Acoustic absement in detail: Quantifying acoustic differences across time-series representations of speech data

Matthew C. Kelley
Department of Linguistics, University of Washington

1. Introduction

- Time-series data types are commonplace in phonetic sciences
- Time often factored out when performing acoustic measurements, e.g., formant values and intensity measurements
- Accounting for temporal aspects when processing acoustic data can benefit analyses
- I demonstrate the use of acoustc absement, the time-integral of acoustic distance, in a speech recognition task

2. Distance and absement

- Acoustic distance quantifies how far apart two measurements are, acoustically
- Euclidean distance on two vectors of acoustic measurements, *x* and *y*, is deinfed as

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{i=1}^{k} |\chi_i - \psi_i|^2} \quad (1)$$

- Acoustic absement is the integral of *d(x, y)* over time

$$a(X, Y) = \int_0^T d(x_t, y_t)\mathrm{d}t \quad (2)$$

- Most common calculation of absement in phonetics is with dynamic time warping [1] (see Fig. 1 for more detail)
- Absement itself originates from mechanics and instrument design [2]

3. Experiment and methods

- Isolated word recognition
- 1,000 words randomly drawn from Massive Auditory Lexical Decision data set [4]
- Each word recorded by three speakers: a young male, a young female, and an older male
- Recordings converted to MFCC-by-time using MFCC.jl
- 25 ms window length, 10 ms advance, 13 coefficients w/ 0th replaced by log energy
- Young female and older male MFCCs averaged together using dynamic barycenter averaging [3]
- Absement between young male MFCCs and averaged MFCCs calculated using dynamic time warping (represented in Fig. 2)
- Lower absement indicates better match
- Scaled absement by dividing by square root of length of averaged MFCCs

4. Results and discussion

- 57.9% of words recognized correctly
- 87.9% of words in top 10 matches
- Making comparisons based on instantaneous measurements would have made the word recognition task far more difficult
- Dynamic time warping has generally been abandoned for speech recognition purposes for being too simple
- However, dynamic time warping illustrates the potential use of absement as a concept
- Absement is a general concept and provides a framework to account for temporal structure into phnetic measurements

Link to Git repo:

Link to paper:

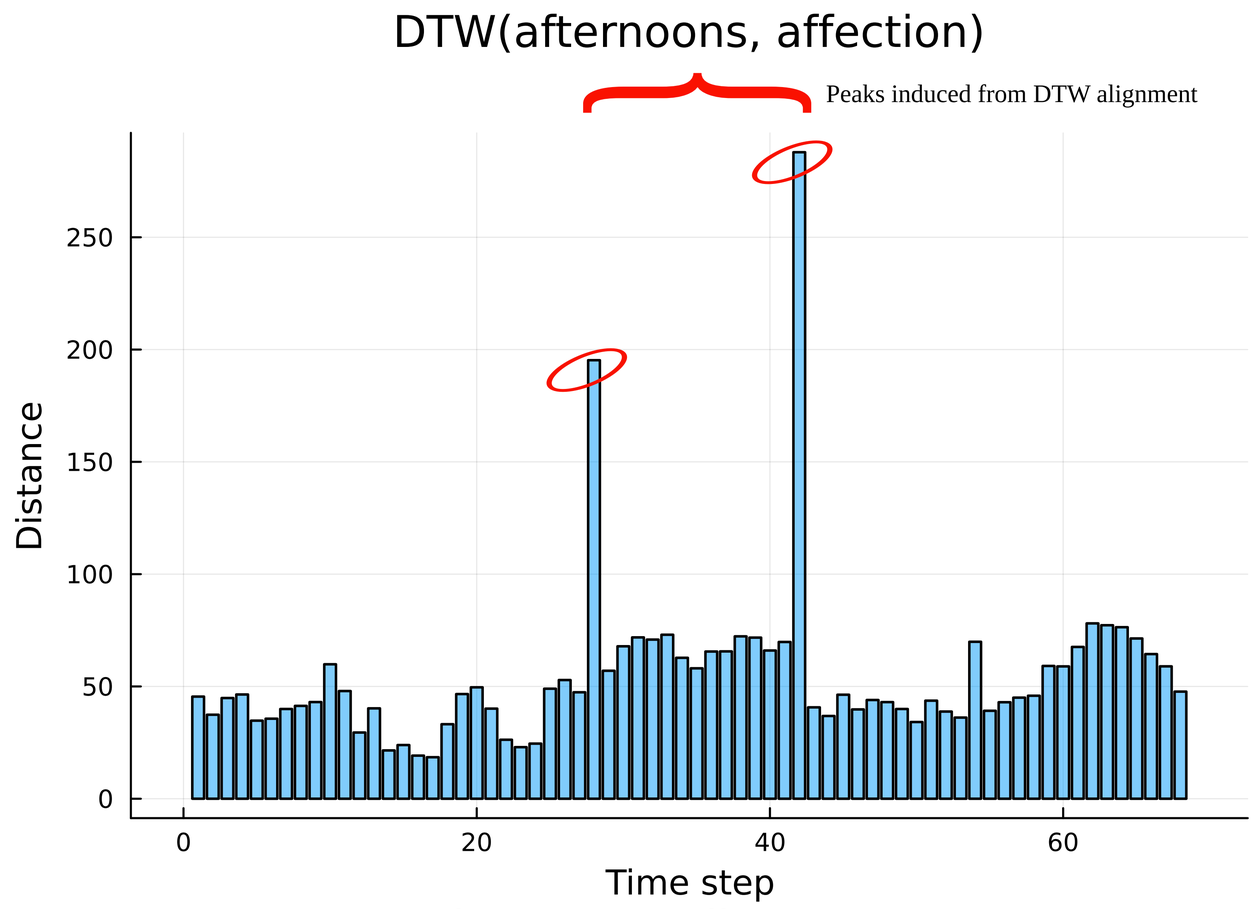

Figure 1. DTW between afternoons and affection, as rectangle rule integration of distance. Some peaks are the result of multiple time steps being aligned together during the DTW process. Absement (as DTW cost) is the total area in all of the rectangles in the plot.

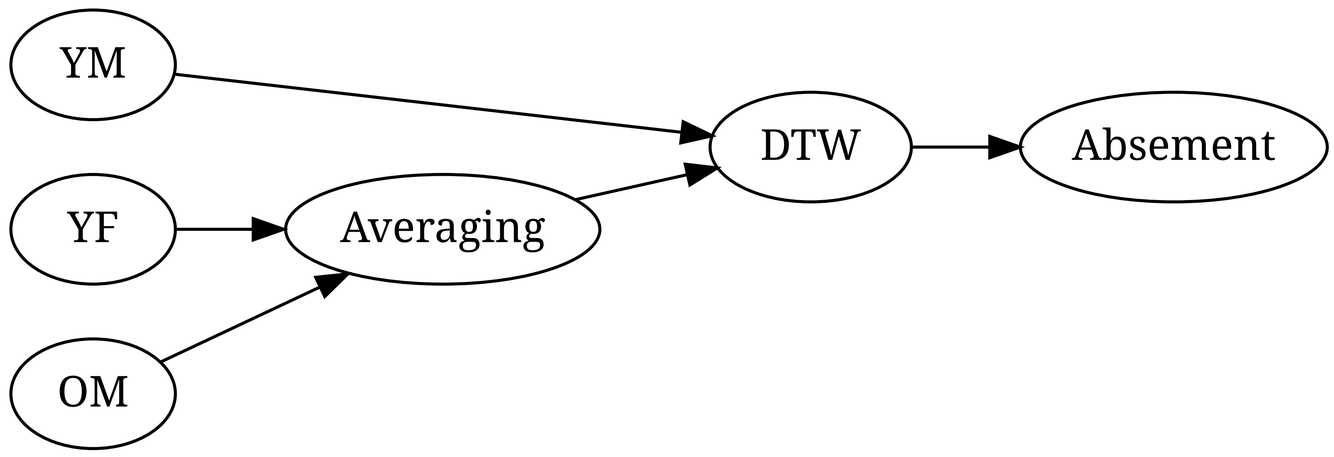

Figure 2. Graph of recognition analysis. "YM" is the young male, "YF" is the young female, and "OM" is the older male. The data at the beginning is the MFCC representation of a recording. "DTW" is the dynamic time warping calculation, yielding the absement value.

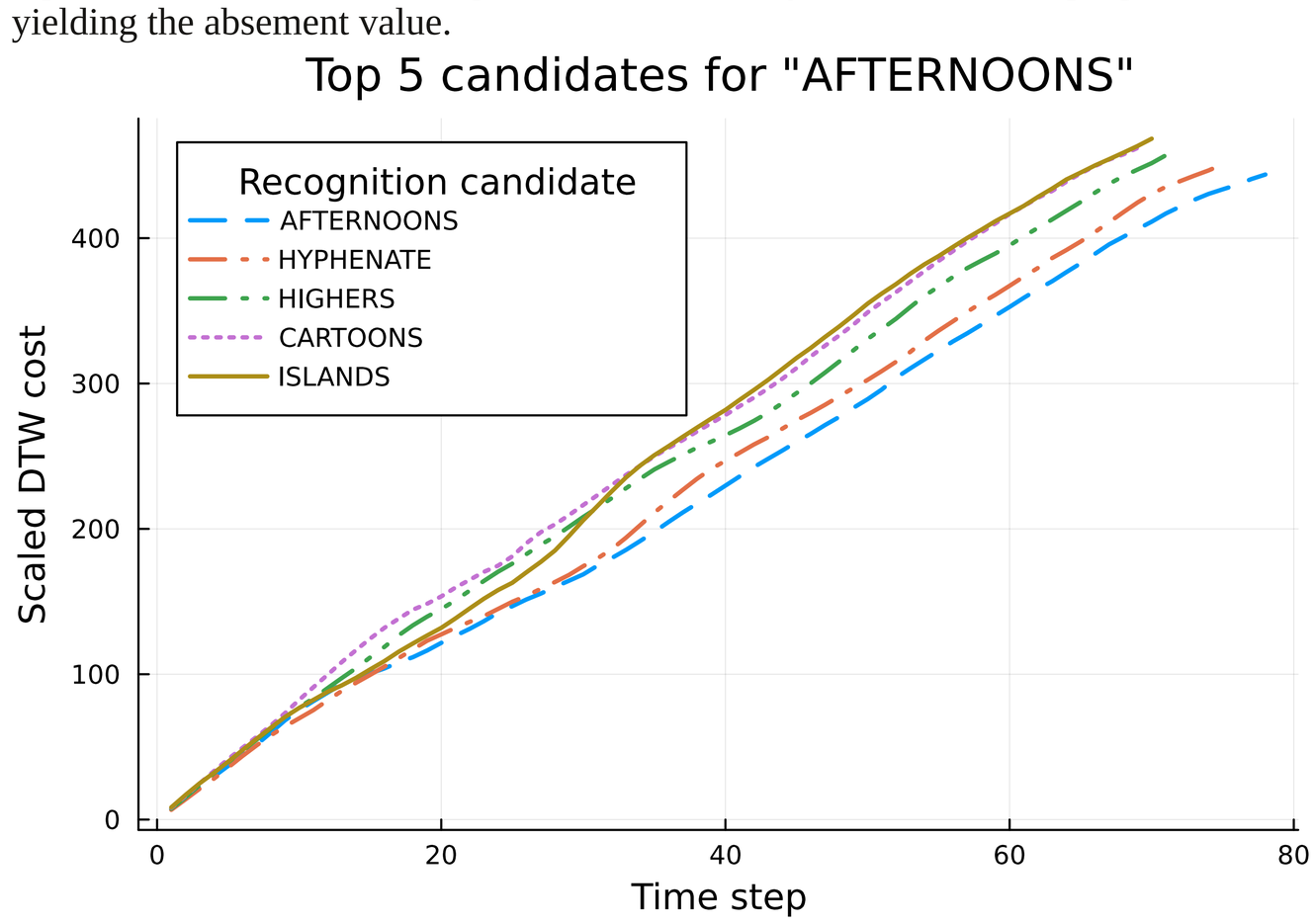

Figure 3. Example of top 5 recognition candidates when recognizing *afternoons*. For each candidate, the young male speaker's recording of the word was compared to the averaged recording from the young female and older male speakers. The values on the y-axis are absement represented as the dynamic time warping (DTW) cost scaled by the square root of the length of the candidate.

References

[1] Kelley, M. C., & Tucker, B. V. (2022). Using acoustic distance and acoustic absement to quantify lexical competition. *The Journal of the Acoustical Society of America*, *151*(2), 1367-1379.

[2] Mann, S., Janzen, R., & Post, M. (2006). Hydraulophone design considerations: Absement, displacement, and velocity-sensitive music keyboard in which each key is a water jet. *Proceedings of the 14th ACM International Conference on Multimedia*, 519–528.

[3] Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition*, *44*(3), 678-693.

[4] Tucker, B. V., Brenner, D., Danielson, D. K., Kelley, M. C., Nenadić, F., & Sims, M. (2019). The massive auditory lexical decision (MALD) database. *Behavior research methods*, *51*, 1187-1204.

Paper number 415, presented at The 20th International Congress of Phonetic Sciences on 9 Aug 2023 in Prague, Czech Republic