IndabaX Western Cape, UCT, Apr. 2018

Herman Kamper

E&E Engineering, Stellenbosch University

http://www.kamperh.com/

Success in automatic speech recognition (ASR)

Success in automatic speech recognition (ASR)

CBSNEWS Video | US | World | Politics | Entertainment | Health | MoneyW

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

Microsoft says speech recognition technology reaches "human parity"

Success in automatic speech recognition (ASR)

CBSNEWS Video | US | World | Politics | Entertainment | Health | MoneyW

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

Microsoft says speech recognition technology reaches "human parity"

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

Talk outline

Talk outline

1. State-of-the-art automatic speech recognition (ASR)

Talk outline

1. State-of-the-art automatic speech recognition (ASR)
2. Examples of non-ASR speech processing

Talk outline

1. State-of-the-art automatic speech recognition (ASR)
2. Examples of non-ASR speech processing (the first rant)

Talk outline

1. State-of-the-art automatic speech recognition (ASR)
2. Examples of non-ASR speech processing (the first rant)
3. Examples of local work

Talk outline

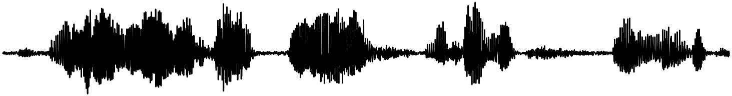
1. State-of-the-art automatic speech recognition (ASR)

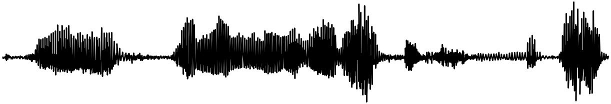
2. Examples of non-ASR speech processing (the first rant)

3. Examples of local work (a second rant)

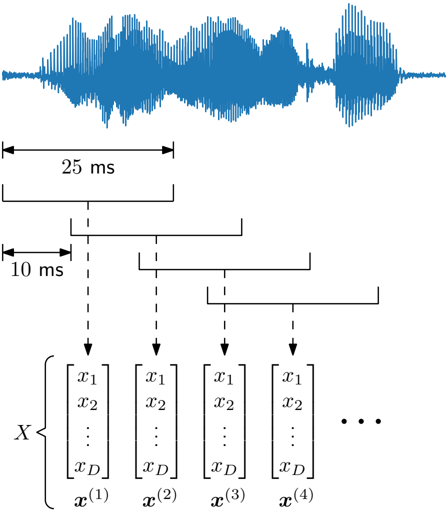
State-of-the-art speech recognition

Supervised speech recognition

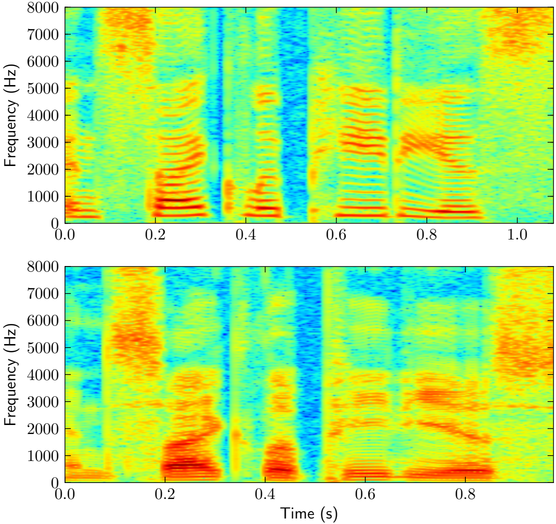




Feature extraction for speech processing



Feature extraction for speech processing



Name these networks

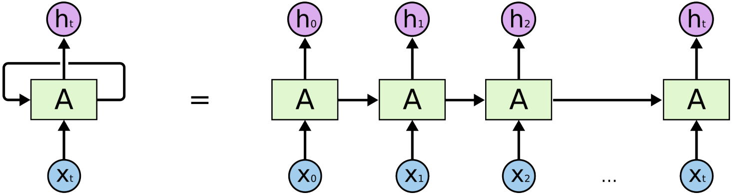
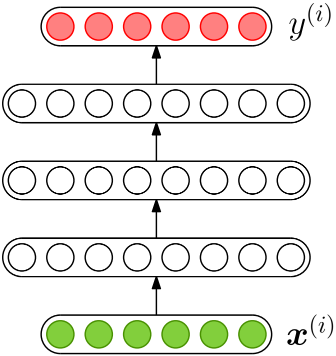


Image: `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

Name these networks



Name these networks

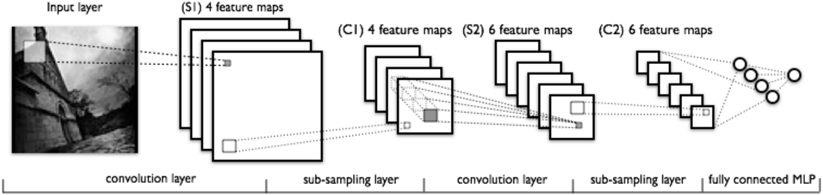
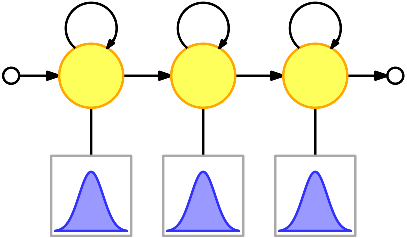


Image: `http://deeplearning.net/tutorial/lenet.html`

Name these networks

$$p(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)})|[\text{ih}])$$



A long time ago in a galaxy far,
far away. . . .

Hidden Markov models (HMMs)

Hidden Markov models (HMMs)

$$W^* = \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)})$$

Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X)
\end{aligned}
$$

Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad [\text{ “without”} = /\text{w ih th aw t}/\ ]
\end{aligned}
$$

Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \dots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad \text{[“without” = /w ih th aw t/]} \\
&= \arg\max_W \sum_U \frac{p(W, U, X)}{p(X)}
\end{aligned}
$$

Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad [\ \text{“without”} = \text{/w ih th aw t/}\ ] \\
&= \arg\max_W \sum_U \frac{p(W, U, X)}{p(X)} \\
&= \arg\max_W \sum_U p(X|W, U) P(U|W) P(W)
\end{aligned}
$$

Hidden Markov models (HMMs)

$$\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad [\text{ “without” } = \text{/w ih th aw t/ }] \\
&= \arg\max_W \sum_U \frac{p(W, U, X)}{p(X)} \\
&= \arg\max_W \sum_U p(X|W, U) P(U|W) P(W) \\
&\approx \arg\max_W \max_U p(X|W, U) P(U|W) P(W)
\end{aligned}$$

Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \ldots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad \text{[“without” = /w ih th aw t/]} \\
&= \arg\max_W \sum_U \frac{p(W, U, X)}{p(X)} \\
&= \arg\max_W \sum_U p(X|W, U) P(U|W) P(W) \\
&\approx \arg\max_W \max_U p(X|W, U) P(U|W) P(W) \\
&\approx \arg\max_W \max_U p(X|U) P(U|W) P(W)
\end{aligned}
$$

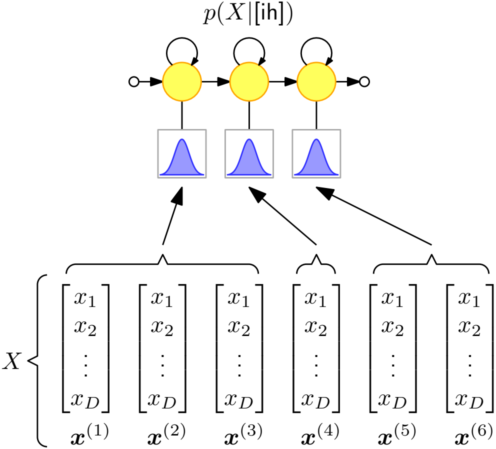
Hidden Markov models (HMMs)

$$
\begin{aligned}
W^* &= \arg\max_W P(W = w^{(1)}, w^{(2)}, \dots w^{(M)} | X = \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}) \\
&= \arg\max_W P(W|X) \\
&= \arg\max_W \sum_U P(W, U|X) \qquad \text{[“without” = /w ih th aw t/]} \\
&= \arg\max_W \sum_U \frac{p(W, U, X)}{p(X)} \\
&= \arg\max_W \sum_U p(X|W, U) P(U|W) P(W) \\
&\approx \arg\max_W \max_U p(X|W, U) P(U|W) P(W) \\
&\approx \arg\max_W \max_U p(X|U) P(U|W) P(W)
\end{aligned}
$$

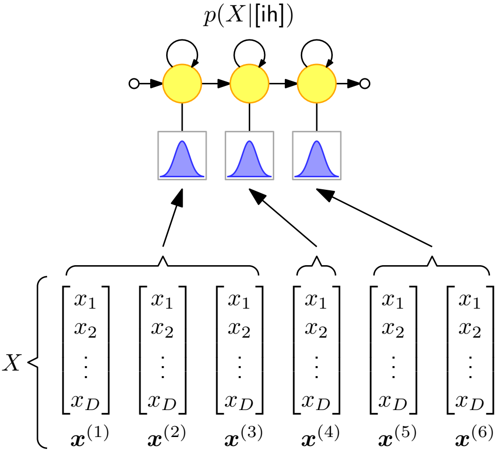
$p(X|U)$: acoustic model $\quad P(U|W)$: pronunciation dictionary

$P(W)$: language model

Hidden Markov models (HMMs)



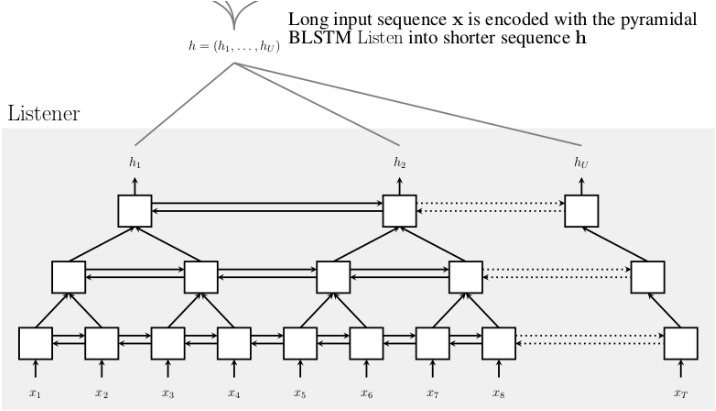
Hidden Markov models (HMMs)



Speech recognition was performed by combining acoustic model (thousands of HMM states) with pronunciation dictionary and language model in (very big) decoder network (finite state machine).

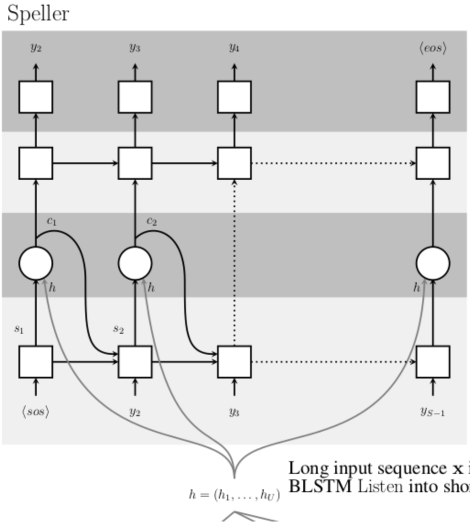
Back to today: End-to-end speech recognition

Back to today: End-to-end speech recognition



[Chan et al., arXiv'15]

End-to-end speech recognition



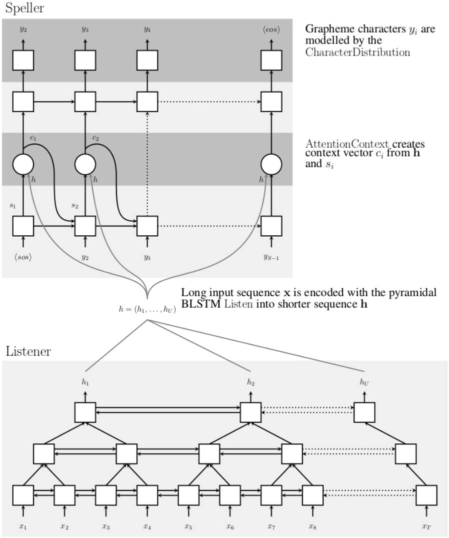
Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from $\mathbf{h}$ and s_i

Long input sequence $\mathbf{x}$ is encoded with the pyramidal BLSTM Listen into shorter sequence $\mathbf{h}$

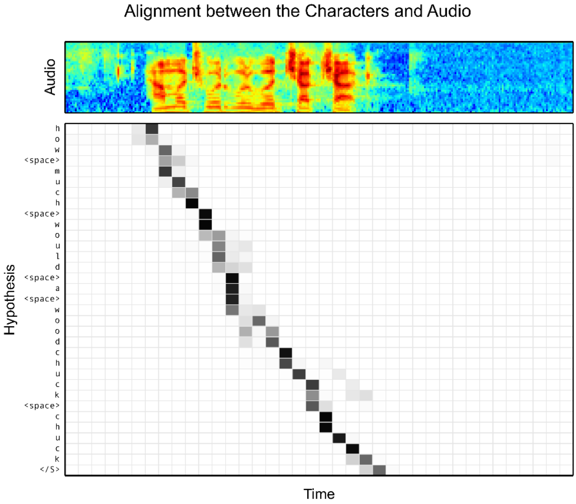
[Chan et al., arXiv'15]

End-to-end speech recognition



[Chan et al., arXiv'15]

End-to-end speech recognition



[Chan et al., arXiv'15]

Why did we talk about HMMs?

- Could we use a standard feedforward deep neural network (DNN) for ASR?

Why did we talk about HMMs?

- Could we use a standard feedforward deep neural network (DNN) for ASR?
- Idea: Use HMM to obtain frame alignments for DNN!

Why did we talk about HMMs?

- Could we use a standard feedforward deep neural network (DNN) for ASR?

- Idea: Use HMM to obtain frame alignments for DNN!

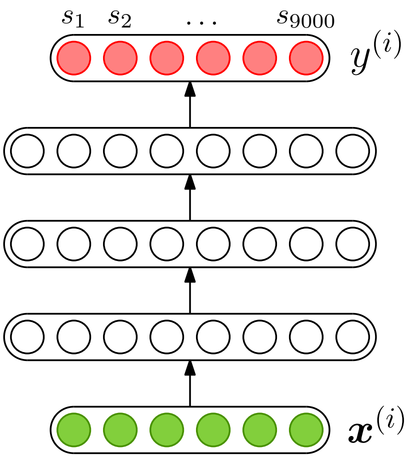
- Hybrid model: DNN-HMM

Why did we talk about HMMs?

- Could we use a standard feedforward deep neural network (DNN) for ASR?
- Idea: Use HMM to obtain frame alignments for DNN!
- Hybrid model: DNN-HMM
- Can be seen as representation learning trained jointly with classifier

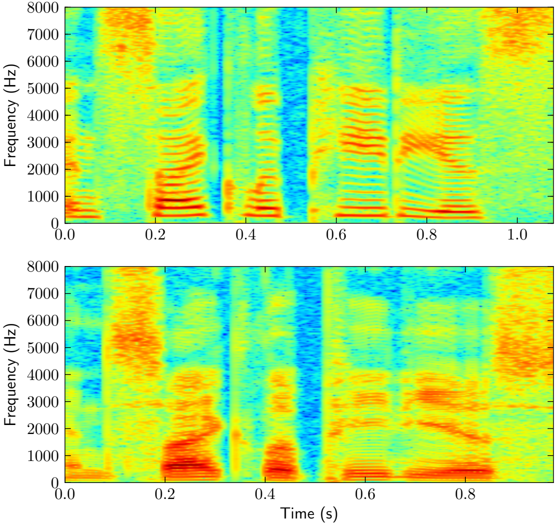
Why did we talk about HMMs?

- Could we use a standard feedforward deep neural network (DNN) for ASR?
- Idea: Use HMM to obtain frame alignments for DNN!
- Hybrid model: DNN-HMM
- Can be seen as representation learning trained jointly with classifier



What about convolutional neural networks?

What about convolutional neural networks?



Is end-to-end the best?

- End-to-end models are easier to implement[1]
- But, do they give state-of-the-art performance?

Is end-to-end the best?

- End-to-end models are easier to implement[1]
- But, do they give state-of-the-art performance?
- What do you think CLDNN-HMM[2] stands for?

Is end-to-end the best?

- End-to-end models are easier to implement[1]
- But, do they give state-of-the-art performance?
- What do you think CLDNN-HMM[2] stands for?

[1] `https://github.com/espnet/espnet` [2] [Sainath et al., ICASSP'15]

Is end-to-end the best?

- End-to-end models are easier to implement[1]
- But, do they give state-of-the-art performance?
- What do you think CLDNN-HMM[2] stands for?

CONVOLUTIONAL, LONG SHORT-TERM MEMORY, FULLY CONNECTED DEEP NEURAL NETWORKS

Tara N. Sainath, Oriol Vinyals, Andrew Senior, Haşim Sak

Google, Inc., New York, NY, USA
{tsainath, vinyals, andrewsenior, hasim}@google.com

[1] `https://github.com/espnet/espnet` [2] [Sainath et al., ICASSP'15]

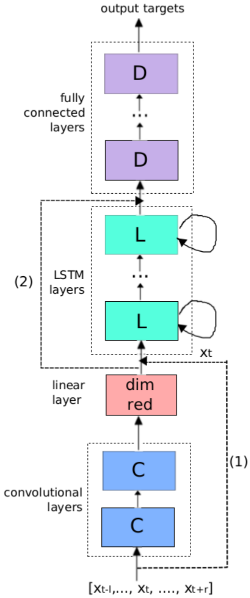
Is end-to-end the best?

- End-to-end models are easier to implement[1]
- But, do they give state-of-the-art performance?
- What do you think CLDNN-HMM[2] stands for?

CONVOLUTIONAL, LONG SHORT-TERM MEMORY, FULLY CONNECTED DEEP NEURAL NETWORKS

Tara N. Sainath, Oriol Vinyals, Andrew Senior, Haşim Sak

Google, Inc., New York, NY, USA
{tsainath, vinyals, andrewsenior, hasim}@google.com



[1] https://github.com/espnet/espnet [2] [Sainath et al., ICASSP'15]

Summary: Speech recognition is important, but. . .

- Very important engineering endeavour:
  information access, illiteracy, assistance for the disabled

Summary: Speech recognition is important, but...

- Very important engineering endeavour:
  information access, illiteracy, assistance for the disabled
- But it is more: speech and language makes us human

Summary: Speech recognition is important, but...

- Very important engineering endeavour:
  information access, illiteracy, assistance for the disabled
- But it is more: speech and language makes us human
- Engineering decisions can tell us something about how we perceive the world: saw how structure helps in speech recognition models

Summary: Speech recognition is important, but...

- Very important engineering endeavour:
  information access, illiteracy, assistance for the disabled
- But it is more: speech and language makes us human
- Engineering decisions can tell us something about how we perceive the world: saw how structure helps in speech recognition models
- And studies about how we perceive the world can tell us something about better engineering decisions

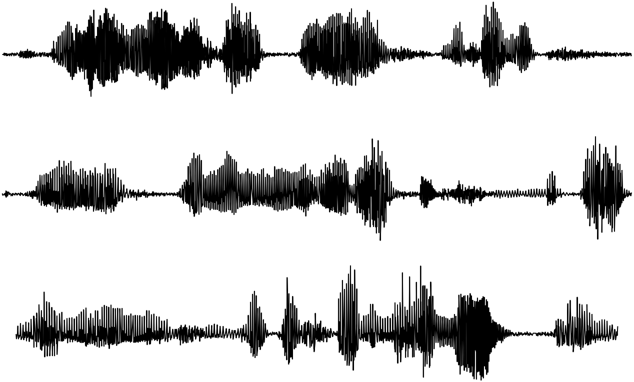
Rant 1: Do we always need/have ASR?

Examples of non-ASR speech processing

What if we do not have supervision?

- Google Voice: English, Spanish, German, ..., Zulu ($\sim$50 languages)
- Data: 2000 hours transcribed speech audio; $\sim$350M/560M words text
- Can we do this for all 7000 languages spoken in the world?
- Many of these languages are endangered and unwritten

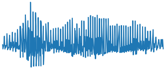
Example 1: Query-by-example search

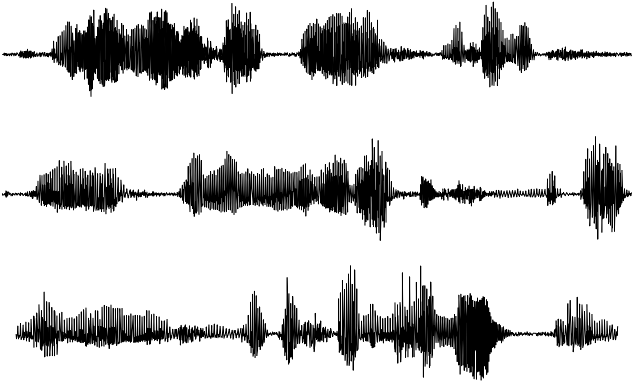


[Jansen and Van Durme, Interspeech'12]

Example 1: Query-by-example search

Spoken query:

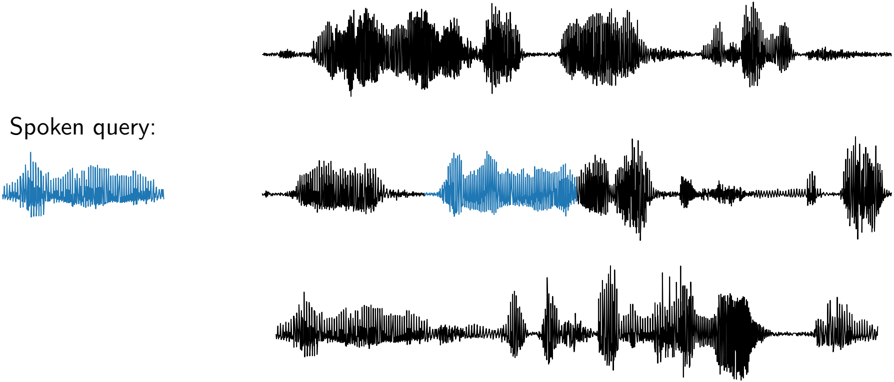




[Jansen and Van Durme, Interspeech'12]

Example 1: Query-by-example search

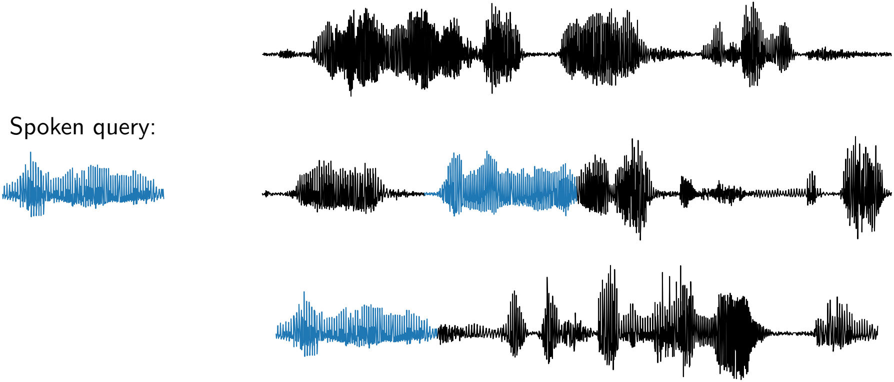
Spoken query:



[Jansen and Van Durme, Interspeech'12]

Example 1: Query-by-example search

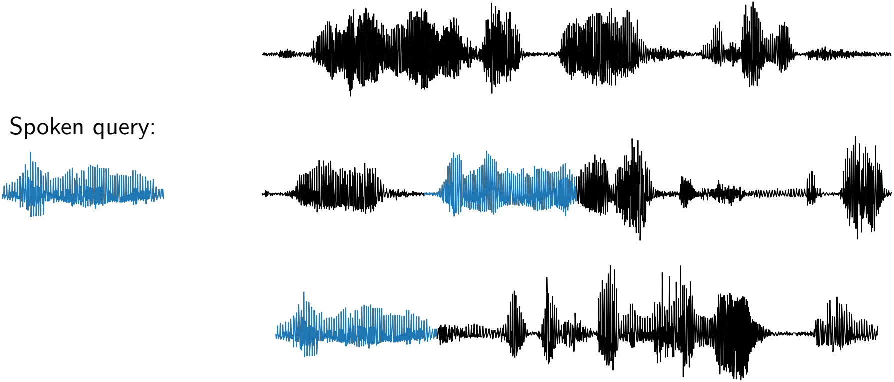
Spoken query:



[Jansen and Van Durme, Interspeech'12]

Example 1: Query-by-example search

Spoken query:



Useful speech system, not requiring any transcribed speech

[Jansen and Van Durme, Interspeech'12]

Example 2: Linguistic and cultural documentation

http://www.stevenbird.net/

Example 2: Linguistic and cultural documentation

Academics team up to save dying languages

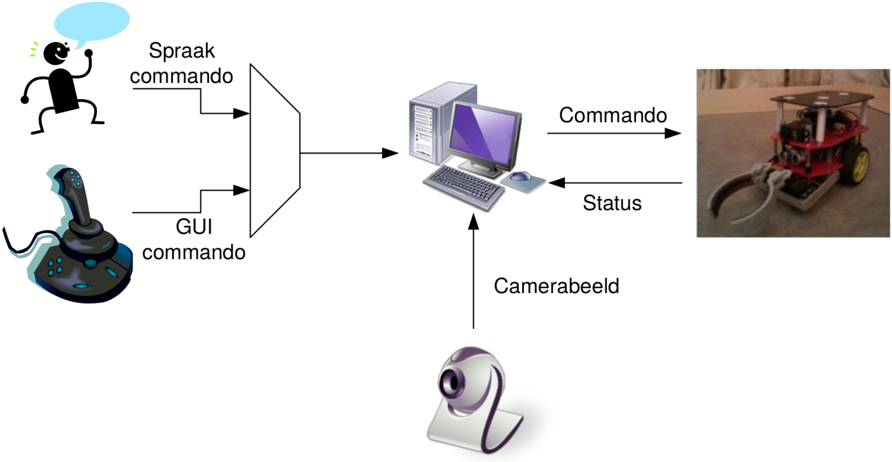
25/3/2014

A beautifully crafted documentary about Aikuma by Thom Cookes which aired on ABC's program *The World*. This video included a segment about Lauren Gawne and her work on Kagate (Nepal).

Like 0 Tweet 0

http://www.stevenbird.net/

Example 3: Learning robots to understand speech

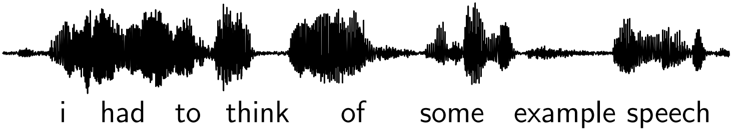


[Janssens and Renkens, 2014]; [Renkens et al., SLT'14]

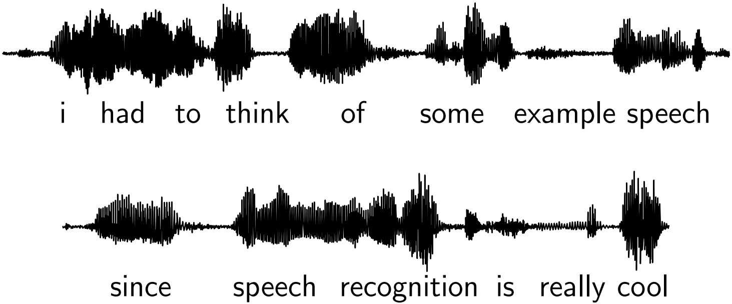
Rant 2: Taking inspiration from humans

Examples of local work

Supervised speech recognition



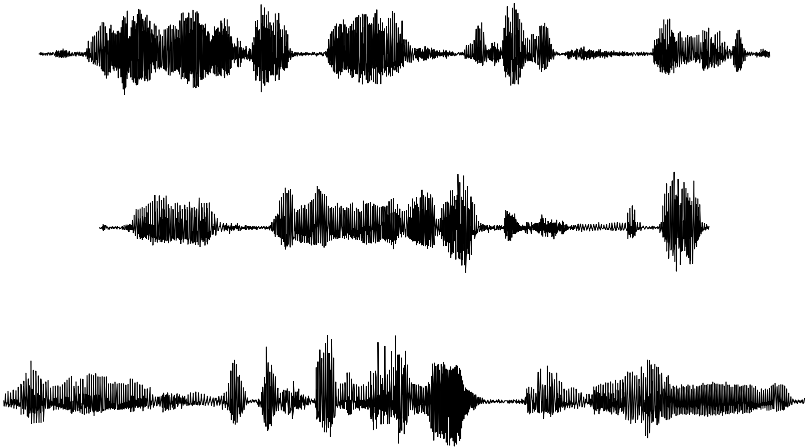
Supervised speech recognition



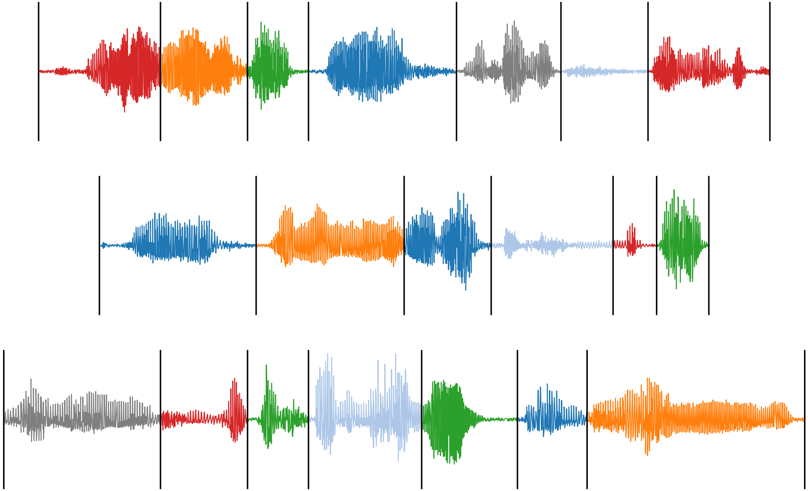
Can we acquire language from audio alone?

Full-coverage segmentation and clustering

Full-coverage segmentation and clustering



Full-coverage segmentation and clustering



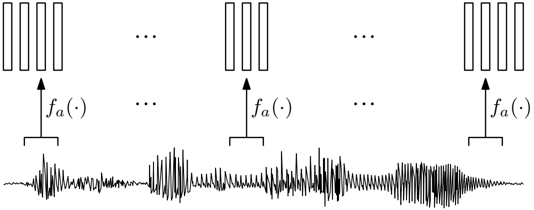
Unsupervised segmental Bayesian model

Speech waveform

Unsupervised segmental Bayesian model

Acoustic frames $\mathbf{y}_{1:M}$

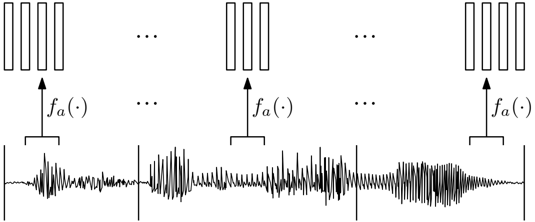
Speech waveform



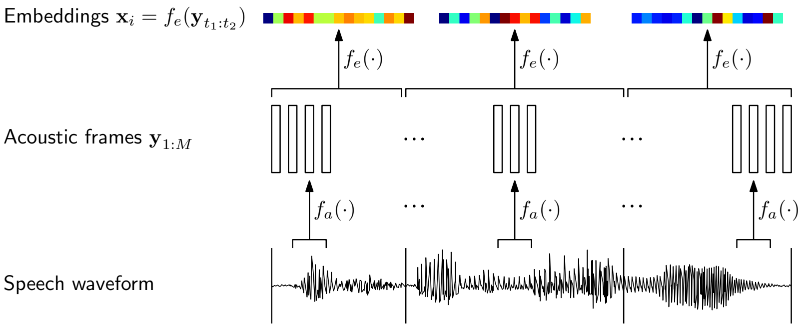
Unsupervised segmental Bayesian model

Acoustic frames $\mathbf{y}_{1:M}$

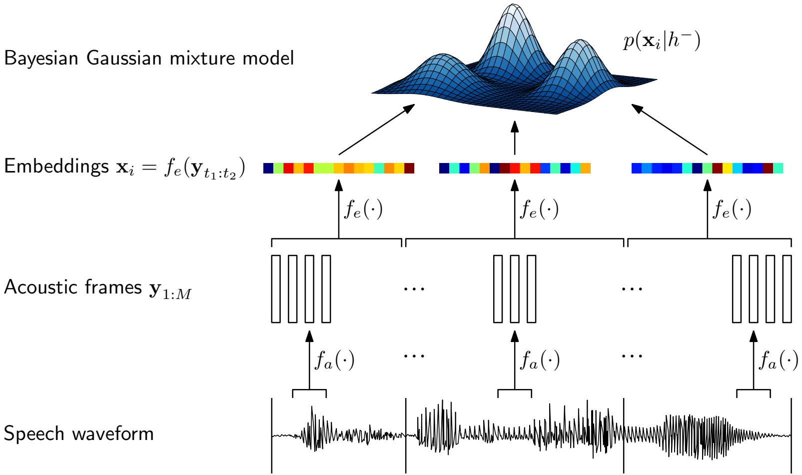
Speech waveform



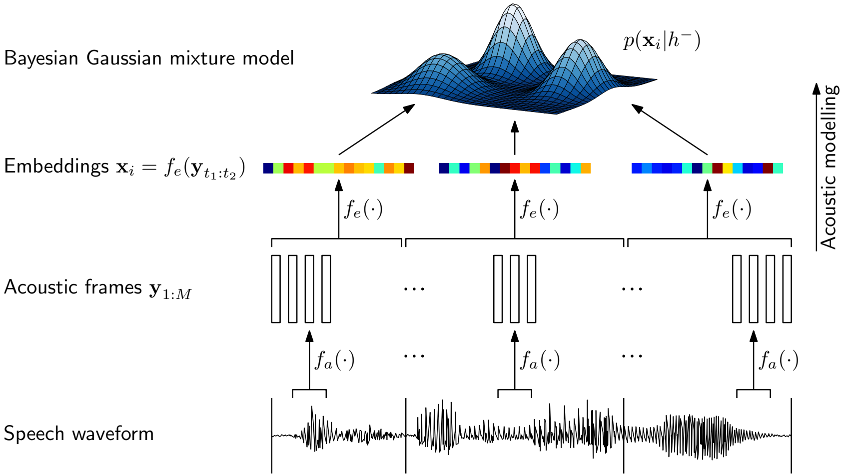
Unsupervised segmental Bayesian model



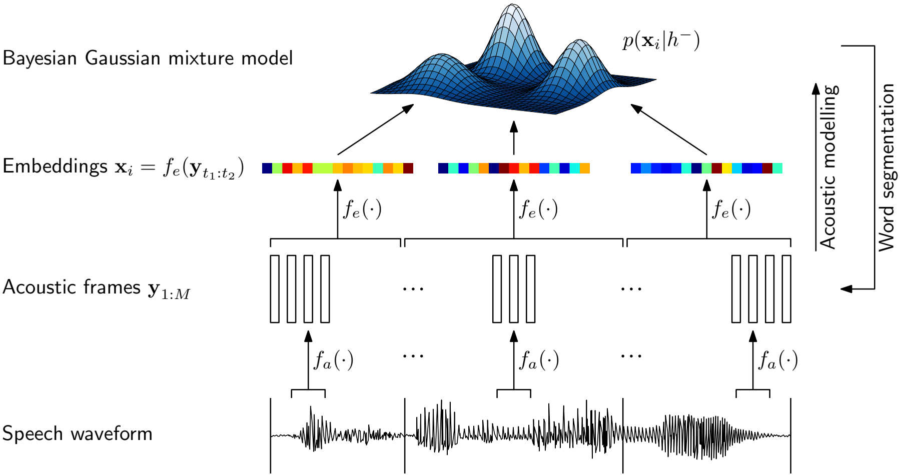
Unsupervised segmental Bayesian model



Unsupervised segmental Bayesian model



Unsupervised segmental Bayesian model



Listen to discovered clusters

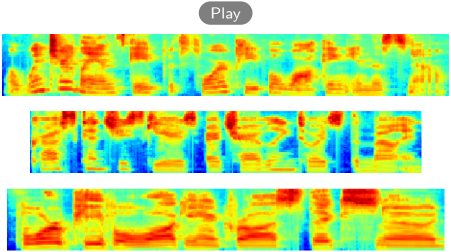
- Small-vocabulary cluster 45: Play
- Large-vocabulary English cluster 1214: Play
- Large-vocabulary Xitsonga cluster 629: Play

Arrival

Using images for grounding language

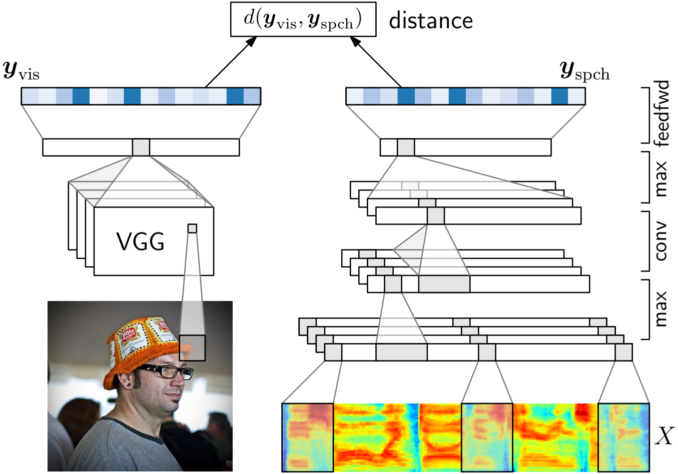
Using images for grounding language

Consider images paired with unlabellel spoken captions:



Map images and speech into common space

Map images and speech into common space



[Harwath et al., NIPS'16]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	Play (one of top 10)	
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind		
bike		
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech’17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	
bike		
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	mistake
bike		
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys		
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	Play	
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large		
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	Play	
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach . . .	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	. . . a rocky cliff overlooking a body of water	
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	Example of matched utterance	Type
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	... a rocky cliff overlooking a body of water	semantic
play		
sitting		
yellow		
young		

[Kamper et al., Interspeech'17]

Visually grounded keyword spotting

Keyword	**Example of matched utterance**	**Type**
beach	a boy in a yellow shirt is walking on a beach ...	correct
behind	a surfer does a flip on a wave	mistake
bike	a dirt biker flies through the air	variant
boys	two children play soccer in the park	semantic
large	... a rocky cliff overlooking a body of water	semantic
play	children playing in a ball pit	variant
sitting	two people are seated at a table with drinks	semantic
yellow	a tan dog jumping over a red and blue toy	mistake
young	a little girl on a kid swing	semantic

[Kamper et al., Interspeech'17]

Summary and conclusion

What did we chat about today?

- Supervised speech recognition: From HMMs all the way to CLDNNs
- Structure is still important in speech recognition
- Saw three examples of models that do not require ASR
- Looked at local work taking inspiration from humans

What's next (specifically for us)?

What's next (specifically for us)?

- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing

What's next (specifically for us)?

- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing
- Building speech search systems for (South) African languages

What's next (specifically for us)?

- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing
- Building speech search systems for (South) African languages
- Can some of these approaches be used in other machine learning domains? E.g. can vision tell us something about speech?

What's next (specifically for us)?

- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing
- Building speech search systems for (South) African languages
- Can some of these approaches be used in other machine learning domains? E.g. can vision tell us something about speech?
- What can we learn about language acquisition in humans?

What's next (specifically for us)?

- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing
- Building speech search systems for (South) African languages
- Can some of these approaches be used in other machine learning domains? E.g. can vision tell us something about speech?
- What can we learn about language acquisition in humans?
- Language acquisition in robots

What's next (specifically for us)?

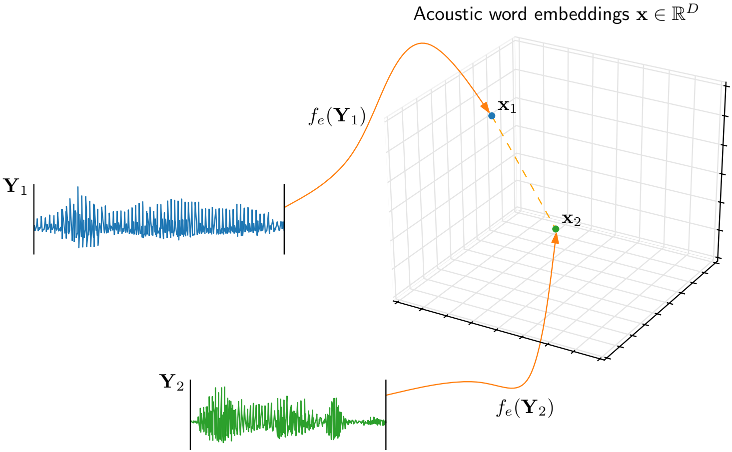
- Still many many unsolved core machine learning problems in unsupervised and low-resource speech processing
- Building speech search systems for (South) African languages
- Can some of these approaches be used in other machine learning domains? E.g. can vision tell us something about speech?
- What can we learn about language acquisition in humans?
- Language acquisition in robots
- **Main take-away:** Look at machine learning problems from different perspectives and angles

http://www.kamperh.com/

https://github.com/kamperh

Backup slides

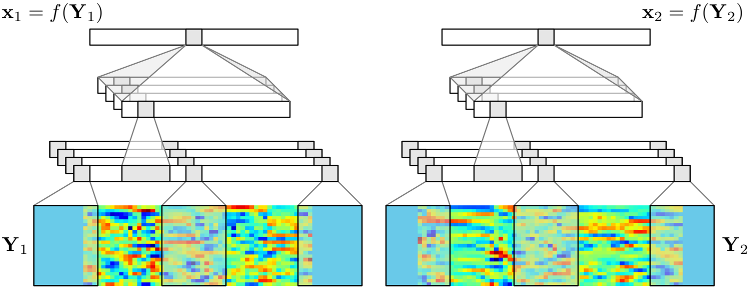
Acoustic word embeddings (AWê)



[Levin et al., ASRU'13]

Word similarity Siamese CNN

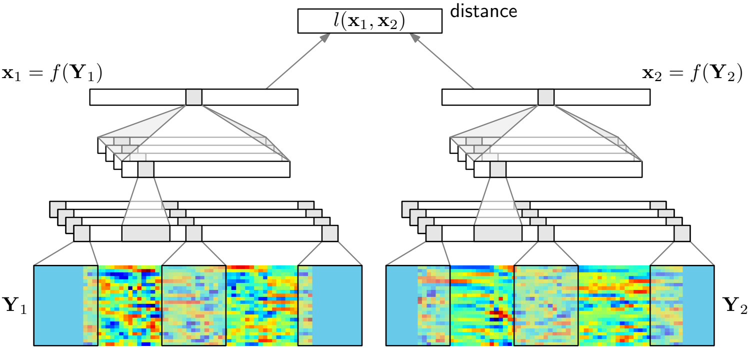
Use idea of **Siamese networks** [Bromley et al., PatRec'93]



[Kamper et al., ICASSP'15]

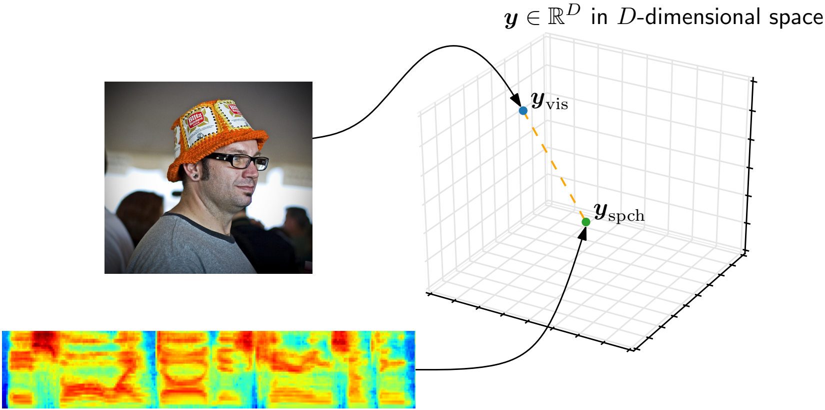
Word similarity Siamese CNN

Use idea of **Siamese networks** [Bromley et al., PatRec'93]



[Kamper et al., ICASSP'15]

Retrieval in common (semantic) space



[Harwath et al., NIPS'16]

Word prediction from images and speech

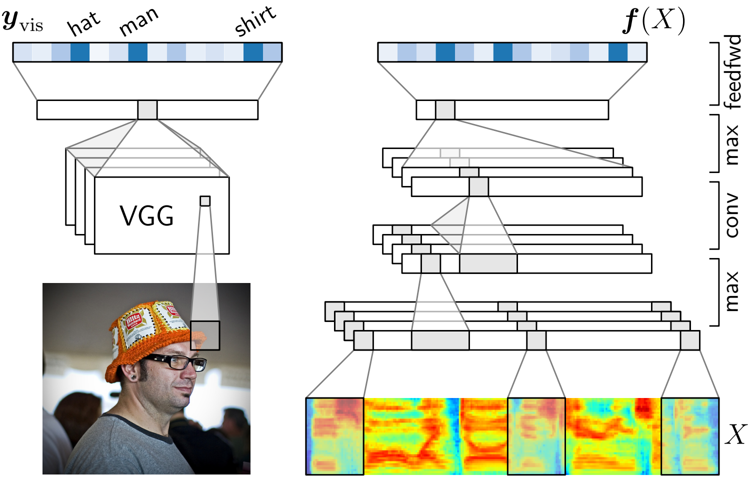
Word prediction from images and speech

[Kamper et al., Interspeech'17]

Word prediction from images and speech

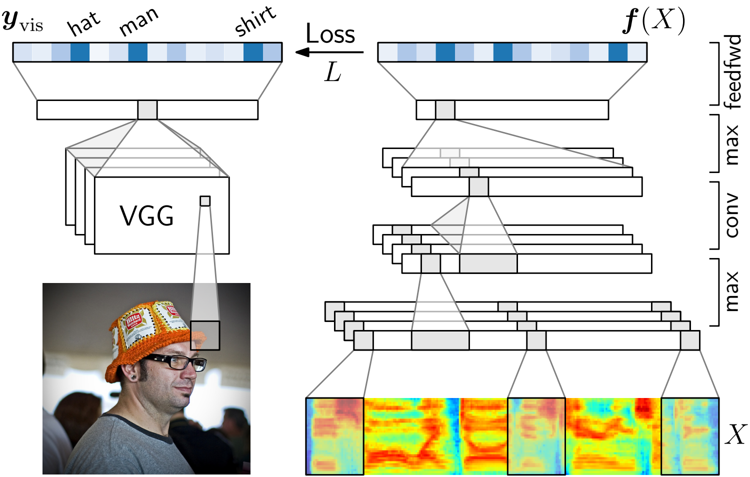
[Kamper et al., Interspeech'17]

Word prediction from images and speech



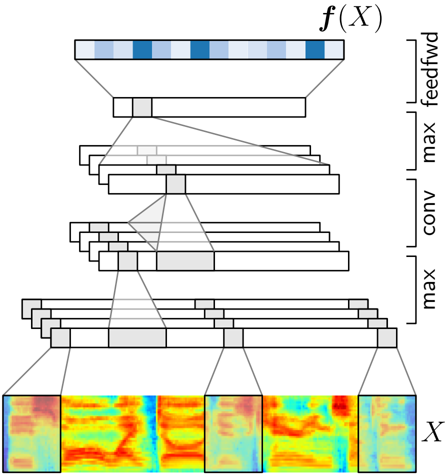
[Kamper et al., Interspeech'17]

Word prediction from images and speech



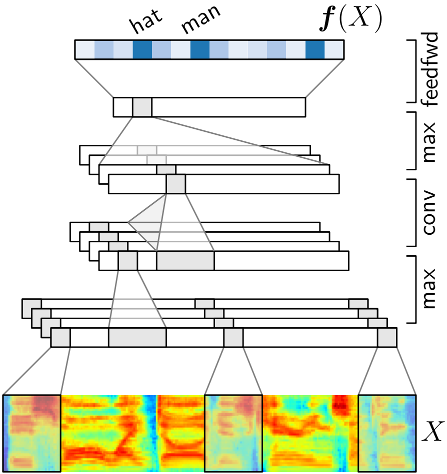
[Kamper et al., Interspeech'17]

Word prediction from images and speech



[Kamper et al., Interspeech'17]

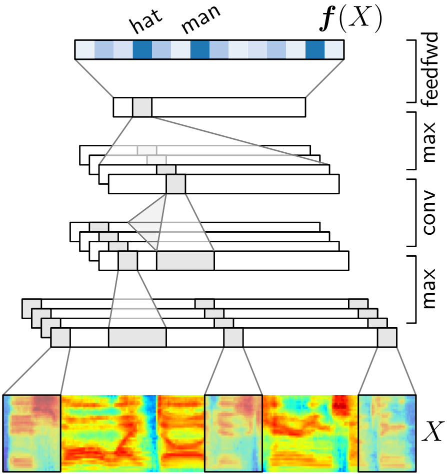
Word prediction from images and speech



[Kamper et al., Interspeech'17]

Word prediction from images and speech

$\boldsymbol{f}(X) \in \mathbb{R}^W$ is vector of word probabilities

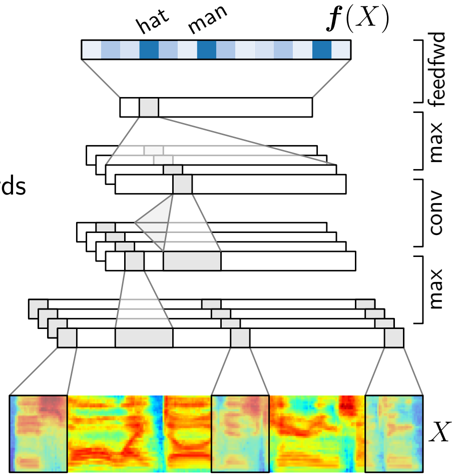


[Kamper et al., Interspeech'17]

Word prediction from images and speech

$\boldsymbol{f}(X) \in \mathbb{R}^W$ is vector of word probabilities

I.e., a spoken bag-of-words (BoW) classifier



[Kamper et al., Interspeech'17]