A PARALLELIZED DYNAMIC PROGRAMMING APPROACH TO ZERO RESOURCE SPOKEN TERM DISCOVERY

Bradley Oosterveld[1], Richard Veale[2], Matthias Scheutz[1]

1. Human-Robot Interaction Laboratory, Tufts University, USA |2. Department of Neurophysiology, Kyoto University, Kyoto, Japan

The Problem

- Finding repeated patterns in acoustic speech signals without any information beyond the signals themselves
- Relevant applications:
  - Foundational work for speech recognition in languages with little to no transcribed data
  - Insight about human development and language acquisition
  - Dealing with OOV speech in open world autonomous systems

Our approach

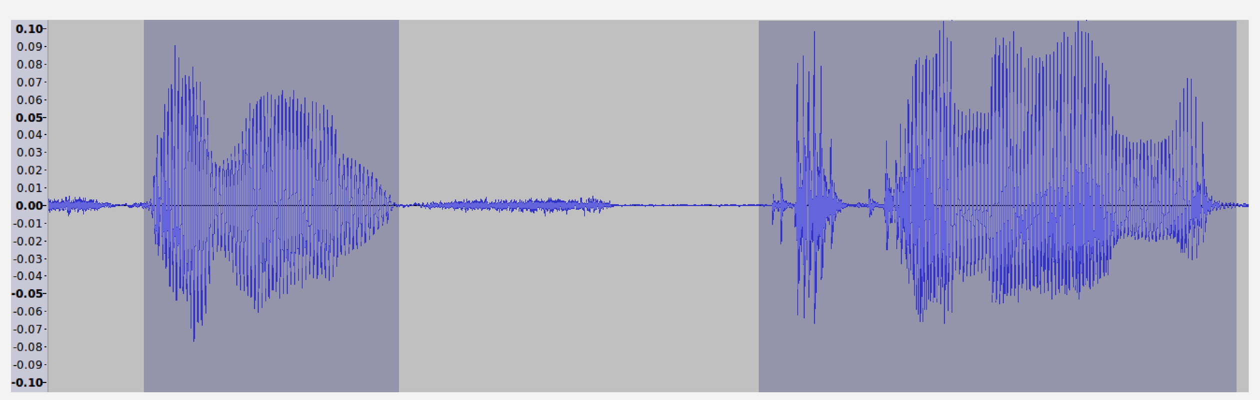
- Main contributions:
  - Adaptation of the Acoustic DP-Ngram Algorithm (DP-Ngrams) [1] to this task
  - Parallelized implantation that enables large scale evaluations:
- Sequence of segmentations, each with increasing computational complexity
- Each segmentation builds upon previous segmentations

Least Complex ⟶ Most Complex

1. Initial Segmentation
2. Feature Extraction
3. Subsequence Discovery
4. Clustering and Boundary Refinement

Initial Segmentation

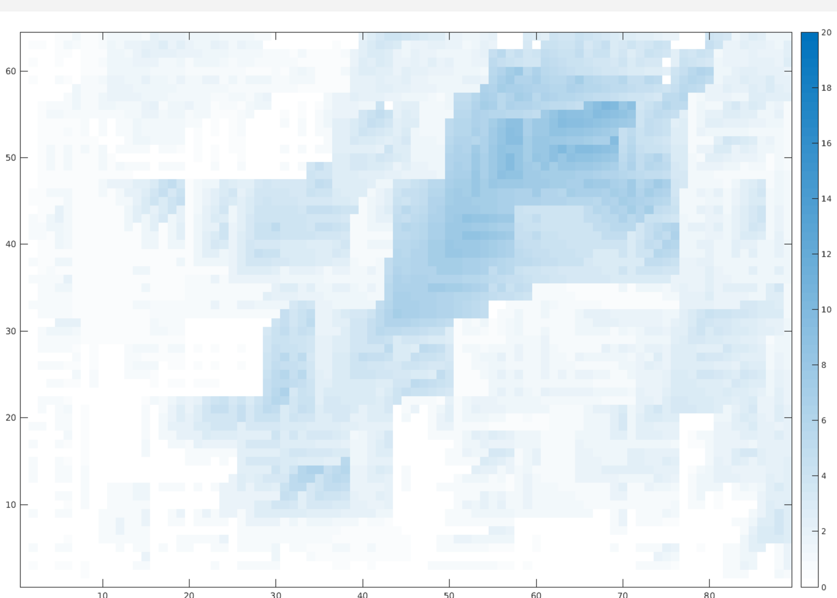
- Amplitude envelope filter
- Splits raw signal into silence-delimited chunks to enable subsequent parallelization



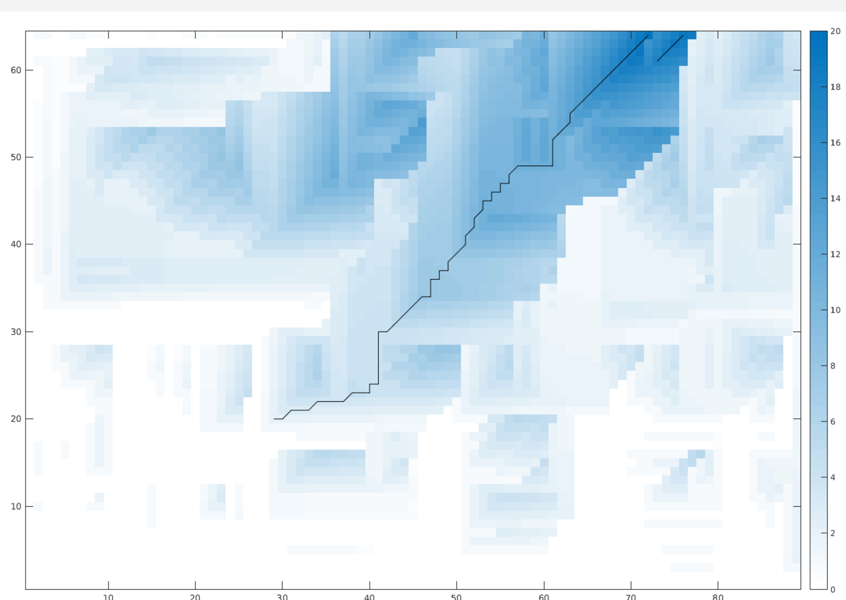
Feature Extraction

- Standard MFCCS
- Smoothed using running average filter
- Reduces the effect of minor dissimilarities in sequence pairs

Raw Features



Smoothed Features



Darker cells represent higher similarity.

Subsequence Discovery

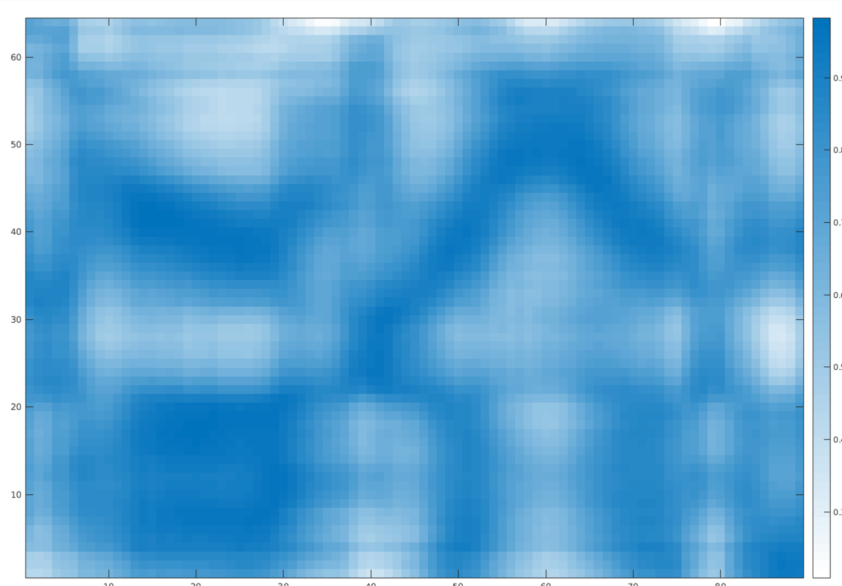
- Adapted from previous applications in sub-word level modeling
  - Modified standard parameterization
  - More aggressive elimination of previously visited cells when considering multiple alignments
- Uses dynamic programming to generate a Quality Matrix based on a Distance Matrix of sequence similarities
- Similar to S-DTW, but does not use predefined alignment start and end points

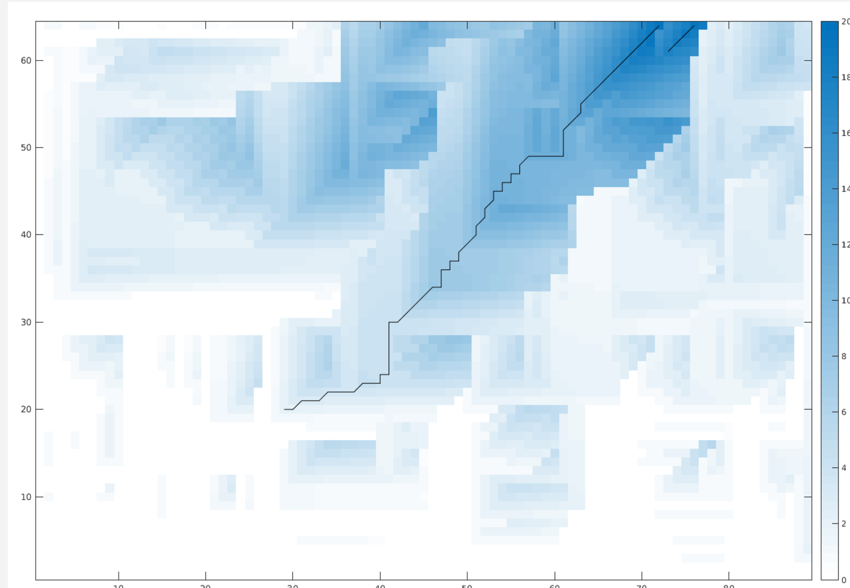
Distance Matrix

$$D_{i,j} = \frac{U_i \cdot V_j}{\| U_i \| \ \| V_j \|}$$

Quality Matrix

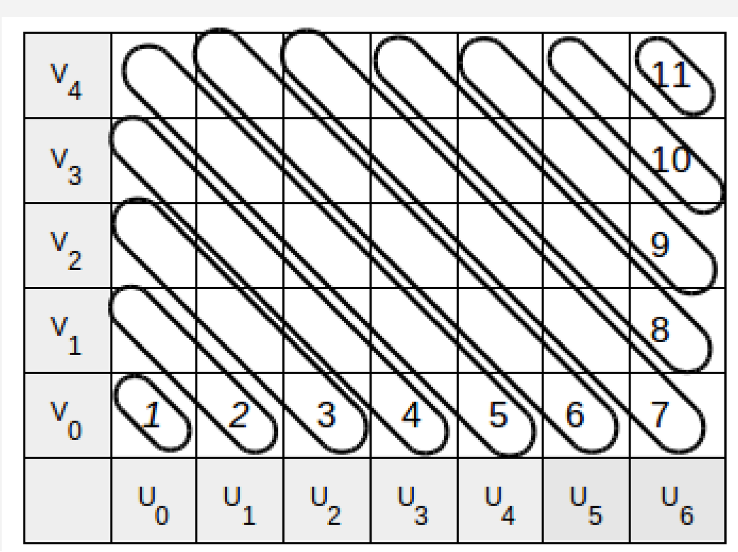
$$Q_{i,j} = max \begin{cases} Q_{i-1,j-1} + b \cdot D_{i,j} \\ Q_{i,j-1} + (p \cdot D_{i,j-1}) \cdot q_{i,j-1} \\ Q_{i-1,j} + (p \cdot D_{i-1,j}) \cdot q_{i-1,j} \\ 0 \end{cases}$$





Parallelization

- System level parallelization: comparing multiple sequences pairs at once
- Algorithm level parallelization: comparisons within DP-Ngrams done in parallel using a GPU
  - Distance matrix calculation: simple Euclidean distance kernel
  - Quality Matrix calculation: sets of cells can be updated in parallel using topology below
- Known segment lengths allow for efficient block filling, so we are able to limit each sequence comparison to a single block
- Minimize memory transfer by containing the entirety of a comparison within a single block
- With blocks performing roughly the same number of comparisons we minimize low usage situations



Clustering and Boundary Refinement

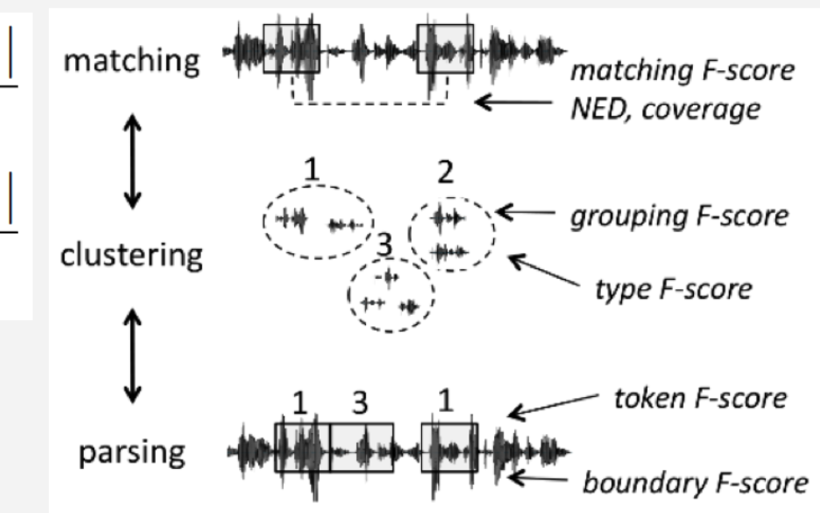
- Connected component clustering using aligned subsequence pairs with common members in order to generate a set of discovered linguistic units
- Averaged start and endpoints across each instance of a segment in a cluster
- Generate final transcription using discovered units, the spaces between them, and the regions of silence form the initial segmentation

Evaluation

- Metrics defined by the 2015 Zero Resource Speech Challenge [2]
- Two copra of spoken language: American English, Tsonga

$$\text{prec} = \frac{|\text{types}(F_{\text{disc}}) \cap \text{types}(F_{\text{goldLex}})|}{|\text{types}(F_{\text{disc}})|}$$

$$\text{recall} = \frac{|\text{types}(F_{\text{disc}}) \cap \text{types}(F_{\text{goldLex}})|}{|\text{types}(F_{\text{goldLex}})|}$$



Figures from [2]

Results

- Our system (O) compared to topline human transcriptions (T), and existing methods

	NED	Cov.	Matching			Grouping			Type			Token			Boundary		
			P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
English																	
T	0.0	100	98.3	18.5	31.1	99.5	100	99.7	50.3	56.2	53.1	68.2	60.8	64.3	88.4	86.7	87.5
[3]	**21.9**	16.3	39.4	1.6	3.1	21.4	84.6	33.3	6.2	1.9	2.9	5.5	0.4	0.8	44.1	4.7	8.6
[4]	70.8	42.4				13.4	15.7	14.2	**14.1**	**12.9**	**13.5**	**22.6**	**6.1**	**9.6**	**75.7**	33.7	**46.7**
[5]	61.2	80.2	6.5	**3.5**	**4.6**				3.1	9.2	4.6	2.4	3.5	2.8	35.4	**38.5**	36.9
O	39.4	**92.1**	**51.8**	0.0	0.0	**76.2**	**100**	**82.7**	5.6	5.1	5.3	10.2	1.9	3.2	71.1	22.5	34.2
Tsonga																	
T	0.0	100	100	6.8	12.7	100	100	100	15.1	18.1	16.5	34.1	49.7	40.4	66.6	91.9	77.2
[3]	**12.0**	16.2	**69.1**	0.3	0.5	**52.1**	77.4	**62.2**	3.2	1.4	2.0	**2.6**	0.5	**0.8**	22.3	5.6	8.9
[4]	63.1	94.7				10.7	3.3	5.0	2.2	6.2	3.3	2.3	3.4	2.7	29.2	39.4	33.5
[5]	43.2	89.4	21.2	**3.8**	**6.5**				**4.9**	**18.8**	**7.8**	2.2	**12.6**	0.8	18.8	**64.0**	29.0
O	39.6	**95.5**	35.7	0.0	0.0	19.1	**100**	31.7	1.6	2.2	1.9	1.5	0.5	0.8	**49.9**	27.6	**35.5**

Discussion

- Improvements on previous results in several categories
- Notably, highest Coverage with relatively low NED
- GPU based implementation allowed our system to run in reasonable amounts of time on these datasets
- Future interests lie in applications related to OOV detection in open-world ASR, especially in human robot interaction contexts

References

[1]G. Aimetti, R. K. Moore, and L. ten Bosch, "Discovering an optimal set of minimally contrasting acoustic speech units: A point of focus for whole-word pattern matching," 2010.

[2]Versteegh, Maarten, Roland Thiolliere, Thomas Schatz, Xuan-Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. "The zero resource speech challenge 2015." In INTERSPEECH, pp. 3169-3173. 2015.

[3]Jansen, Aren, and Benjamin Van Durme. "Efficient spoken term discovery using randomized algorithms." In Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, pp. 401-406. IEEE, 2011.

[4]Räsänen, Okko, Gabriel Doyle, and Michael C. Frank. "Unsupervised word discovery from speech using automatic segmentation into syllable-like units." In INTERSPEECH, pp. 3204-3208. 2015.

[5]VLyzinski, Vince, Gregory Sell, and Aren Jansen. "An evaluation of graph clustering methods for unsupervised term discovery." In INTERSPEECH, pp. 3209-3213. 2015.

Acknowledgements

This work was funded in part by US Office of Naval Research grant #N00014-14-1-0751. We thank Roger K Moore for the helpful discussion.