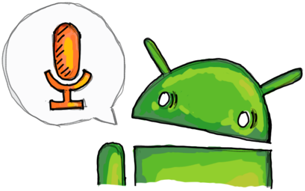
Unsupervised speech processing using acoustic word embeddings

Herman Kamper

School of Informatics, University of Edinburgh → TTI at Chicago

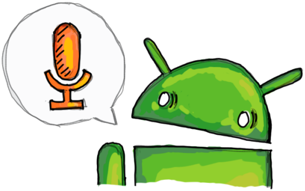
MLSLP 2016: Spotlight invited talk

Unsupervised speech processing



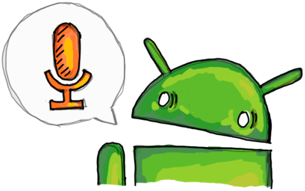
- Speech recognition applications are becoming wide-spread
- Google Voice Search already supports more than 50 languages: English, Spanish, German, . . . , Afrikaans, Zulu

Unsupervised speech processing



- Speech recognition applications are becoming wide-spread
- Google Voice Search already supports more than 50 languages: English, Spanish, German, . . . , Afrikaans, Zulu
- But there are roughly 7000 languages spoken in the world!
- Audio data are becoming available, even for languages spoken by only a few speakers, but generally unlabelled

Unsupervised speech processing



- Speech recognition applications are becoming wide-spread
- Google Voice Search already supports more than 50 languages: English, Spanish, German, . . . , Afrikaans, Zulu
- But there are roughly 7000 languages spoken in the world!
- Audio data are becoming available, even for languages spoken by only a few speakers, but generally unlabelled
- Goal: Unsupervised learning of linguistic structure directly from raw speech audio, in order to develop zero-resource speech technology

Motivation for unsupervised speech processing

Criticism:

- Always some labelled data to start with (e.g. related language)
- Small set of labelled data: semi-supervised problem

Motivation for unsupervised speech processing

Criticism:

- Always some labelled data to start with (e.g. related language)
- Small set of labelled data: semi-supervised problem

Reasons for focusing on purely unsupervised case:

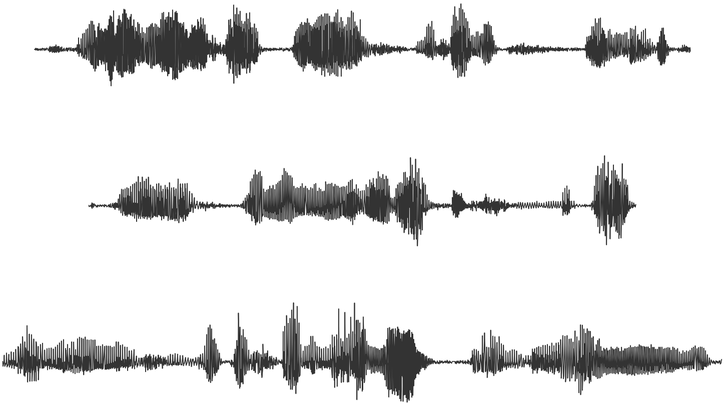
- Modelling infant language acquisition [Räsänen, 2012]
- Language acquisition in robotics [Renkens and Van hamme, 2015]
- Practical use of zero-resource technology: Allow linguists to analyze and investigate unwritten languages [Besacier et al., 2014]
- New insights and models for speech processing: E.g. unsupervised methods can improve supervised systems [Jansen et al., 2012]

Unsupervised segmentation and clustering

Full-coverage segmentation:

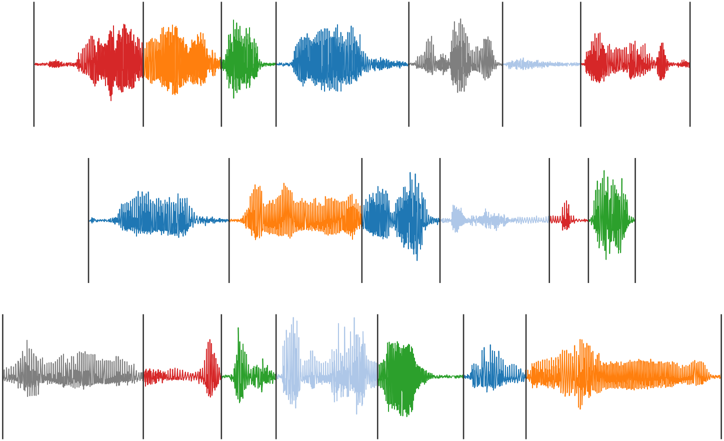
Unsupervised segmentation and clustering

Full-coverage segmentation:



Unsupervised segmentation and clustering

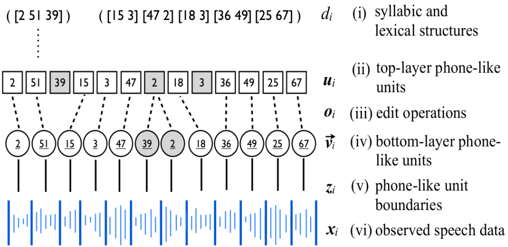
Full-coverage segmentation:



Segmental modelling for full-coverage segmentation

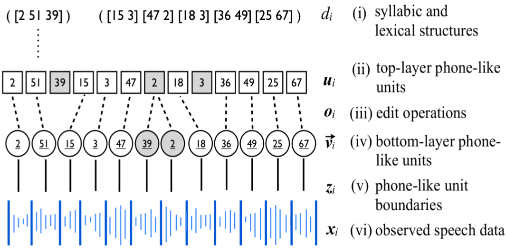
Segmental modelling for full-coverage segmentation

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]:

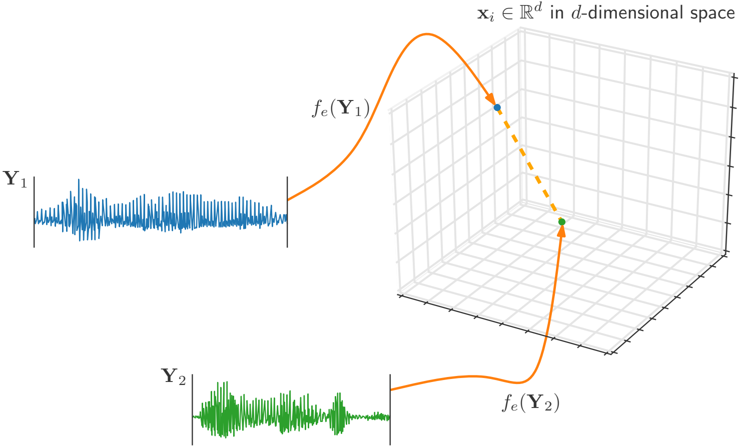


Our approach uses whole-word segmental representations, i.e. acoustic word embeddings

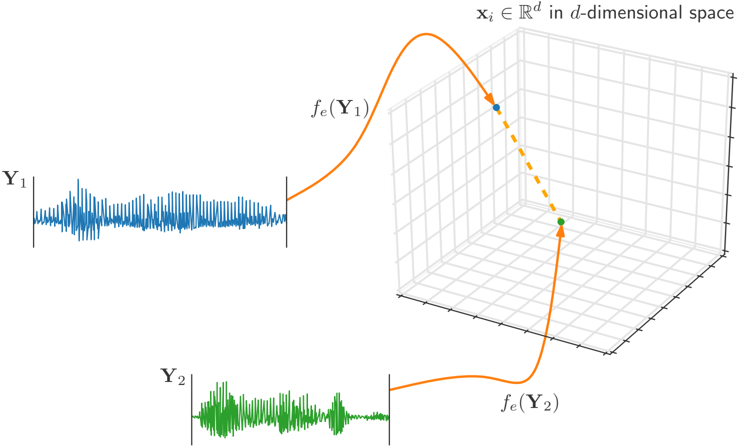
[Kamper et al., IS'15; Kamper et al., TASLP'16]

Acoustic word embeddings

Acoustic word embeddings



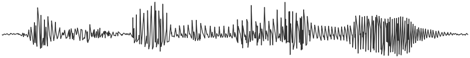
Acoustic word embeddings



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

An unsupervised segmental Bayesian model

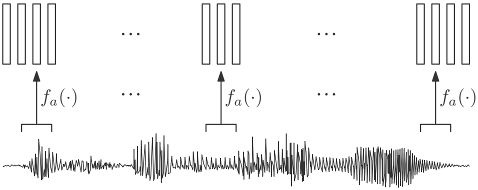
Speech waveform



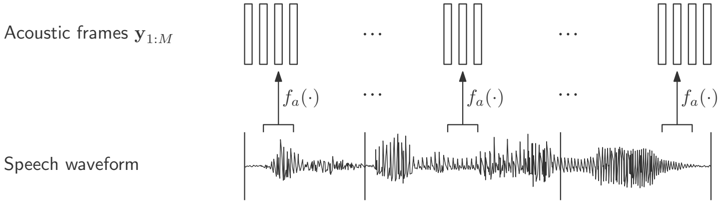
An unsupervised segmental Bayesian model

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

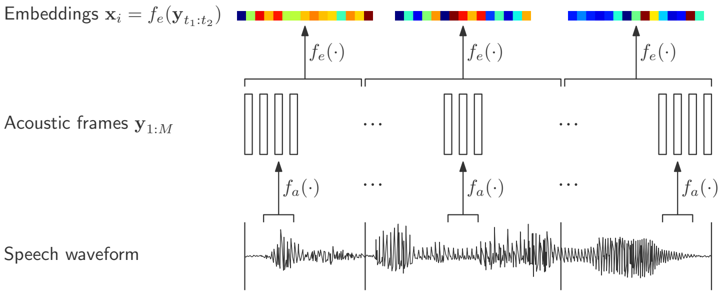
Speech waveform



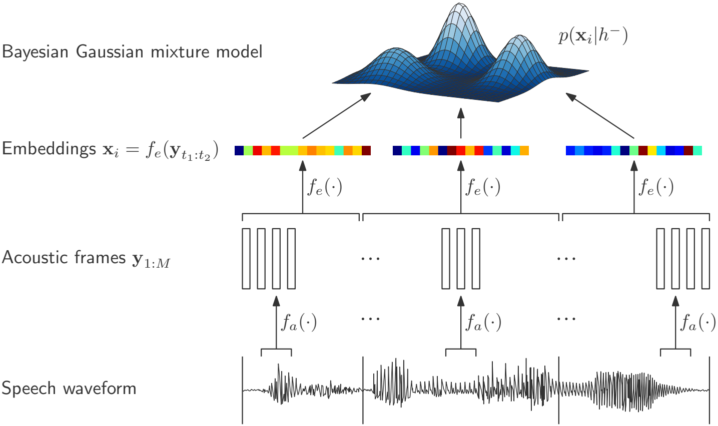
An unsupervised segmental Bayesian model



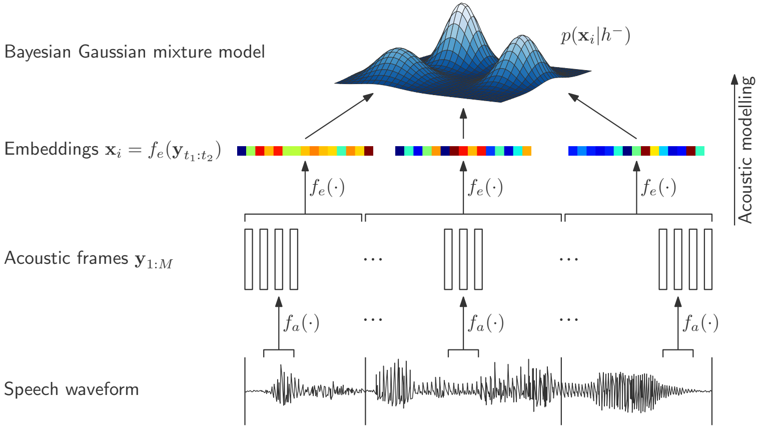
An unsupervised segmental Bayesian model



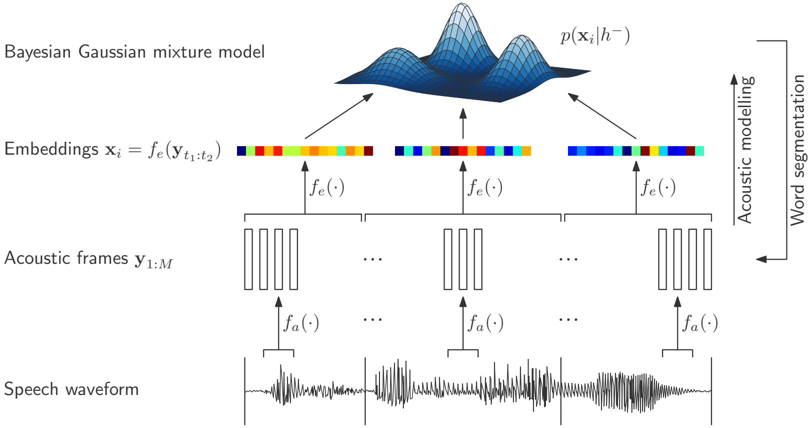
An unsupervised segmental Bayesian model



An unsupervised segmental Bayesian model

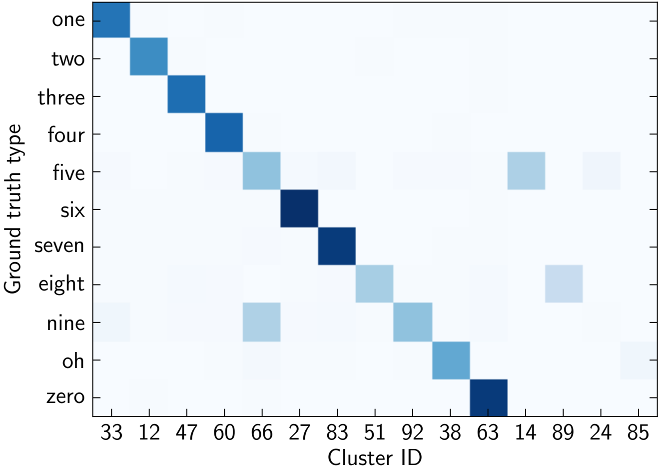


An unsupervised segmental Bayesian model



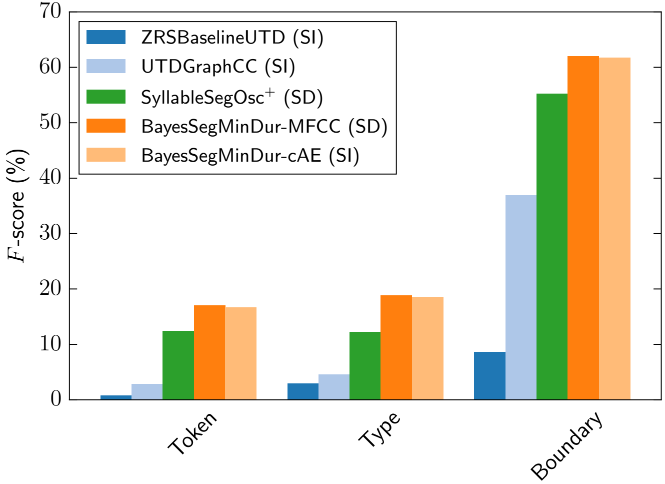
Applied to a small-vocabulary task

Applied to a small-vocabulary task



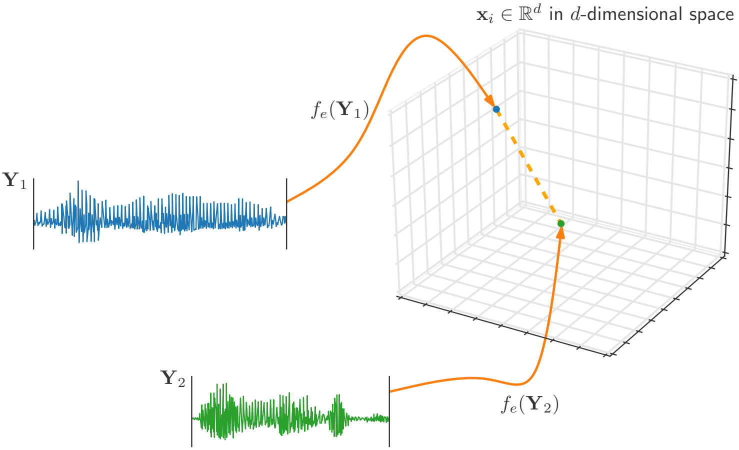
Applied to a large-vocabulary task

Applied to a large-vocabulary task



ZRSBaselineUTD: [Versteegh et al., 2015]; UTDGraphCC: [Lyzinski et al., 2015]; SyllableSegOsc$^+$: [Räsänen et al., 2015]

Acoustic word embeddings



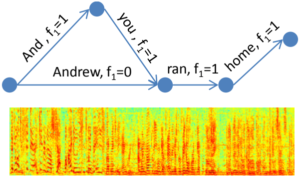
Acoustic word embeddings

Useful for more than just unsupervised modelling

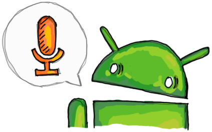
Acoustic word embeddings

Useful for more than just unsupervised modelling

- Segmental conditional random field ASR [Maas et al., 2012]:



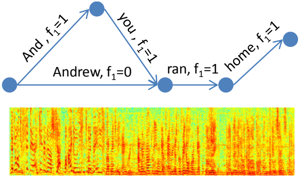
- Whole-word lattice rescoring [Bengio and Heigold, 2014]
- Query-by-example search, e.g. [Chen et al., 2015] for "Okay Google":



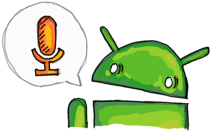
Acoustic word embeddings

Useful for more than just unsupervised modelling

- Segmental conditional random field ASR [Maas et al., 2012]:



- Whole-word lattice rescoring [Bengio and Heigold, 2014]
- Query-by-example search, e.g. [Chen et al., 2015] for "Okay Google":



Word classification CNN

Fully supervised approach

[Bengio and Heigold, 2014]

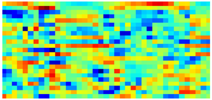
Word classification CNN

Fully supervised approach

[Bengio and Heigold, 2014]

$$w_i$$

$$\boxed{0\ 0\ 0\ \cdots\ 1\ \cdots\ 0\ 0}$$



$\mathbf{Y}_i$

Word classification CNN

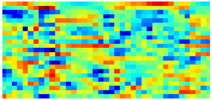
Fully supervised approach

[Bengio and Heigold, 2014]

softmax

w_i

$\boxed{0\ 0\ 0\ \cdots\ 1\ \cdots\ 0\ 0}$



$\mathbf{Y}_i$

Word classification CNN

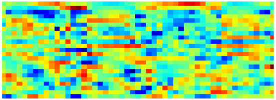
Fully supervised approach

[Bengio and Heigold, 2014]

softmax

w_i

$\boxed{0\ 0\ 0\ \cdots\ 1\ \cdots\ 0\ 0}$



$\mathbf{Y}_i$

Word classification CNN

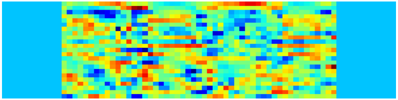
Fully supervised approach

[Bengio and Heigold, 2014]

softmax

$$w_i$$

$$\boxed{0\ 0\ 0\ \cdots\ 1\ \cdots\ 0\ 0}$$

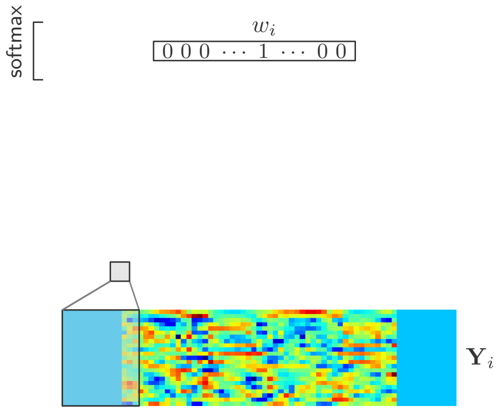


$$\mathbf{Y}_i$$

Word classification CNN

Fully supervised approach

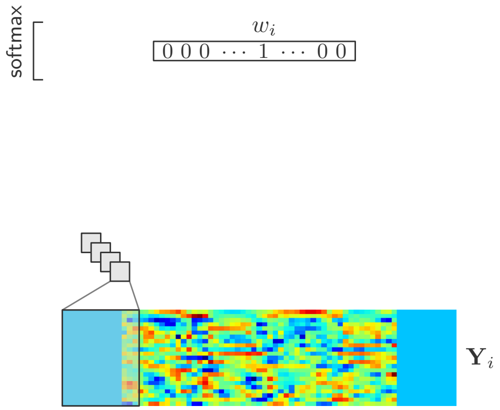
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

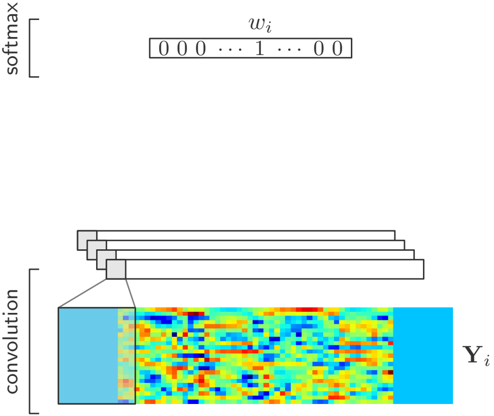
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

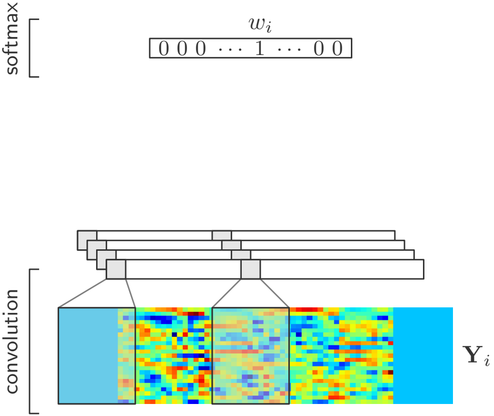
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

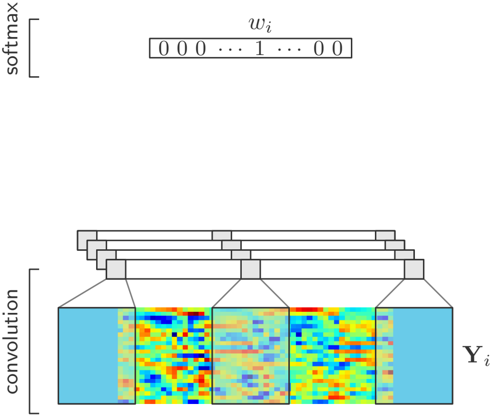
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

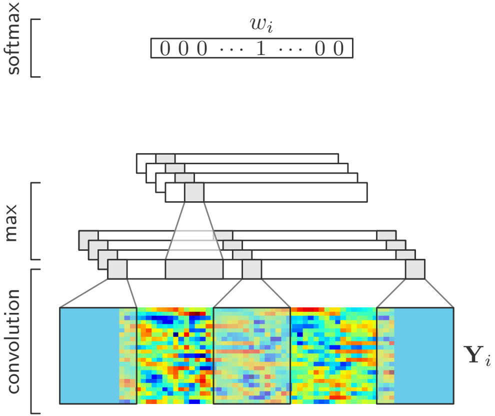
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

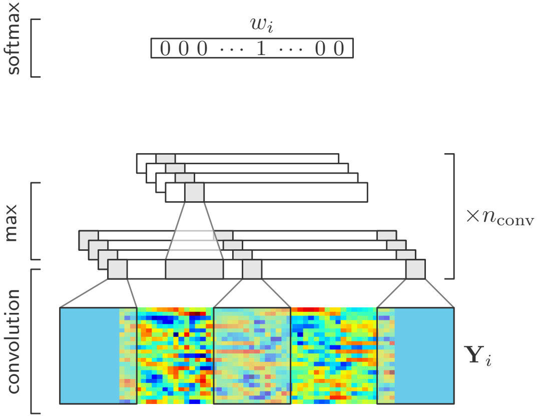
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

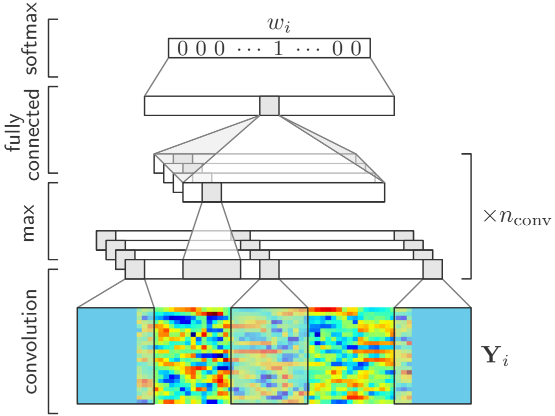
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

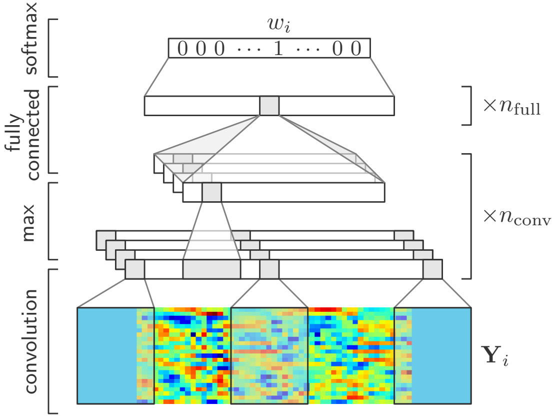
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

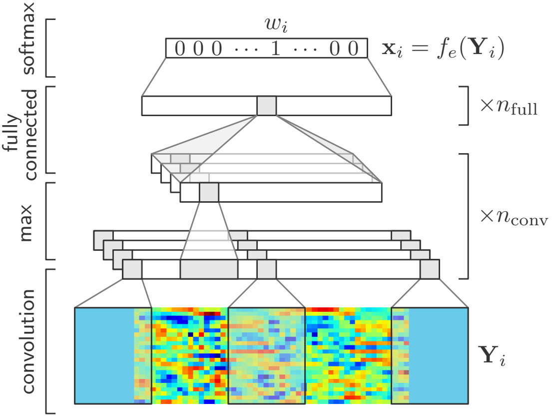
[Bengio and Heigold, 2014]



Word classification CNN

Fully supervised approach

[Bengio and Heigold, 2014]



Word similarity Siamese CNN

Weak supervision we sometimes have [Thiollière et al., 2015] are known word pairs: $\mathcal{S}_{\text{train}} = \{(m, n) : (Y_m, Y_n) \text{ are of the same type}\}$

Word similarity Siamese CNN

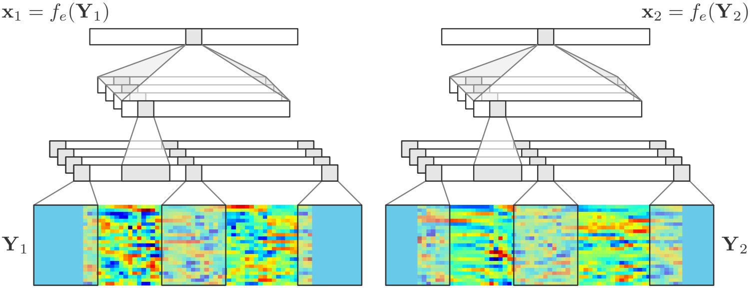
Weak supervision we sometimes have [Thiollière et al., 2015] are known word pairs: $\mathcal{S}_{\text{train}} = \{(m, n) : (Y_m, Y_n) \text{ are of the same type}\}$

Use idea of Siamese networks [Bromley et al., 1993]

Word similarity Siamese CNN

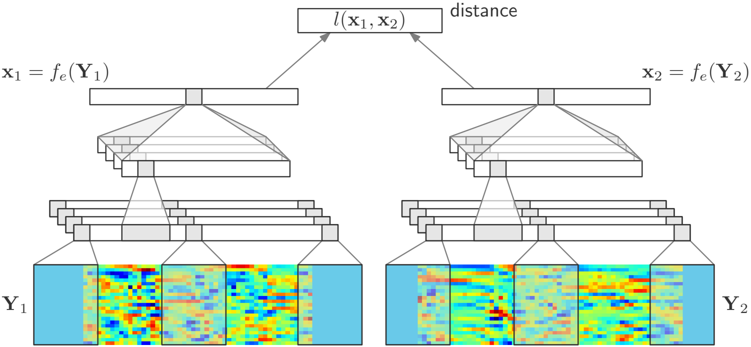
Weak supervision we sometimes have [Thiollière et al., 2015] are known word pairs: $\mathcal{S}_{\text{train}} = \{(m, n) : (Y_m, Y_n) \text{ are of the same type}\}$

Use idea of Siamese networks [Bromley et al., 1993]

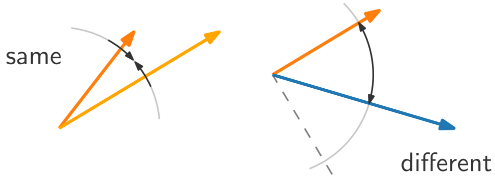


Weak supervision we sometimes have [Thiollière et al., 2015] are known word pairs: $\mathcal{S}_{\text{train}} = \{(m, n) : (Y_m, Y_n) \text{ are of the same type}\}$

Use idea of Siamese networks [Bromley et al., 1993]



Triplet margin-based loss

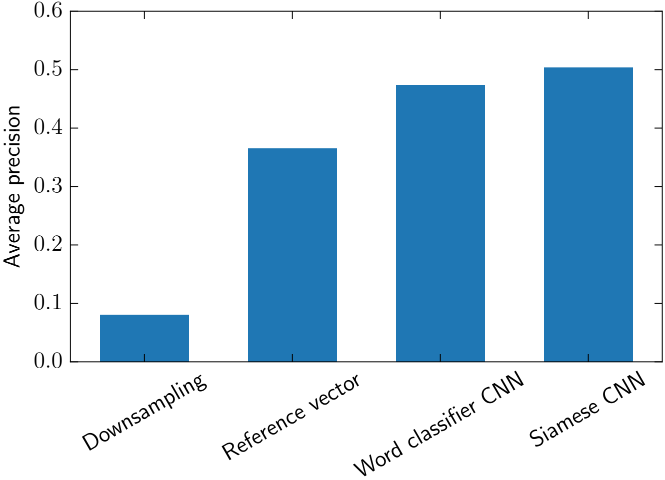


Margin-based triplet hinge loss [Mikolov, 2013]:

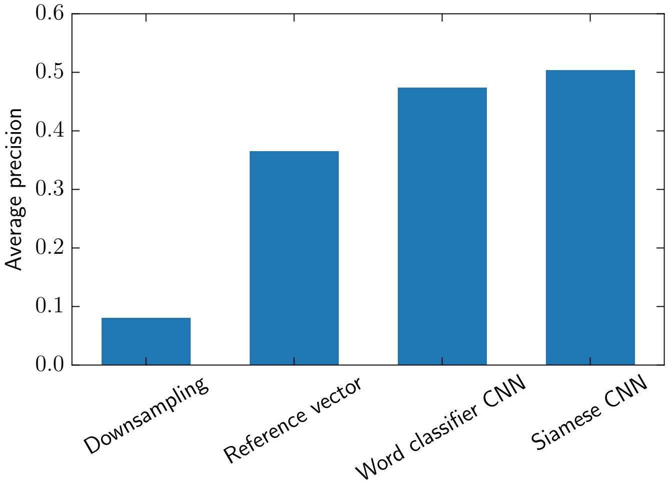
$$l_{\text{triplets}} = \max\{0, m + d_{\cos}(\mathbf{x}_1, \mathbf{x}_2) - d_{\cos}(\mathbf{x}_1, \mathbf{x}_3)\}$$

where $d_{\cos}(\mathbf{x}_1, \mathbf{x}_2) = \frac{1-\cos(\mathbf{x}_1, \mathbf{x}_2)}{2}$ is the cosine distance between $\mathbf{x}_1$ and $\mathbf{x}_2$, and m is a margin parameter. Pair $(\mathbf{x}_1, \mathbf{x}_2)$ is same, $(\mathbf{x}_1, \mathbf{x}_3)$ is different.

Evaluation of acoustic word embeddings



Evaluation of acoustic word embeddings



But Siamese CNN still uses weak supervision. Still work to be done for unsupervised case, e.g. [Chung et al., IS'16].

Looking forward

Looking forward

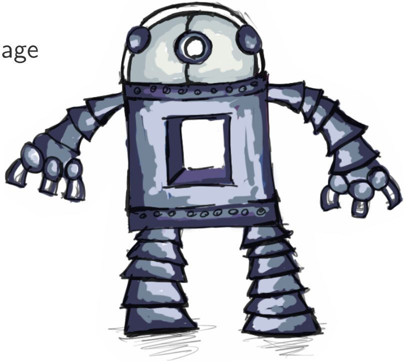
- Much to be done in zero-resource speech processing
- Core issues: evaluation; what do we want to discover?

Looking forward

- Much to be done in zero-resource speech processing
- Core issues: evaluation; what do we want to discover?
- Do these models allow us to model language acquisition in human infants?

Looking forward

- Much to be done in zero-resource speech processing
- Core issues: evaluation; what do we want to discover?
- Do these models allow us to model language acquisition in human infants?
- Can these models be used for language acquisition in robotic applications?
- Extensions to multiple modalities



Take-aways

- Unsupervised, or zero-resource, speech processing is an important and cool problem
- Segmental acoustic word embeddings is a sensible way to approach unsupervised segmentation and clustering, and is cool in general
- Interesting to look at speech problems from a different perspective: allows you to play around with cool models, and get new insights

Poster: Better features using the correspondence autoencoder

Two problems in zero-resource speech processing:

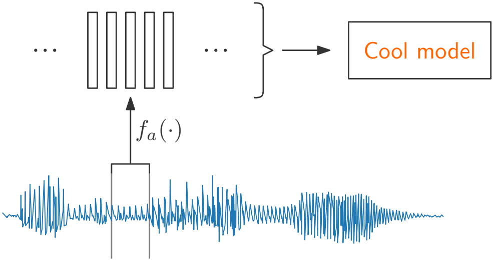
1. Unsupervised segmentation and clustering

Poster: Better features using the correspondence autoencoder

Two problems in zero-resource speech processing:

1. Unsupervised segmentation and clustering

2. Unsupervised frame-level representation learning:



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

References I

- Abdel-Hamid, O., Deng, L., Yu, D., and Jiang, H. (2013).
  Deep segmental neural networks for speech recognition.
  In *Proc. Interspeech*.
- Bengio, S. and Heigold, G. (2014).
  Word embeddings for speech recognition.
  In *Proc. Interspeech*.
- Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014).
  Automatic speech recognition for under-resourced languages: A survey.
  Speech Commun., 56:85–100.
- Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993).
  Signature verification using a 'Siamese' time delay neural network.
  Int. J. Pattern Rec., 7(4):669–688.

References II

- Chen, G., Parada, C., and Sainath, T. N. (2015).
  Query-by-example keyword spotting using long short-term memory networks.
  In *Proc. ICASSP*.
- Chung, Y.-A., Wu, C.-C., Shen, C.-H., and Lee, H.-Y. (2016).
  Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks.
  Proc. Interspeech.
- Jansen, A. et al. (2013).
  A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition.
  In *Proc. ICASSP*.

References III

- Kamper, H., Goldwater, S. J., and Jansen, A. (2015).
  Fully unsupervised small-vocabulary speech recognition using a segmental Bayesian model.
  In *Proc. Interspeech*.
- Kamper, H., Jansen, A., and Goldwater, S. J. (2016a).
  Unsupervised word segmentation and lexicon discovery using acoustic word embeddings.
  IEEE/ACM Trans. Audio, Speech, Language Process., 24(4):669–679.
- Kamper, H., Wang, W., and Livescu, K. (2016b).
  Deep convolutional acoustic word embeddings using word-pair side information.
  In *Proc. ICASSP*.

References IV

- Lee, C.-y., O'Donnell, T., and Glass, J. R. (2015).
  Unsupervised lexicon discovery from acoustic input.
  Trans. ACL, 3:389–403.
- Levin, K., Henry, K., Jansen, A., and Livescu, K. (2013).
  Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings.
  In *Proc. ASRU*.
- Levin, K., Jansen, A., and Van Durme, B. (2015).
  Segmental acoustic indexing for zero resource keyword search.
  In *Proc. ICASSP*.
- Lyzinski, V., Sell, G., and Jansen, A. (2015).
  An evaluation of graph clustering methods for unsupervised term discovery.
  In *Proc. Interspeech*.

References V

- Maas, A. L., Miller, S. D., O'Neil, T. M., Ng, A. Y., and Nguyen, P. (2012).
  Word-level acoustic modeling with convolutional vector regression.
  In *Proc. ICML Workshop Representation Learn.*
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
  Efficient estimation of word representations in vector space.
  arXiv preprint arXiv:1301.3781.
- Räsänen, O. J. (2012).
  Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions.
  Speech Commun., 54:975–997.

References VI

- Räsänen, O. J., Doyle, G., and Frank, M. C. (2015).
  Unsupervised word discovery from speech using automatic segmentation into syllable-like units.
  In *Proc. Interspeech*.
- Renkens, V. and Van hamme, H. (2015).
  Mutually exclusive grounding for weakly supervised non-negative matrix factorisation.
  In *Proc. Interspeech*.
- Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., and Dupoux, E. (2015).
  A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.
  In *Proc. Interspeech*.

References VII

- Versteegh, M., Thiollière, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., and Dupoux, E. (2015).
  The Zero Resource Speech Challenge 2015.
  In *Proc. Interspeech*.