PHONEME BASED EMBEDDED SEGMENTAL K-MEANS FOR UNSUPERVISED TERM DISCOVERY

Saurabhchand Bhati*, Herman Kamper† and K. Sri Rama Murty*
ee12b1044@iith.ac.in, kamperh@gmail.com, ksrm@iith.ac.in

*Department of Electrical Engineering, IIT Hyderabad, India
†Electrical and Electronic Engineering, Stellenbosch University, South Africa

Objectives

- Zero resource speech processing refers to a scenario where no or minimal transcribed data is available
- To identify and group the frequently occurring word-like patterns from raw acoustic waveforms

Overview

- **ESK-Means:** Iteratively eliminates initial subword boundaries to arrive at longer word-like units
- **Issue:** Performance critically depends on the initial subword boundaries
- **Proposal:** Use phoneme boundaries rather than syllable boundaries
- **Advantage:** Better resolution in word search
- **Drawback:** Higher computational complexity

Embedded Segmental K-Means

$$\min_{Q,z} \sum_{c=1}^{K} \sum_{x \in \mathcal{Y}_c \cap \mathcal{Y}(Q)} \text{Length}(y) \|y - \mu_c\|^2$$

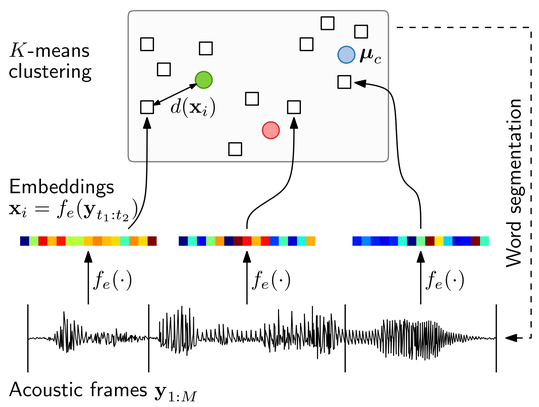

Figure 1: Illustration of ESK-means algorithm for unsupervised word detection/clustering

Phoneme Segmentation

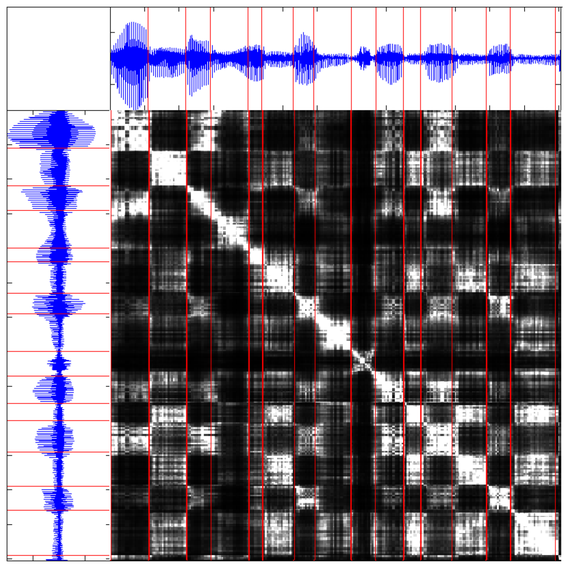

Figure 2: Kernel Gram distance matrix with Manually marked phone boundaries

Acoustic Segment Representation

- Smaller subword units lead to increased computational complexity
- Varying-length segments are uniformly divided into fixed number of frames.
- Frame averages are concatenated to arrive a high dimensional fixed-length vector
- Autoencoder is trained to reduce dimensionality of these vectors.

Table 1: Effect of bottleneck dimension
Performance (F-Score) vs Complexity trade-off

Embedding	Type	Token	Boundary	speed-up
MFCC - 390	12.1	10.6	55.5	1
Dim - 20	12.4	10.7	55.6	10.3
Dim - 15	10.3	8.6	54.9	13.3
Dim - 10	11.5	10.1	55.6	15.4

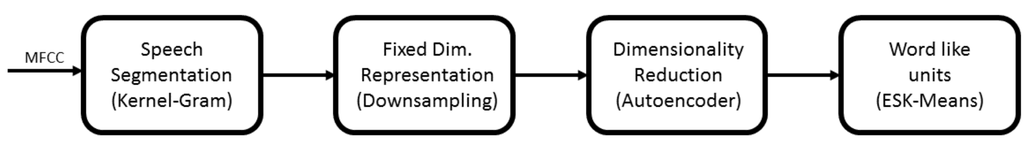

Figure 3: Overview of Phoneme ESK-Means

Choice of Initial Boundaries Syllable vs Phoneme

- Phoneme boundaries are closer to the true word boundaries
- Initial segmentation performance has a direct effect on the final token/type accuracy

Table 2: Effect of initial segmentation on the quality discovered words

Language	Boundary (F)	type (F)	Token (F)
English (phn)	34.6	6.1	6.2
syl	38.6	11.1	13.5
French (phn)	33.0	5.5	5.9
syl	24.3	4.2	3.7
Mandarin (phn)	43.9	8.8	8.7
syl	39.9	3.1	2.9

Figure 4: Comparison of phonetic, syllable and true boundaries on Mandarin dataset

Evaluation on Zerospeeech - 2017 Challenge Dataset

Table 3: Performance Comparison: Baseline, ESK-Means Syllable and ESK-Means Phoneme systems

Language	System	F-Score		
		Boundary	Type	Token
English (45 hours) (69 speakers)	Baseline	0.2	0.1	1.8
	Syllable	**11.1**	**13.5**	**52.7**
	Phoneme	6.1	6.2	32.2
French (24 hours) (28 speakers)	Baseline	0.3	0.1	1.1
	Syllable	4.2	3.7	**39.6**
	Phoneme	5.5	5.9	30.6
Mandarin (2.5 hours) (12 speakers)	Baseline	0.2	0.1	1.8
	Syllable	3.1	2.9	41.1
	Phoneme	**8.8**	**8.7**	**52.9**

- Shorter acoustic segments, like phonemes, allow finer adjustments during word discovery

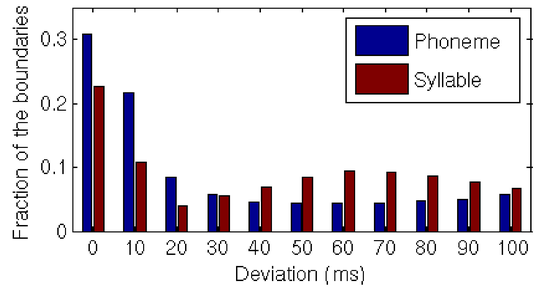

Figure 5: Histogram of deviations of detected boundaries from actual boundaries

Conclusion

- Better initial segmentation yields higher performance
- Learning a finite dimensional embedding from varying length acoustic segments
- Automatically learning the word distribution instead of a fixed minimum word length across languages

References

[1] S. Bhati, S. Nayak, and K. S. R. Murty, "Unsupervised speech signal to symbol transformation for zero resource speech applications," *Proc. Interspeech 2017*, pp. 2133–2137, 2017.

[2] H. Kamper, K. Livescu, and S. Goldwater, "An embedded segmental k-means model for unsupervised segmentation and clustering of speech," *arXiv preprint arXiv:1703.08135*, 2017.

[3] K. Levin, K. Henry, A. Jansen, and K. Livescu, "Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, IEEE, 2013.

[4] S. Bhati, S. Nayak, and K. S. R. Murty, "Unsupervised segmentation of speech signals using kernel-gram matrices," in *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, Springer, 2017.