Unsupervised neural network based feature extraction using weak top-down constraints

Herman Kamper[1,2], Micha Elsner[3], Aren Jansen[4], Sharon Goldwater[2]

[1]CSTR and [2]ILCC, School of Informatics, University of Edinburgh, UK
[3]Department of Linguistics, The Ohio State University, USA
[4]HLTCOE and CLSP, Johns Hopkins University, USA

ICASSP 2015

Introduction

- Huge amounts of speech audio data are becoming available online.
- Even for severely under-resourced and endangered languages (e.g. unwritten), data is being collected.
- Generally this data is unlabelled.
- We want to build speech technology on available unlabelled data.

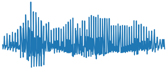
Introduction

- Huge amounts of speech audio data are becoming available online.
- Even for severely under-resourced and endangered languages (e.g. unwritten), data is being collected.
- Generally this data is unlabelled.
- We want to build speech technology on available unlabelled data.
- Need unsupervised speech processing techniques.

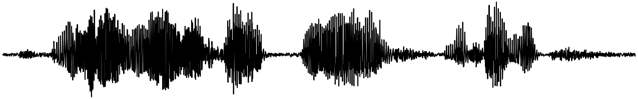
Example application: query-by-example search

Example application: query-by-example search

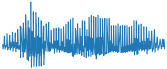
Spoken query:



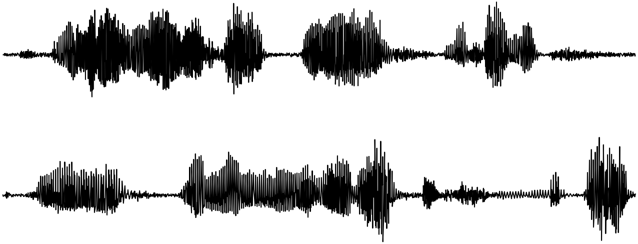
Example application: query-by-example search



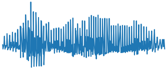
Spoken query:



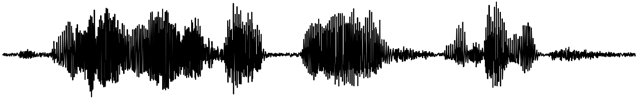
Example application: query-by-example search



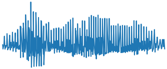
Spoken query:

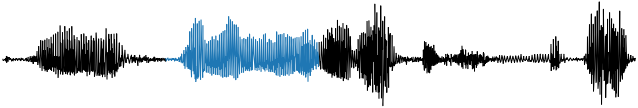


Example application: query-by-example search

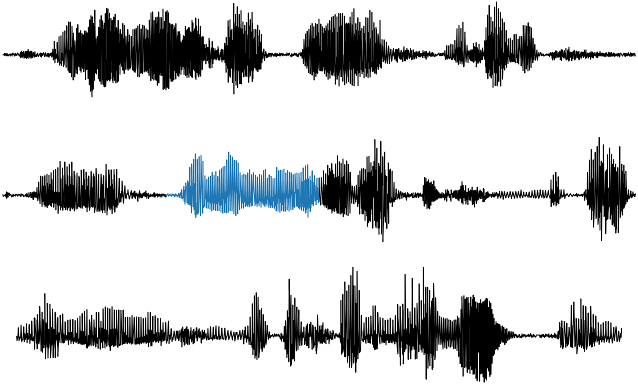


Spoken query:

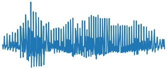




Example application: query-by-example search

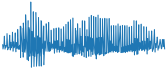


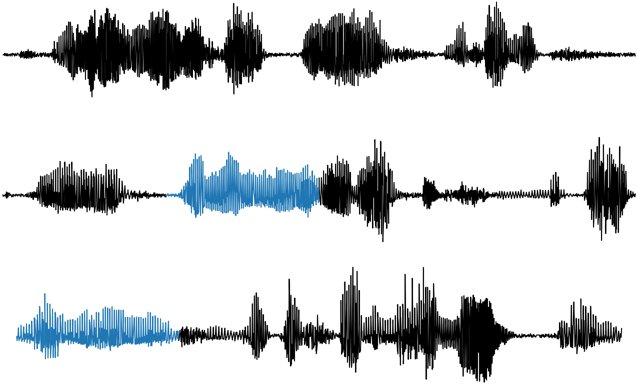
Spoken query:



Example application: query-by-example search

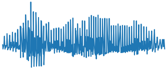
Spoken query:

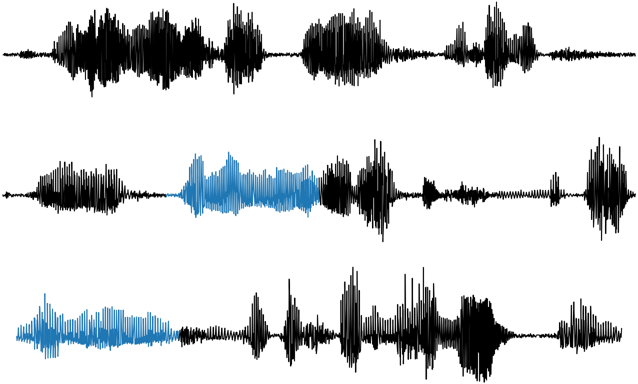




Example application: query-by-example search

Spoken query:

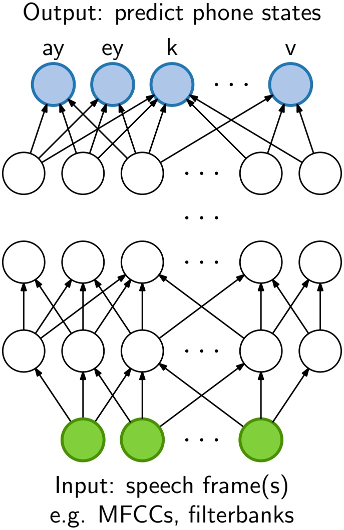




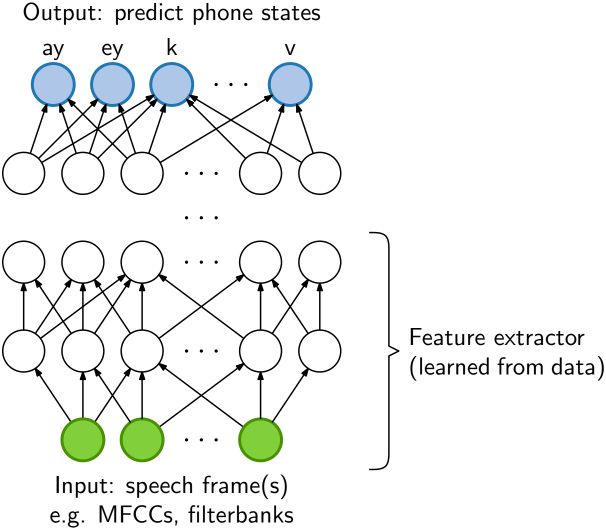
What features should we use to represent the speech for such unsupervised tasks?

Supervised neural network feature extraction

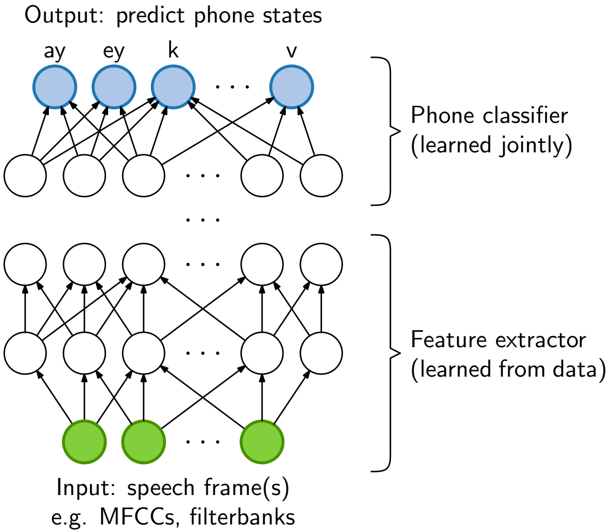
Supervised neural network feature extraction



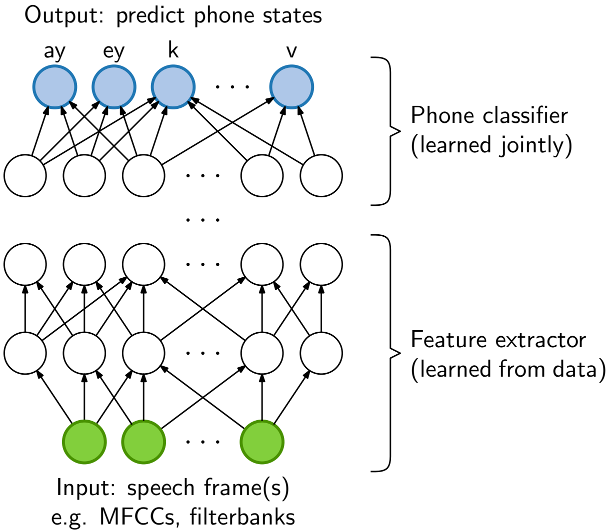
Supervised neural network feature extraction



Supervised neural network feature extraction



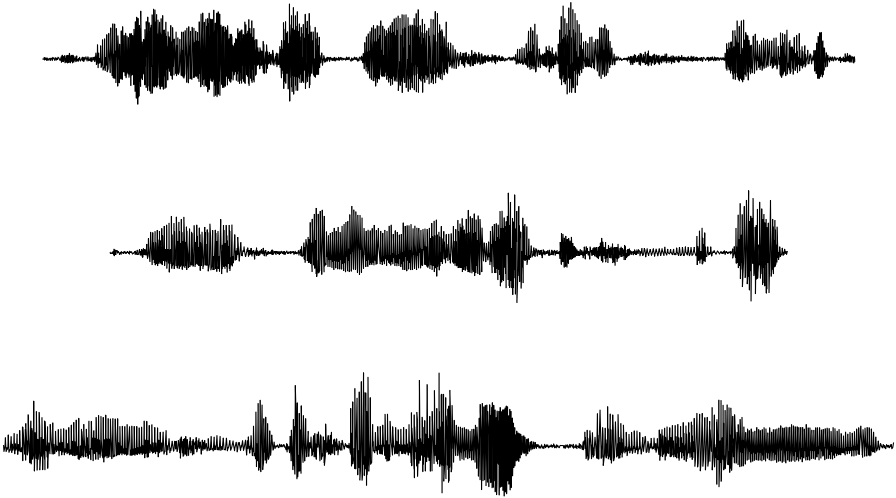
Supervised neural network feature extraction



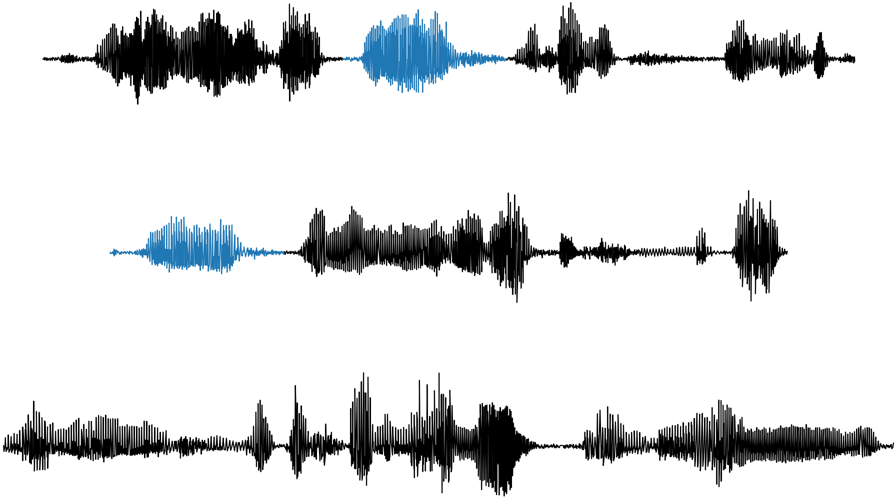
But what if we do not have phone class targets to train our network?

Weak supervision: unsupervised term discovery

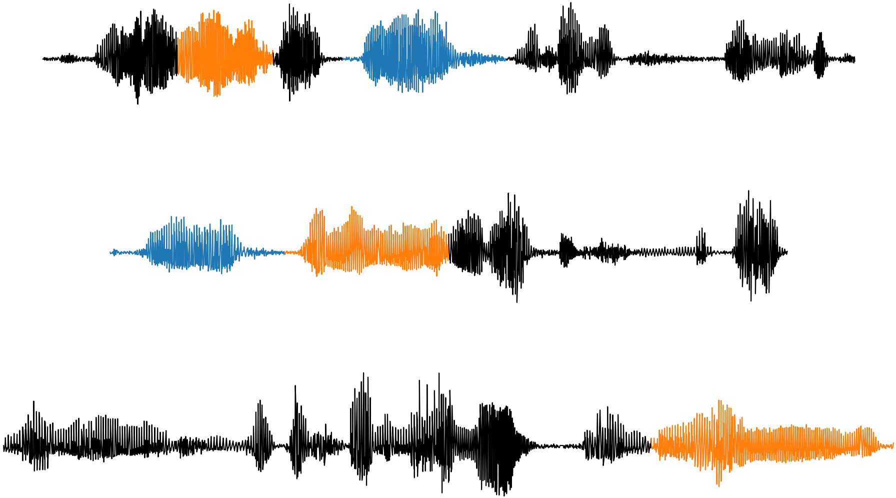
Weak supervision: unsupervised term discovery



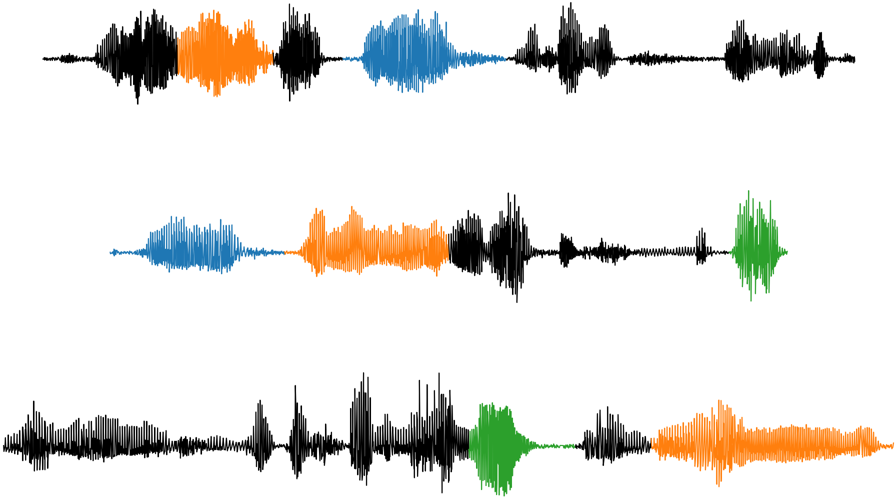
Weak supervision: unsupervised term discovery



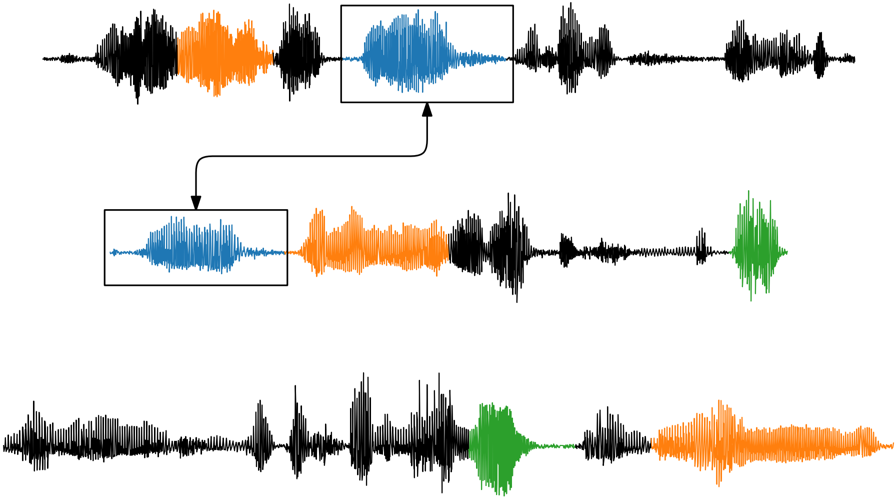
Weak supervision: unsupervised term discovery



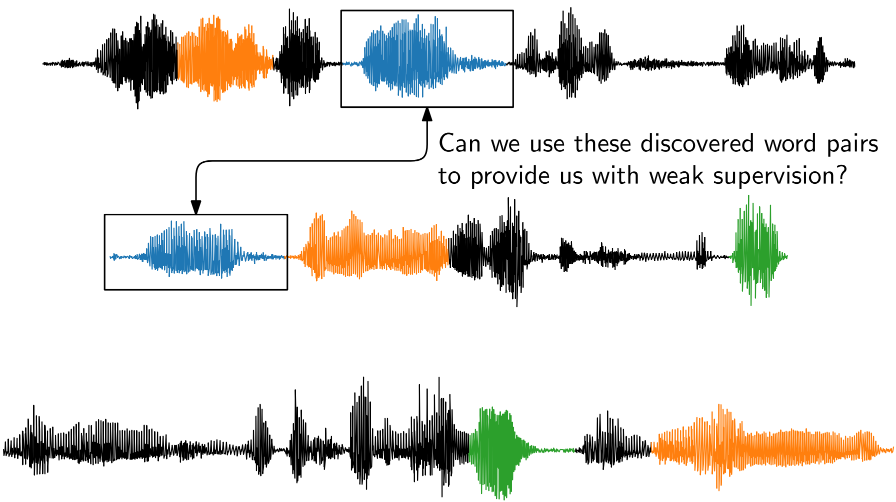
Weak supervision: unsupervised term discovery



Weak supervision: unsupervised term discovery



Weak supervision: unsupervised term discovery

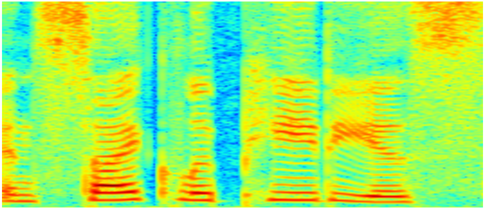


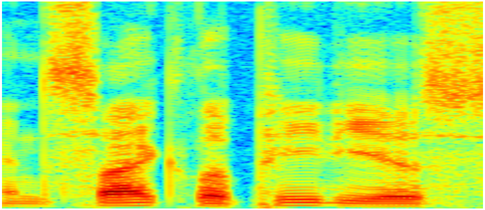
Weak supervision: align the discovered word pairs

Use correspondence idea from [Jansen et al., 2013]

Weak supervision: align the discovered word pairs

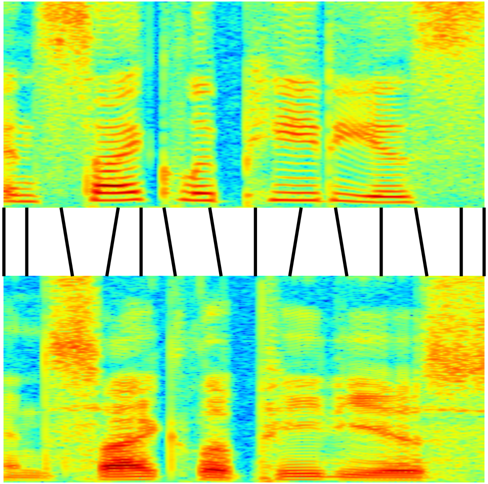
Use correspondence idea from [Jansen et al., 2013]:





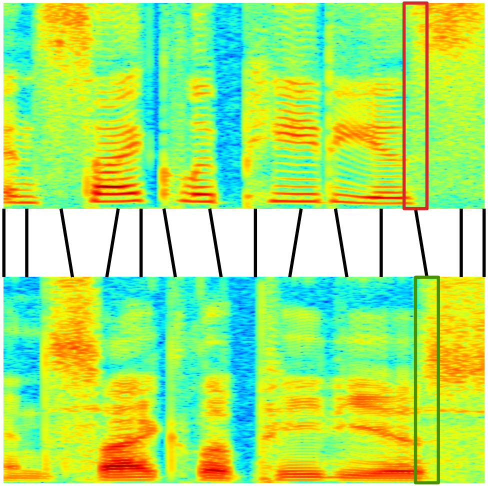
Weak supervision: align the discovered word pairs

Use correspondence idea from [Jansen et al., 2013]:



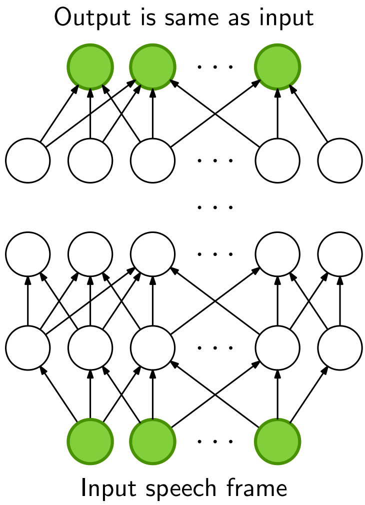
Weak supervision: align the discovered word pairs

Use correspondence idea from [Jansen et al., 2013]:



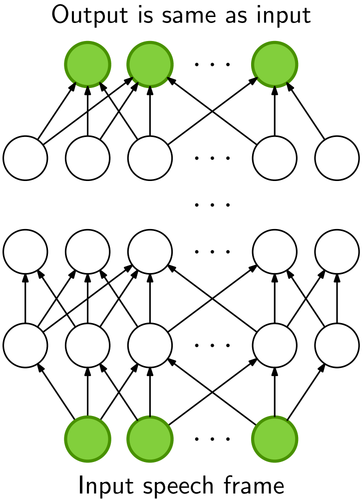
Autoencoder (AE) neural network

Autoencoder (AE) neural network



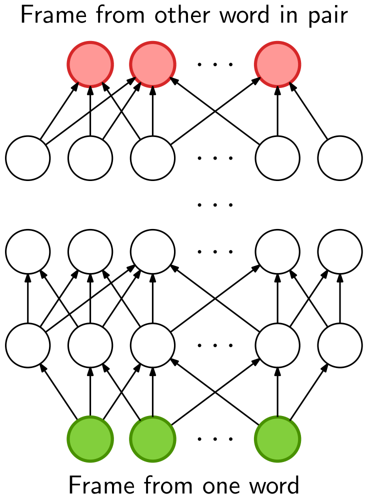
A normal autoencoder neural network is trained to reconstruct its input.

Autoencoder (AE) neural network



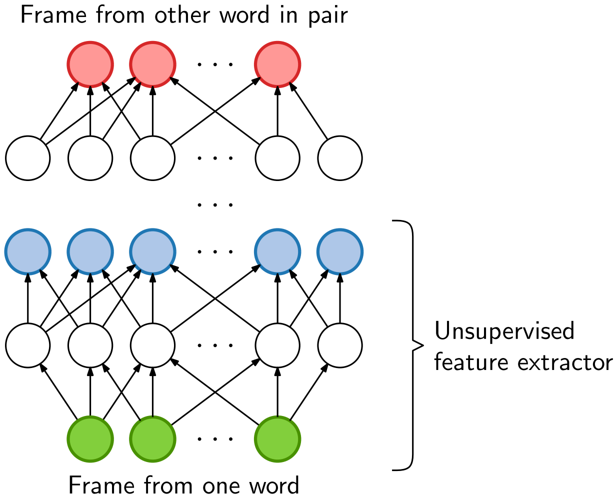
This reconstruction criterion can be used to pretrain a deep neural network.

The correspondence autoencoder (cAE)



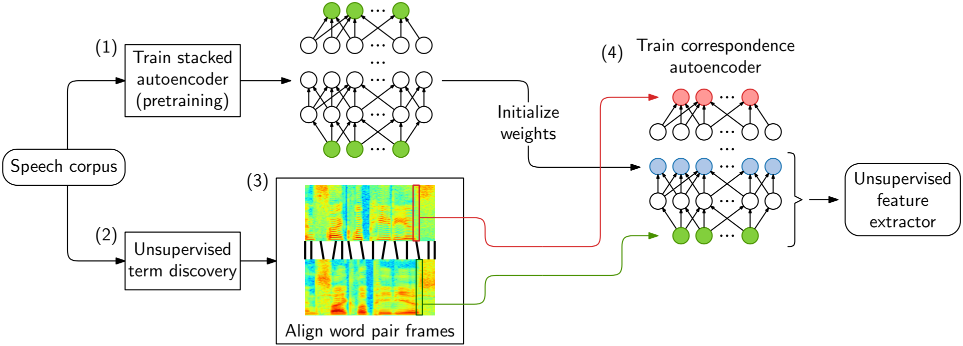
The correspondence autoencoder (cAE) takes a frame from one word, and tries to reconstruct the corresponding frame from the other word in the pair.

The correspondence autoencoder (cAE)



In this way we learn an unsupervised feature extractor using the weak word-pair supervision.

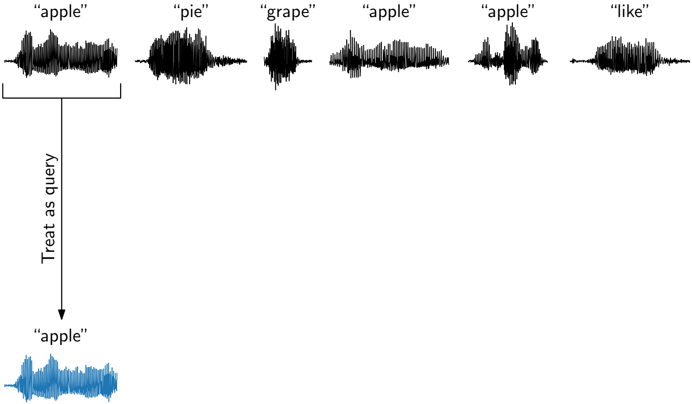
Complete unsupervised cAE training algorithm



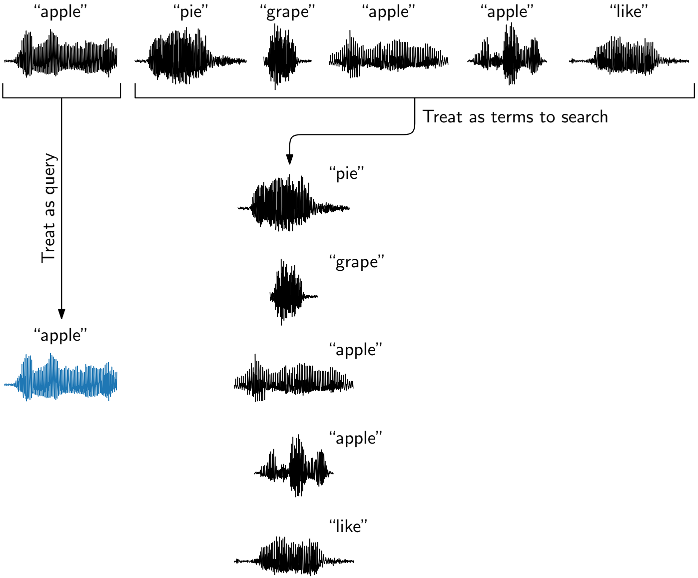
Evaluation of features: the same-different task

Evaluation of features: the same-different task

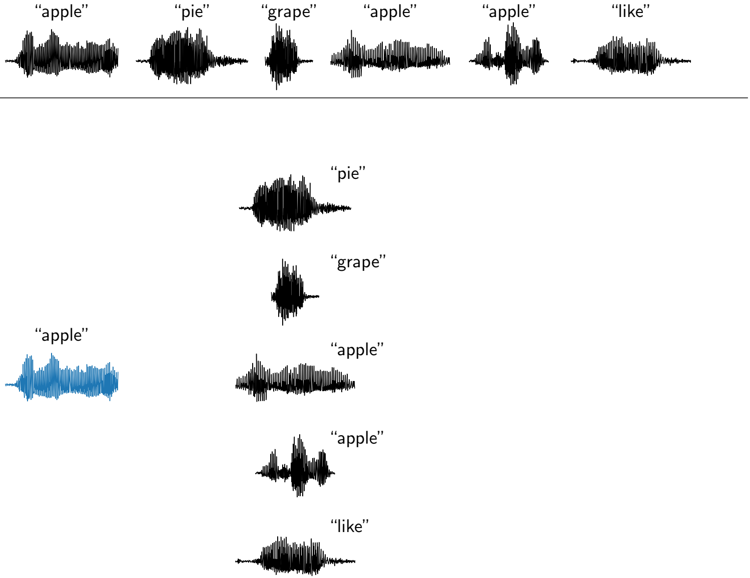
Evaluation of features: the same-different task



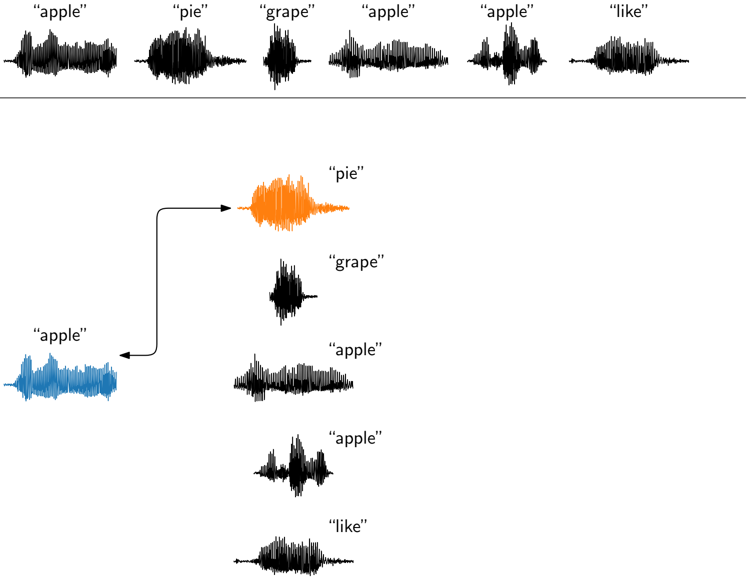
Evaluation of features: the same-different task



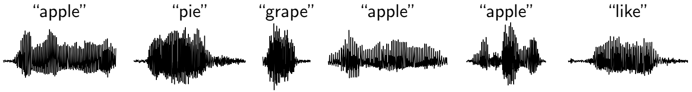
Evaluation of features: the same-different task

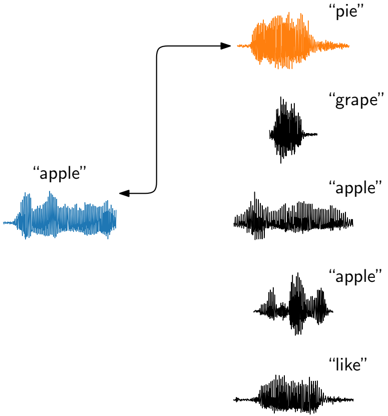


Evaluation of features: the same-different task



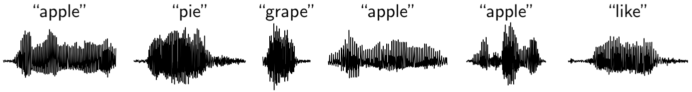
Evaluation of features: the same-different task

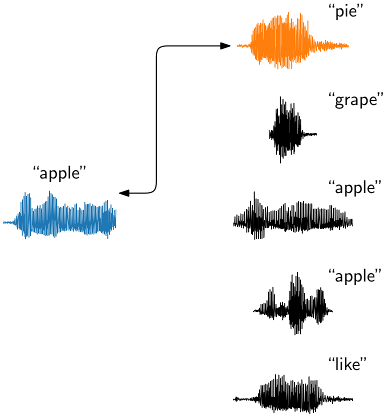




DTW distance:	
d_1	

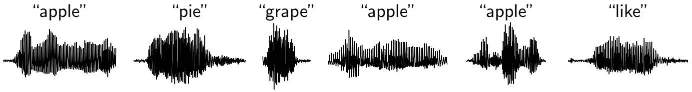
Evaluation of features: the same-different task

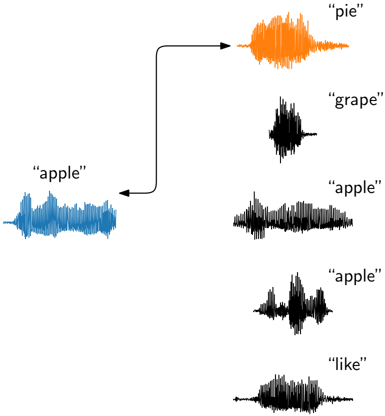




DTW distance:	$d_i <$ threshold? predict:
d_1	different

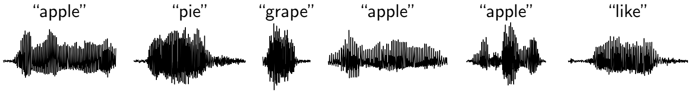
Evaluation of features: the same-different task

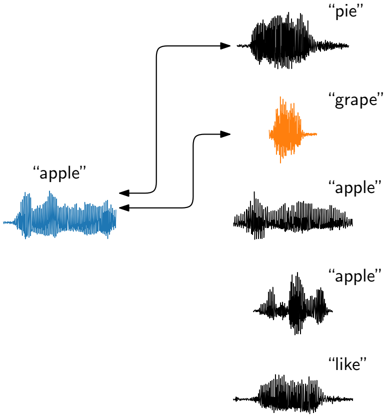




DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓

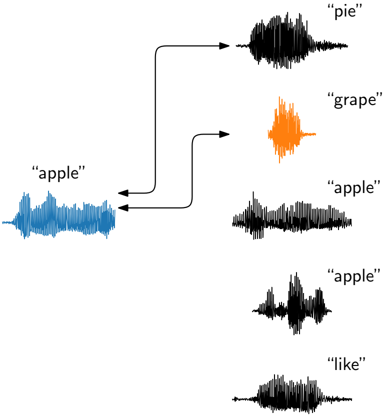
Evaluation of features: the same-different task





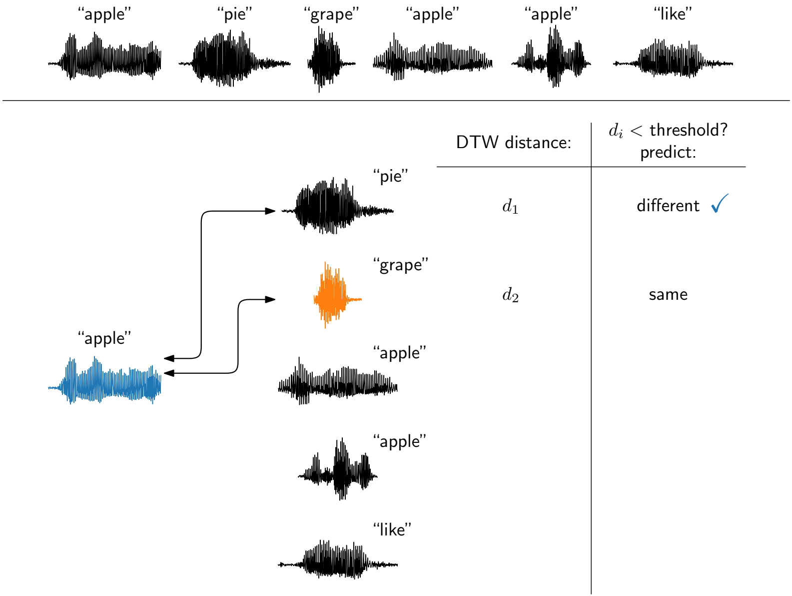
DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓

Evaluation of features: the same-different task

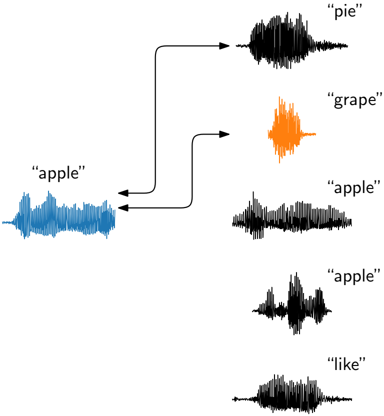


DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	

Evaluation of features: the same-different task

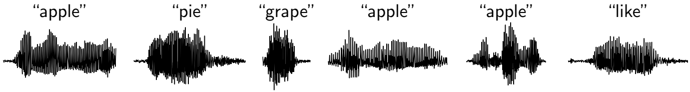


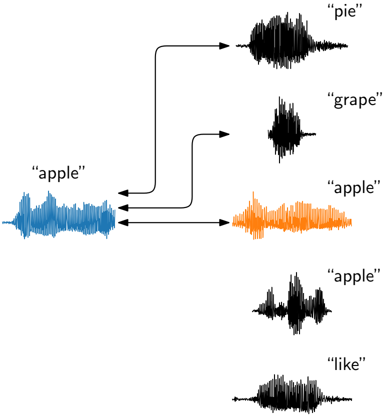
Evaluation of features: the same-different task



DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ×

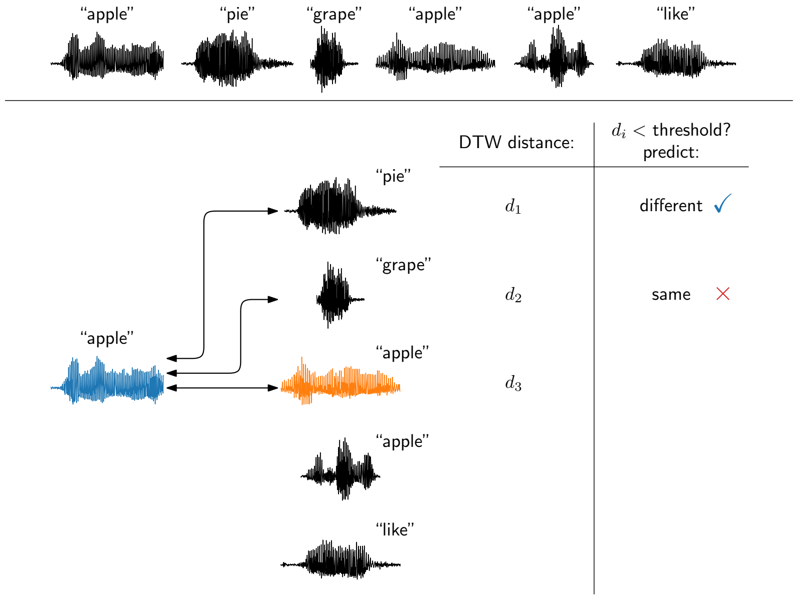
Evaluation of features: the same-different task



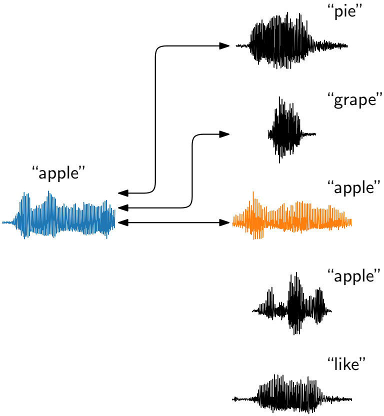


DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ✗

Evaluation of features: the same-different task

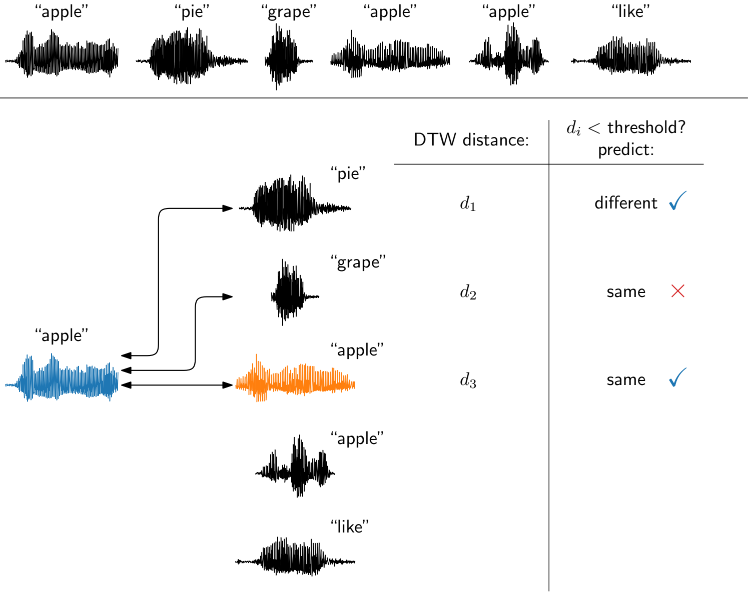


Evaluation of features: the same-different task

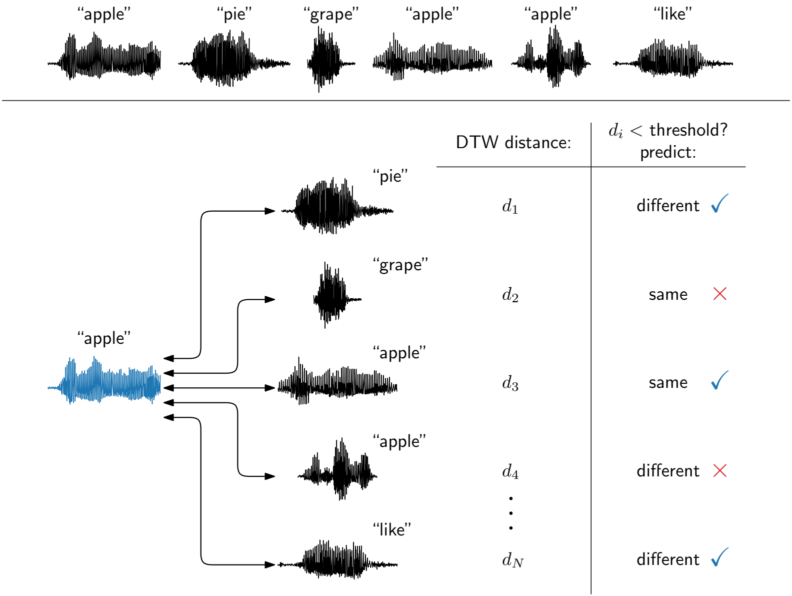


DTW distance:	$d_i <$ threshold? predict:
d_1	different ✓
d_2	same ×
d_3	same

Evaluation of features: the same-different task



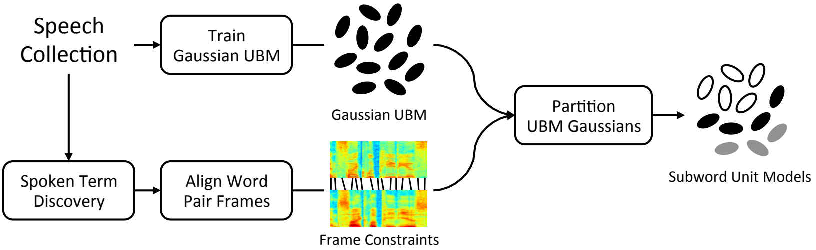
Evaluation of features: the same-different task



Evaluation of features: the same-different task

- Each term is treated in turn as the query.
- The threshold is varied to obtain a precision-recall curve.
- The area under the precision-recall curve is used as the final evaluation metric, referred to as average precision (AP).
- AP is higher for feature representations which are better able to associate words of the same type and discriminate between words of different types.
- AP has been shown to correlate well with phone recognition error rates [Carlin et al., 2011] and has been used in several other unsupervised studies.

Baseline: partitioned universal background model



Use posteriorgram features from the partitioned universal background model (UBM) as baseline [Jansen et al., 2013].

Evaluation

- Speech from Switchboard is used for evaluation.
- Pretraining data: 23 hours of untranscribed speech.
- We consider two sets of word pairs for training the cAE:
  1. 100k gold standard word pairs.
  2. 80k word pairs discovered using unsupervised term discovery (UTD).
- Test set for same-different evaluation: 11k word tokens, 60.7M pairs, 3% produced by same speaker.

Evaluation

- Speech from Switchboard is used for evaluation.
- Pretraining data: 23 hours of untranscribed speech.
- We consider two sets of word pairs for training the cAE:
  1. 100k gold standard word pairs.
  2. 80k word pairs discovered using unsupervised term discovery (UTD).
- Test set for same-different evaluation: 11k word tokens, 60.7M pairs, 3% produced by same speaker.
- Neural network architecture (optimized on development set):
  39-dimensional single-frame MFCC input features, 13 layers, 100 hidden units per layer, take features from the fourth-last encoding layer.

Comparison with baseline: gold standard word pairs

Features	Average precision
MFCCs with CMVN	0.214
UBM with 1024 components [Jansen et al., 2013]	0.222
1024-UBM partitioned 100 components [Jansen et al., 2013]	0.286
100-unit, 13-layer stacked autoencoder	0.215
100-unit, 13-layer correspondence autoencoder	**0.469**
Supervised NN, 10 hours [Carlin et al., 2011]	0.439
Supervised NN, 100 hours [Carlin et al., 2011]	0.516

Evaluation using terms from unsupervised term discovery

Features	Average precision
MFCCs with CMVN	0.214
Best of [Jansen et al., 2013] using gold standard word pairs	0.286
Correspondence autoencoder trained on gold standard word pairs	0.469
Correspondence autoencoder trained on UTD pairs	**0.341**
Supervised NN, 10 hours [Carlin et al., 2011]	0.439
Supervised NN, 100 hours [Carlin et al., 2011]	0.516

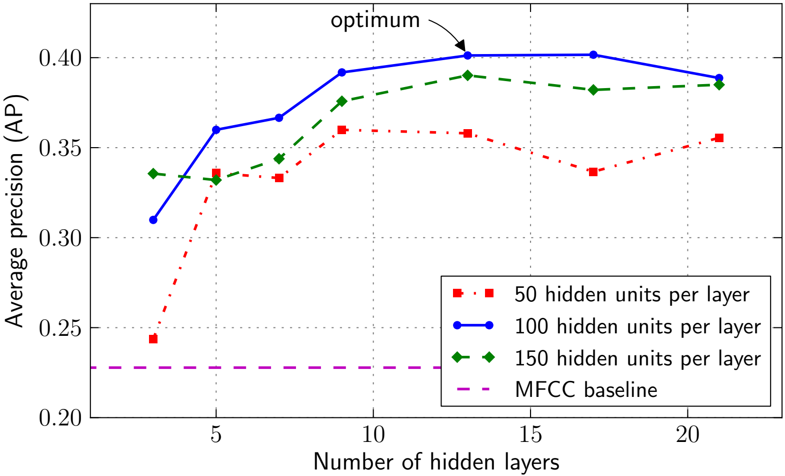
Summary and conclusion

- Introduced the correspondence autoencoder (cAE), a novel neural network which can be trained unsupervised on unlabelled speech data.
- Evaluated the network in a word discrimination task.
- Showed 64% relative improvement over a previous state-of-the-art GMM system.
- Come to within 23% of supervised baseline.
- Future work: apply in further unsupervised speech processing tasks; how can the correspondence idea be used in other neural network structures?

Code

https://github.com/kamperh/speech_correspondence/

Choosing the network architecture



Development set cAE performance using gold standard word pairs. Features were taken from the fourth-last to second-last encoding layers.