Unsupervised neural and Bayesian models for zero-resource speech processing

MIT CSAIL, 15 Nov. 2016

Herman Kamper

University of Edinburgh; TTI at Chicago

http://www.kamperh.com

Speech recognition success

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

Speech recognition success

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

Speech recognition success

Speech recognition success

[Xiong et al., arXiv'16]

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

Speech recognition success

[Xiong et al., arXiv'16]

- **Google Voice:** English, Spanish, German, . . . , Zulu ($\sim$50 languages)
- **Data:** 2000 hours of labelled speech audio; $\sim$350M words of text

Speech recognition success

[Xiong et al., arXiv'16]

- **Google Voice:** English, Spanish, German, . . . , Zulu (∼50 languages)
- **Data:** 2000 hours of labelled speech audio; ∼350M words of text
- **But:** Can we do this for all 7000 languages spoken in the world?

Unsupervised speech processing

Developing unsupervised methods that can learn structure directly from raw speech audio, i.e. zero-resource technology

Unsupervised speech processing

Developing unsupervised methods that can learn structure directly from raw speech audio, i.e. zero-resource technology

Criticism: Always some data; semi-supervised problem

Unsupervised speech processing

Developing unsupervised methods that can learn structure directly from raw speech audio, i.e. zero-resource technology

Criticism: Always some data; semi-supervised problem

Reasons for purely unsupervised case:

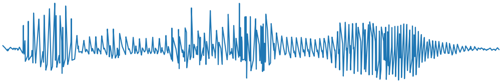
- Modelling infant language acquisition [Räsänen, SpecCom'12]
- Language acquisition in robotics [Renkens and Van hamme, IS'15]
- Analysis of audio for unwritten languages [Besacier et al., SpecCom'14]
- New insights and models for speech processing [Jansen et al., ICASSP'13]

Unsupervised speech processing: Two problems

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

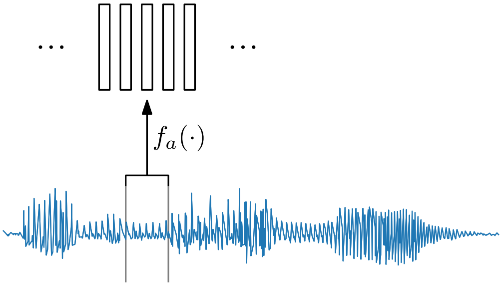
Unsupervised speech processing: Two problems

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



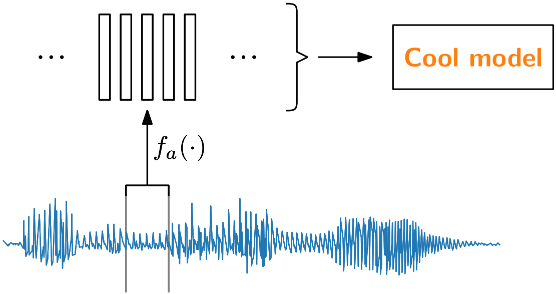
Unsupervised speech processing: Two problems

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



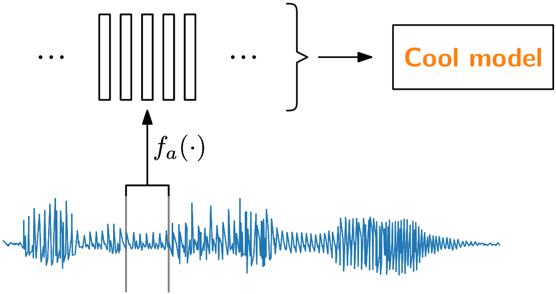
Unsupervised speech processing: Two problems

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



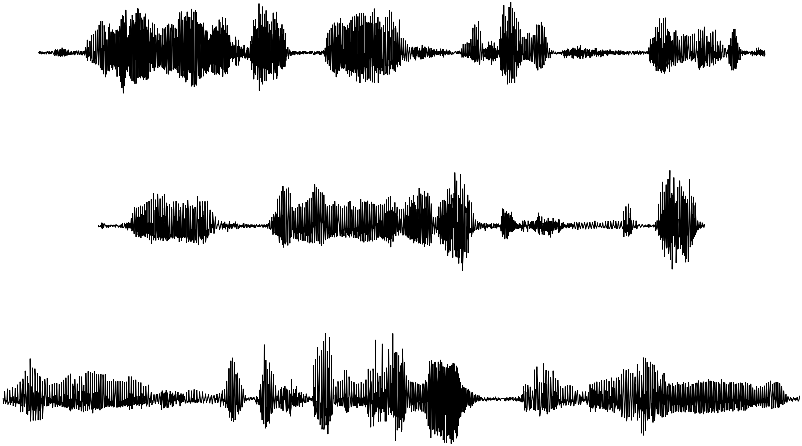
Unsupervised speech processing: Two problems

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



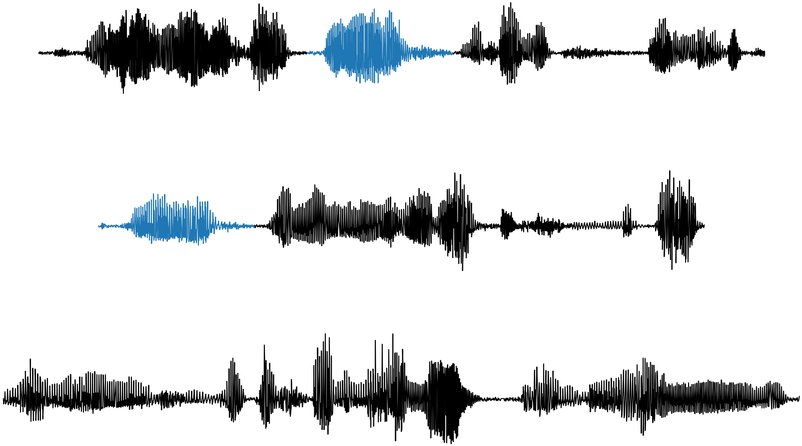
2. Unsupervised **segmentation** and **clustering**:
   How do we discover meaningful units in unlabelled speech?

Unsupervised term discovery (UTD)



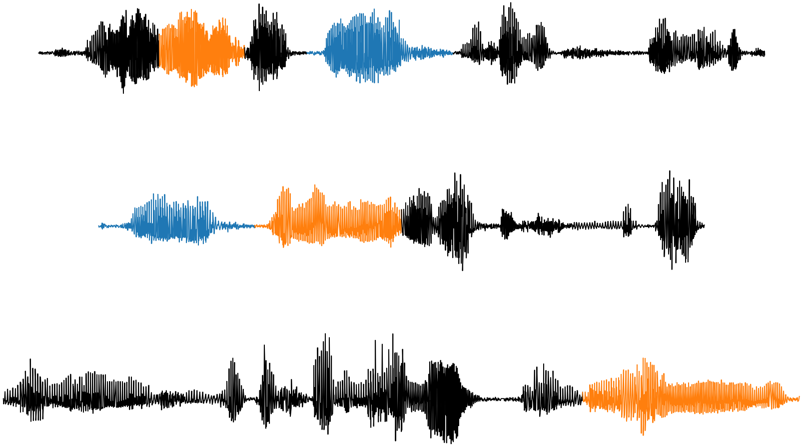
[Park and Glass, TASLP'08]

Unsupervised term discovery (UTD)



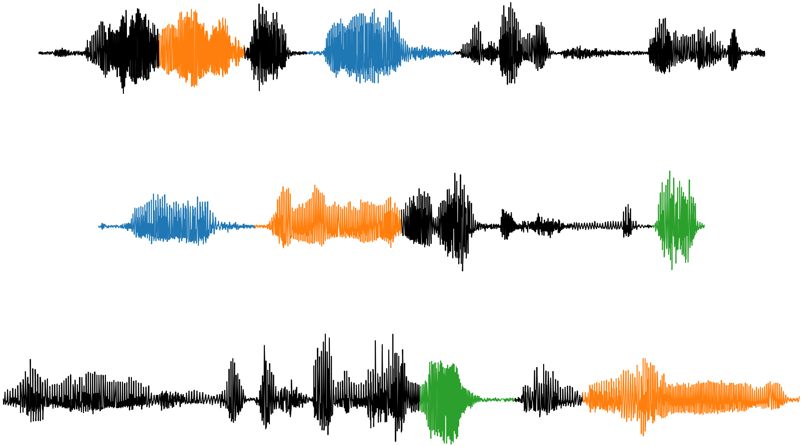
[Park and Glass, TASLP'08]

Unsupervised term discovery (UTD)



[Park and Glass, TASLP'08]

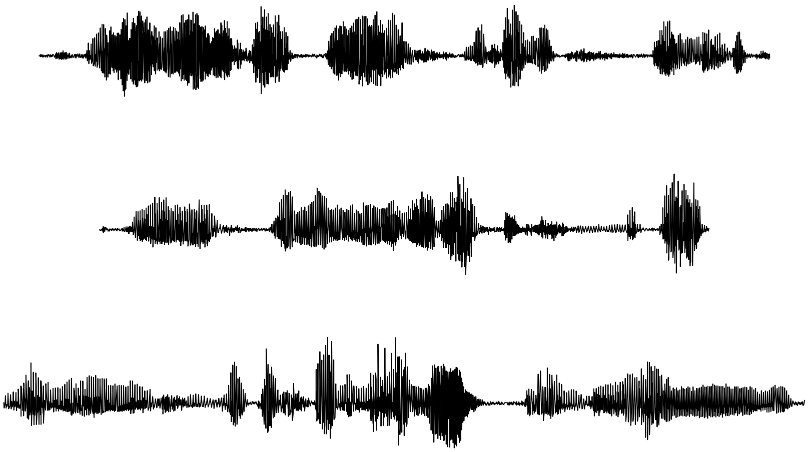
Unsupervised term discovery (UTD)



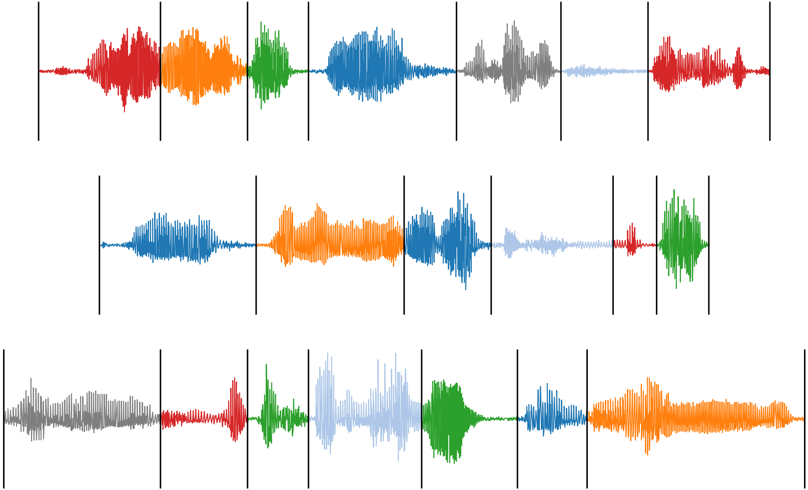
[Park and Glass, TASLP'08]

Full-coverage segmentation and clustering

Full-coverage segmentation and clustering

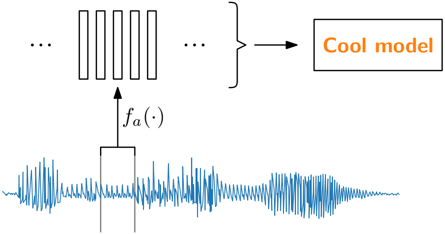


Full-coverage segmentation and clustering



Unsupervised speech processing: Two problems

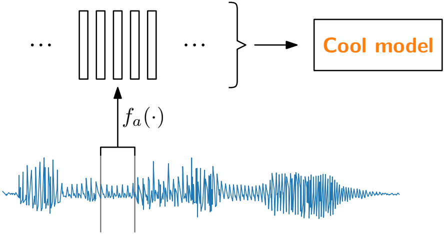
1. Unsupervised frame-level **representation learning**:



2. Unsupervised **segmentation** and **clustering**:
   We focus on full-coverage segmentation and clustering

Unsupervised speech processing: Two problems

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



2. Unsupervised **segmentation** and **clustering**:
   We focus on full-coverage segmentation and clustering

Our claim: Unsupervised speech processing benefits from both top-down and bottom-up modelling

Top-down and bottom-up modelling

Top-down: Use knowledge of higher-level units to learn about lower-level parts

Bottom-up: Piece together lower-level parts to get more complex higher-level structures

[Feldman et al., CCSS'09]

Unsupervised frame-level representation learning:

The Correspondence Autoencoder

Unsupervised frame-level representation learning:

The Correspondence Autoencoder

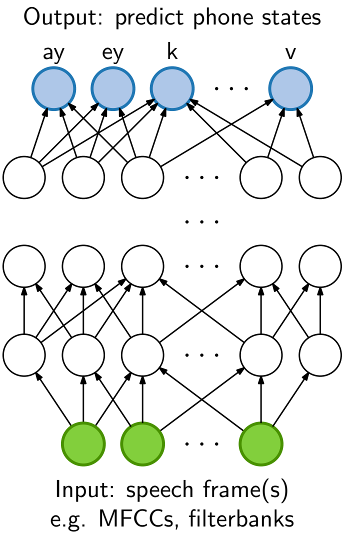

Micha Elsner


Daniel Renshaw

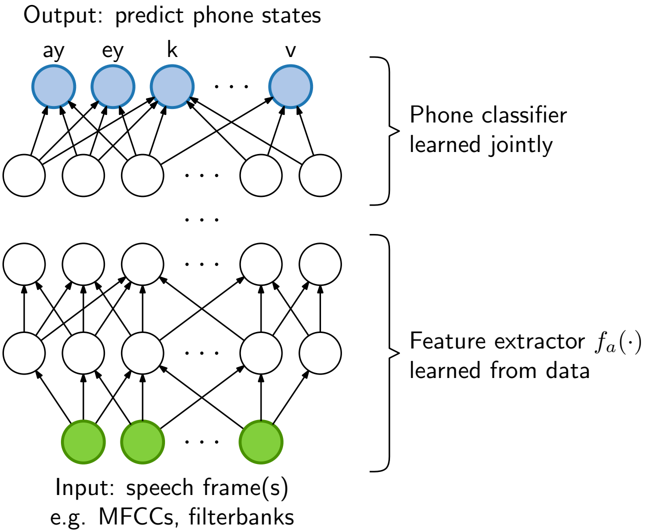

Aren Jansen


Sharon Goldwater

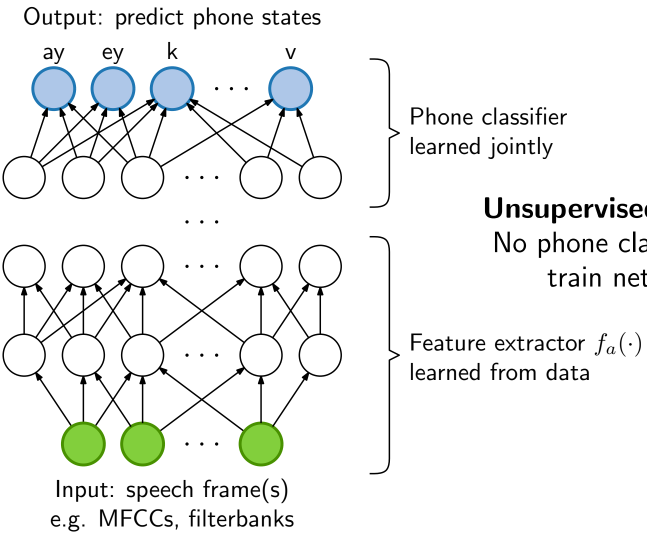
Supervised representation learning using DNN



Supervised representation learning using DNN

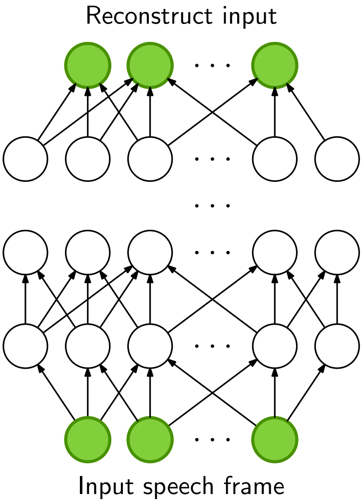


Supervised representation learning using DNN



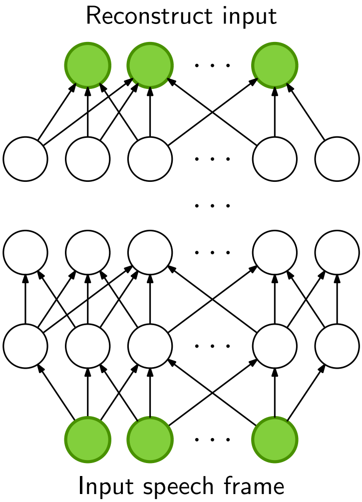
Unsupervised modelling:
No phone class targets to train network on

Autoencoder (AE) neural network



[Badino et al., ICASSP'14]

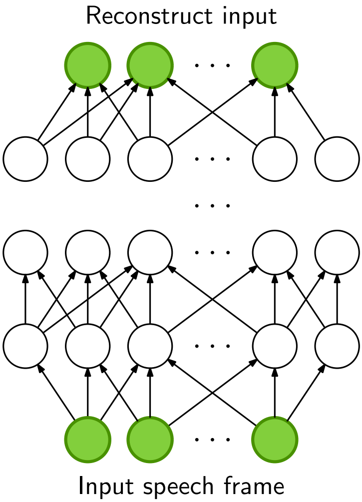
Autoencoder (AE) neural network



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

[Badino et al., ICASSP'14]

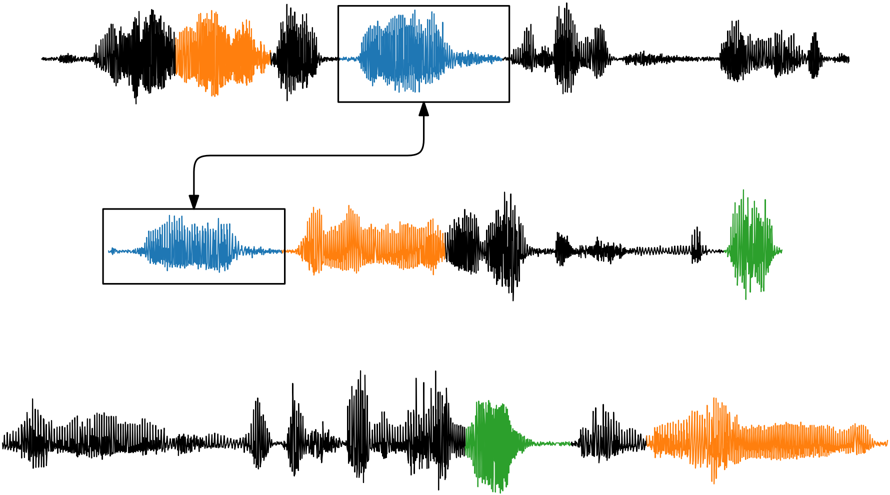
Autoencoder (AE) neural network



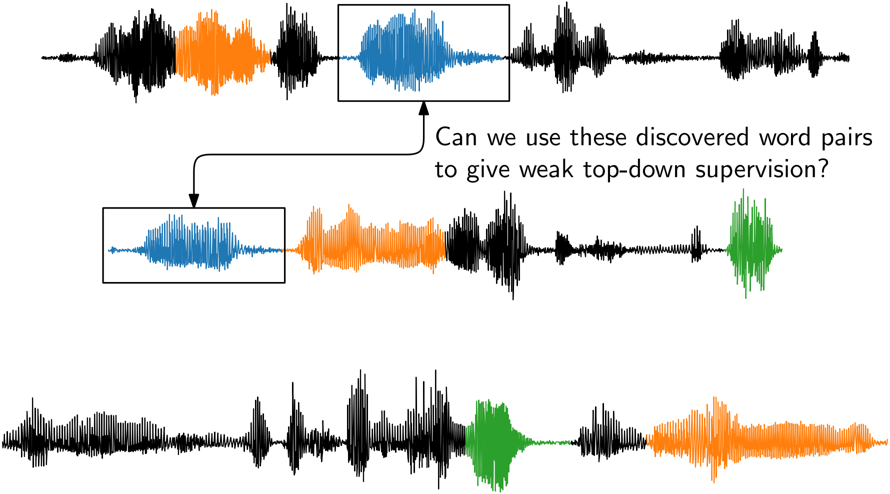
- Completely unsupervised
- But purely bottom-up
- Can we use top-down information?
- **Idea:** Unsupervised term discovery

[Badino et al., ICASSP'14]

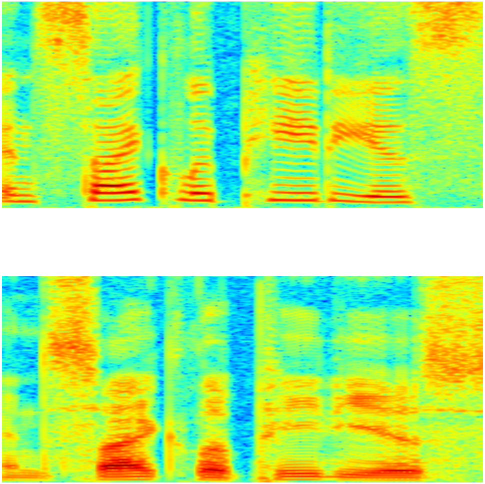
Unsupervised term discovery (UTD)



Unsupervised term discovery (UTD)

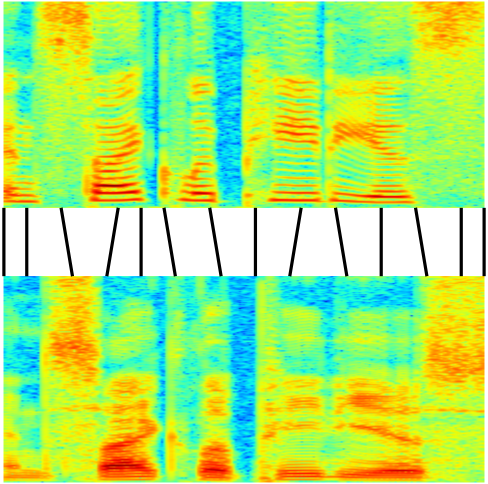


Weak top-down supervision: Align frames



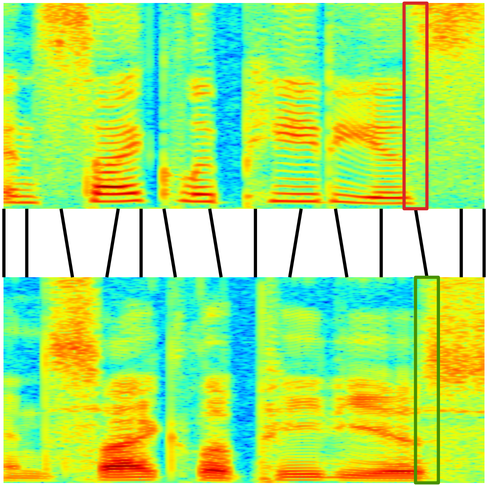
[Jansen et al., ICASSP'13]

Weak top-down supervision: Align frames



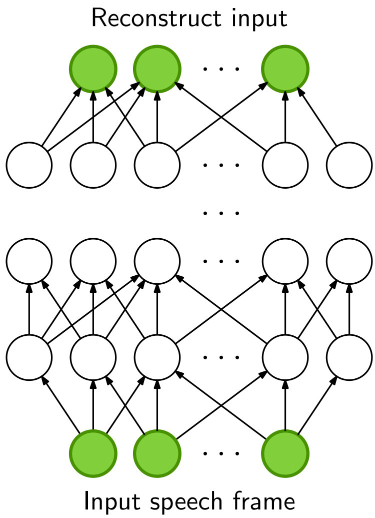
[Jansen et al., ICASSP'13]

Weak top-down supervision: Align frames

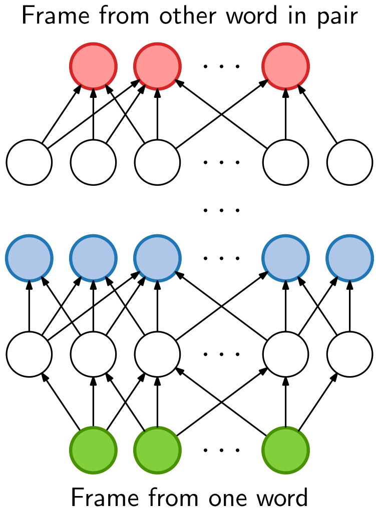


[Jansen et al., ICASSP'13]

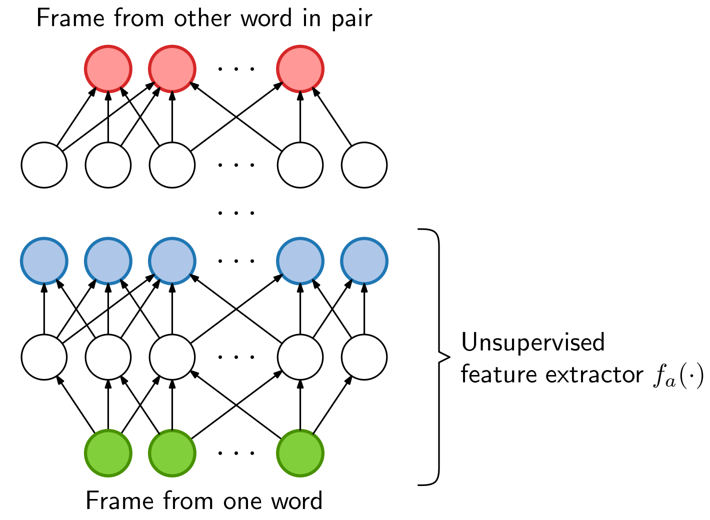
Autoencoder (AE)



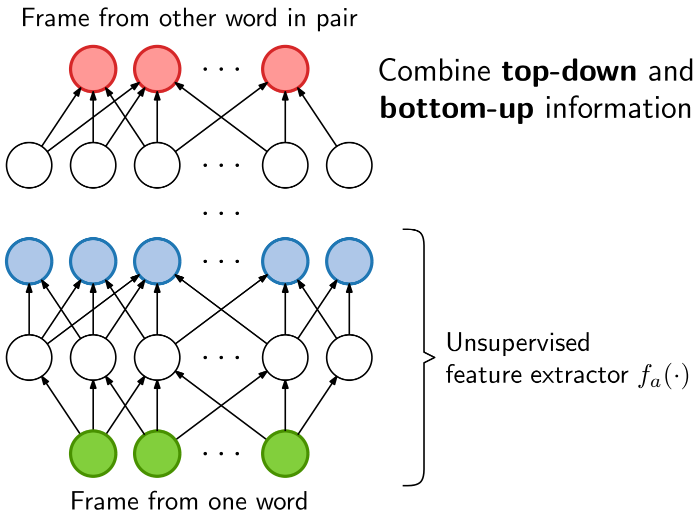
Correspondence autoencoder (cAE)



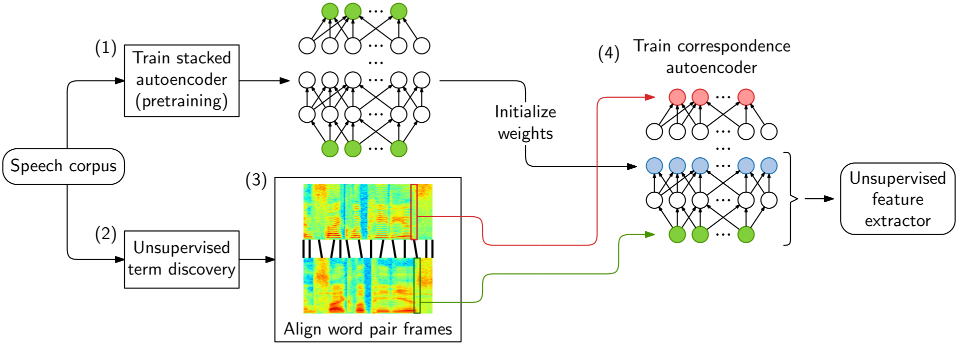
Correspondence autoencoder (cAE)



Correspondence autoencoder (cAE)

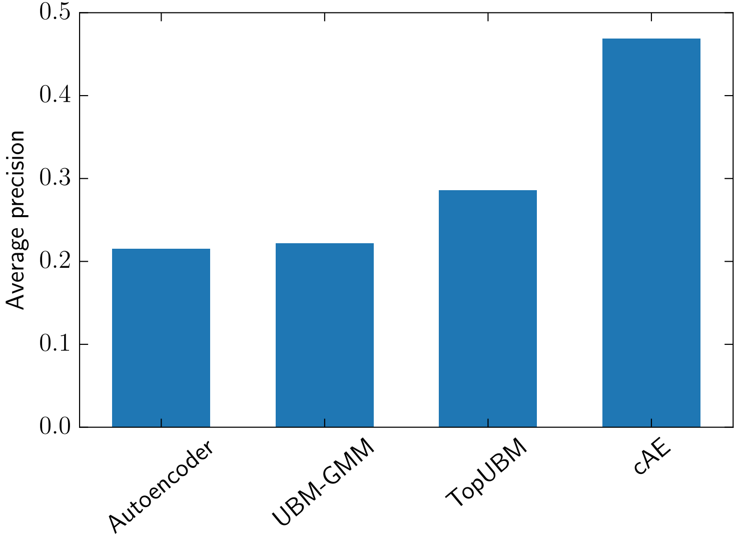


Correspondence autoencoder (cAE)

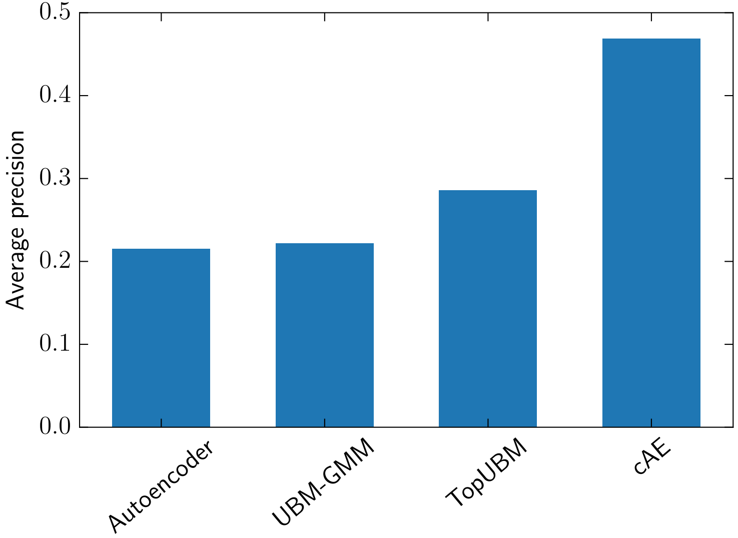


[Kamper et al., ICASSP'15]

Intrinsic evaluation: Isolated word query task



Intrinsic evaluation: Isolated word query task



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

Unsupervised segmentation and clustering:

The Segmental Bayesian Model

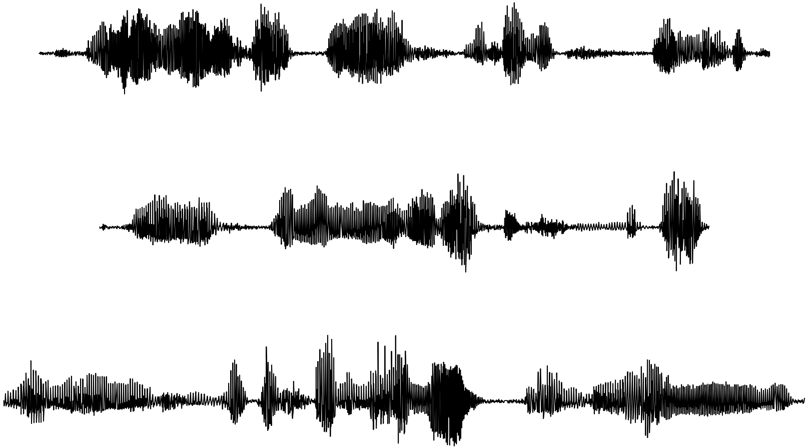
Unsupervised segmentation and clustering:

The Segmental Bayesian Model

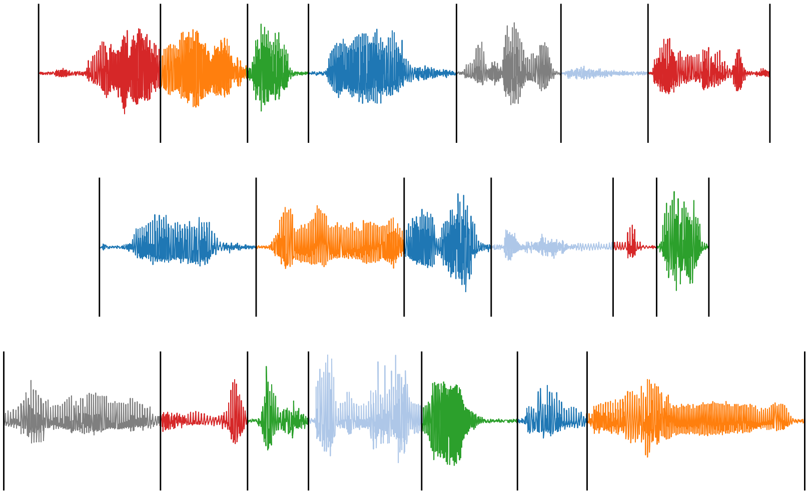
Aren Jansen

Sharon Goldwater

Full-coverage segmentation and clustering

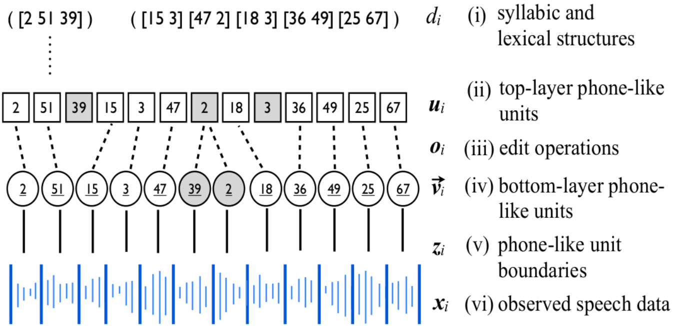


Full-coverage segmentation and clustering



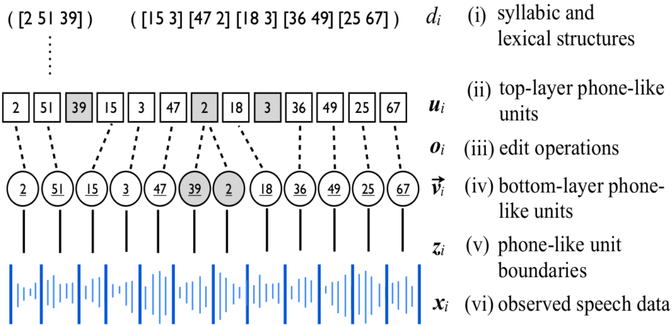
Segmental modelling for full-coverage segmentation

Previous models use explicit subword discovery directly on speech features, e.g. [Lee et al., 2015]:



Segmental modelling for full-coverage segmentation

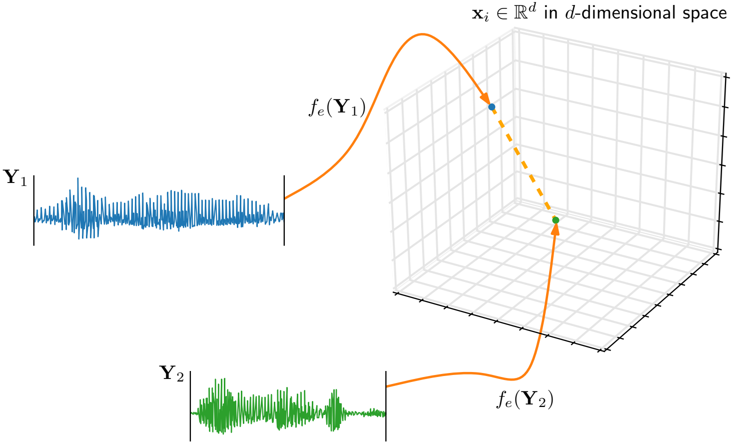
Previous models use explicit subword discovery directly on speech features, e.g. [Lee et al., 2015]:



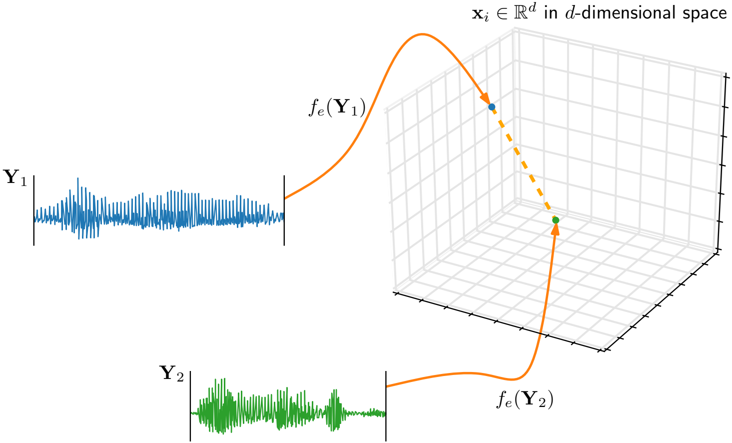
Our approach uses whole-word segmental representations, i.e. acoustic word embeddings [Kamper et al., IS'15; Kamper et al., TASLP'16]

Acoustic word embeddings

Acoustic word embeddings



Acoustic word embeddings

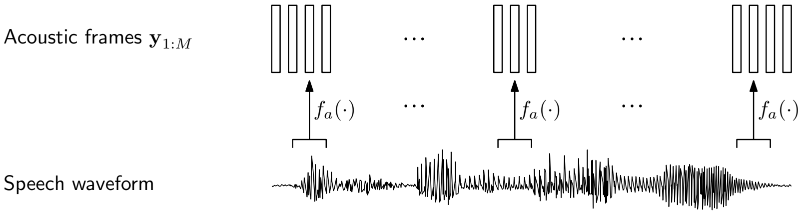


Dynamic programming alignment has quadratic complexity, while embedding comparison is linear time. Can use standard clustering.

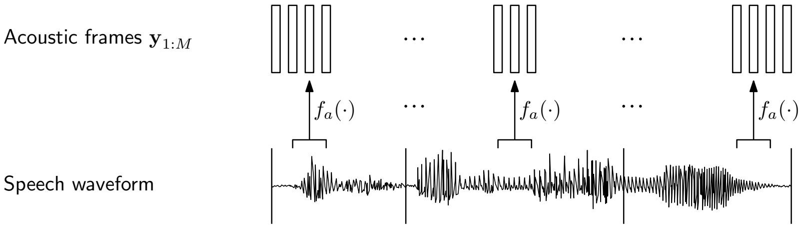
Unsupervised segmental Bayesian model

Speech waveform

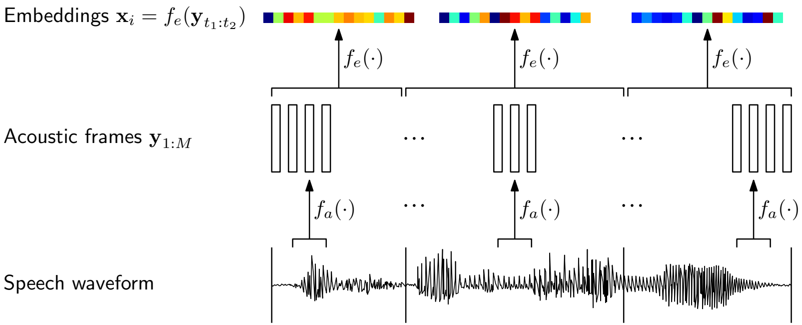
Unsupervised segmental Bayesian model



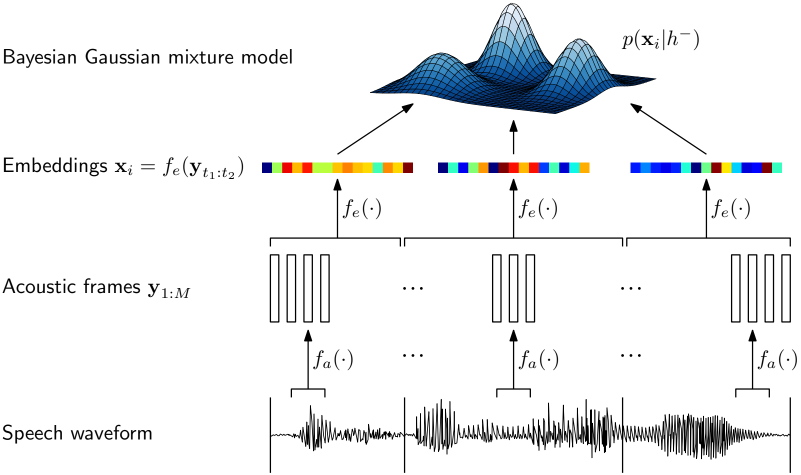
Unsupervised segmental Bayesian model



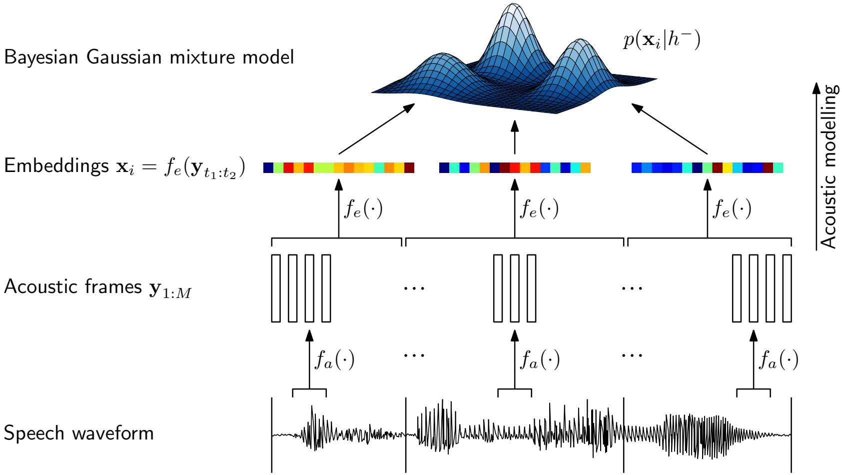
Unsupervised segmental Bayesian model



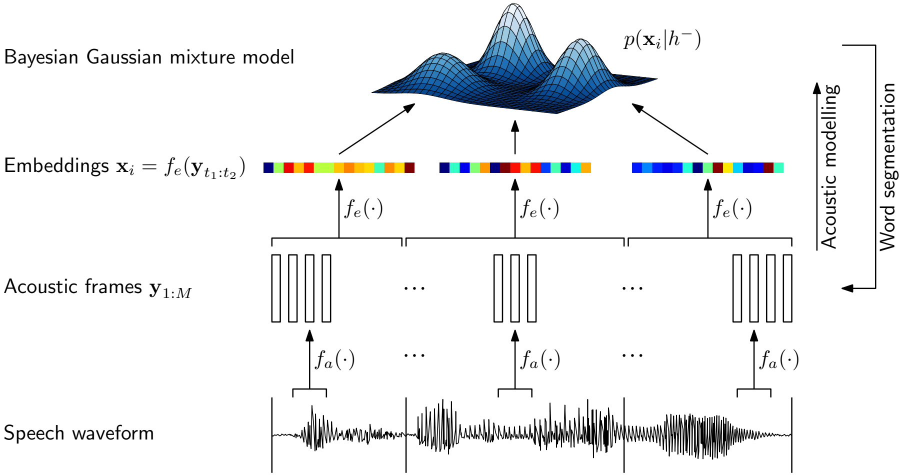
Unsupervised segmental Bayesian model



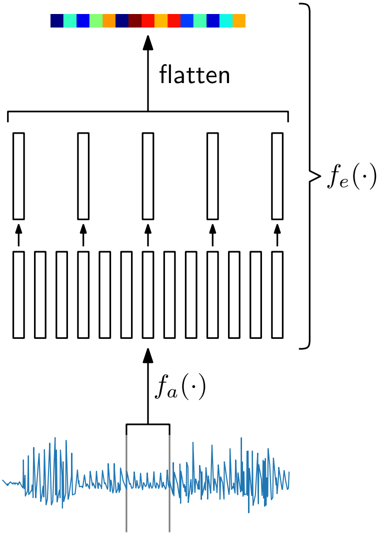
Unsupervised segmental Bayesian model



Unsupervised segmental Bayesian model

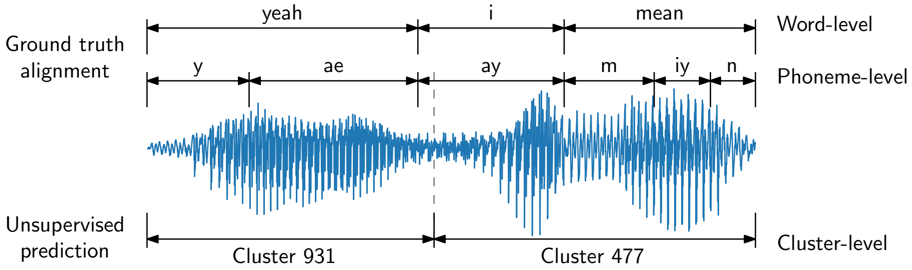


Acoustic word embeddings: Downsampling

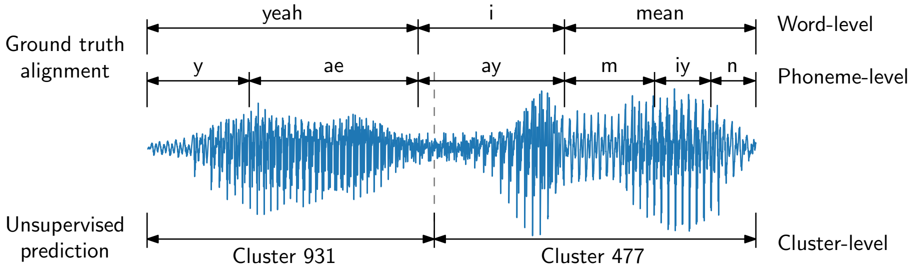


- Simple embedding approach also used in other studies
  e.g. [Abdel-Hamid et al., 2013]
- Consider both MFCCs and cAE features as frame-level function $f_a(\cdot)$
- cAE combines top-down learned feature representations with segmentation and clustering

Evaluation


yeah
i
mean
Word-level
Ground truth
alignment
y
ae
ay
m
iy
n
Phoneme-level
Unsupervised
prediction
Cluster 931
Cluster 477
Cluster-level

Evaluation

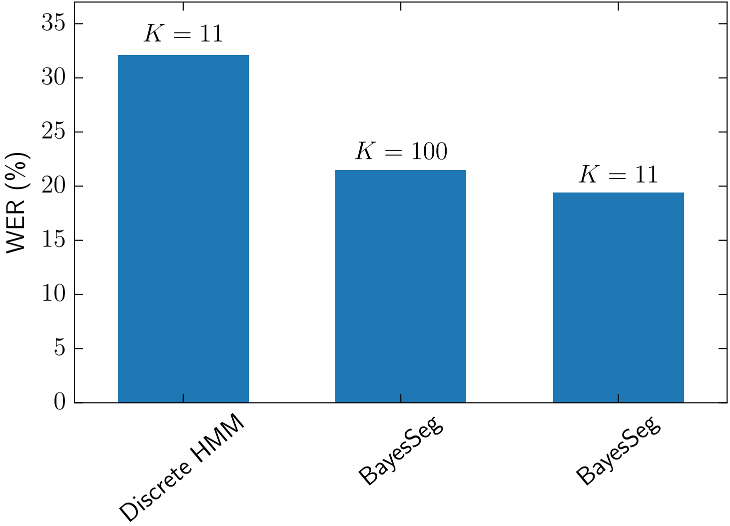


Metrics:

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

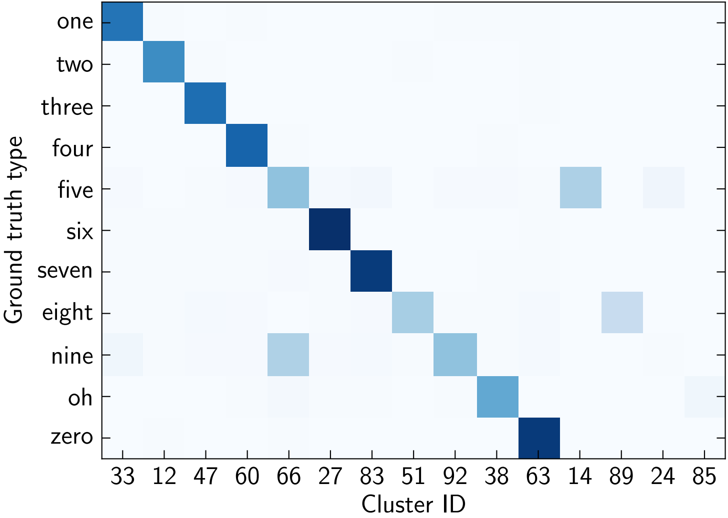
Small-vocabulary segmentation and clustering

Small-vocabulary segmentation and clustering



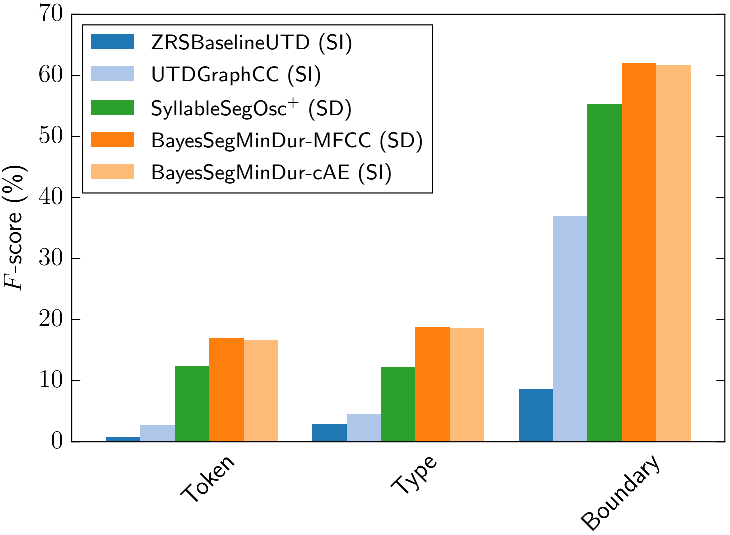
Discrete HMM: [Walter et al., ASRU'13]. BayesSeg: [Kamper et al., TASLP'16].

Small-vocabulary segmentation and clustering



[Kamper et al., TASLP'16]

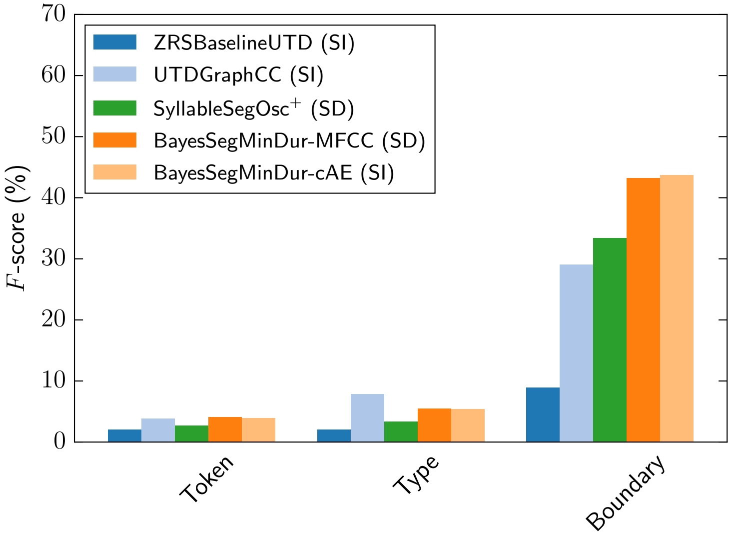
Large-vocabulary: English



ZRSBaselineUTD: [Versteegh et al., IS'15]. UTDGraphCC: [Lyzinski et al., IS'15].

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

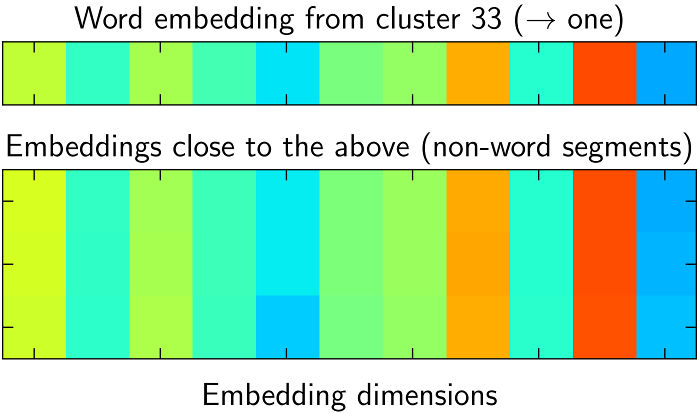
Large-vocabulary: Xitsonga



ZRSBaselineUTD: [Versteegh et al., IS'15]. UTDGraphCC: [Lyzinski et al., IS'15].

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

The true (less rosy) picture

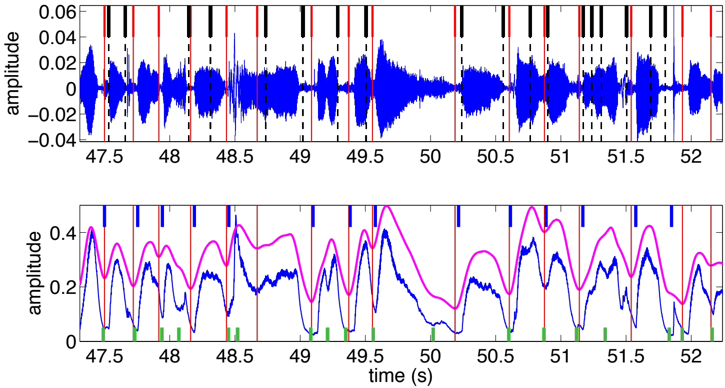


Bottom-up constraints

- Minimum and maximum duration constraints

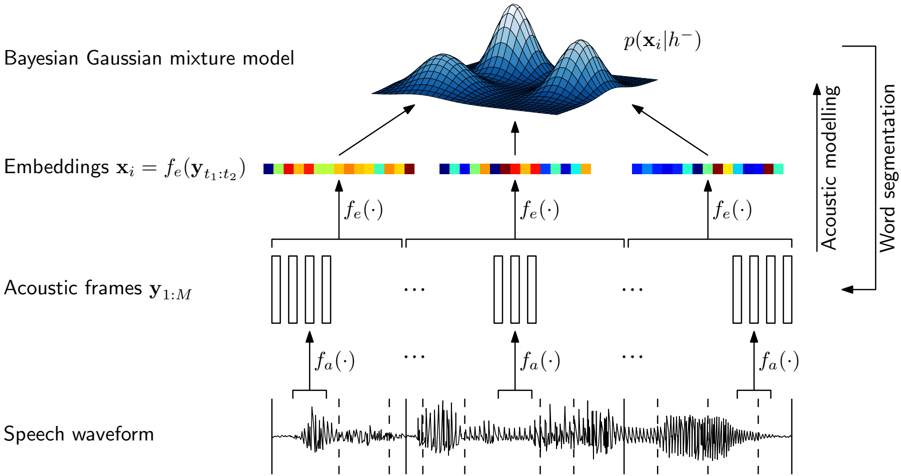
Bottom-up constraints

- Minimum and maximum duration constraints
- Use unsupervised syllable boundary detection:

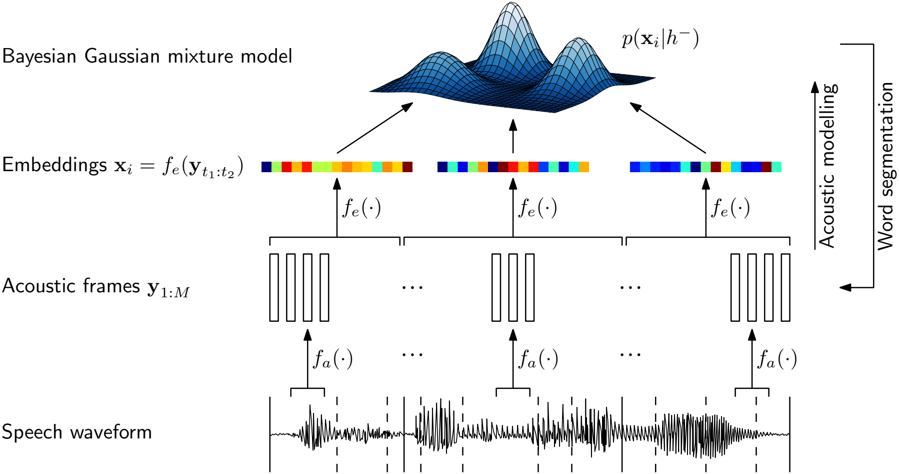


[Räsänen et al., IS'15]

Bottom-up constraints



Bottom-up constraints



Performs **top-down** segmentation while adhering to **bottom-up** constraints

Effect of using cAE features

	English (%)			Xitsonga (%)		
Embeds.	Cluster	Speaker	Gender	Cluster	Speaker	Gender
MFCC	29.9	55.9	87.6	24.5	43.1	87.1
cAE	**30.0**	**35.7**	**73.8**	**33.1**	**29.3**	**76.6**

Summary and Conclusions

Conclusions

Unsupervised speech processing benefits from both top-down and bottom-up modelling

Conclusions

Unsupervised speech processing benefits from both top-down and bottom-up modelling

- **Correspondence autoencoder:** Use top-down constraints with bottom-up initialization to improve frame-level representations
- **Segmental Bayesian model:** Top-down segmentation taking bottom-up constraints into account
- **English and Xitsonga:** Large-vocabulary multi-speaker data
- **cAE in BayesSeg:** Improves cluster, speaker and gender purity

Extending this work

- Improve cAE using UTD and vice versa (with Sameer Bansal)
- Improve unsupervised acoustic word embeddings [Chung et al., IS'16]
- Simplify BayesSeg so that it can be applied to larger corpora
- Frame-based vs. segmental unsupervised models
- Evaluation: What do we want to discover?

Looking forward

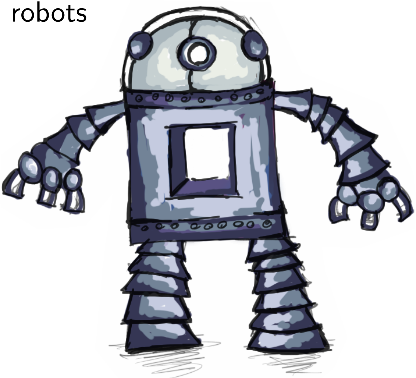
- Building audio analysis tools for field linguists

Looking forward

- Building audio analysis tools for field linguists
- Using weak labels, e.g. translations [Bansal et al., arXiv'16]
  (with Sameer Bansal, Adam Lopez, Sharon Goldwater)

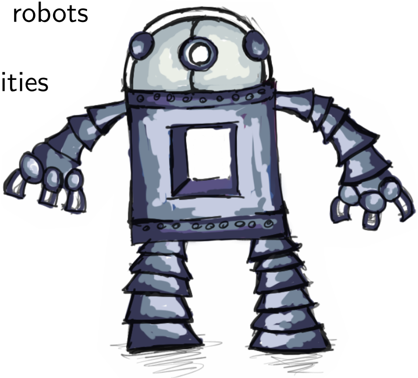
Looking forward

- Building audio analysis tools for field linguists
- Using weak labels, e.g. translations [Bansal et al., arXiv'16]
  (with Sameer Bansal, Adam Lopez, Sharon Goldwater)
- Language acquisition in humans and robots



Looking forward

- Building audio analysis tools for field linguists
- Using weak labels, e.g. translations [Bansal et al., arXiv'16]
  (with Sameer Bansal, Adam Lopez, Sharon Goldwater)
- Language acquisition in humans and robots
- Extending models to multiple modalities
  (with Shane Settle, Karen Livescu,
  Greg Shakhnarovich)



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.
- L. Badino, C. Canevari, L. Fadiga, and G. Metta, “An auto-encoder based approach to unsupervised learning of subword units,” in *Proc. ICASSP*, 2014.
- S. Bansal, H. Kamper, S. J. Goldwater, and A. Lopez, “Weakly supervised spoken term discovery using cross-lingual side information,” *arXiv preprint arXiv:1609.06530*, 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.
- Y.-A. Chung, C.-C. Wu, C.-H. Shen, and H.-Y. Lee, “Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks,” *Proc. Interspeech*, 2016.
- N. H. Feldman, T. L. Griffiths, and J. L. Morgan, “Learning phonetic categories by learning a lexicon,” in *Proc. CCSS*, 2009.
- A. Jansen, S. Thomas, and H. Hermansky, “Weak top-down constraints for unsupervised acoustic model training,” in *Proc. ICASSP*, 2013.
- 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.

References II

- 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.
- 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, 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.
- 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.
- A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 186–197, 2008.

References III

- O. J. Räsänen, "Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions," *Speech Commun.*, vol. 54, pp. 975–997, 2012.
- 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.
- V. Renkens and H. Van hamme, "Mutually exclusive grounding for weakly supervised non-negative matrix factorisation," in *Proc. Interspeech*, 2015.
- D. Renshaw, H. Kamper, A. Jansen, and S. J. Goldwater, "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge," in *Proc. Interspeech*, 2015.
- 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. 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.

References IV

- Y. Yuan, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Learning neural network representations using cross-lingual bottleneck features with word-pair information," in *Proc. Interspeech*, 2016.