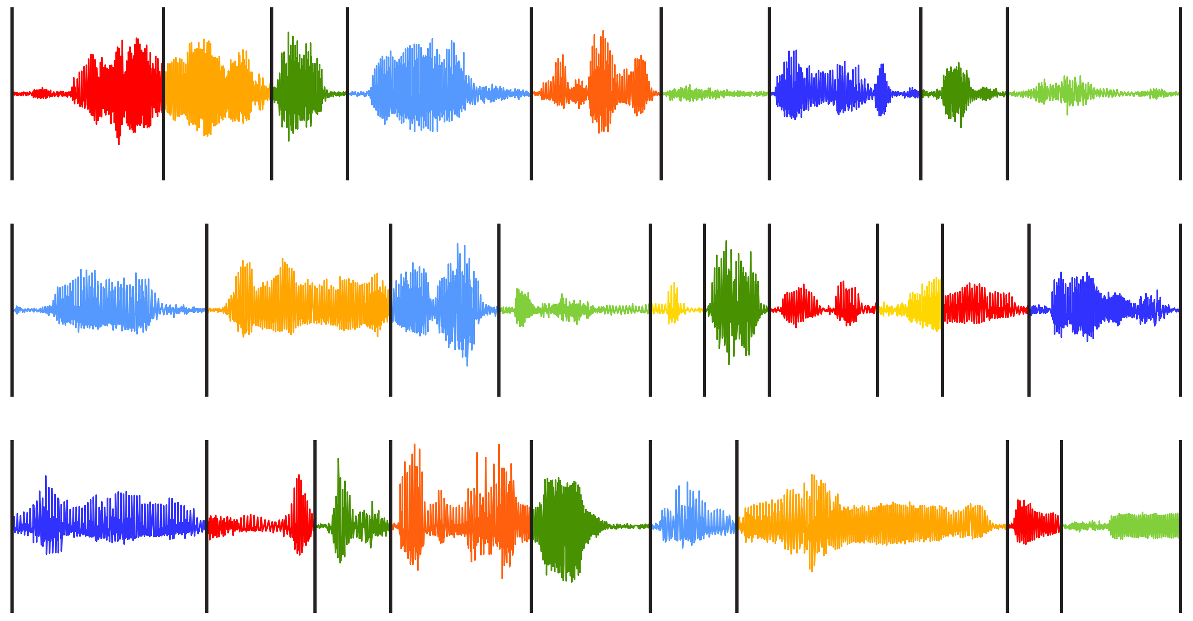
Unsupervised Lexical Clustering of Speech Segments using Fixed-Dimensional Acoustic Embeddings

Herman Kamper[1], Aren Jansen[2], Simon King[1], Sharon Goldwater[1]

[1]School of Informatics, University of Edinburgh, UK & [2]HLTCOE/CLSP, Johns Hopkins University, USA

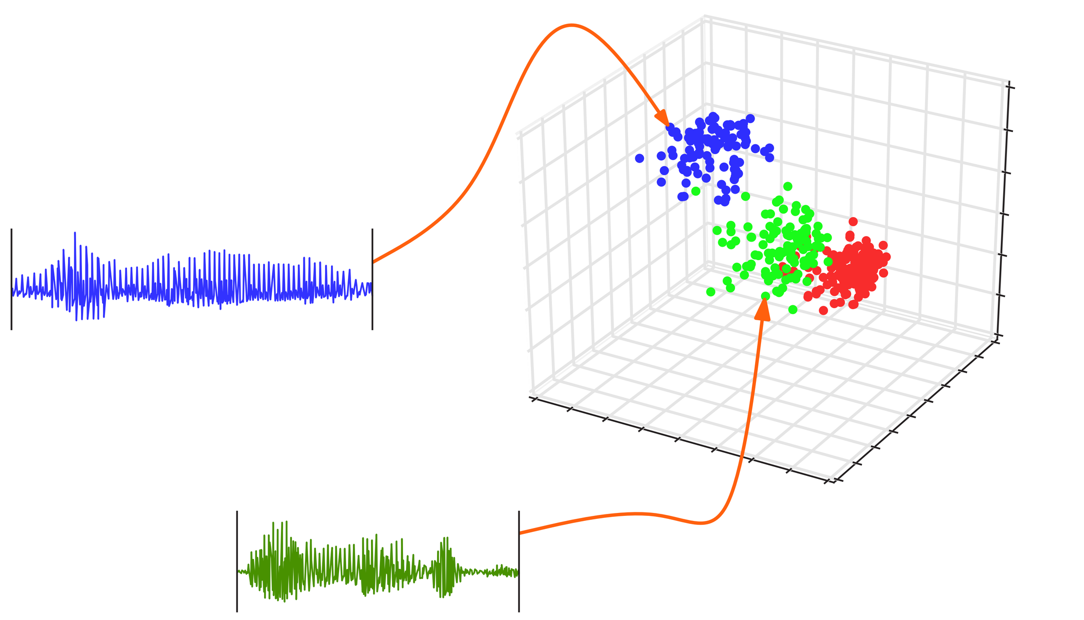
Big picture

- Interested in unsupervised learning of structure directly from raw speech.
- Envisioned architecture will:
  1. Hypothesize complete lexical segmentation of input speech.
  2. Learn word categories of segments and relate these to underlying acoustics.
  3. Estimate language model over the discovered word categories.



This work: main question

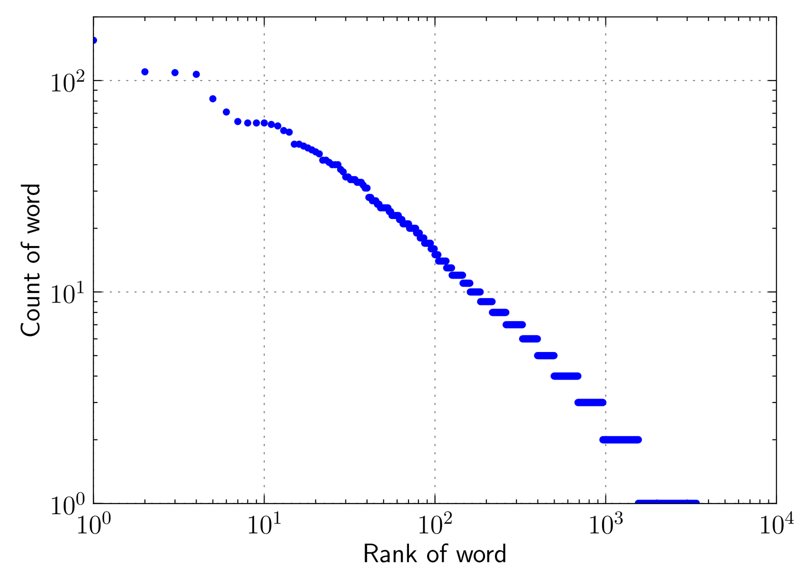
- Here we focus only on task (2) above: learning lexical categories.



- Levin et al. (ASRU '13) showed that embedding variable-length speech segments in a fixed-dimensional space is a viable alternative to dynamic time warping.
- We use these embeddings of word tokens as input.

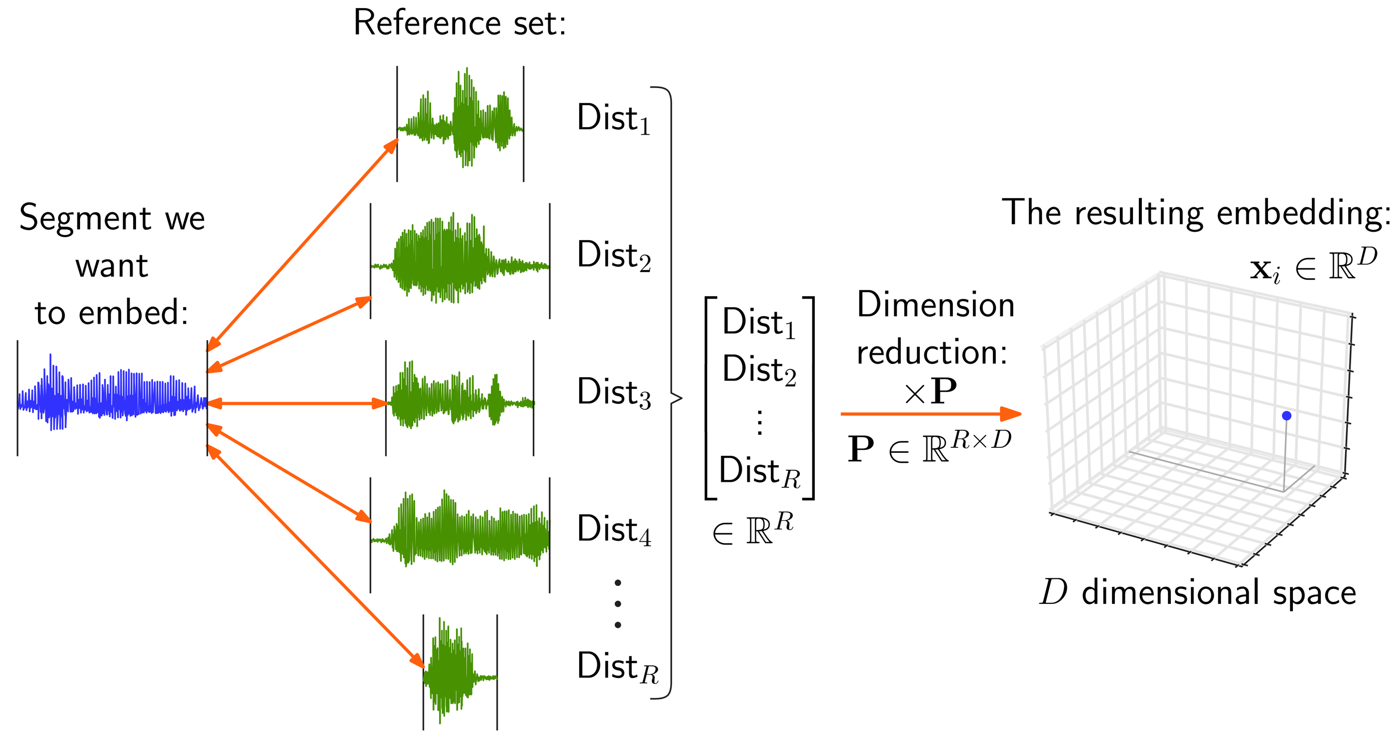
Main question: Can we cluster these acoustic embeddings of speech segments, and which clustering method works best?

Dataset



- We use the same dataset as that used by Levin et al.
- Content words extracted from Switchboard corpus.
- 3392 word types, 11 024 word tokens, std. deviation of no. of tokens per type: 7.05.

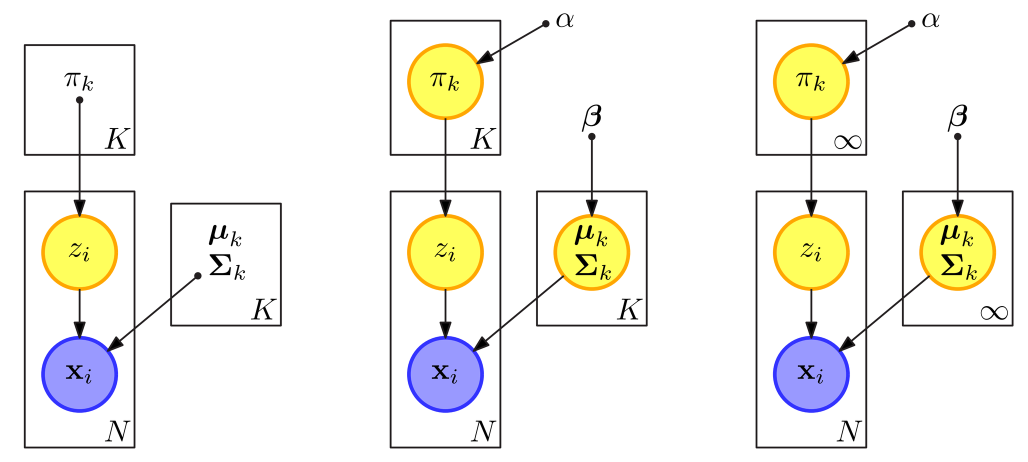
How acoustic embeddings are calculated



Considered two scenarios for the reference set:

1. UnsupTrain: Only have word exemplars $\mathcal{Y}_{\text{train}} = \{Y_i\}_{i=1}^{N_{\text{train}}}$.
2. SupTrain: Also know the word identities $\mathcal{W}_{\text{train}} = \{w_i\}_{i=1}^{N_{\text{train}}}$ of exemplars.

Clustering approaches



(1) EMGMM (2) FBGMM (3) IGMM

Probabilistic approaches:

1. EMGMM: GMM trained using expectation maximisation.
2. FBGMM: Finite Bayesian GMM using Gibbs sampling.
3. IGMM: Infinite GMM using Gibbs sampling for inference.

Non-probabilistic approaches:

1. K-means clustering
2. Hierarchical clustering: Greedy agglomerative clustering using average linkage.
3. Chinese whispers algorithm: Randomized graph clustering.

Quantitative evaluation measures

- Cluster purity
- One-to-one mapping accuracy: A greedy mapping from clusters to true classes.
- Adjusted rand index (ARI): Considers all pairs of tokens and compare the true labelling and the predicted labelling for these pairs.
- Standard deviation of cluster sizes: Desire large variance across cluster sizes, as is the case for natural language (power-law).

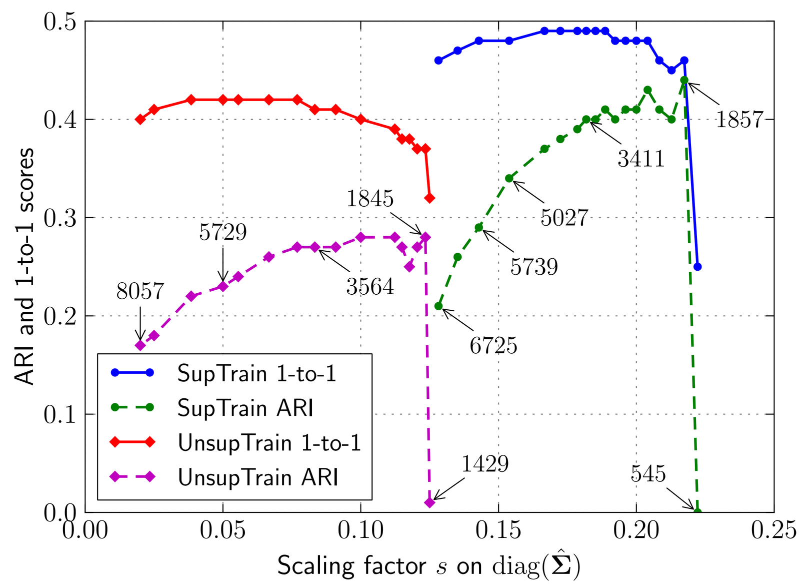
Experimental results

	Results on SupTrain:					Results on UnsupTrain:				
Algorithm	Purity	ARI	1-to-1	Std. size	K	Purity	ARI	1-to-1	Std. size	K
DTW hier.	0.66	0.36	0.48	4.61	3392	0.66	0.36	0.48	4.61	3392
EMGMM	0.67	0.17	0.42	2.43	3392	0.59	0.19	0.38	3.37	3392
FBGMM	0.67	0.34	0.47	3.89	3199	0.59	0.23	0.40	4.09	3379
IGMM	0.67	0.40	0.49	5.63	3411	0.60	0.27	0.41	5.54	3564
K-means	0.66	0.17	0.41	2.49	3392	0.59	0.17	0.37	3.22	3392
Hierarchical	0.69	0.48	0.54	5.39	3392	0.59	0.32	0.44	6.87	3392
Chinese w.	0.70	0.44	0.53	8.73	3756	0.61	0.25	0.43	11.26	3941

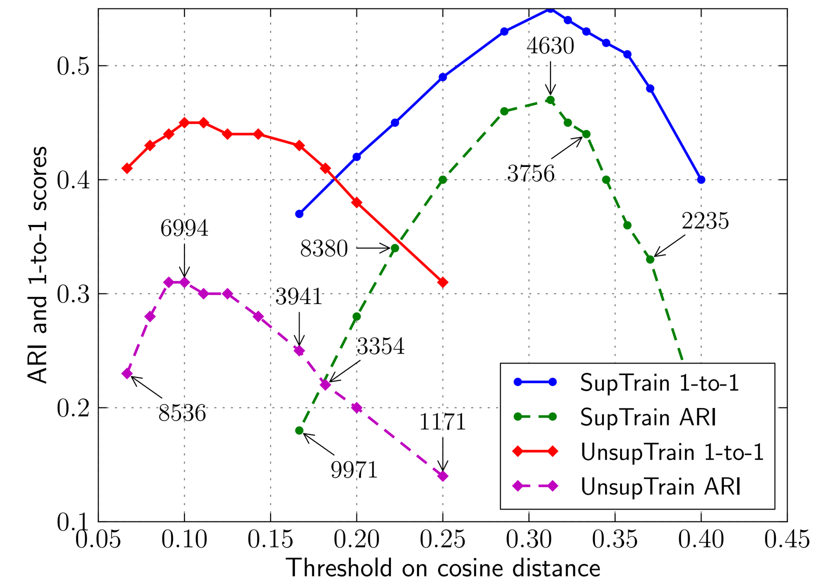
Std. size = standard deviation of cluster sizes; K = number of clusters obtained.

Adjusting hyper-parameters of IGMM and Chinese whispers

IGMM:



Chinese whispers algorithm:



Qualitative evaluation: Biggest clusters from IGMM

recycle: 62, recycled: 28, recycles: 4, recyclable: 4, recyclables: 2, recyclings: 1, recycler: 1, medical: 1, residual: 1, hypothetical: 1

people: 50, default: 1

vacation: 43, medication: 3, dedication: 1, changing: 1

program: 36, programs: 10

expenses: 28, expensive: 14, experimented: 1, experiment: 1, inexpensive: 1

probably: 41, prevalent: 1, highway: 1

society: 29, society's: 2, societies: 1, provide: 1

situation: 26, situations: 5, circulation: 1, subscriptions: 1

really: 26, rarely: 2, partly: 1

education: 7, reputation: 4, execution: 4, application: 2, executions: 1, electrocution: 1, electrocutions: 1, limitations: 1, modifications: 1, occupations: 1, obligations: 1, educational: 1, indication: 1, gratification: 1, ramifications: 1, obligation: 1, recognition: 1, restitution: 1

punishment: 28

- Left: No. of tokens for each type in biggest clusters obtained using IGMM on SupTrain.
- Righthand cluster: Overclusters several -tion(s) word types.
- Despite noise from variations in surface forms in conversational speech, qualitatively the clusters are reasonable.

Conclusions

- Best clustering methods allow for large variation in cluster sizes.
- Best probabilistic approach is infinite Gaussian mixture model (IGMM).
- Best overall approach is hierarchical clustering algorithm.
- Future: Use IGMM on fixed-dimensional embeddings for segmentation.