Learning from unlabelled speech, with and without visual cues

Ohio State University, May 2017

Herman Kamper

Toyota Technological Institute at Chicago

http://www.kamperh.com/

Success in speech recognition

Success in speech recognition

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

ars TECHNICA BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

TECHNOLOGY LAB —

Microsoft relea

source toolkit u

human-level sp

recognition

Microsoft wants to put machine l

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

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

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life

DIGITS

Speech Recognition Gets Conversational

By ROBERT MCMILLAN

May 28, 2015 12:54 pm ET

Success in speech recognition

Success in speech recognition

[Xiong et al., arXiv'16]; [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)
- Data: 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 (∼50 languages)
- Data: 2000 hours transcribed speech audio; ∼350M/560M words text
- Can we do this for all 7000 languages spoken in the world?

Learning from raw speech with no or weak labels

Learning from raw speech with no or weak labels

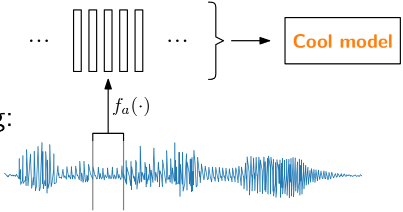
Unsupervised, or zero-resource, speech processing:

- What can we learn directly from raw speech?

Learning from raw speech with no or weak labels

Unsupervised, or zero-resource, speech processing:

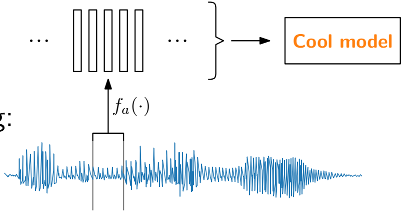
- What can we learn directly from raw speech?
- Unsupervised representation learning:



Learning from raw speech with no or weak labels

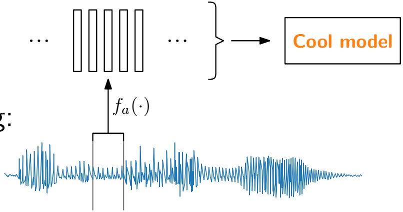
Unsupervised, or zero-resource, speech processing:

- What can we learn directly from raw speech?
- Unsupervised representation learning:
- Query-by-example search



Learning from raw speech with no or weak labels

Unsupervised, or zero-resource, speech processing:

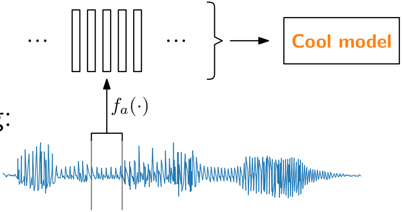


- What can we learn directly from raw speech?
- Unsupervised representation learning:
- Query-by-example search
- Unsupervised segmentation and clustering (word discovery)

Learning from raw speech with no or weak labels

Unsupervised, or zero-resource, speech processing:

- What can we learn directly from raw speech?
- Unsupervised representation learning:
- Query-by-example search
- Unsupervised segmentation and clustering (word discovery)

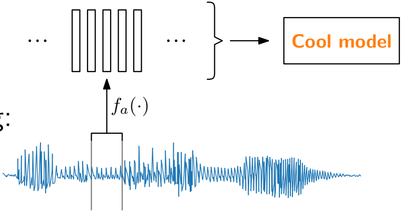


Learning from weak (distant) labels:

Learning from raw speech with no or weak labels

Unsupervised, or zero-resource, speech processing:

- What can we learn directly from raw speech?



- Unsupervised representation learning:
- Query-by-example search
- Unsupervised segmentation and clustering (word discovery)

Learning from weak (distant) labels:

- What can we learn from speech paired with another modality?
- E.g. translations or images

Why learn with no or weak labels?

- **Criticism:** You always have some labelled data

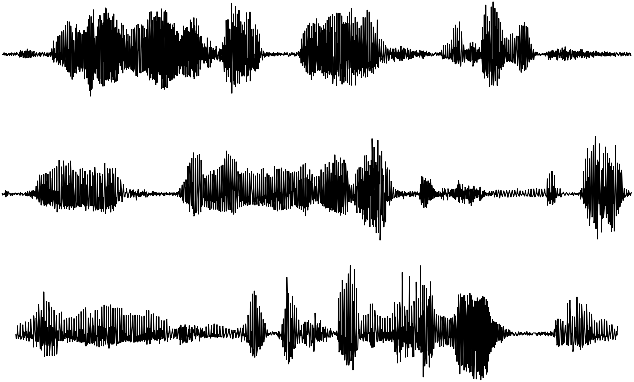
Why learn with no or weak labels?

- **Criticism:** You always have some labelled data, but...
- 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 with no or weak labels?

- **Criticism:** You always have some labelled data, but...
- 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]

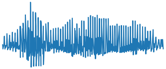
Example: Query-by-example search

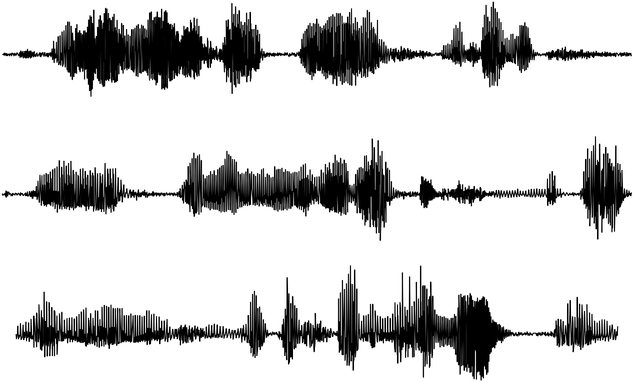


[Jansen and Van Durme, IS'12]

Example: Query-by-example search

Spoken query:

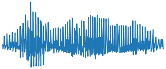


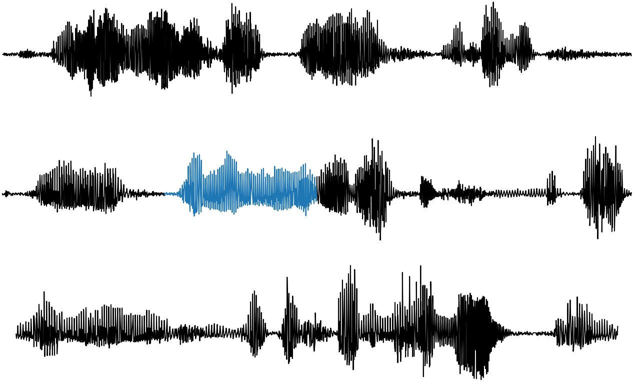


[Jansen and Van Durme, IS'12]

Example: Query-by-example search

Spoken query:

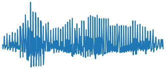


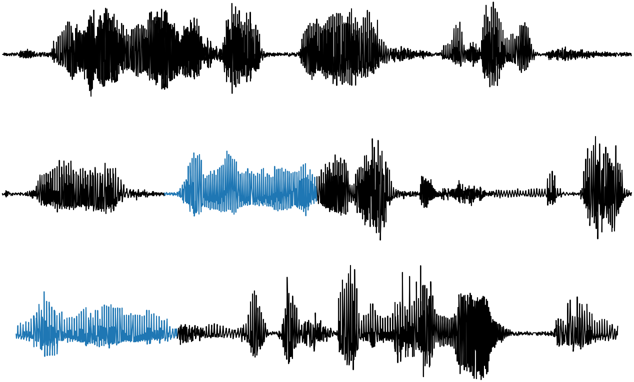


[Jansen and Van Durme, IS'12]

Example: Query-by-example search

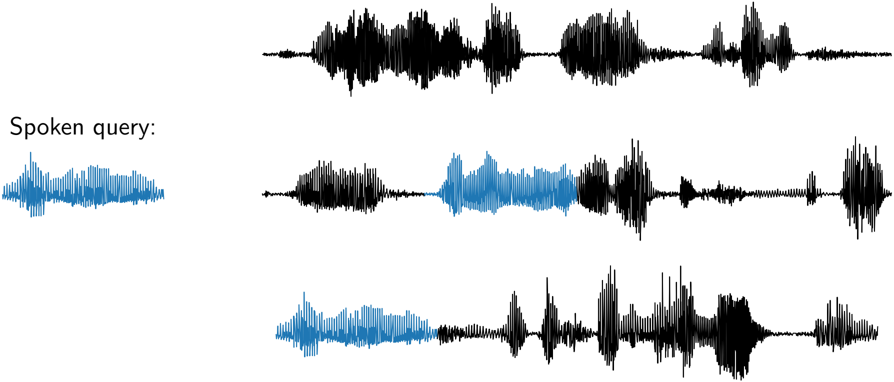
Spoken query:





[Jansen and Van Durme, IS'12]

Example: Query-by-example search



Useful speech system, not requiring any transcribed speech

[Jansen and Van Durme, IS'12]

Learning from unlabelled speech **with** and **without** visual cues

Learning from unlabelled speech **with** and **without** visual cues

Talk outline:

1. Unsupervised segmentation and clustering of speech (**without**)

Learning from unlabelled speech **with** and **without** visual cues

Talk outline:

1. Unsupervised segmentation and clustering of speech (**without**)
2. Using images to visually ground untranscribed speech (**with**)

Unsupervised segmentation and clustering:

Segmental Bayesian Speech Model

Unsupervised segmentation and clustering:

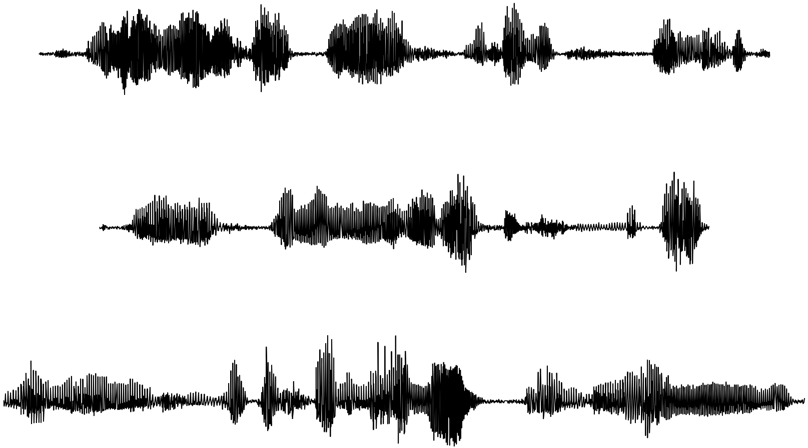
Segmental Bayesian Speech Model

Aren Jansen

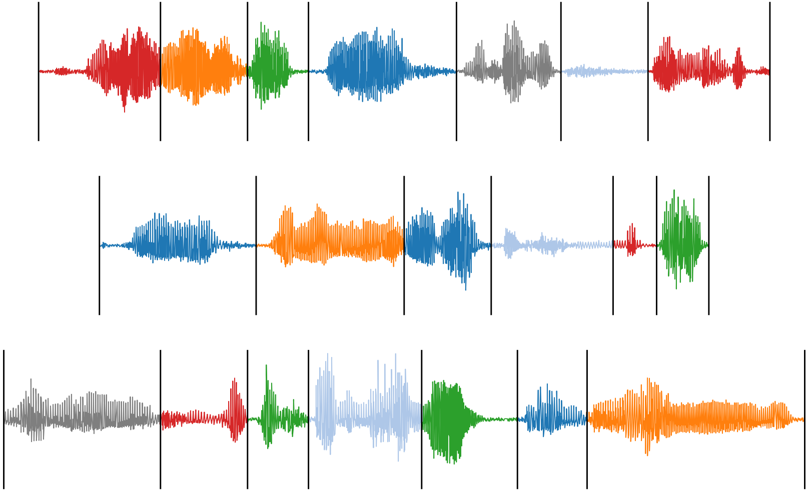
Sharon Goldwater

Full-coverage segmentation and clustering

Full-coverage segmentation and clustering

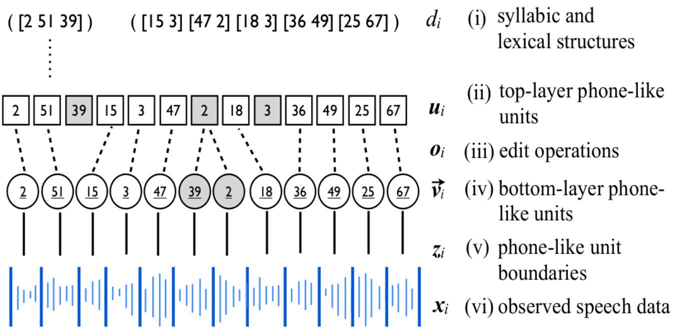


Full-coverage segmentation and clustering



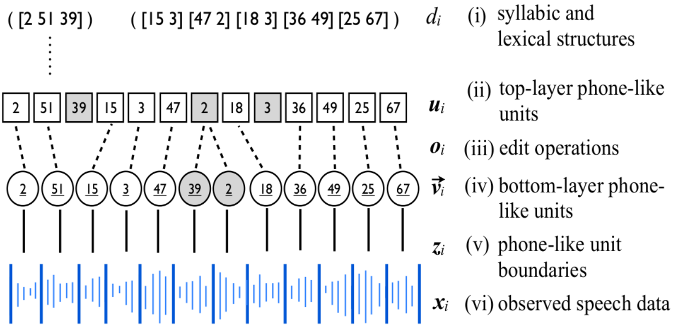
Bayesian models for full-coverage segmentation

Previous models use explicit subword discovery directly on speech features, e.g. [Lee et al., TACL'15]:



Bayesian models for full-coverage segmentation

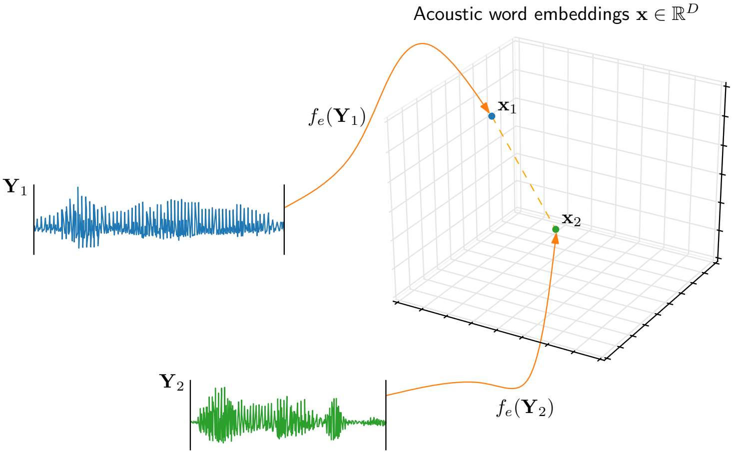
Previous models use explicit subword discovery directly on speech features, e.g. [Lee et al., TACL'15]:



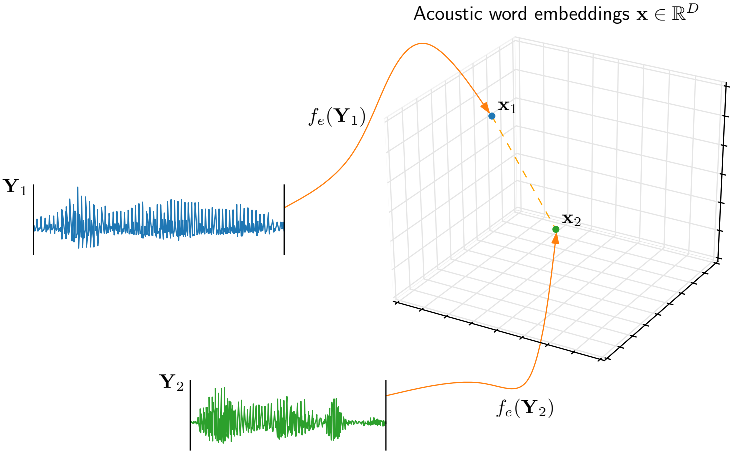
Our approach uses whole-word segmental representations, i.e. acoustic word embeddings [Kamper et al., TASLP'16]

Acoustic word embeddings

Acoustic word embeddings



Acoustic word embeddings



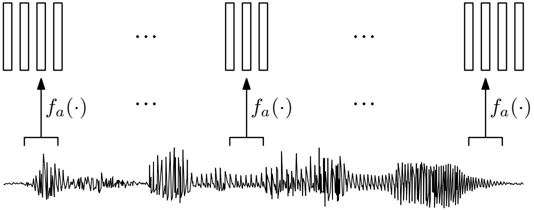
Dynamic programming alignment has quadratic complexity, while embedding comparison is linear time. Can use standard clustering.

Unsupervised segmental Bayesian model

Speech waveform

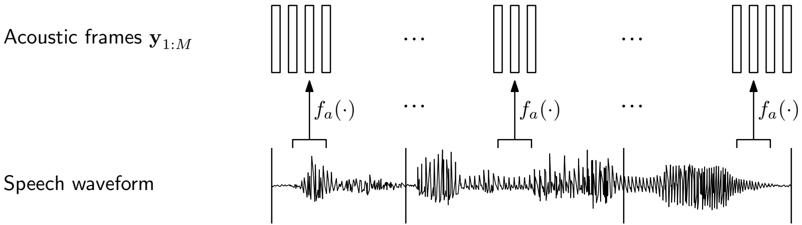
Unsupervised segmental Bayesian model

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

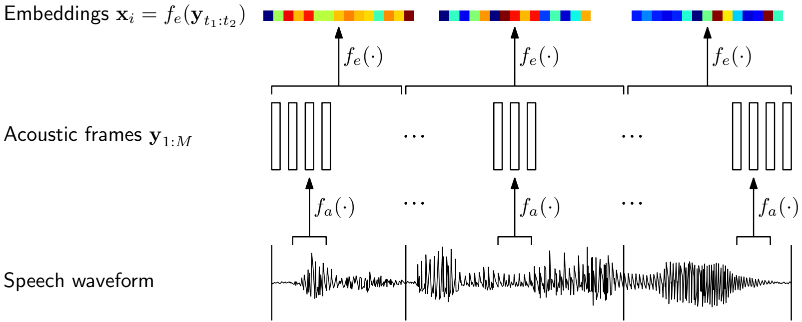


Speech waveform

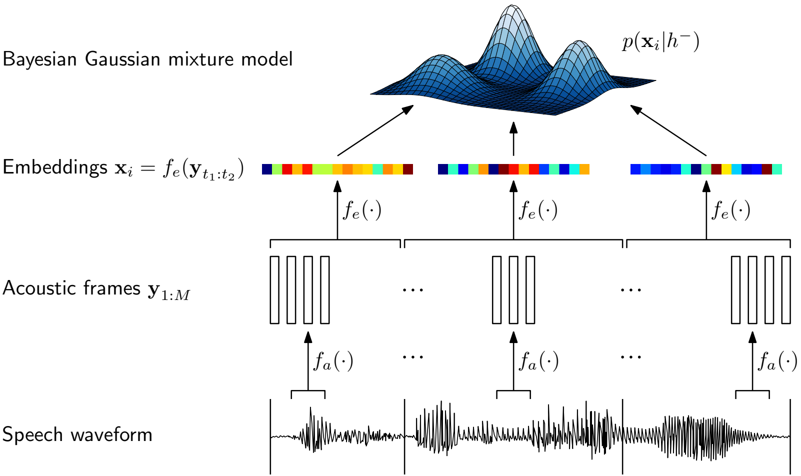
Unsupervised segmental Bayesian model



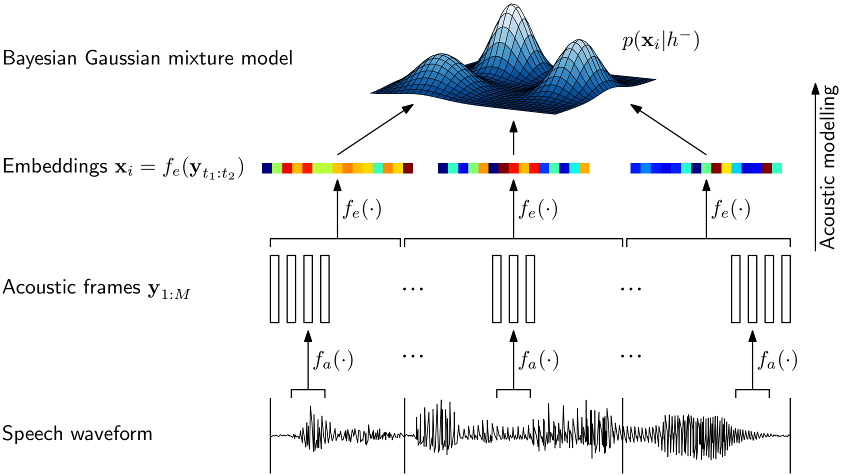
Unsupervised segmental Bayesian model



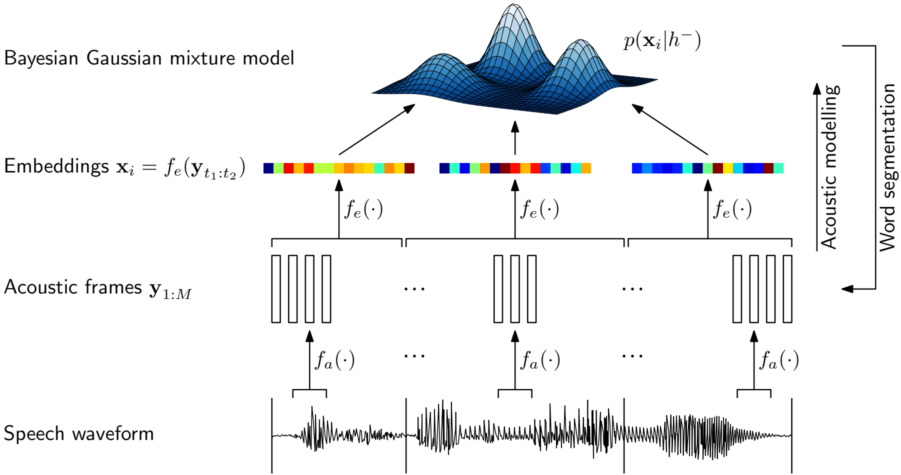
Unsupervised segmental Bayesian model



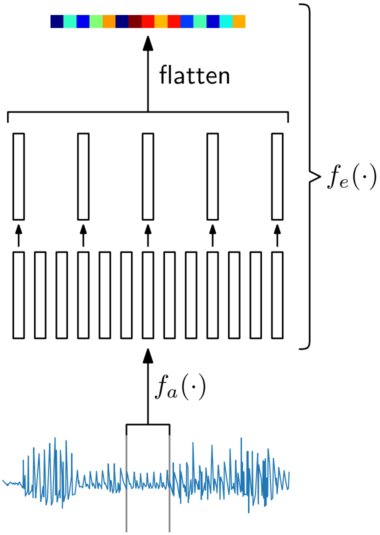
Unsupervised segmental Bayesian model



Unsupervised segmental Bayesian model

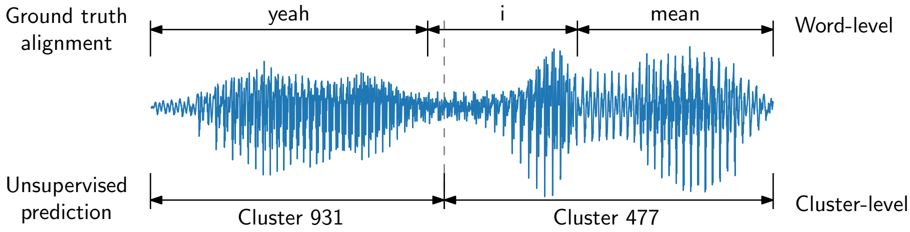


Acoustic word embeddings: Downsampling

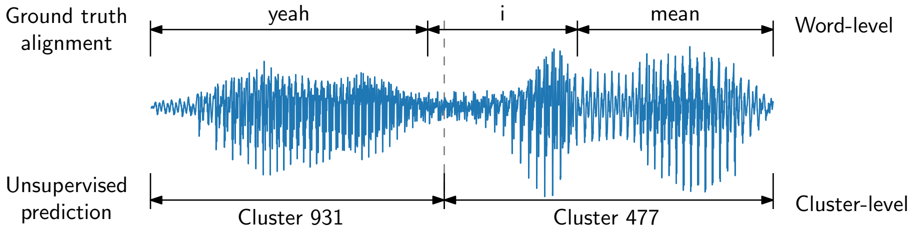


- Simple embedding approach also used in other studies
  e.g. [Abdel-Hamid et al., 2013]
- Downsampling is simple, but actually hard to beat (unsupervised)
- Ongoing work, e.g.,
  [Levin et al., ASRU'13]; [Kamper et al., ICASSP'16]; [Settle and Livescu, SLT'16]

Evaluation


Ground truth alignment
yeah
i
mean
Word-level
Unsupervised prediction
Cluster 931
Cluster 477
Cluster-level

Evaluation

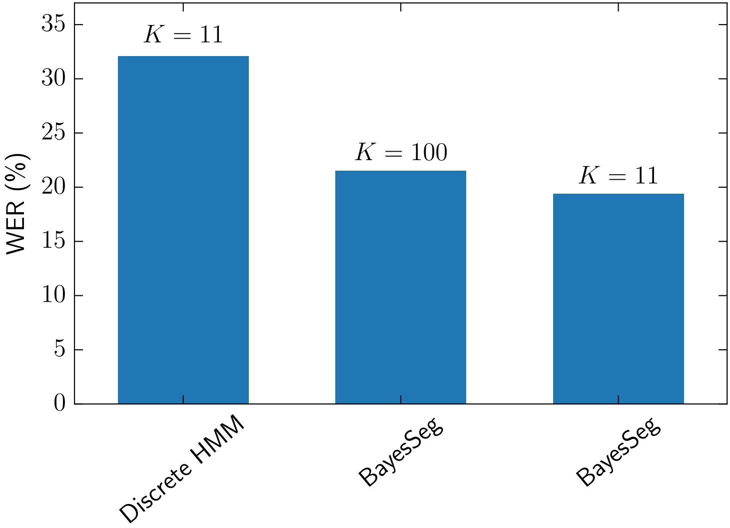


Metrics:

- Unsupervised word error rate (WER)
- Word token precision, recall, F-score
- Word type precision, recall, F-score
- Word boundary precision, recall, F-score

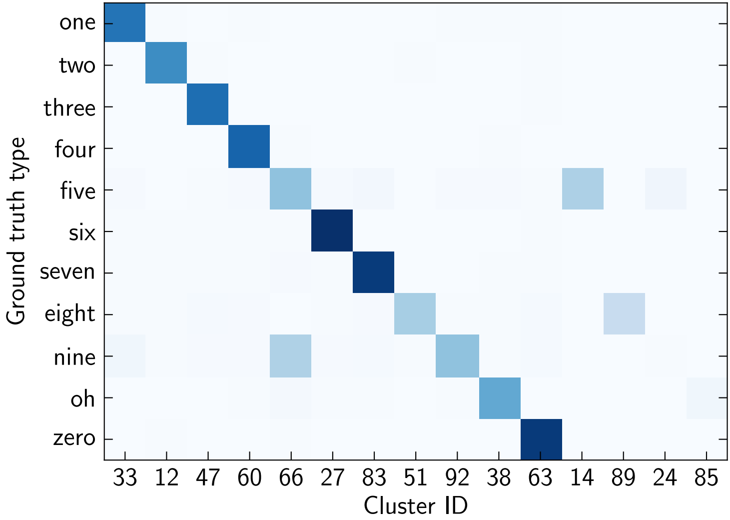
Small-vocabulary segmentation and clustering

Small-vocabulary segmentation and clustering



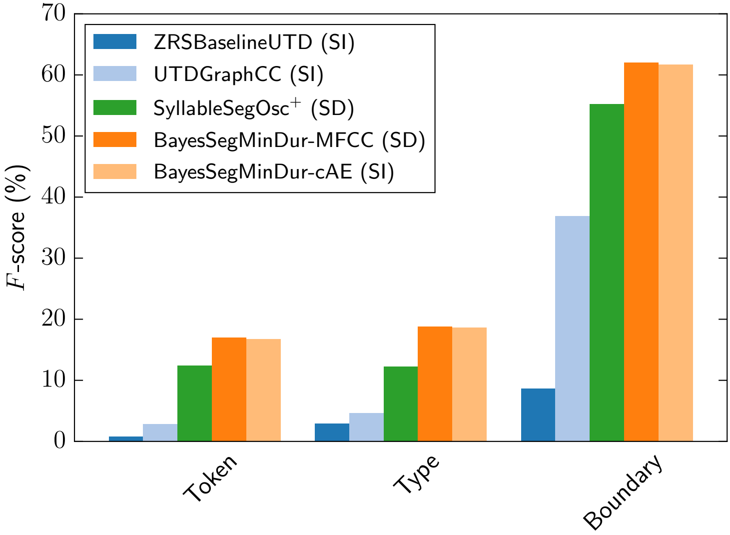
Discrete HMM: [Walter et al., ASRU'13]. BayesSeg: [Kamper et al., TASLP'16].

Small-vocabulary segmentation and clustering



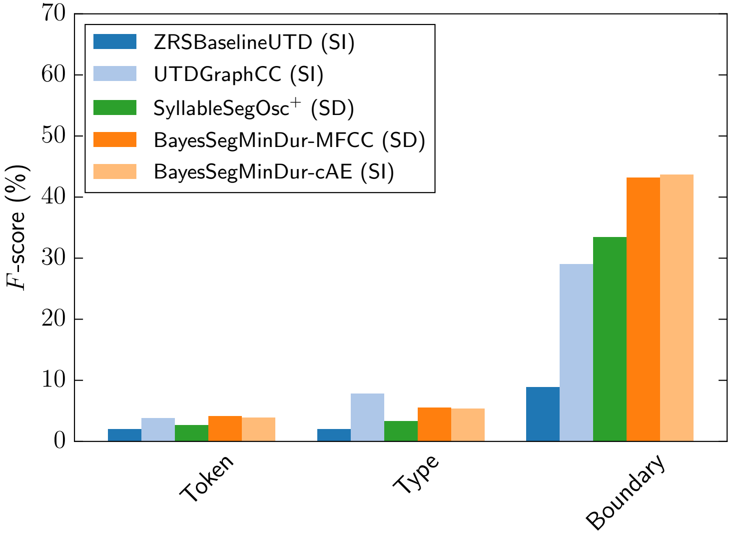
[Kamper et al., TASLP'16]

Large-vocabulary: English



ZRSBaselineUTD: [Versteegh et al., IS'15]. UTDGraphCC: [Lyzinski et al., IS'15].
SyllableSegOsc$^+$: [Räsänen et al., IS'15]. BayesSeg: [Kamper et al., arXiv'16].

Large-vocabulary: Xitsonga

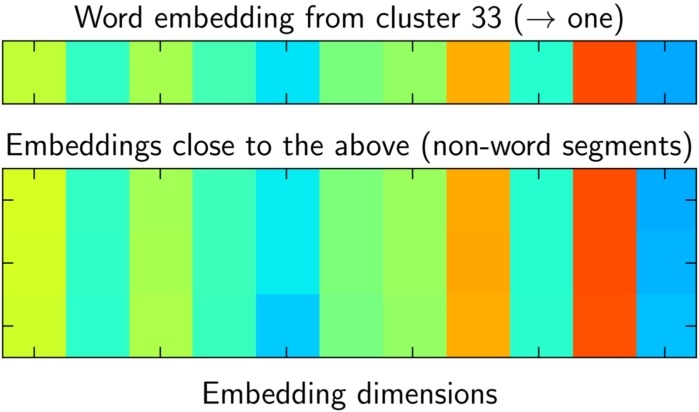


ZRSBaselineUTD: [Versteegh et al., IS'15]. UTDGraphCC: [Lyzinski et al., IS'15]. SyllableSegOsc^{+}: [Räsänen et al., IS'15]. BayesSeg: [Kamper et al., arXiv'16].

Listen to discovered clusters

- Data for small-vocabulary experiments: Play
- Small-vocabulary cluster 45: Play
- Large-vocabulary English cluster 1214: Play
- Large-vocabulary Xitsonga cluster 629: Play

The true (less rosy) picture



[Levin et al., ASRU'13]; [Kamper et al., ICASSP'16]; [Settle and Livescu, SLT'16]

Using visual cues to learn from untranscribed speech:

Visually Grounded Keyword Prediction

Using visual cues to learn from untranscribed speech:

Visually Grounded Keyword Prediction

Shane Settle

Greg Shakhnarovich

Karen Livescu


Arrival

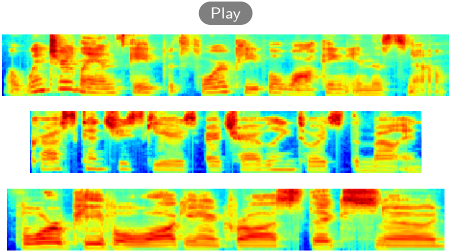
Using images for grounding language

Using images for grounding language

- Image captioning: Generate written natural language description of a given image [Vinyals et al., CVPR'15]
- Grounding written language using images [Bernardi et al., JAIR'16]

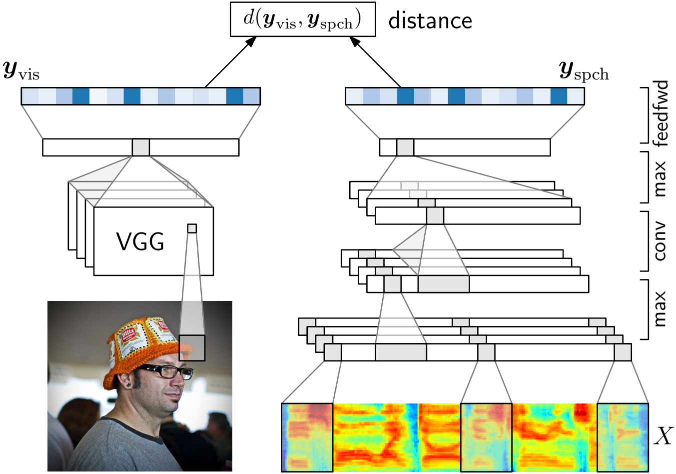
Using images for grounding language

- Image captioning: Generate written natural language description of a given image [Vinyals et al., CVPR'15]
- Grounding written language using images [Bernardi et al., JAIR'16]
- We consider images paired with unlabellel spoken captions:



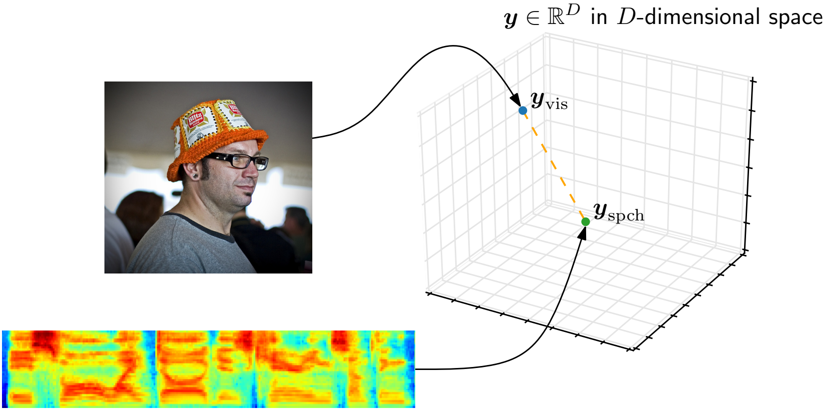
Map images and speech into common space

Map images and speech into common space



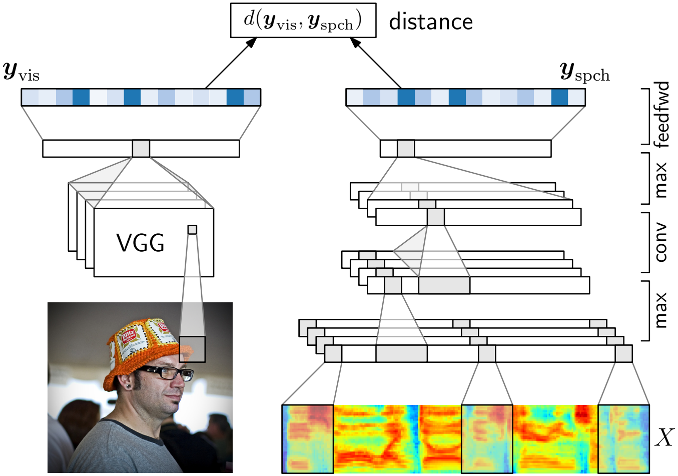
[Harwath et al., NIPS'16]

Retrieval in common (semantic) space



[Harwath et al., NIPS'16]

Can we use (supervised) vision model to get labels?



Cannot obtain textual labels for the speech using this model

Word prediction from images and speech

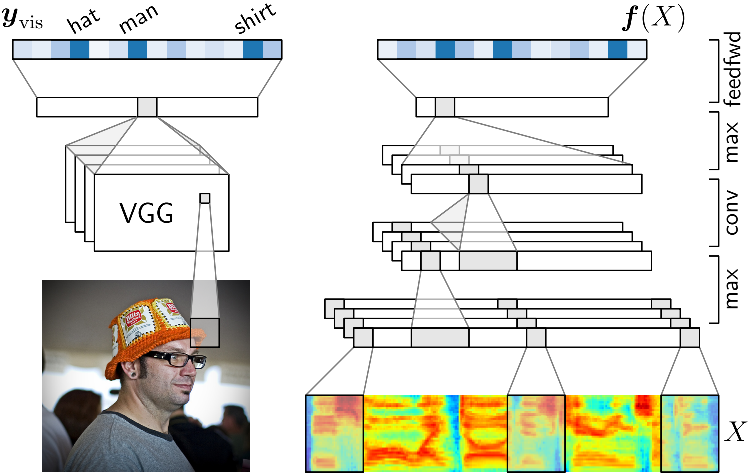
Word prediction from images and speech

[Kamper et al., arXiv'17]

Word prediction from images and speech

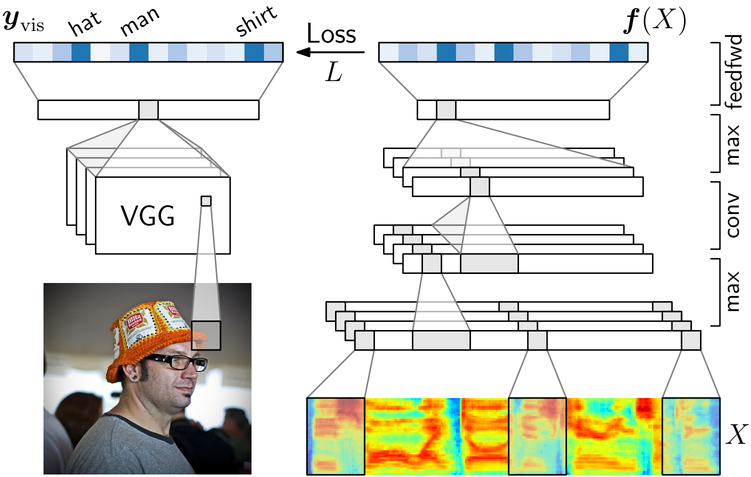
[Kamper et al., arXiv'17]

Word prediction from images and speech



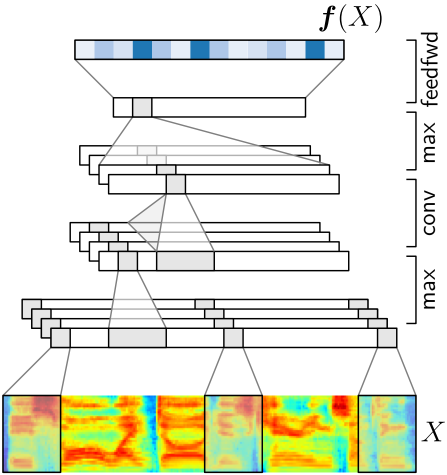
[Kamper et al., arXiv'17]

Word prediction from images and speech



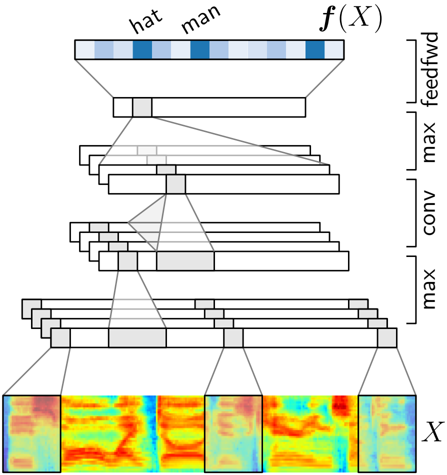
[Kamper et al., arXiv'17]

Word prediction from images and speech



[Kamper et al., arXiv'17]

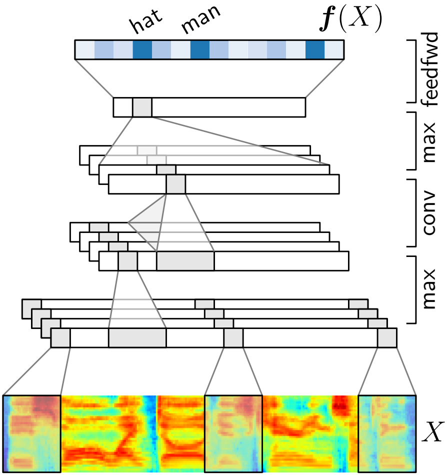
Word prediction from images and speech



[Kamper et al., arXiv'17]

Word prediction from images and speech

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

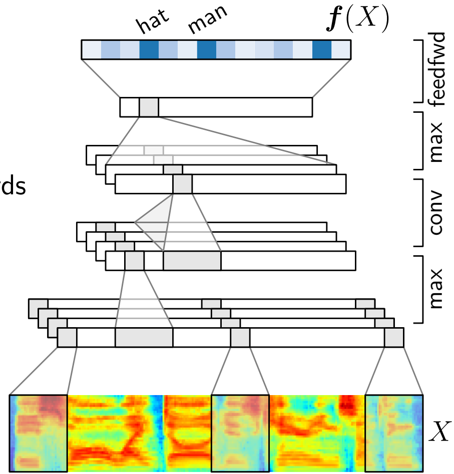


[Kamper et al., arXiv'17]

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., arXiv'17]

Word prediction from images and speech

Vision system outputs $\boldsymbol{y}_{\text{vis}}$, giving probability of word w for image I:

$$y_{\text{vis},w} = P(w|I, \boldsymbol{\gamma})$$

Word prediction from images and speech

Vision system outputs $\boldsymbol{y}_{\text{vis}}$, giving probability of word w for image I:

$$y_{\text{vis},w} = P(w|I, \boldsymbol{\gamma})$$

Interpret dimension w of the speech network output $\boldsymbol{f}(X)$ as:

$$f_w(X) = P(w|X, \boldsymbol{\theta})$$

Word prediction from images and speech

Vision system outputs $\boldsymbol{y}_{\text{vis}}$, giving probability of word w for image I:

$$y_{\text{vis},w} = P(w|I, \boldsymbol{\gamma})$$

Interpret dimension w of the speech network output $\boldsymbol{f}(X)$ as:

$$f_w(X) = P(w|X, \boldsymbol{\theta})$$

Train using cross-entropy loss (i.e. soft targets):

$$L(\boldsymbol{f}(X), \boldsymbol{y}_{\text{vis}}) = -\sum_{w=1}^{W} \left\{ y_{\text{vis},w} \log f_w(X) + (1 - y_{\text{vis},w}) \log \left[1 - f_w(X)\right] \right\}$$

Word prediction from images and speech

Vision system outputs $\boldsymbol{y}_{\text{vis}}$, giving probability of word w for image I:

$$y_{\text{vis},w} = P(w|I, \boldsymbol{\gamma})$$

Interpret dimension w of the speech network output $\boldsymbol{f}(X)$ as:

$$f_w(X) = P(w|X, \boldsymbol{\theta})$$

Train using cross-entropy loss (i.e. soft targets):

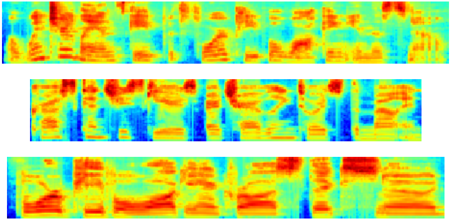
$$L(\boldsymbol{f}(X), \boldsymbol{y}_{\text{vis}}) = -\sum_{w=1}^{W} \left\{ y_{\text{vis},w} \log f_w(X) + (1 - y_{\text{vis},w}) \log \left[1 - f_w(X)\right] \right\}$$

If $y_{\text{vis},w} \in \{0, 1\}$, this is summed log loss of W binary classifiers.

[Kamper et al., arXiv'17]

Images paired with untranscribed speech

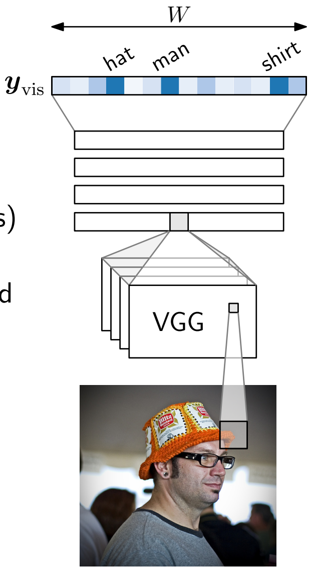
We are still in this setting:



- I.e., we do not use any of the speech transcriptions during model training (only for evaluation)
- But our resulting model can make bag-of-words (BoW) predictions

The vision system

- VGG-16 input layers (1.3M images)
  [Simonyan and Zisserman, arXiv'14]
- Train on Flickr30k (caption BoW labels)
- Targets: $W = 1000$ most common word types after removing stop words
- Note: Vision system could be seen as language independent (future work)



Experimental details

- **Data:** 8000 images with 5 spoken captions, divided into train, development and test sets [Harwath and Glass, ASRU'15]
- **Prediction:** Output words w where $f_w(X) > \alpha$
- **Tasks:** Spoken bag-of-words prediction; keyword spotting
- **Evaluation:** Compare to words in transcriptions of test data

Task 1: Spoken bag-of-words prediction

Input utterance	Predicted BoW labels
Play	

Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
Play	**bicycle**, bike, **man**, riding, wearing

Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing

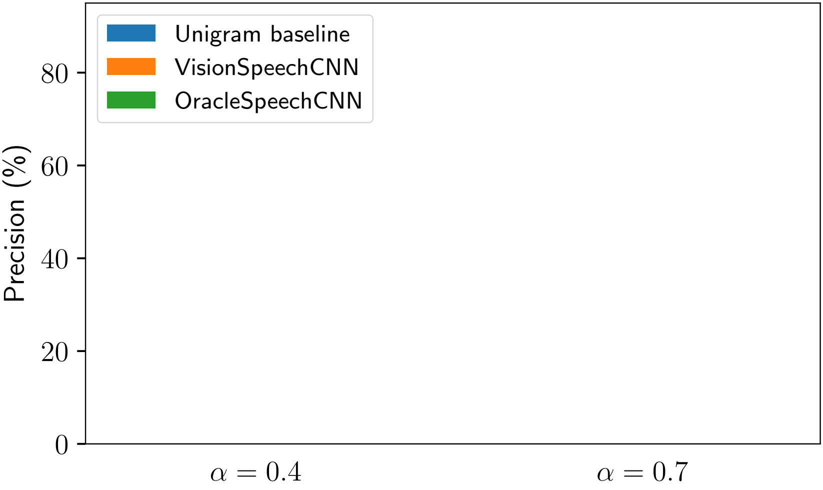
Task 1: Spoken bag-of-words prediction

Input utterance	Predicted BoW labels
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing
a little girl is climbing a ladder	child, **girl**, **little**, young
a rock climber standing in a crevasse	climbing, man, **rock**
a dog running in the grass around sheep	**dog**, field, **grass**, **running**
a man in a miami basketball uniform looking to the right	ball, **basketball**, **man**, player, **uniform**, wearing

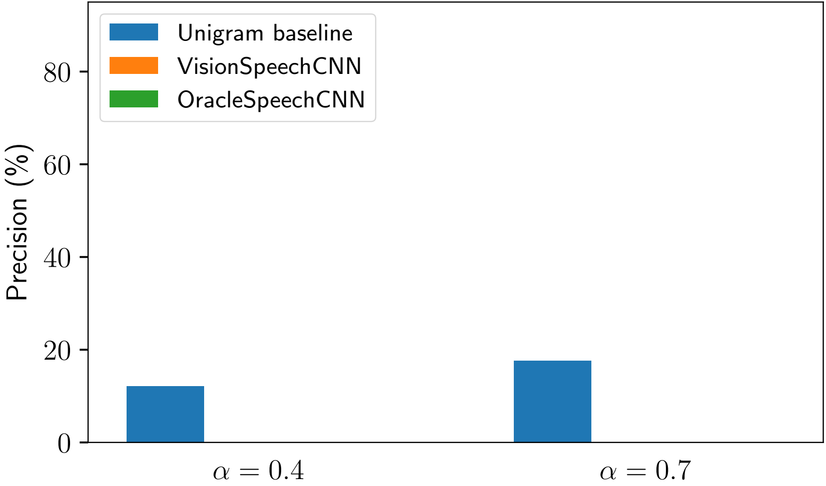
Task 1: Spoken bag-of-words prediction

Input utterance	**Predicted BoW labels**
man on bicycle is doing tricks in an old building	**bicycle**, bike, **man**, riding, wearing
a little girl is climbing a ladder	child, **girl**, **little**, young
a rock climber standing in a crevasse	climbing, man, **rock**
a dog running in the grass around sheep	**dog**, field, **grass**, **running**
a man in a miami basketball uniform looking to the right	ball, **basketball**, **man**, player, **uniform**, wearing

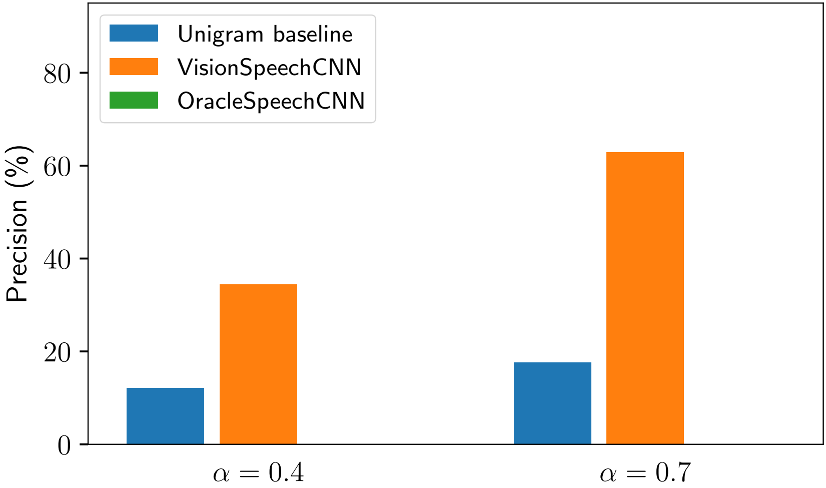
Task 1: Spoken bag-of-words prediction



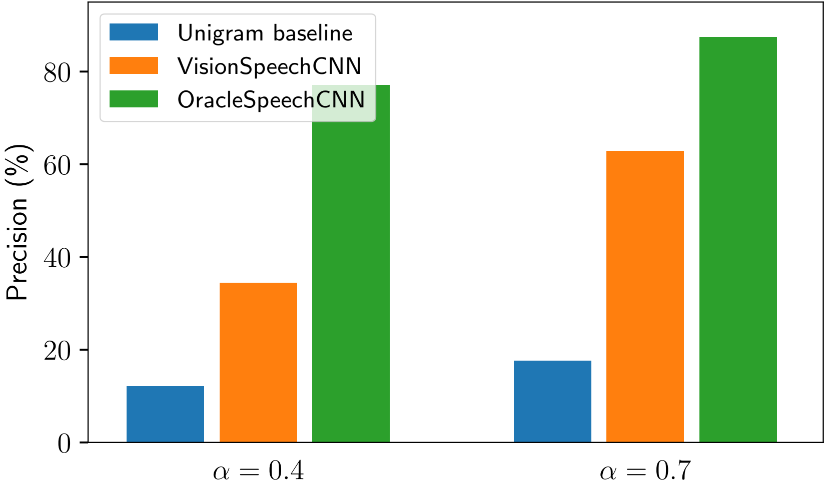
Task 1: Spoken bag-of-words prediction



Task 1: Spoken bag-of-words prediction



Task 1: Spoken bag-of-words prediction

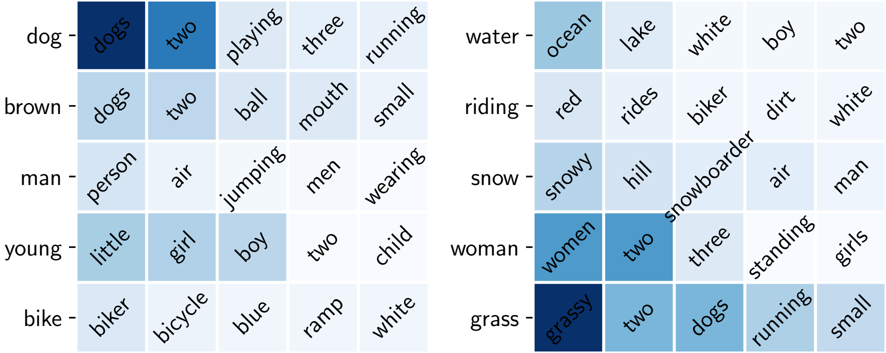


Task 1: Spoken bag-of-words prediction

False alarm keywords and words in corresponding utterances

Task 1: Spoken bag-of-words prediction

False alarm keywords and words in corresponding utterances:



Task 2: Keyword spotting

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

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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		

Task 2: 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

Task 2: Keyword spotting

Model	$P@10$	$P@N$	EER
Unigram baseline	5.0	3.5	50.0
VisionSpeechCNN	54.5	33.1	22.3
OracleSpeechCNN	96.5	83.0	4.1

Task 3: (Towards) semantic keyword spotting

Retrieve all utterances in a set containing content **related in meaning** to a given textual keyword

Task 3: (Towards) semantic keyword spotting

Retrieve all utterances in a set containing content **related in meaning** to a given textual keyword

Model	$P@10$
Unigram baseline	10.0
VisionSpeechCNN	82.5
OracleSpeechCNN	99.5

Task 3: (Towards) semantic keyword spotting

Retrieve all utterances in a set containing content **related in meaning** to a given textual keyword

Model	$P@10$
Unigram baseline	10.0
VisionSpeechCNN	82.5
OracleSpeechCNN	99.5

Thoughts on this task are very welcome!

Conclusions and Future Work

Summary and conclusion

- We are able to discover (some) structure directly from raw speech audio (segmentation and clustering) [Kamper et al., TASLP'16; arXiv'16]
- Visual grounding makes it possible to develop a word prediction model without any parallel speech and text [Kamper et al., arXiv'17]
- Useful to look at speech processing from a different perspective

Looking forward

- Thorough analysis of VisionSpeech models to see if they learn something about semantics; multi-lingual aspects

Looking forward

- Thorough analysis of VisionSpeech models to see if they learn something about semantics; multi-lingual aspects
- BayesSeg learns from acoustics, VisionSpeech captures something about semantics: can we combine these?

Looking forward

- Thorough analysis of VisionSpeech models to see if they learn something about semantics; multi-lingual aspects
- BayesSeg learns from acoustics, VisionSpeech captures something about semantics: can we combine these?
- Building audio analysis tools for field linguists

Looking forward

- Thorough analysis of VisionSpeech models to see if they learn something about semantics; multi-lingual aspects
- BayesSeg learns from acoustics, VisionSpeech captures something about semantics: can we combine these?
- Building audio analysis tools for field linguists
- What can we learn about language acquisition in humans?

Looking forward

- Thorough analysis of VisionSpeech models to see if they learn something about semantics; multi-lingual aspects
- BayesSeg learns from acoustics, VisionSpeech captures something about semantics: can we combine these?
- Building audio analysis tools for field linguists
- What can we learn about language acquisition in humans?
- Language acquisition in robots

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

References I

- O. Abdel-Hamid, L. Deng, D. Yu, and H. Jiang, "Deep segmental neural networks for speech recognition," in *Proc. Interspeech*, 2013.
- R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, 2016.
- L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, 2014.
- D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. NIPS*, 2016.
- D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. ASRU*, 2015.
- A. Jansen and B. Van Durme, "Indexing raw acoustic features for scalable zero resource search," in *Proc. Interspeech*, 2012.
- A. Jansen *et al.*, "A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition," in *Proc. ICASSP*, 2013.
- H. Kamper, M. Elsner, A. Jansen, and S. J. Goldwater, "Unsupervised neural network based feature extraction using weak top-down constraints," in *Proc. ICASSP*, 2015.

References II

- H. Kamper, A. Jansen, and S. J. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, 2016.
- H. Kamper, W. Wang, and K. Livescu, "Deep convolutional acoustic word embeddings using word-pair side information," in *Proc. ICASSP*, 2016.
- H. Kamper, S. J. Goldwater, and A. Jansen, "Fully unsupervised small-vocabulary speech recognition using a segmental Bayesian model," in *Proc. Interspeech*, 2015.
- H. Kamper, A. Jansen, and S. J. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *arXiv preprint arXiv:1606.06950*, 2016.
- H. Kamper, S. Settle, G. Shakhnarovich, and K. Livescu, "Visually grounded learning of keyword prediction from untranscribed speech," *arXiv preprint arXiv:1703.08136*, 2017.
- C.-y. Lee, T. O'Donnell, and J. R. Glass, "Unsupervised lexicon discovery from acoustic input," *Trans. ACL*, vol. 3, pp. 389–403, 2015.
- K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Proc. ASRU*, 2013.
- V. Lyzinski, G. Sell, and A. Jansen, "An evaluation of graph clustering methods for unsupervised term discovery," in *Proc. Interspeech*, 2015.

References III

- D. Palaz, G. Synnaeve, and R. Collobert, "Jointly learning to locate and classify words using convolutional networks," in *Proc. Interspeech*, 2016.
- O. J. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proc. Interspeech*, 2015.
- O. Räsänen and H. Rasilo, "A joint model of word segmentation and meaning acquisition through cross-situational learning," *Psychol. Rev.*, vol. 122, no. 4, pp. 792–829, 2015.
- V. Renkens and H. Van hamme, "Mutually exclusive grounding for weakly supervised non-negative matrix factorisation," in *Proc. Interspeech*, 2015.
- D. Roy, "Learning from sights and sounds: A computational model," Ph.D. dissertation, Learning from Sights and Sounds: A Computational Model, Cambridge, MA, 1999.
- G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.
- S. Settle and K. Livescu, "Discriminative acoustic word embeddings: Recurrent neural network-based approaches," in *Proc. SLT*, 2016.
- K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

References IV

- M. Versteegh, R. Thiollière, T. Schatz, X. N. Cao, X. Anguera, A. Jansen, and E. Dupoux, "The Zero Resource Speech Challenge 2015," in *Proc. Interspeech*, 2015.
- O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. CVPR*, 2015.
- O. Walter, T. Korthals, R. Haeb-Umbach, and B. Raj, "A hierarchical system for word discovery exploiting DTW-based initialization," in *Proc. ASRU*, 2013.
- W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.