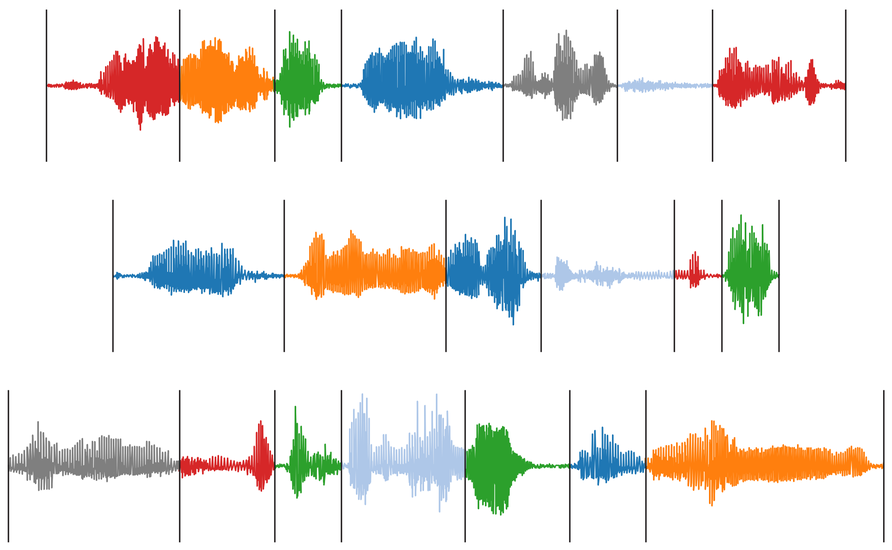
Fully Unsupervised Small-Vocabulary Speech Recognition Using a Segmental Bayesian Model

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

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

Introduction

- Current supervised speech technology is built using hundreds of hours of transcribed speech data and pronunciation dictionaries.
- For many languages, these resources are simply not available.
- We present an unsupervised Bayesian model which segments speech into word-like segments and clusters these into hypothesized word types.



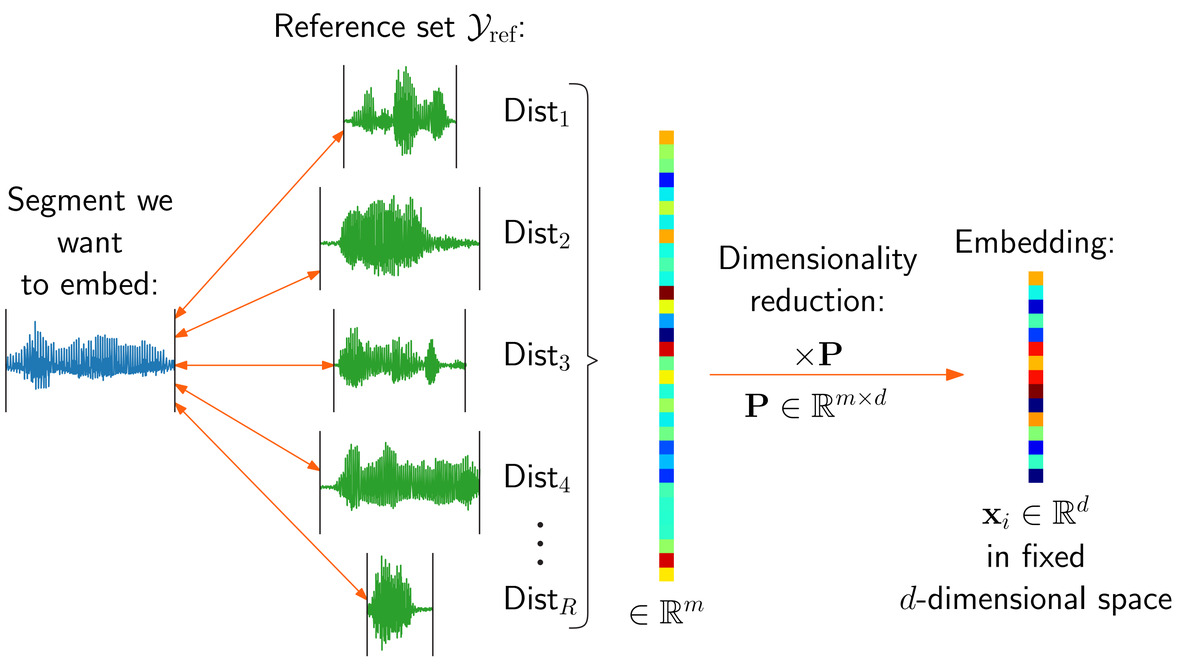
Dataset

- We evaluate our model in a connected digit recognition task.
- Use the TIDigits corpus. Development and test sets each contain: 112 speakers (male and female), 77 digit sequences per speaker.
- The corpus contains 11 word types: 'oh' and 'zero' through 'nine'.

Evaluation

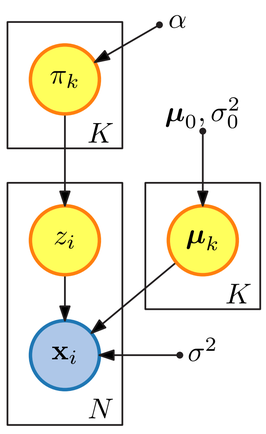
- Compare unsupervised decoding output to ground truth transcriptions: map each discovered cluster to a ground truth label.
- From this we can calculate unsupervised word error rate (WER).
- Compare to a previous study by Walter et al. [ASRU, 2013]: discrete hidden Markov models (HMMs) were trained unsupervised.

Fixed-dimensional embedding of speech segments



This relies on a reference set of unidentified exemplars $\mathcal{Y}_{\text{ref}}$.

Acoustic modelling: discovering word types

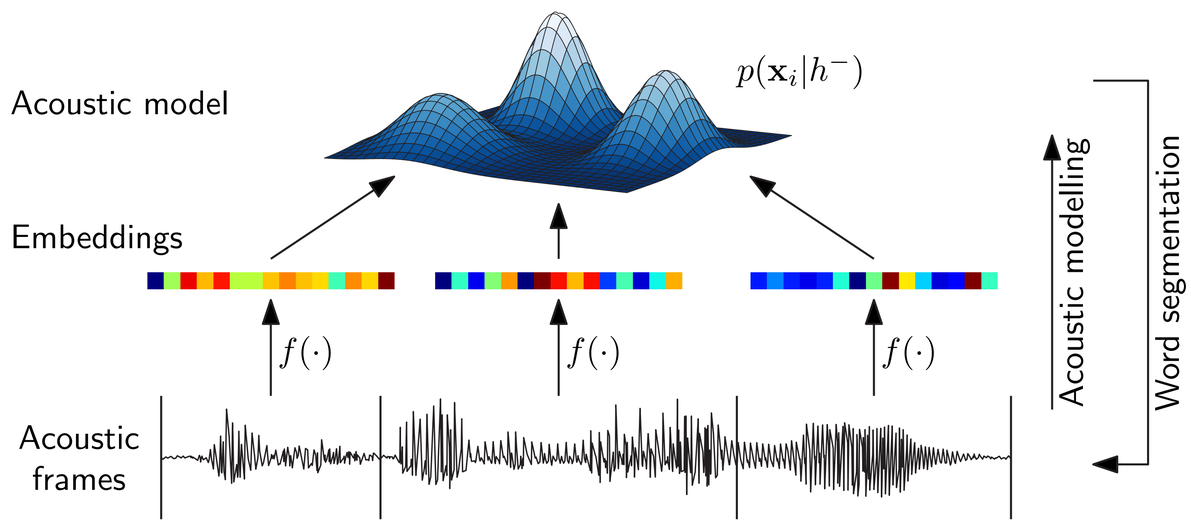


Every word type is modelled as a mixture component of a Bayesian Gaussian mixture model (GMM) with fixed spherical covariance $\sigma^2\mathbf{I}$.

Consider two settings for the number of components K:

1. Constrained: $K = 11$ is true number of word types.
2. Unconstrained: Model left to discover the number of word types up to a maximum of $K = 100$.

Word segmentation of speech



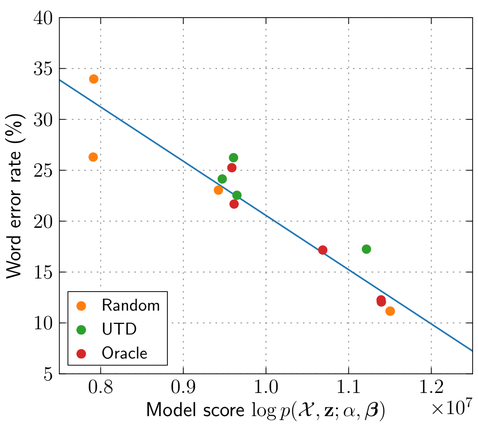
- Acoustic modelling and segmentation are performed jointly: Bayesian GMM provides likelihood terms for segmentation; segmentation hypothesizes the boundaries for the word segments which are clustered.
- Implemented as a blocked Gibbs sampler with dynamic programming.

Results

Table: Development and test set WERs (%).

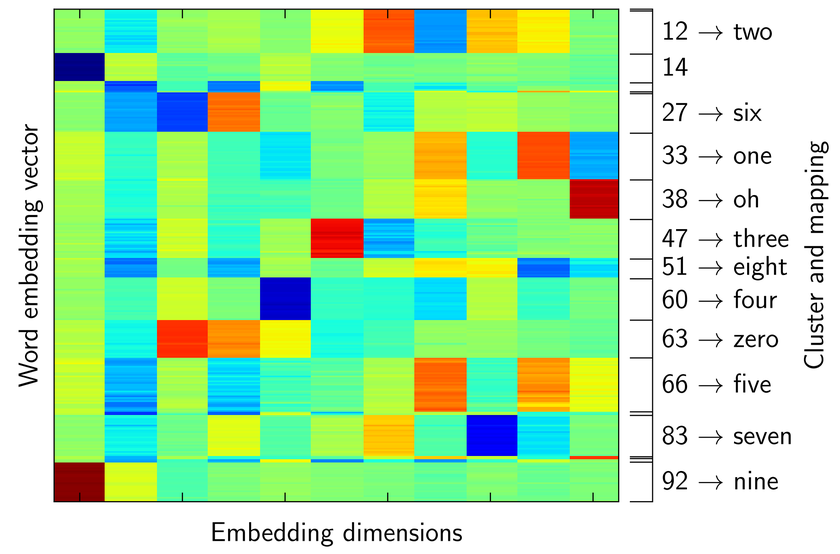
Model	Dev.	Test
Constrained* discrete HMM [Walter et al., ASRU 2013]	32.1	-
Average constrained Bayes	21.1	27.2
Highest prob. constr. Bayes	11.2	20.8
Avg. unconstrained* Bayes	20.7	32.3
Highest prob. unconstr. Bayes	20.6	32.3

constrained refers to models limited to $K = 11$ clusters; *unconstrained* allows up to $K = 100$



Embeddings in discovered clusters for single speaker

Figure: Embedding vectors ordered and stacked by discovered type along the y-axis, with embedding values coloured along the x-axis.



Mapping between clusters and ground truth digits

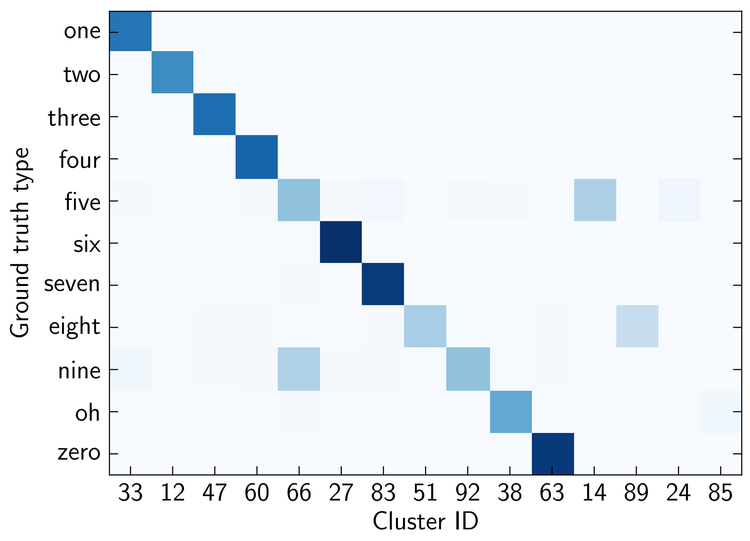


Figure: Colouring indicates the number of frames from a ground truth digit that overlaps with a particular cluster. 15 biggest clusters from an unconstrained model over all speakers are shown. The model used all $K = 100$ components, but it's 13 biggest clusters cover more than 90% of the data.

Conclusions

- Presented a novel Bayesian model for segmenting and clustering unlabelled speech into hypothesized word-sized units.
- Achieved improvements over previous study using unsupervised HMM.

Interspeech 2015, Dresden, Germany **Contact:** h.kamper@sms.ed.ac.uk **Code development:** https://github.com/kamperh/