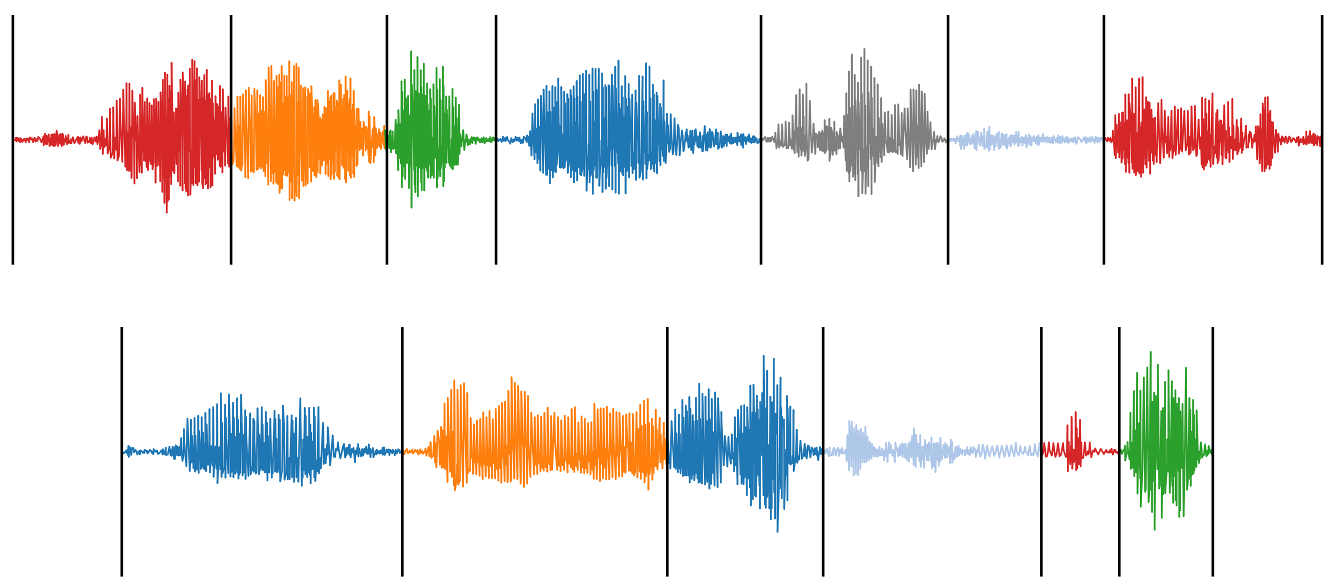
An embedded segmental K-means model for unsupervised segmentation and clustering of speech

Herman Kamper[1], Karen Livescu[2], Sharon Goldwater[3]

[1]E&E Engineering, Stellenbosch University, [2]TTI-Chicago, [3]School of Informatics, University of Edinburgh

Introduction

Zero-resource speech processing aims to develop unsupervised methods that can learn directly from raw speech, without transcriptions, lexicons or LM text. We consider full-coverage **segmentation** and **clustering**:



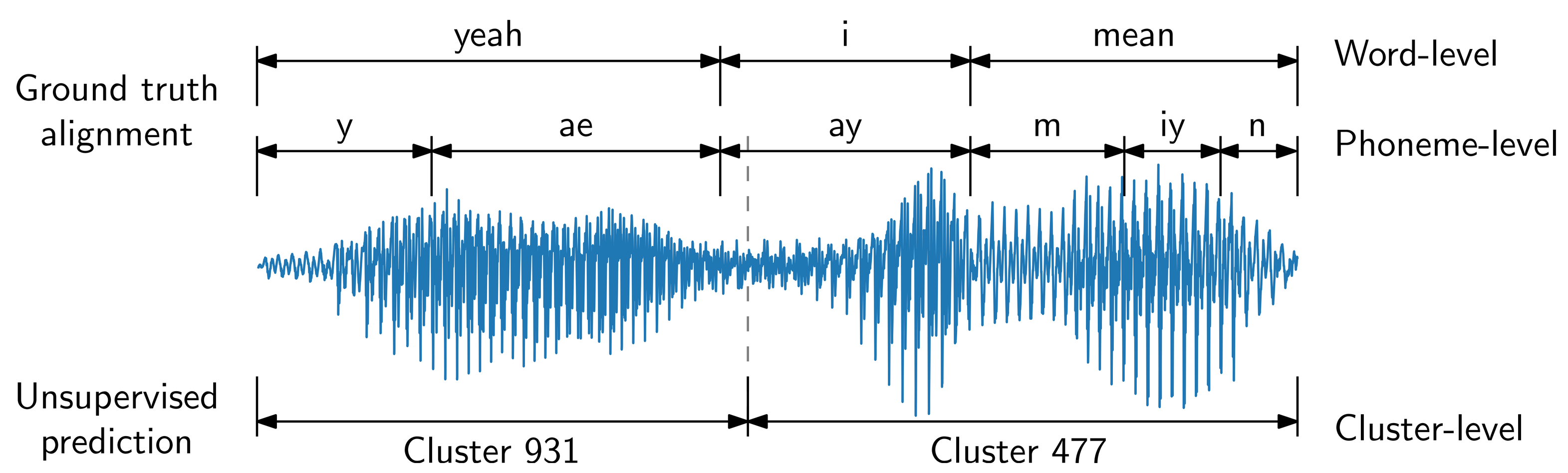
The *Bayesian embedded segmental Gaussian mixture model* (BES-GMM) was proposed before (Kamper et al., 2016). Here we develop a new model: *Embedded segmental K-means* (ES-KMeans). It also relies on fixed-dimensional segmental representations, but uses hard clustering and segmentation rather than Bayesian inference as in BES-GMM. We study the effects of these hard assignments on **performance**, **speed** and **scalability**.

Data sets

We use the Zero Resource Speech Challenge (ZRSC) 2015 and 2017 data:

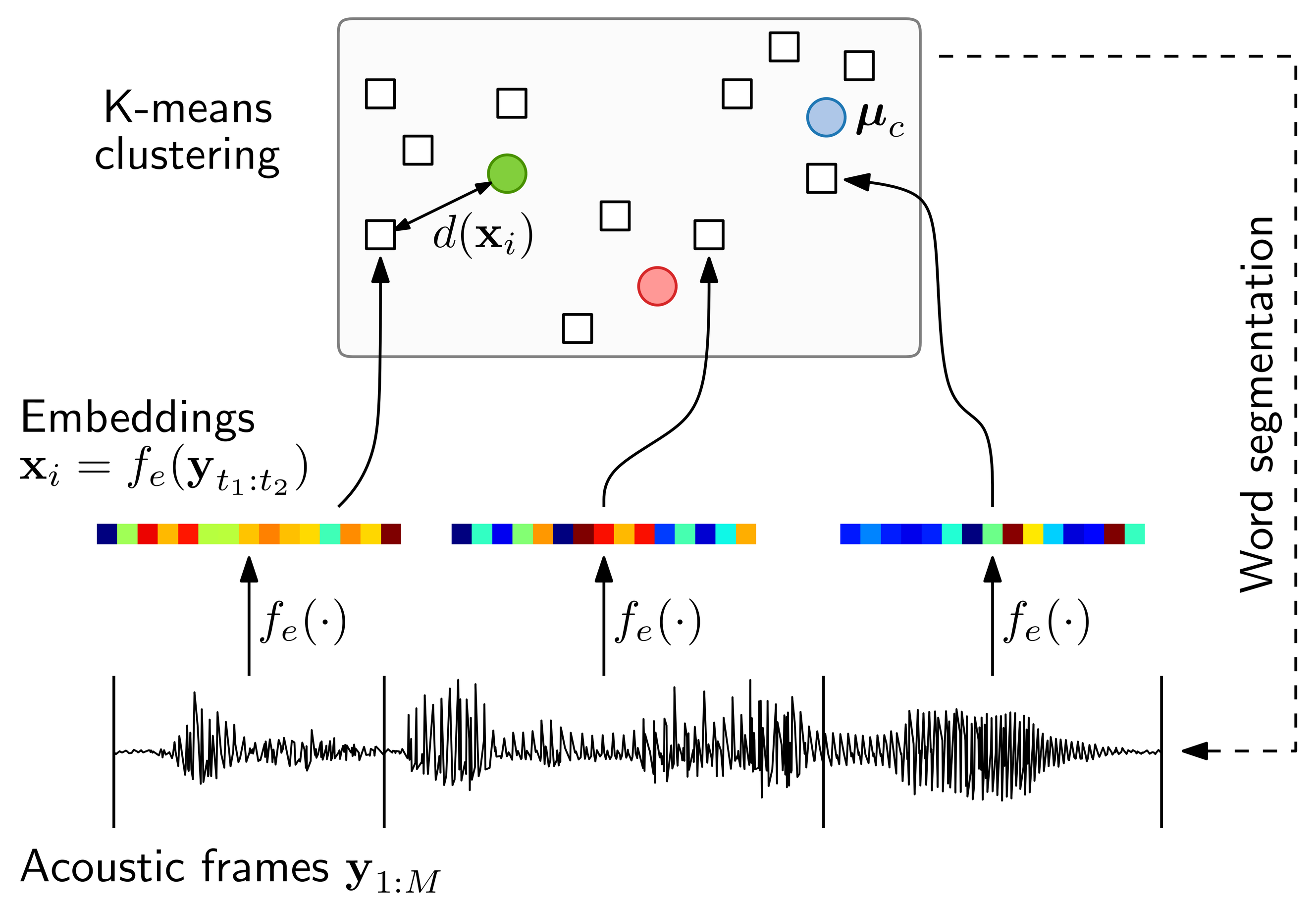
- **ZRSC'15:** English (5 h, 12 speakers), Xitsonga (2.5 h, 24 speakers)
- **ZRSC'17:** English (45 h, 69 spk), French (24 h, 28 spk), Mandarin (2.5 h, 12 spk), Surprise 1 (25 h, 30 spk), Surprise 2 (10 h, 24 spk)

Evaluation



- Boundary, token, type F-score: Compare unsupervised segmentation to ground truth alignment.
- Normalised edit distance (NED): Map discovered tokens to phoneme sequence with max overlap and calculate edit distance between pairs.

Embedded segmental K-means

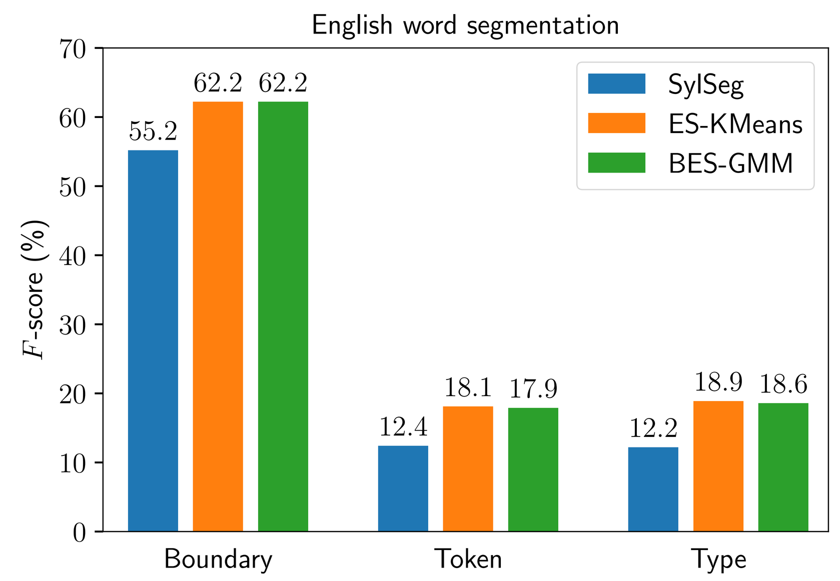


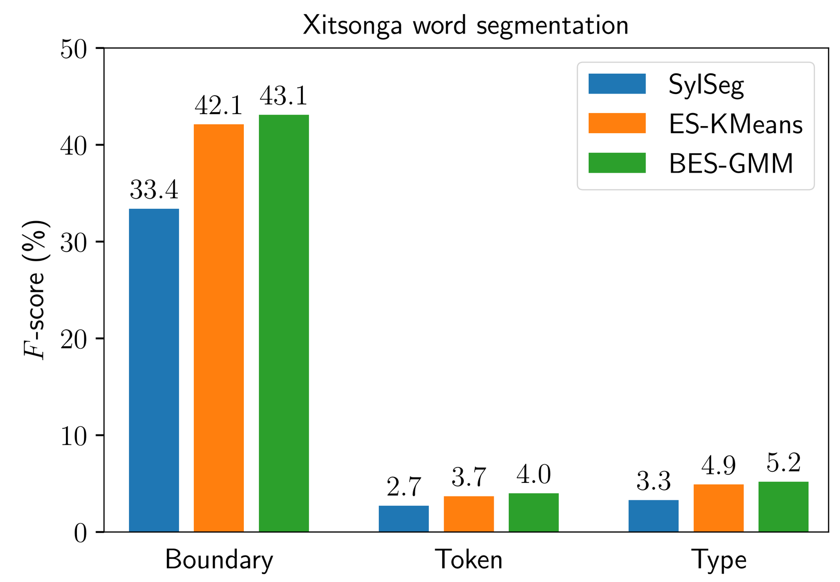
ES-KMeans optimises the cost:

$$\min_{\mathcal{Q},\mathbf{z}} \sum_{c=1}^{K} \sum_{\mathbf{x} \in \mathcal{X}_c \cap \mathcal{X}(\mathcal{Q})} \text{len}(\mathbf{x}) \, ||\mathbf{x} - \boldsymbol{\mu}_c||^2$$

$\mathcal{X}_c \cap \mathcal{X}(\mathcal{Q})$ are embeddings assigned to cluster c under segmentation $\mathcal{Q}$, and $\text{len}(\mathbf{x})$ is the no. of frames in the sequence on which $\mathbf{x}$ is calculated.

Comparison and analysis (ZRSC'15)

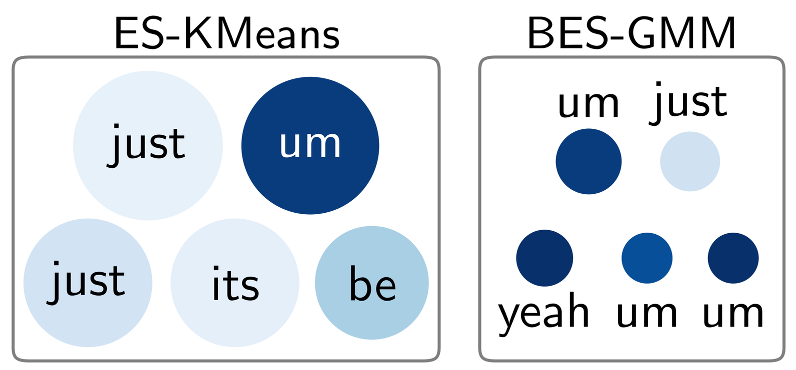




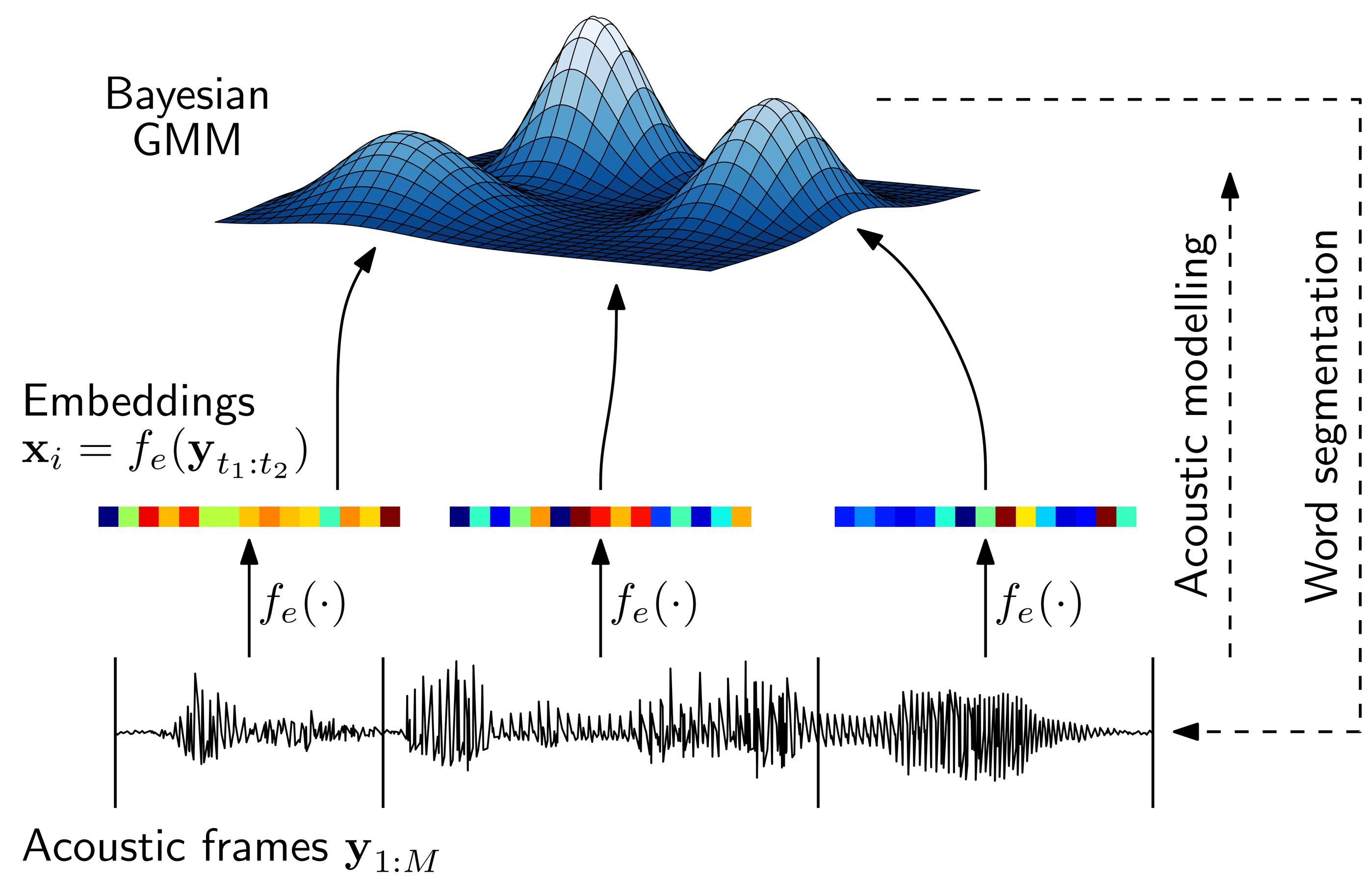
Model	Runtime (sec) English	Xitsonga
SylSeg	100	20
ES-KMeans	193	44
BES-GMM	1052	196

We use the ZRSC'15 data to compare to previous results. SylSeg is an heuristic approach developed by Räsänen et al. (2015).

Right: Five biggest clusters for ES-KMeans and BES-GMM. Circle radii indicate size, shading gives purity. The cluster sizes for BES-GMM can be controlled using hyperparameters.



Bayesian embedded segmental GMM



BES-GMM uses Gibbs sampling for joint segmentation and clustering. Downsampling is used as embedding function $f_e(\cdot)$ for ES-KMeans and BES-GMM. ES-KMeans results in the limit from BES-GMM when the variance $\rightarrow 0$.

Zero Resource Speech Challenge 2017

Language	Model	Coverage	NED	F-score (%) Boundary	Token	Type
English	JHU-PLP	7.9	**33.9**	5.7	0.5	1.2
	ES-KMeans	**100**	72.6	**52.7**	**13.5**	**11.1**
French	JHU-PLP	1.6	**25.4**	1.1	0.1	0.3
	ES-KMeans	**97.2**	67.3	**39.6**	**3.7**	**4.2**
Mandarin	JHU-PLP	2.9	**30.7**	1.8	0.1	0.2
	ES-KMeans	**100**	88.1	**41.1**	**2.9**	**3.1**
Surprise 1	JHU-PLP	3.0	**30.5**	2.3	0.2	0.6
	ES-KMeans	**100**	66.4	**48.6**	**12.0**	**7.5**
Surprise 2	JHU-PLP	5.9	**30.8**	2.0	0.1	0.2
	ES-KMeans	**100**	72.2	**43.3**	**5.0**	**6.3**

The JHU-PLP term discovery system aims for high-precision clusters (good NED), but does not cover all the data (low coverage, segmentation recall).

Conclusions

ES-KMeans performs slightly worse than BES-GMM, but is much faster. This allows it to be applied to corpora of reasonable size (ZRSC'17). In contrast to heuristic methods, ES-KMeans still has a clear objective function. Future work will focus on improving the acoustic word embedding method.