Intonational tunes in imitative speech: Are phonological contrasts maintained?

Jeremy Steffman & Jennifer Cole

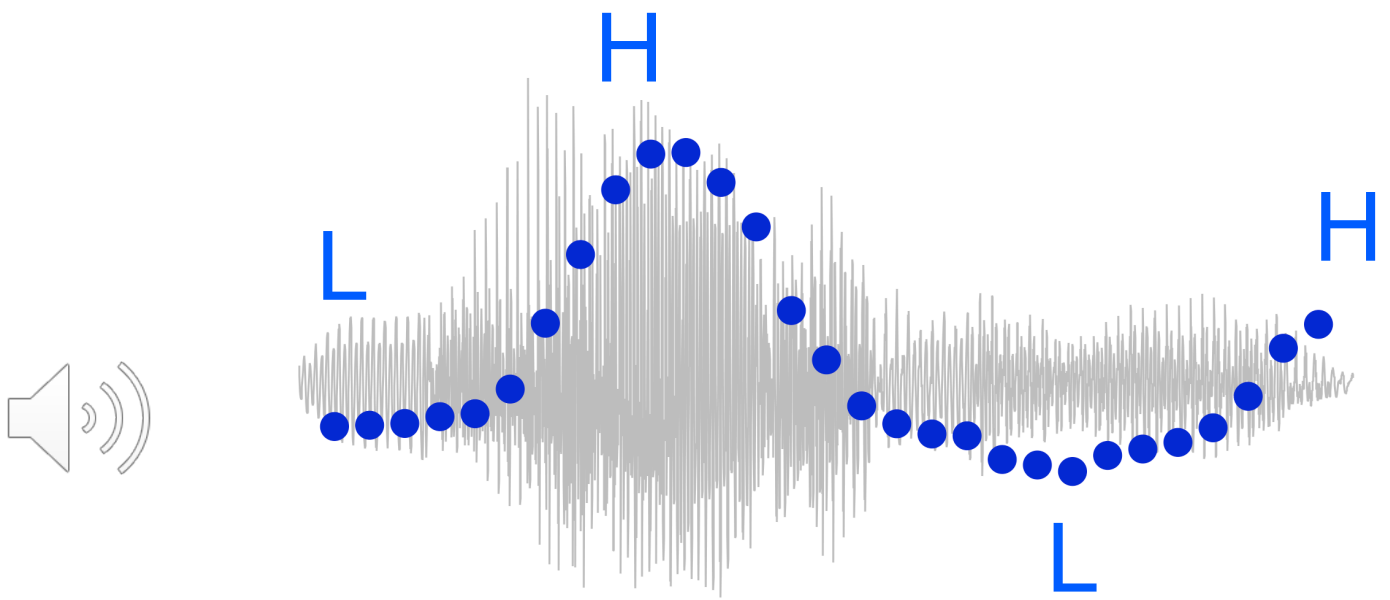
Northwestern University

MidPhon 2021 – September 18, 2021

jeremy.steffman@northwestern.edu

Introduction

- Prevalent theory of Intonational Phonology: intonational contours are phonologically specified as high (H) and low (L) tone targets
- Phonological inventories consist of tones which link to prominent positions and domain edges



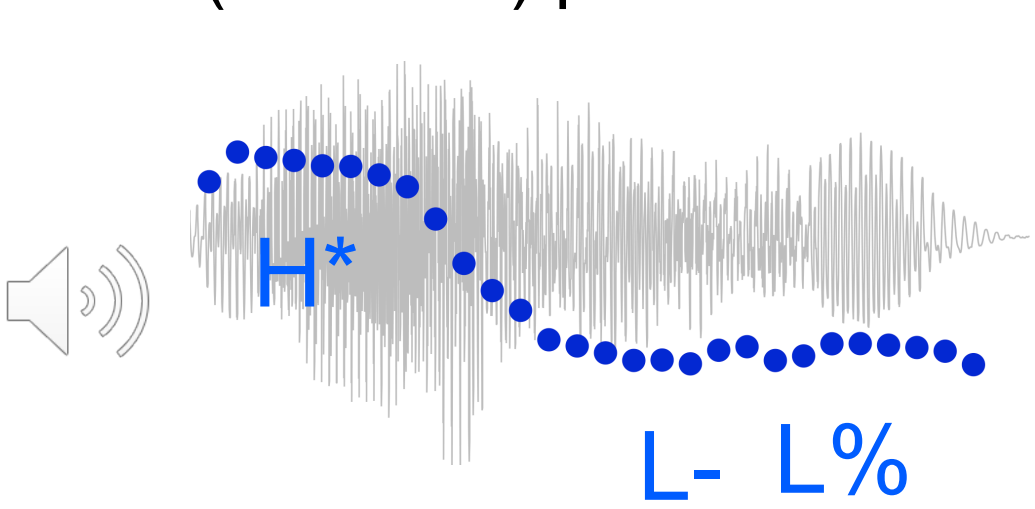
= L+H* L- H%

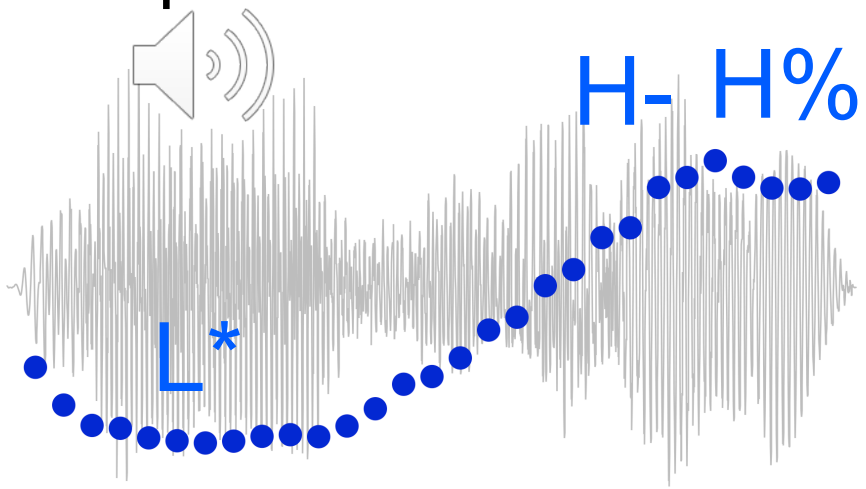
American English: Nuclear Tunes

The final (nuclear) pitch accent + boundary tones = **the nuclear tune**, which may be composed of…

- 2 *monotonal* pitch accents: H*, L* (ignoring other pitch accents)
- 2 phrase accents: H-,L- (smaller phrase edge)
- 2 boundary tones: H%, L% (larger phrase edge)

8 (2 x 2 x 2) possible nuclear tunes composed of the above





The present study

Though 8 distinct tunes are predicted, we lack clear empirical evidence for an 8-way distinction

- Esp. from naïve participants - i.e. not trained intonational phonologists

- Some distinctions (e.g., rise vs. fall) are well studied - others not
- Some nuclear tunes are barley attested in labeled corpora (Dainora 2009)

Research question:

Do speakers evidence a robust 8-way distinction in nuclear tune shape, as predicted by the tonal inventory?

Do speakers vary in this regard?

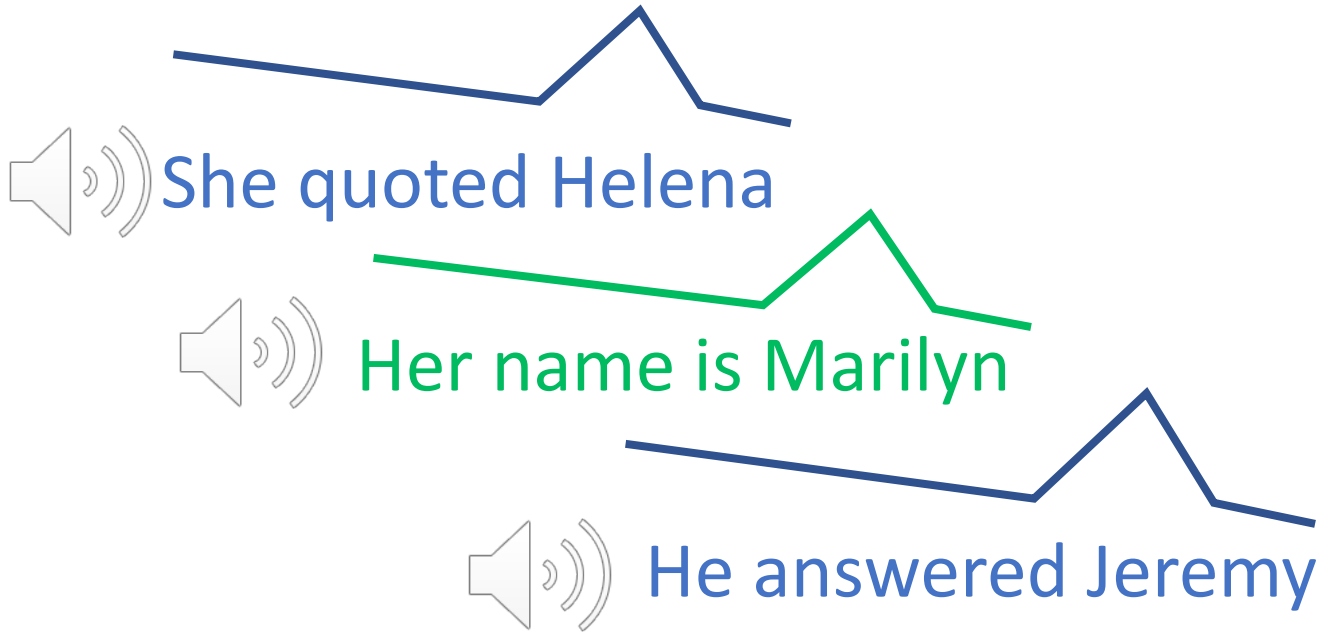
Method: Imitative speech production paradigm

- Method from Chodroff & Cole (2019)
- Participants hear auditory models with resynthesized f0 which exemplify a tune

- They reproduce the tune on a **new** sentence and are instructed to do so in a way “that is familiar to you”

Method: Imitative speech production paradigm

Model sentences



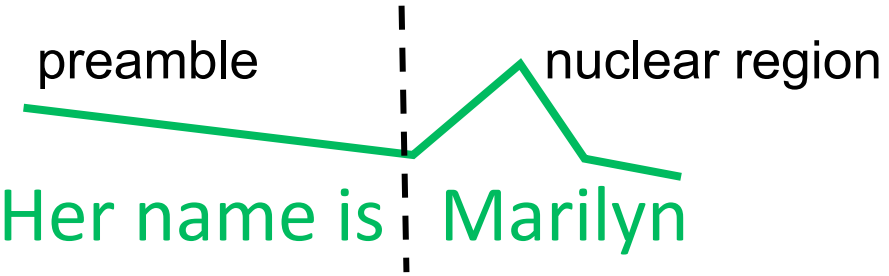
Target sentence

They honored Melanie

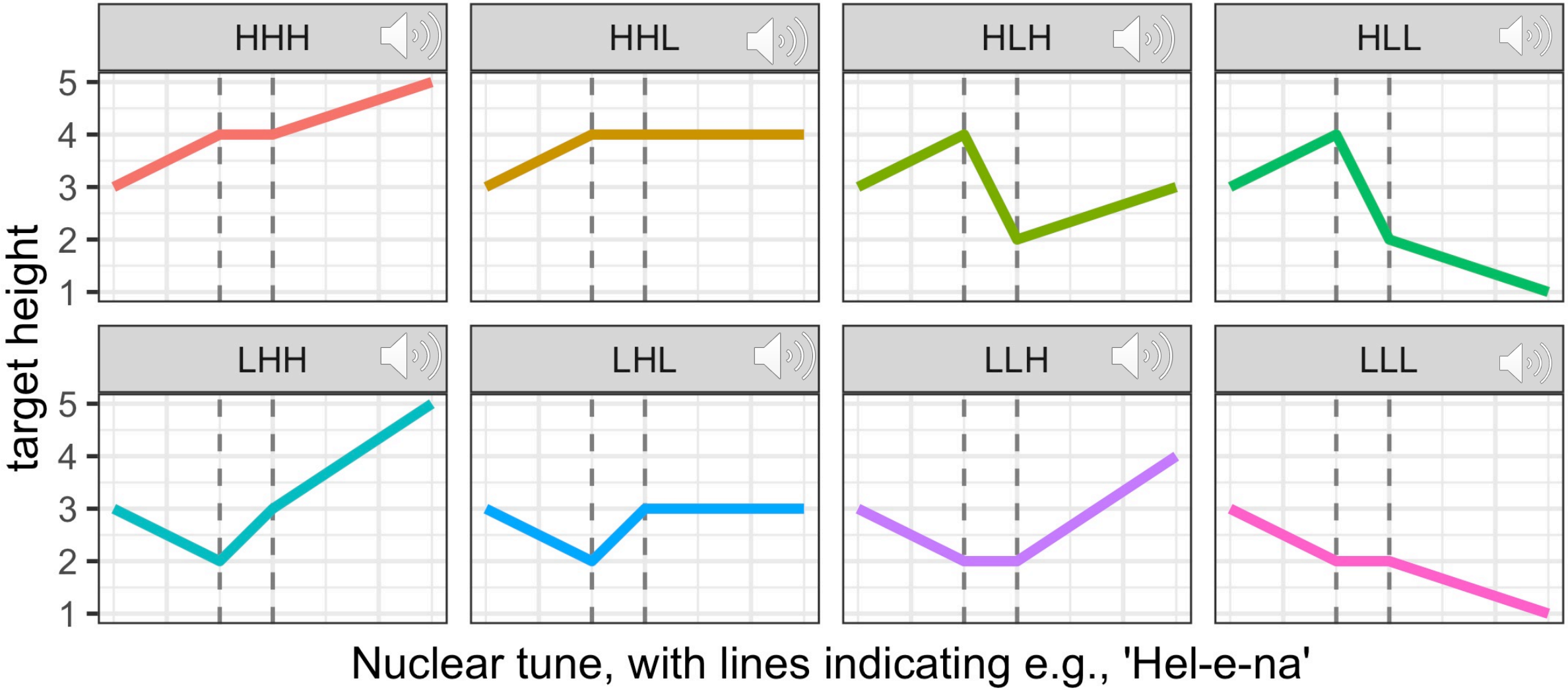
Method contd.

- 30 participants, 144 trials varying model sentence and speaker order
- **8 model nuclear tunes**
  - based on straight-line approximations in ToBI training materials

    (MIT OpenCourseWare: Veilleux et al., 2006 - based on Pierrehumbert, 1980)
  - 5 target heights spread across model speakers' pitch range
  - "preamble" f0 declines until nuclear region (same for all tunes)
  - nuclear region: always a trisyllabic, initial-stress, name



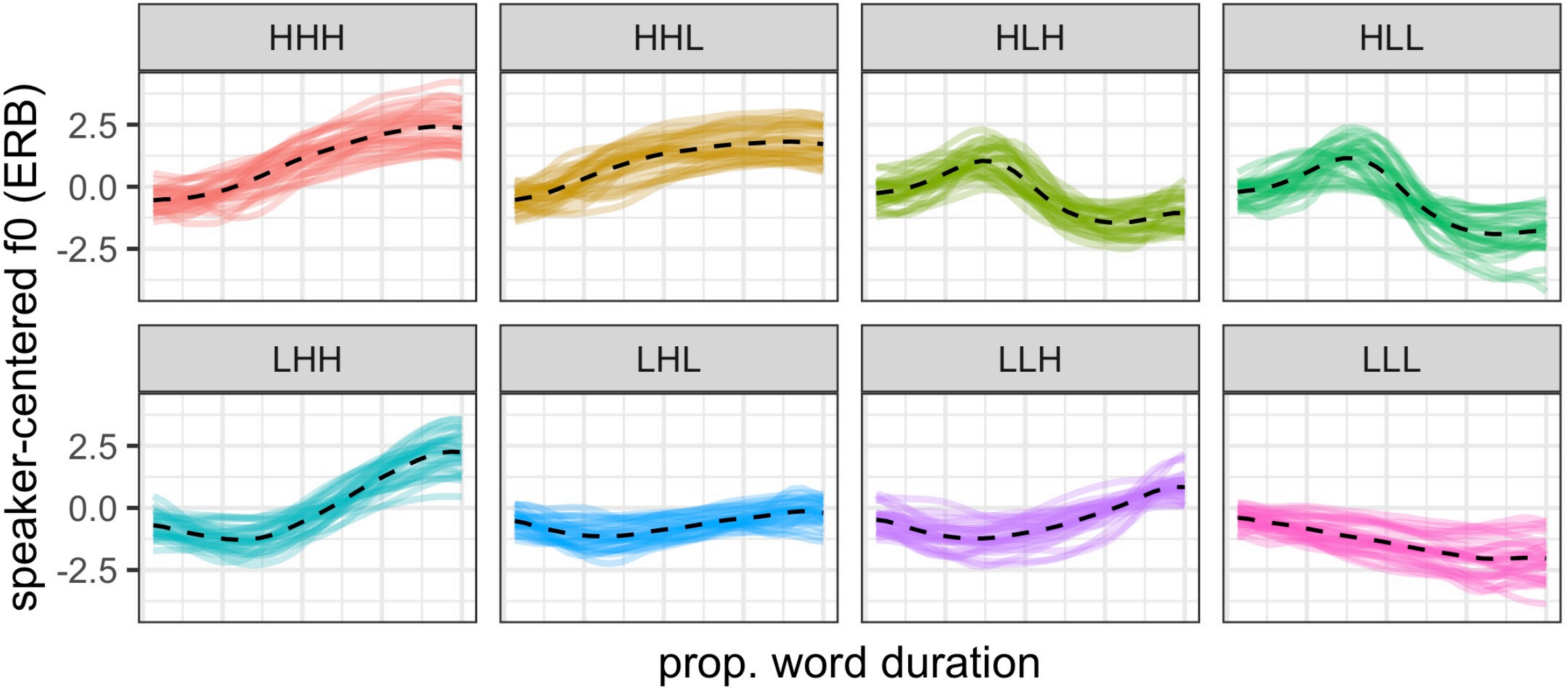
Model tunes: nuclear region



Measures

- Time-normalized f0
  - For the nuclear region only
  - Measured via STRAIGHT, using VoiceSauce
    (Kawahara et al. 2005 ; Shue 2010)
  - 30 samples per word
  - Converted to ERB, and centered on speaker mean values

Data: speaker means



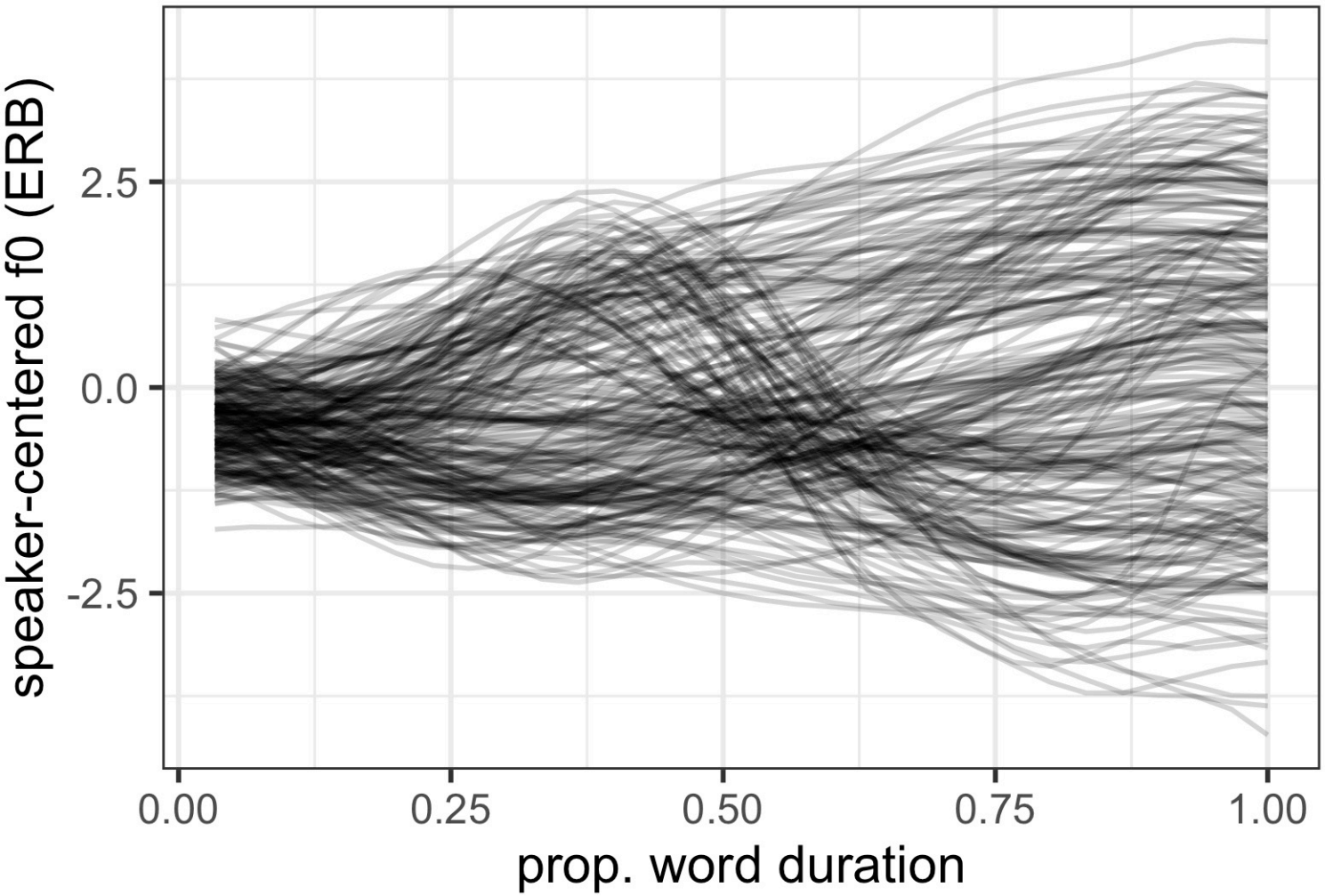
Analysis

K-means clustering for longitudinal data (Genolini et al. 2015)

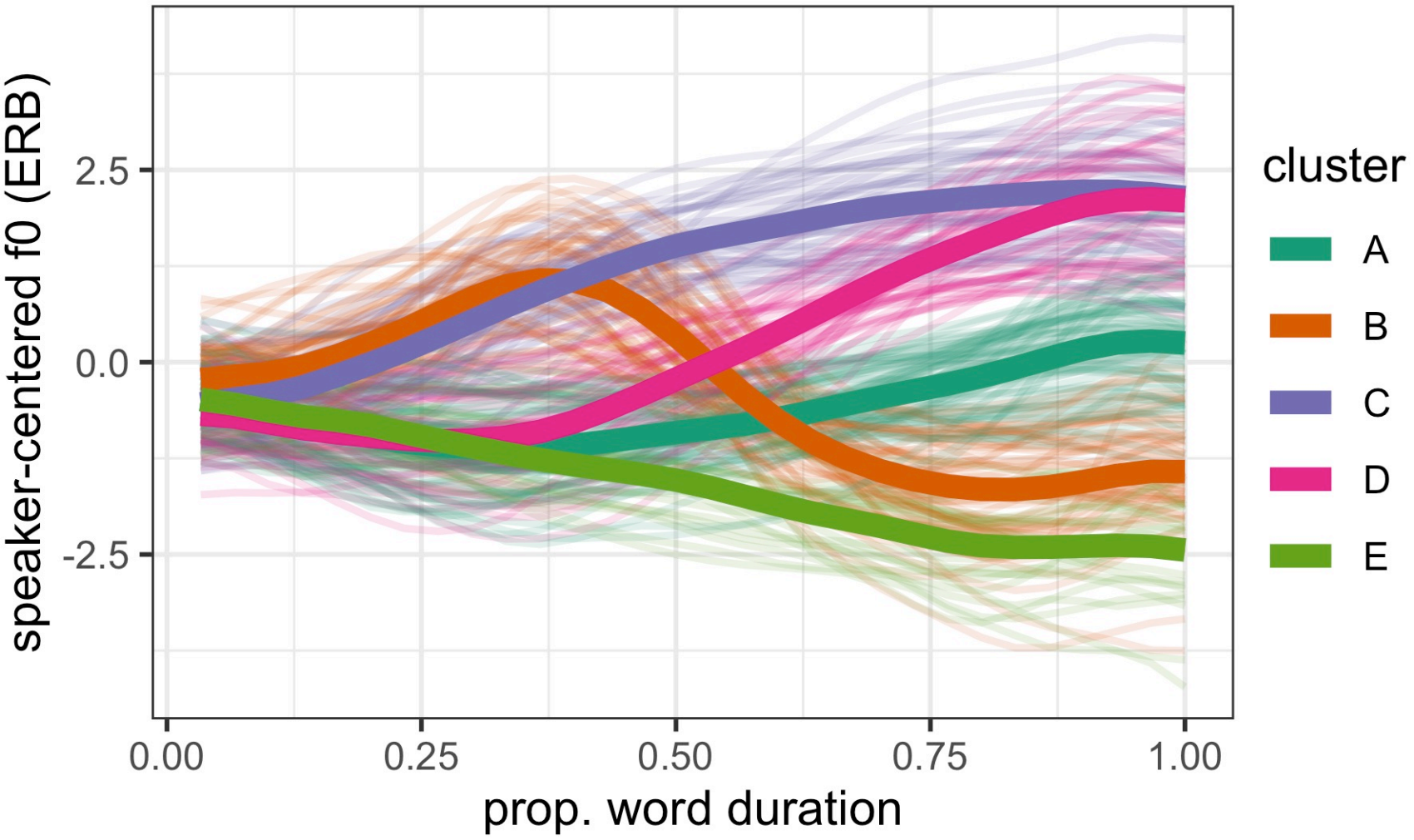
- Assigns trajectories to clusters, and iteratively optimizes cluster membership to minimize within-cluster variation
- We tested 2 through 10 clusters, and assessed what number of clusters best partitions the data (Calinski-Harabasz Criterion)

Results

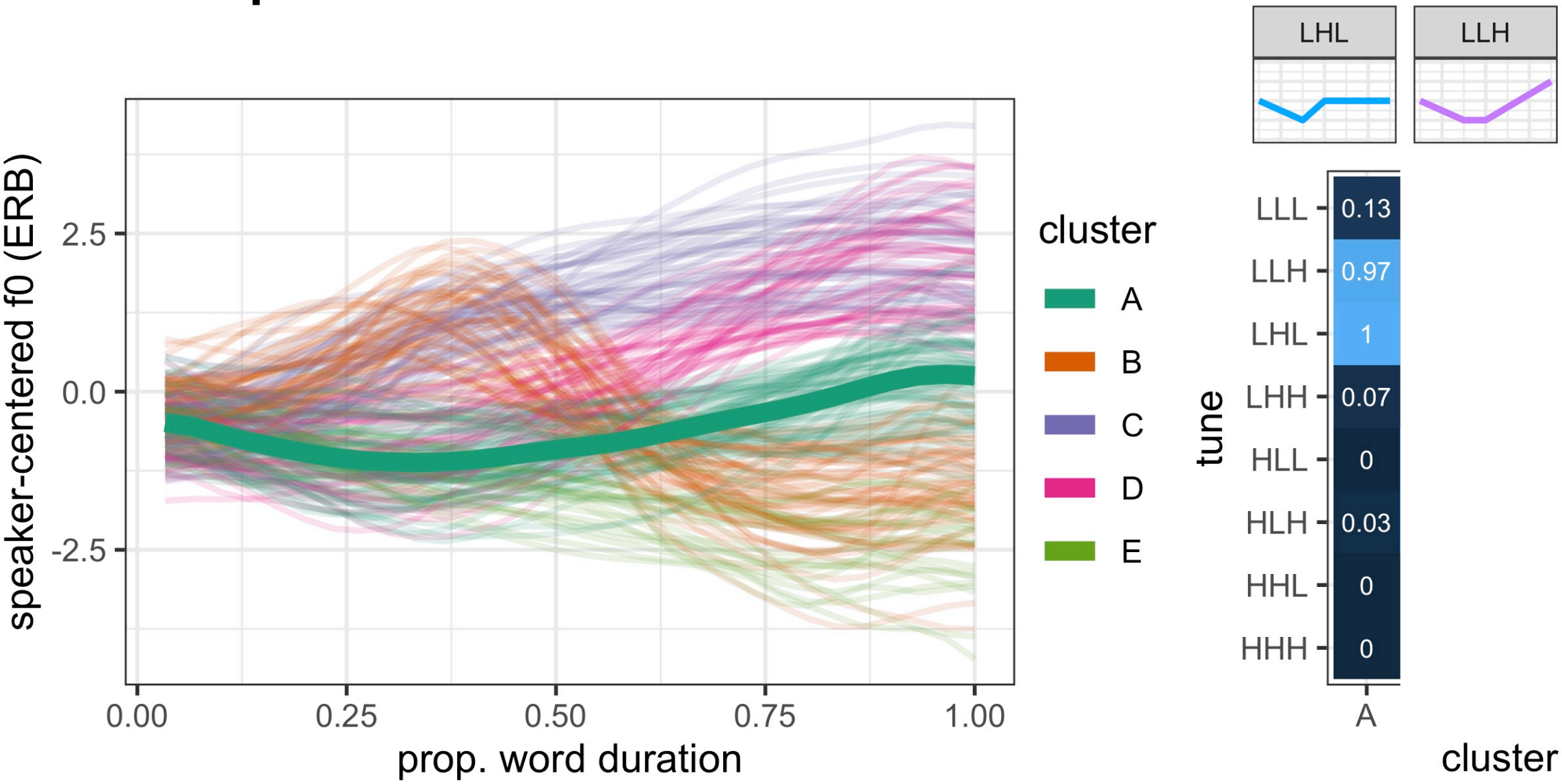
Input: unlabeled speaker means



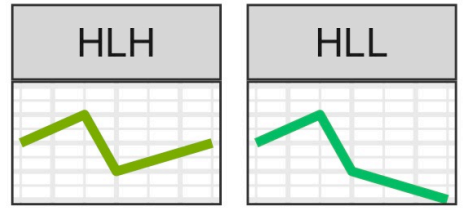
Output: 5 clusters

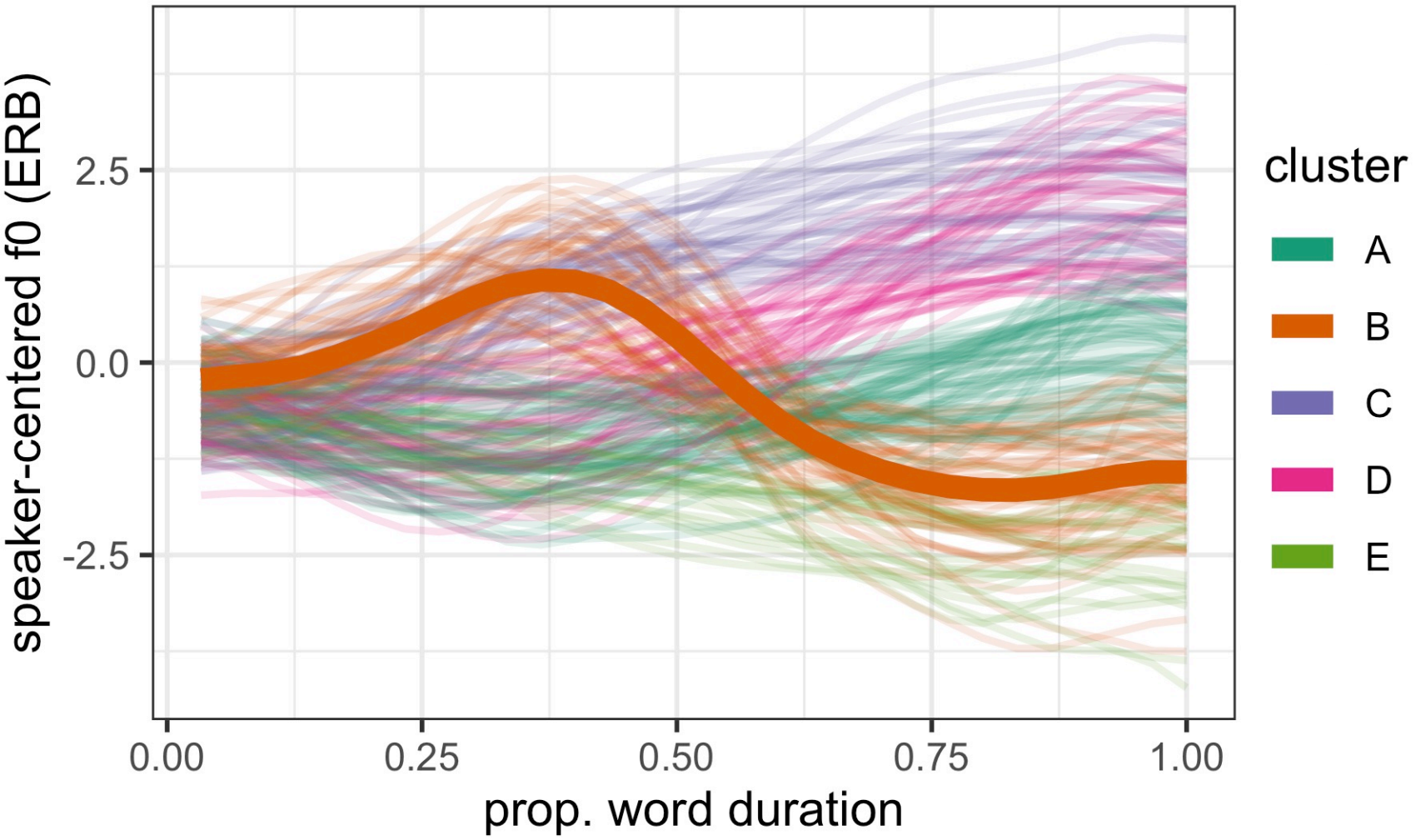


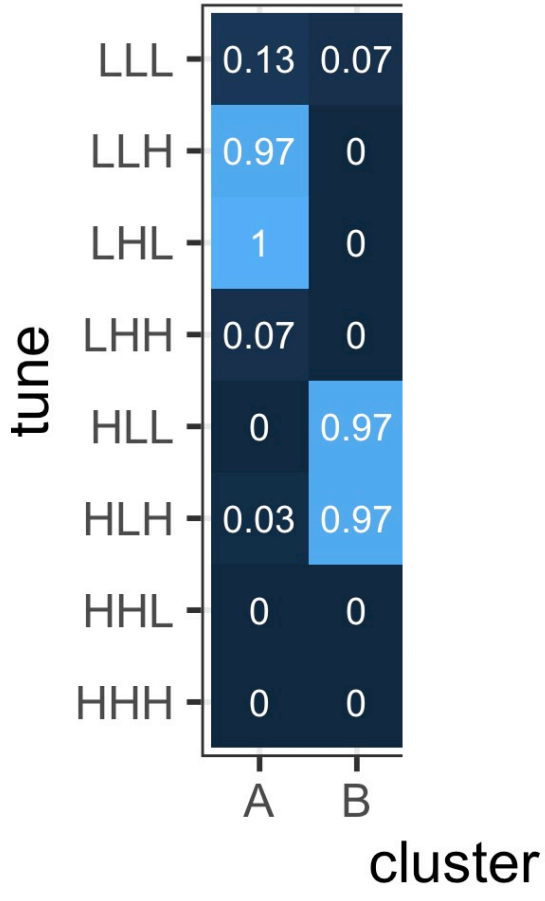
Output: 5 clusters



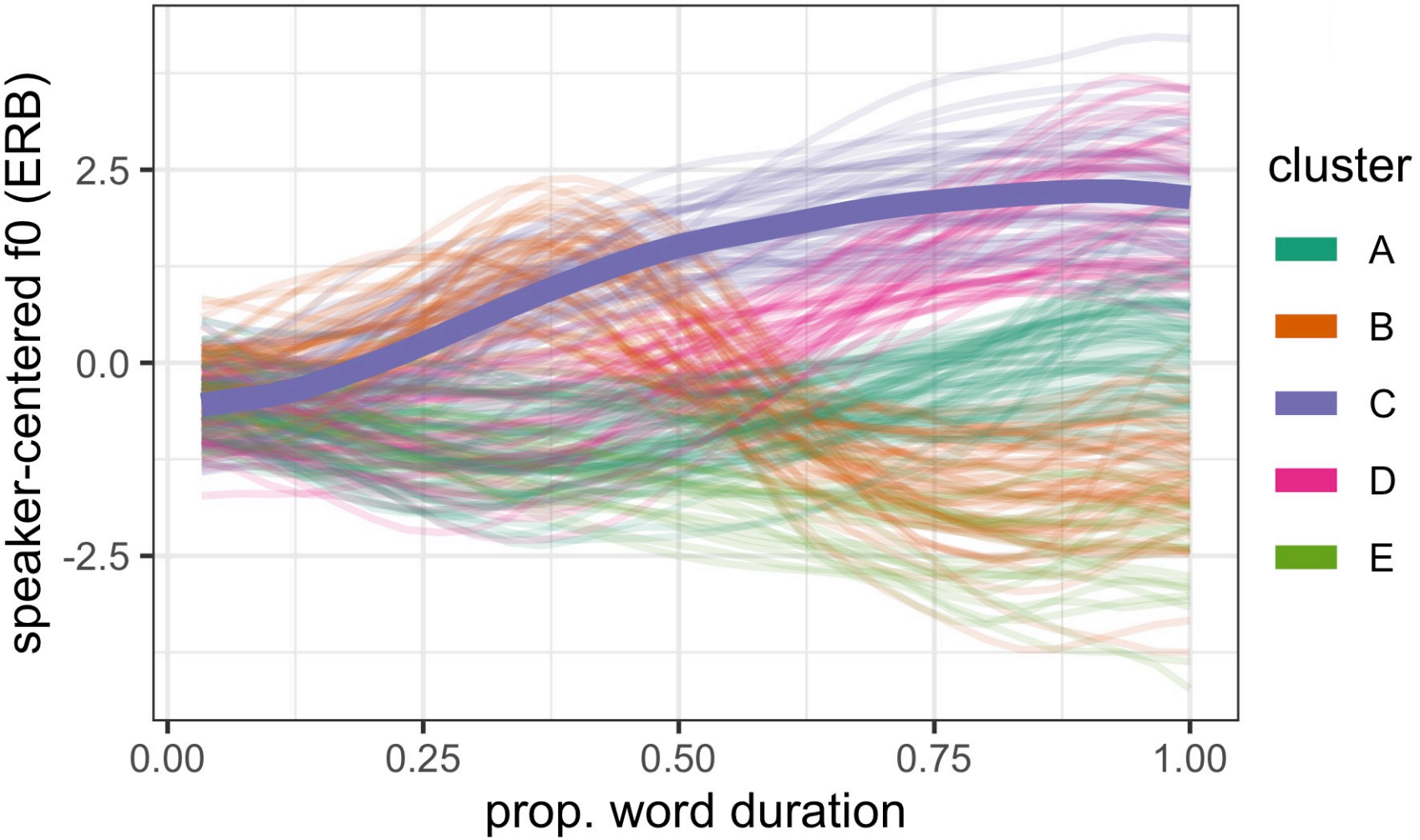
Output: 5 clusters

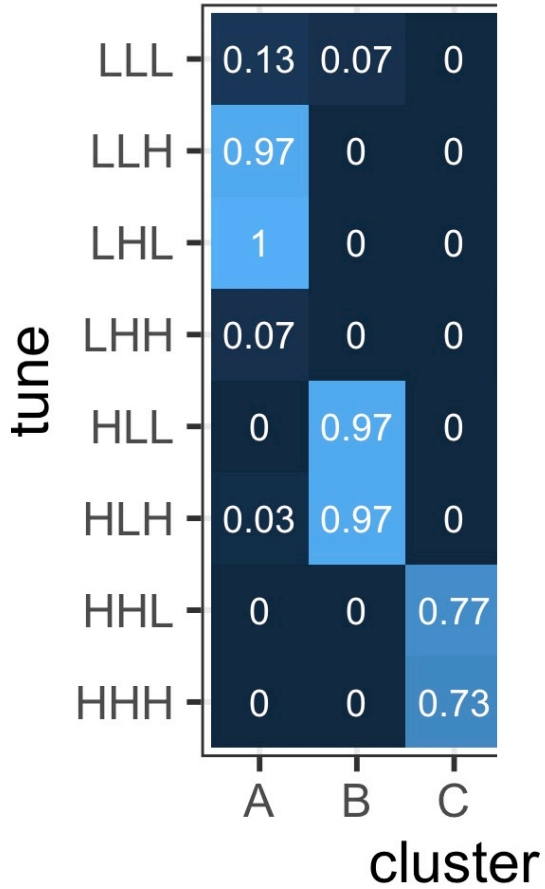




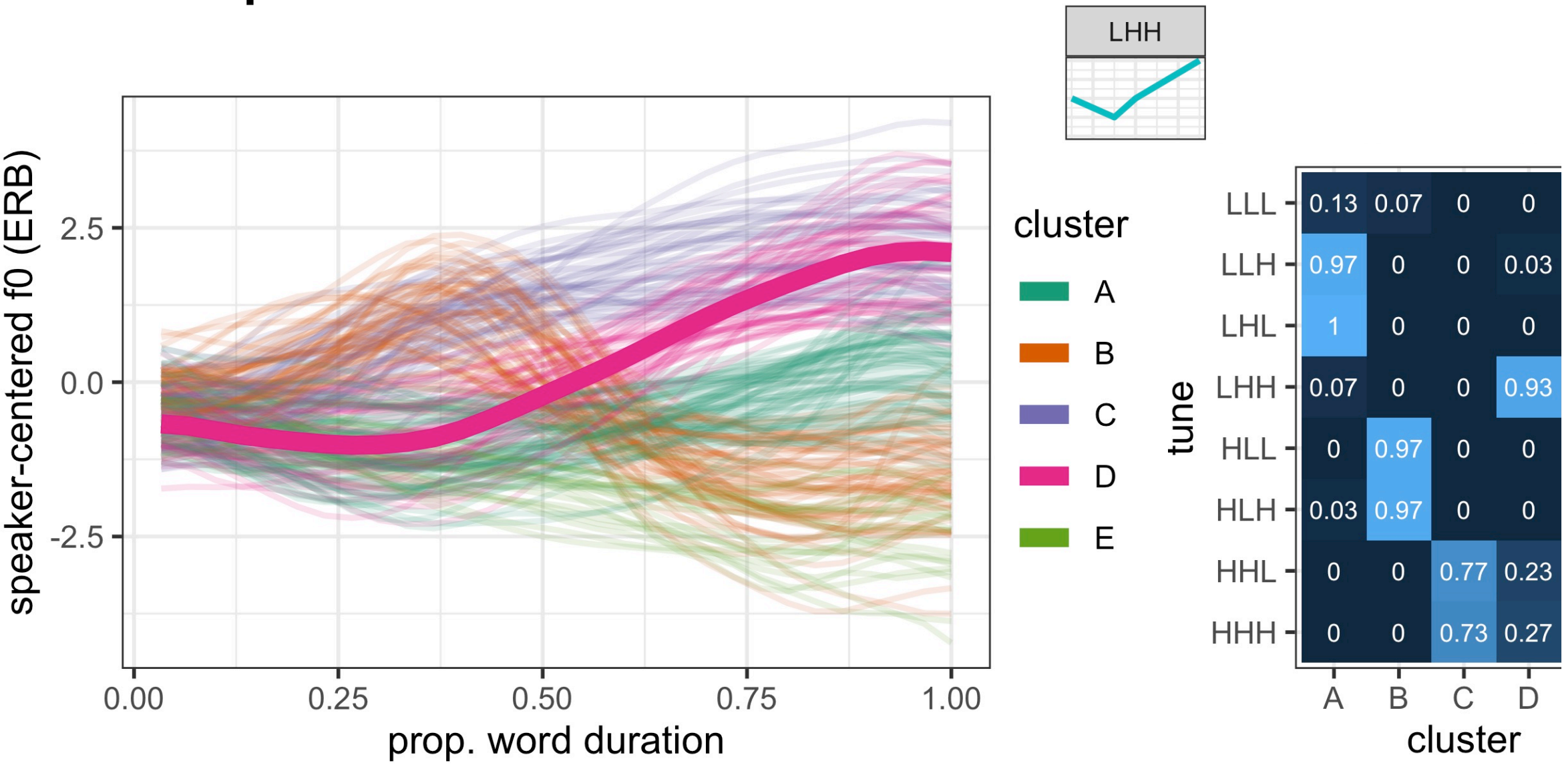


Output: 5 clusters

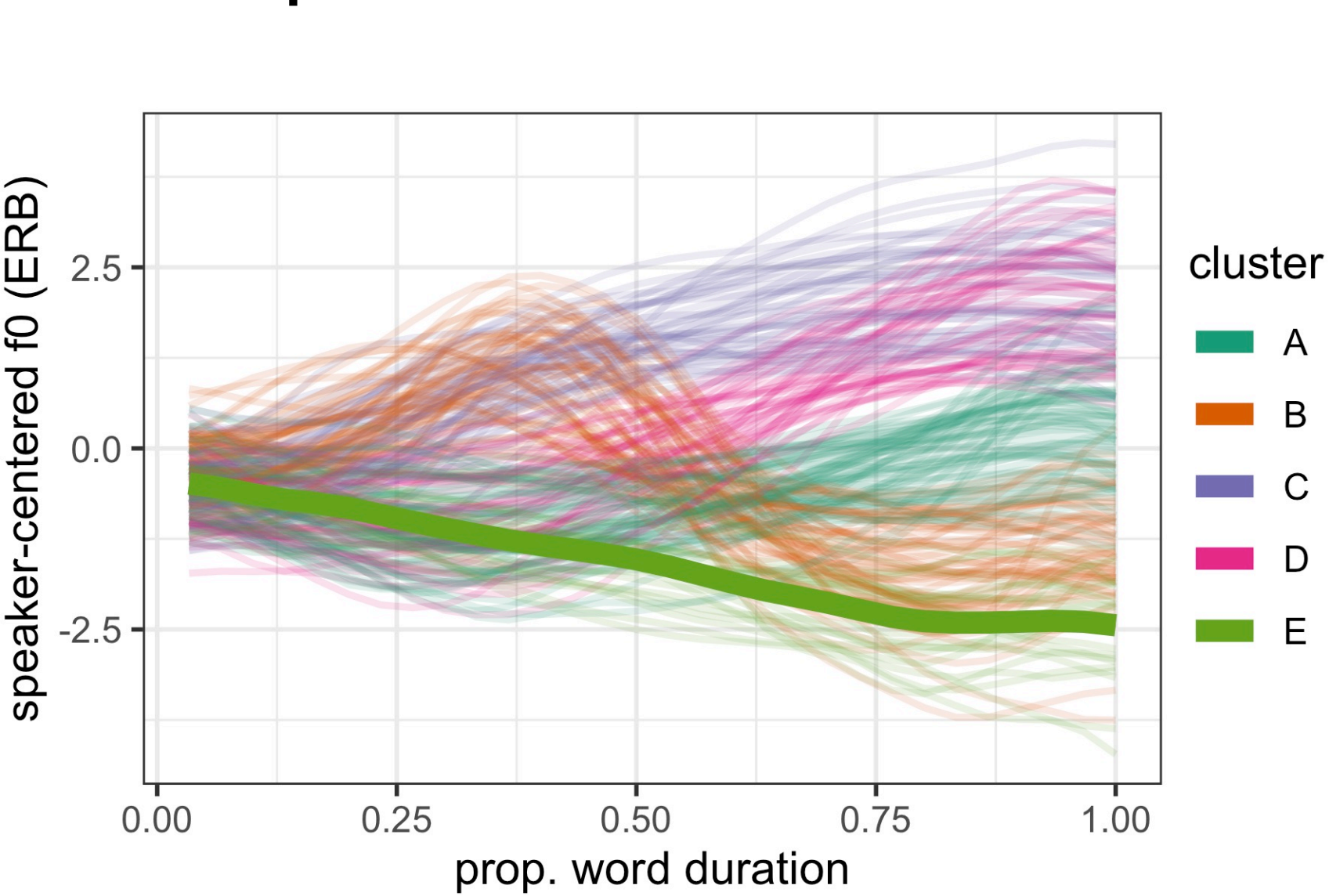


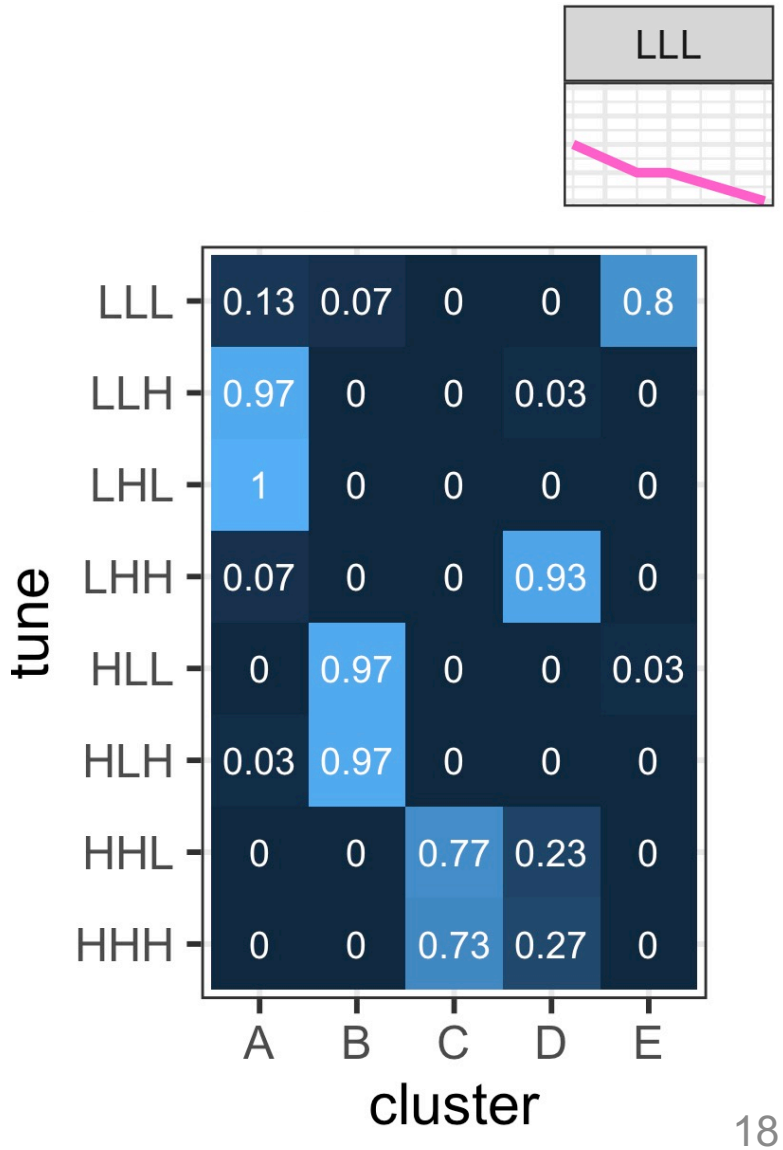


Output: 5 clusters

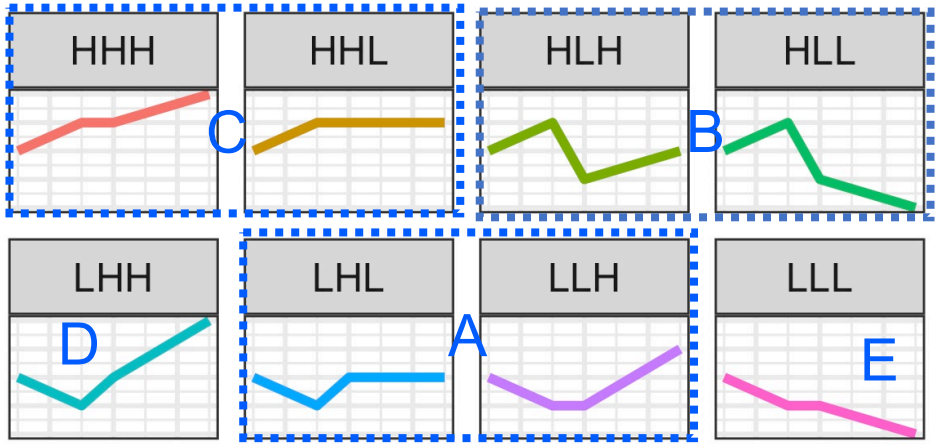


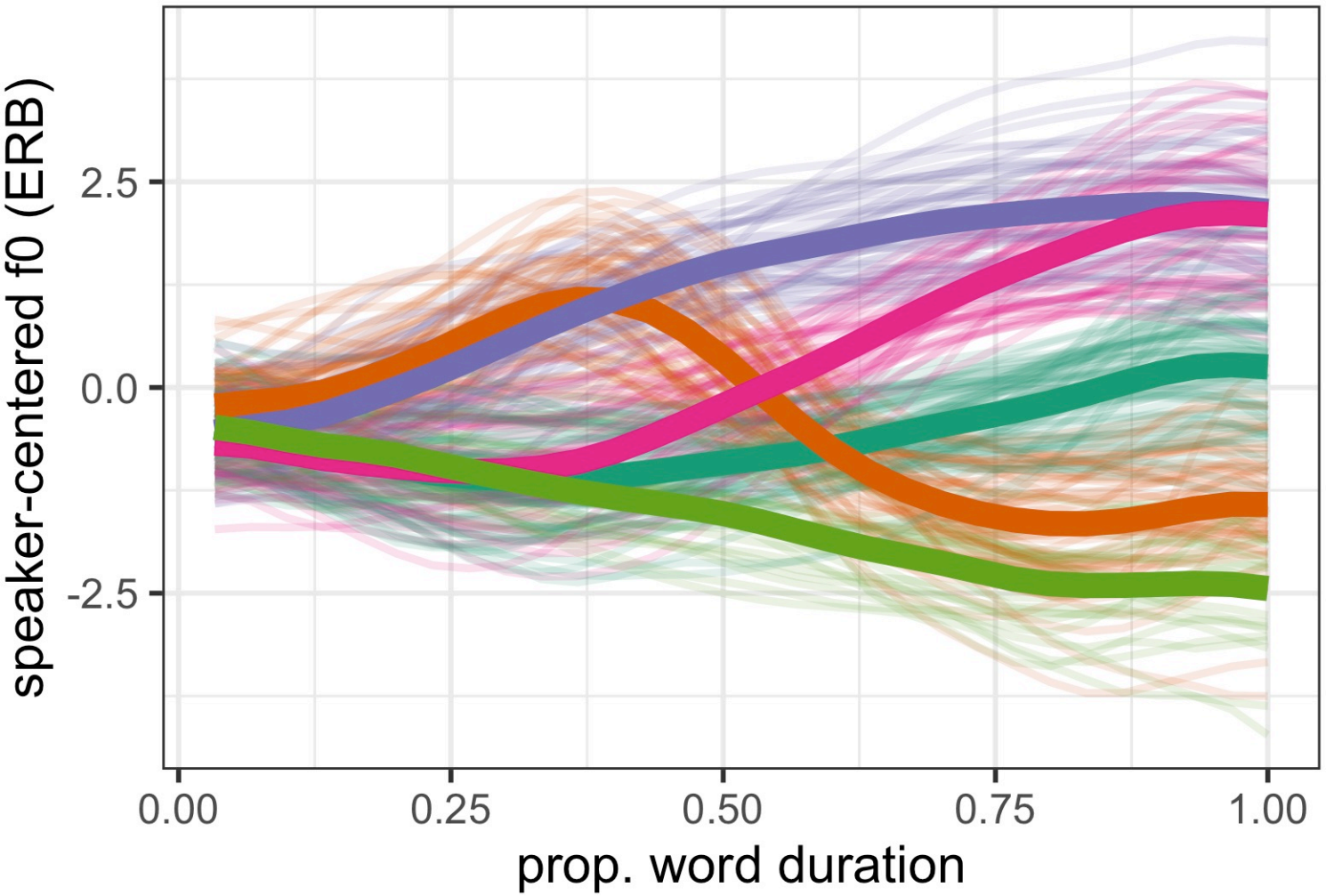
Output: 5 clusters

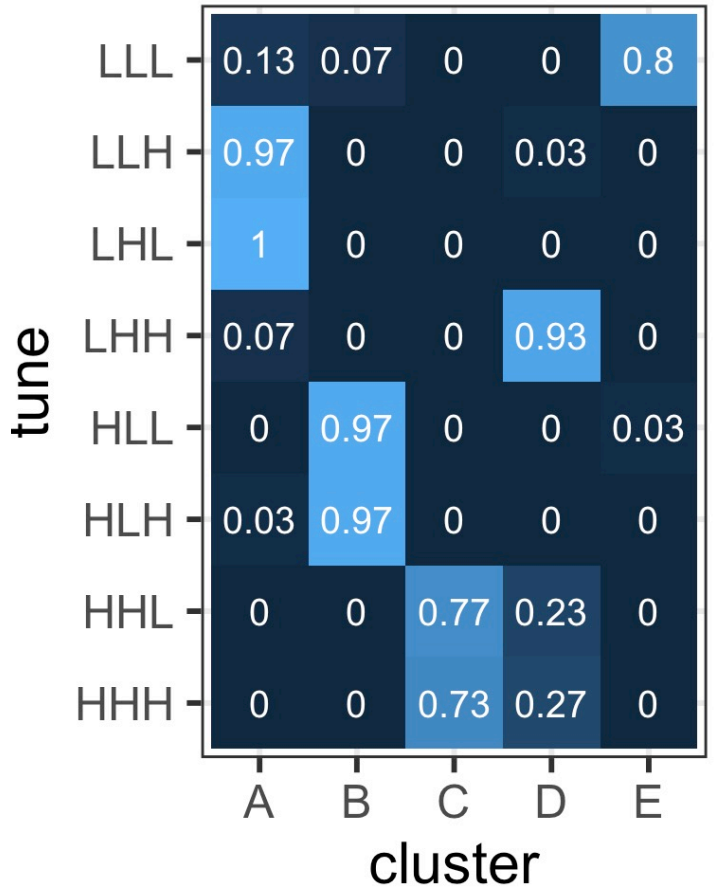




Output: 5 clusters





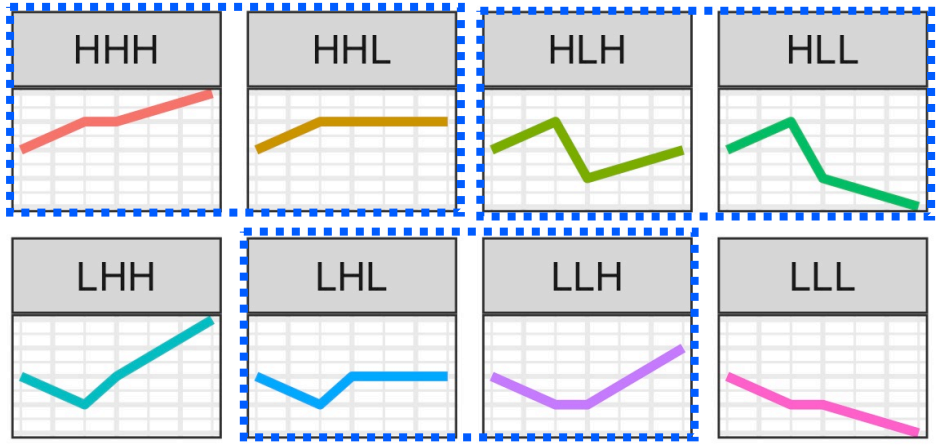


What about individual speakers?

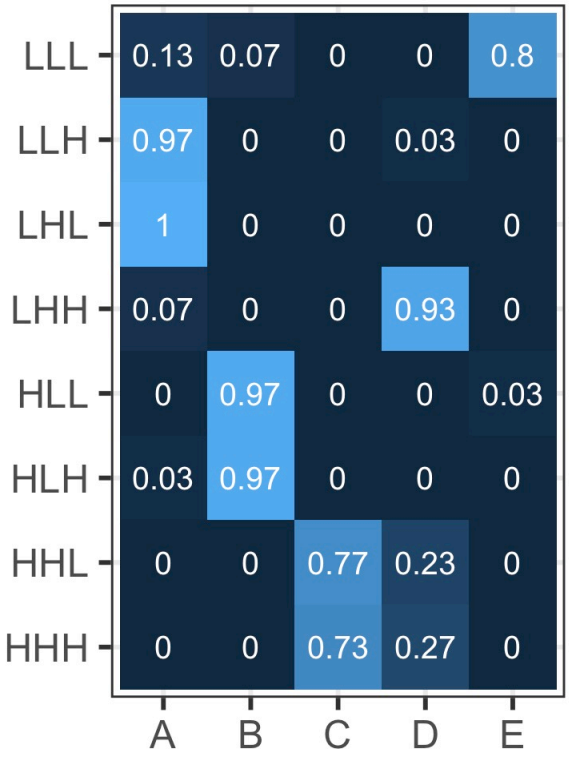
- Are any speakers capturing all distinctions?
- Is there variation?

- To address this: clustering for each speaker on individual trajectories (not speaker means: 144 trajectories per speaker)
  - What are the optimal number of clusters for each speaker?

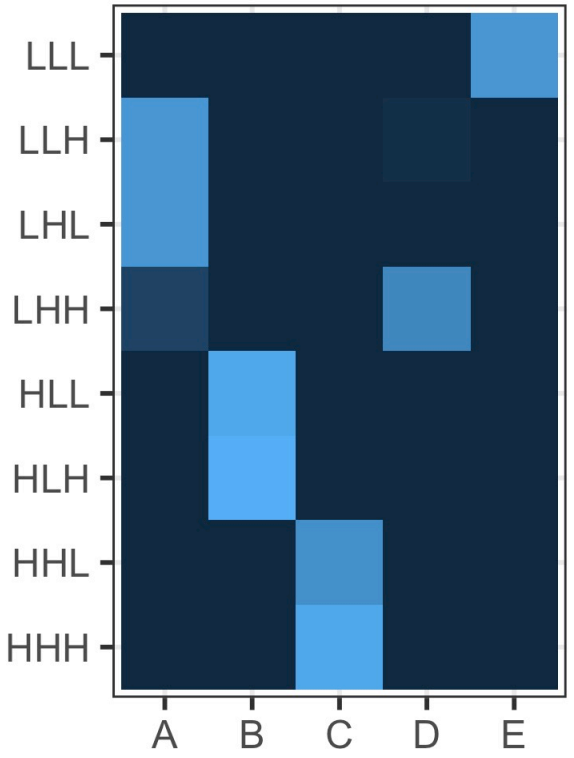
Individual clustering: some speakers resemble the group data

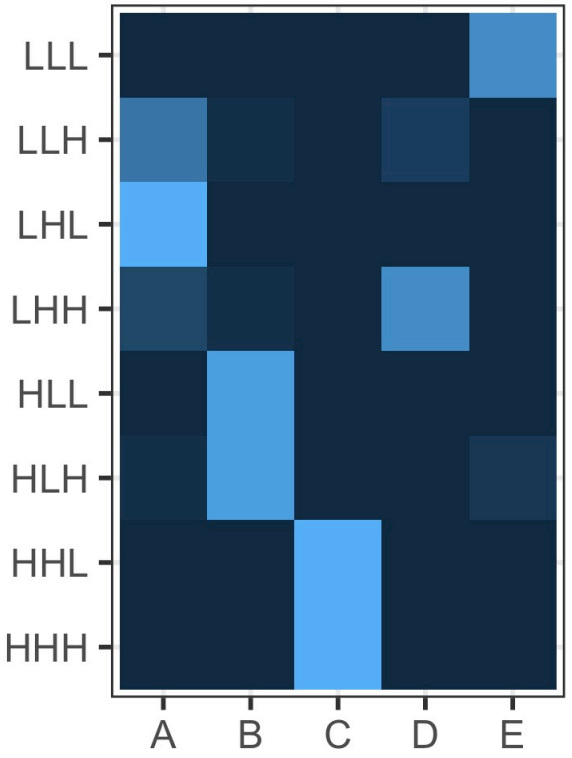


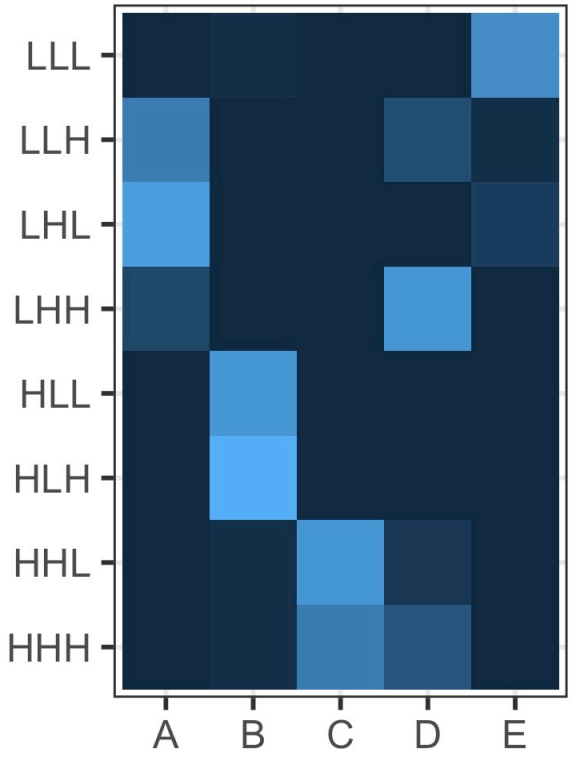
Group data



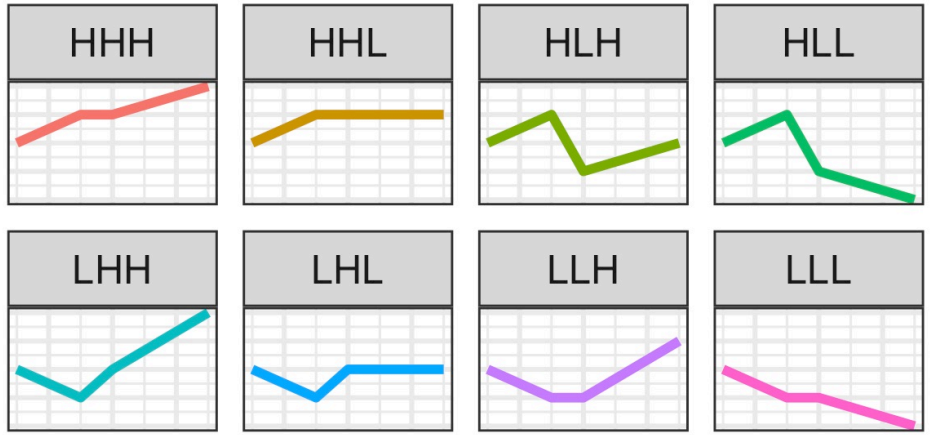
Individual speakers

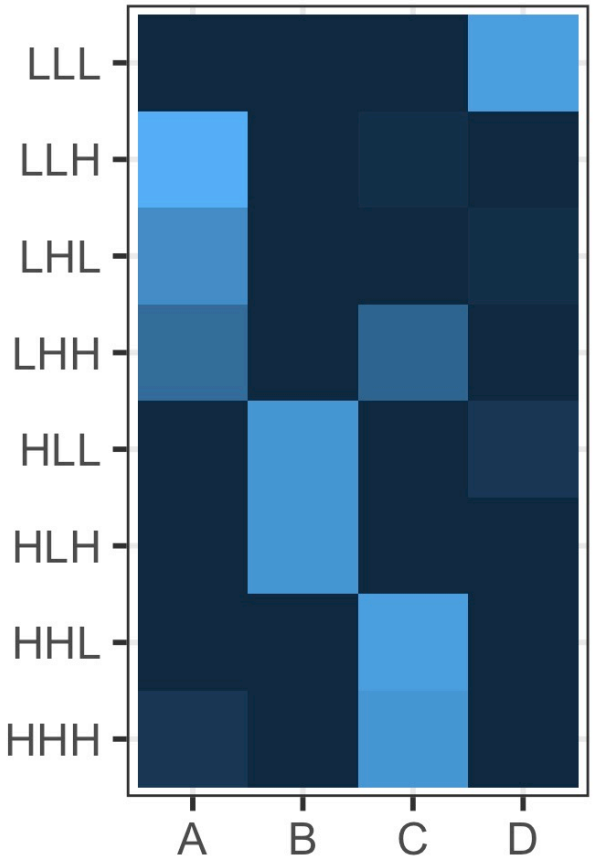


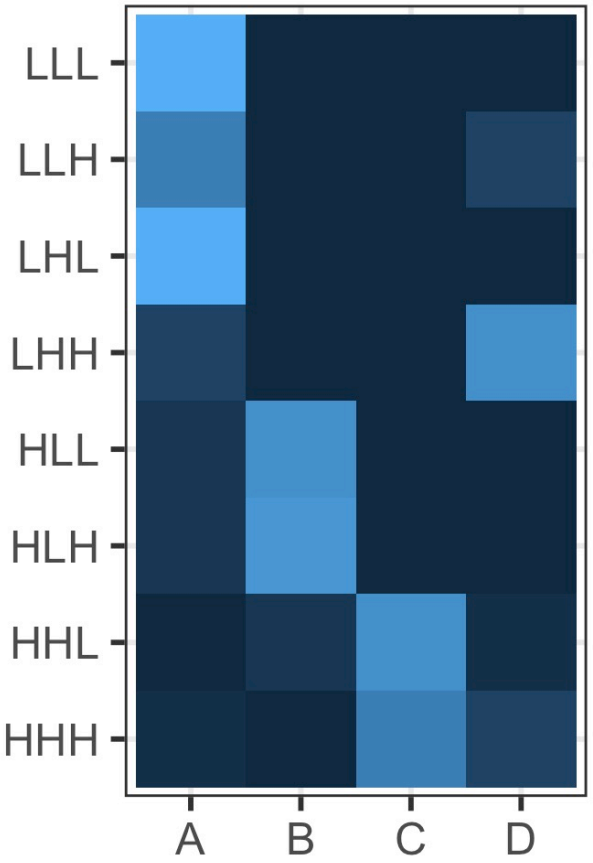


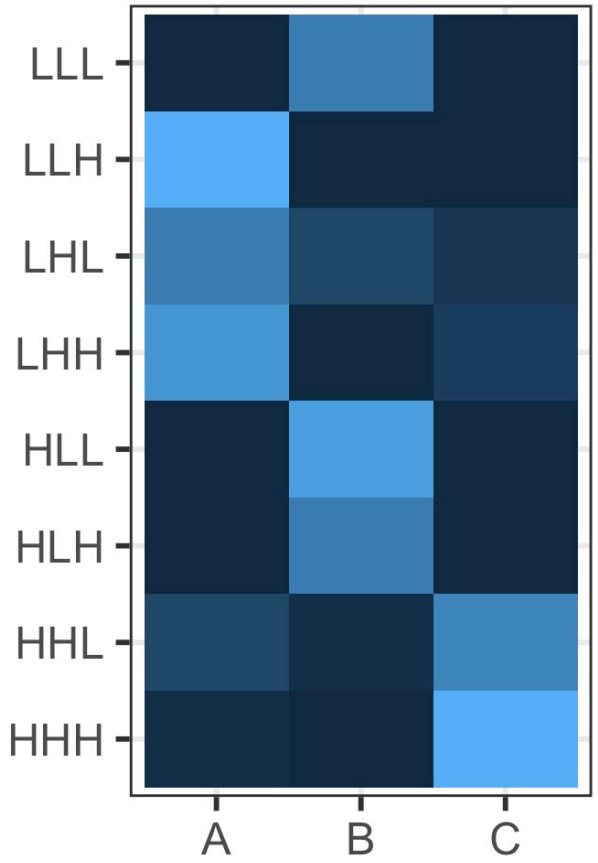


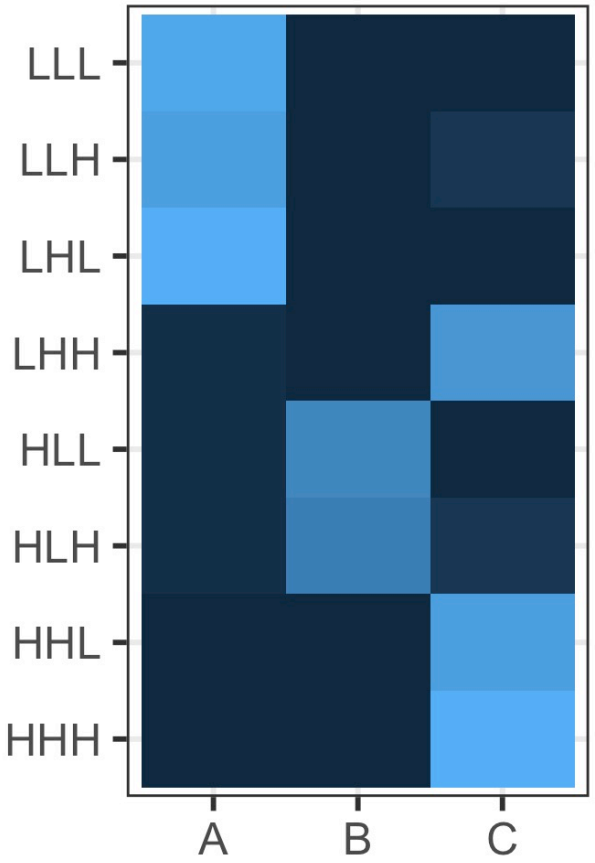
Individual clustering: many speakers do not resemble the group data



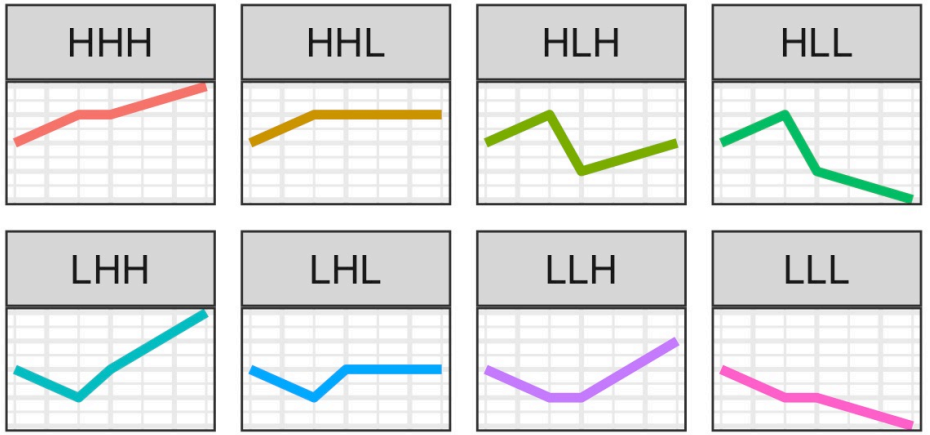


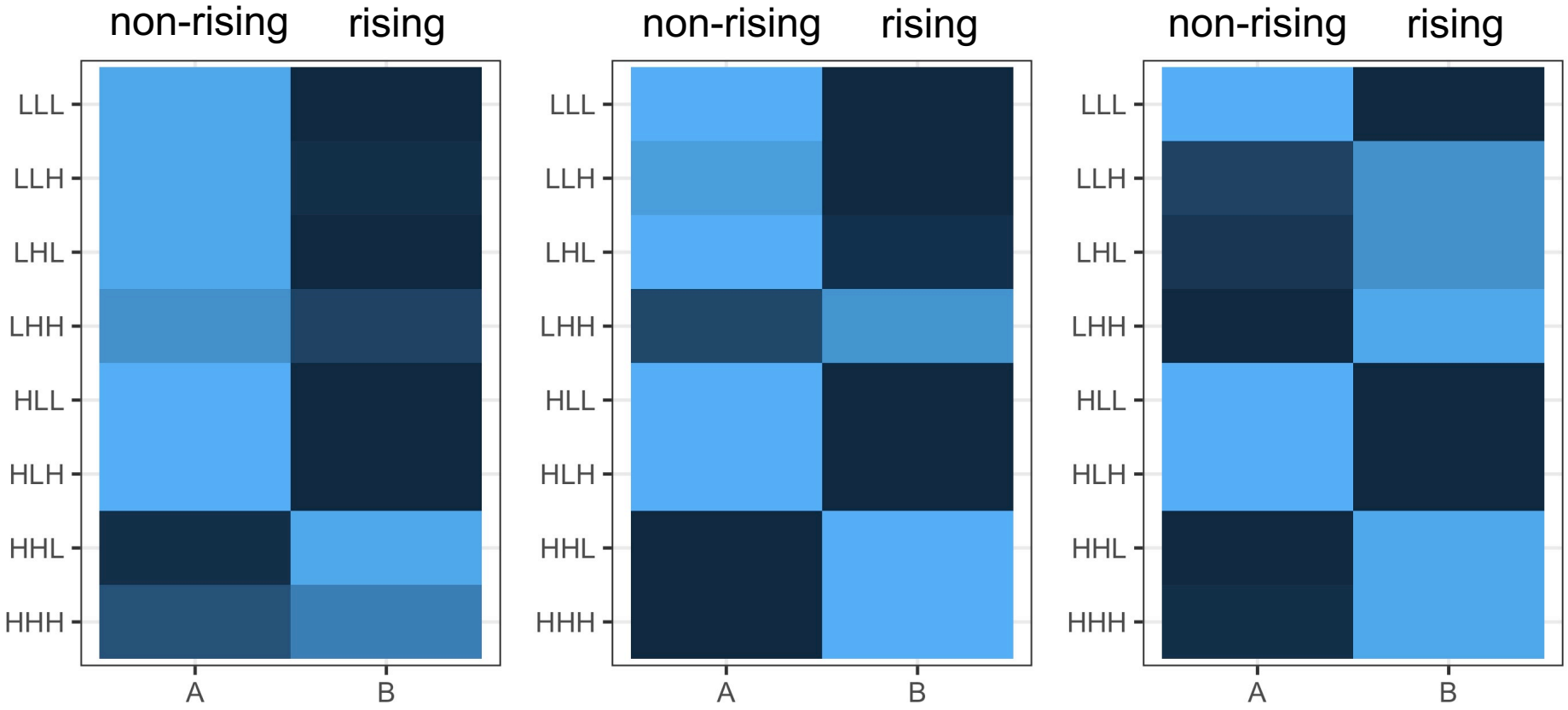




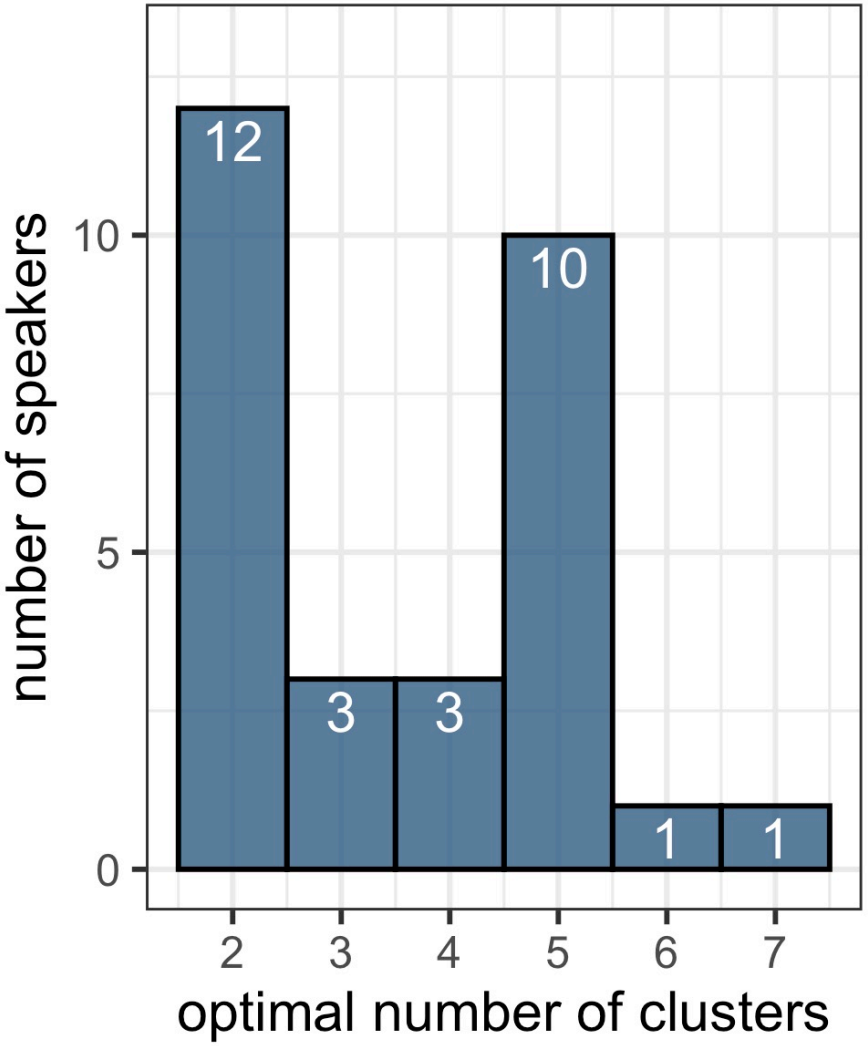


Individual clustering: a prevalent 2 cluster “rising”/ “non-rising” dichotomy





Individual clustering: distribution of optimal clustering solutions



Conclusions

Research question:

Do speakers evidence a robust 8-way distinction in nuclear tune shape, as predicted by the tonal inventory?

Do speakers vary in this regard?

- Group level data: evidence for 5 distinct tune shapes

- Individual level data
  - Lots of variation
  - Many speakers are best characterized as having a **rising**/**non-rising** distinction, with various partitions of the 8 tunes into these categories

Take home message

- Evidence for a **hierarchy of distinctions**
    - all speakers differentiate some tunes from one another, some make additional distinctions
    - not predicted by the current AM theory
- Some *types* of differences are clearly more easily accessible to speakers
    - (in this paradigm, in the absence of context)

Many thanks!

We are grateful to

- The NSF
- The Northwestern ProSD Lab
- Stefanie Shattuck-Hufnagel
- Lisa Cox

References

Chodroff, E., & Cole, J. (2019, September). Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English. In *Proceedings of INTERSPEECH 2019* (pp. 1966-1970). International Speech Communication Association.

Dainora, A. (2009). Modeling intonation in English: A probabilistic approach to phonological competence. In *Laboratory Phonology 8* (pp. 107-132). De Gruyter Mouton.

Genolini, C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. Journal of Statistical Software, 65(4), 1-34. http://www.jstatsoft.org/v65/i04/.

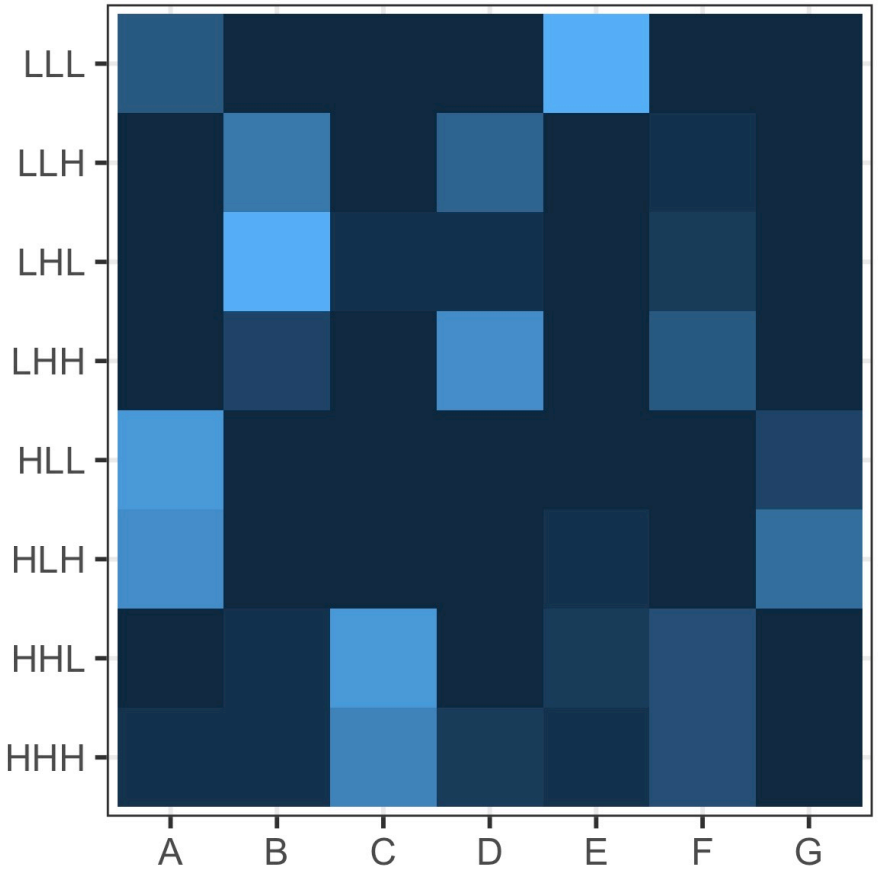
Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T., & Irino, T. (2005). Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Ninth European Conference on Speech Communication and Technology*.

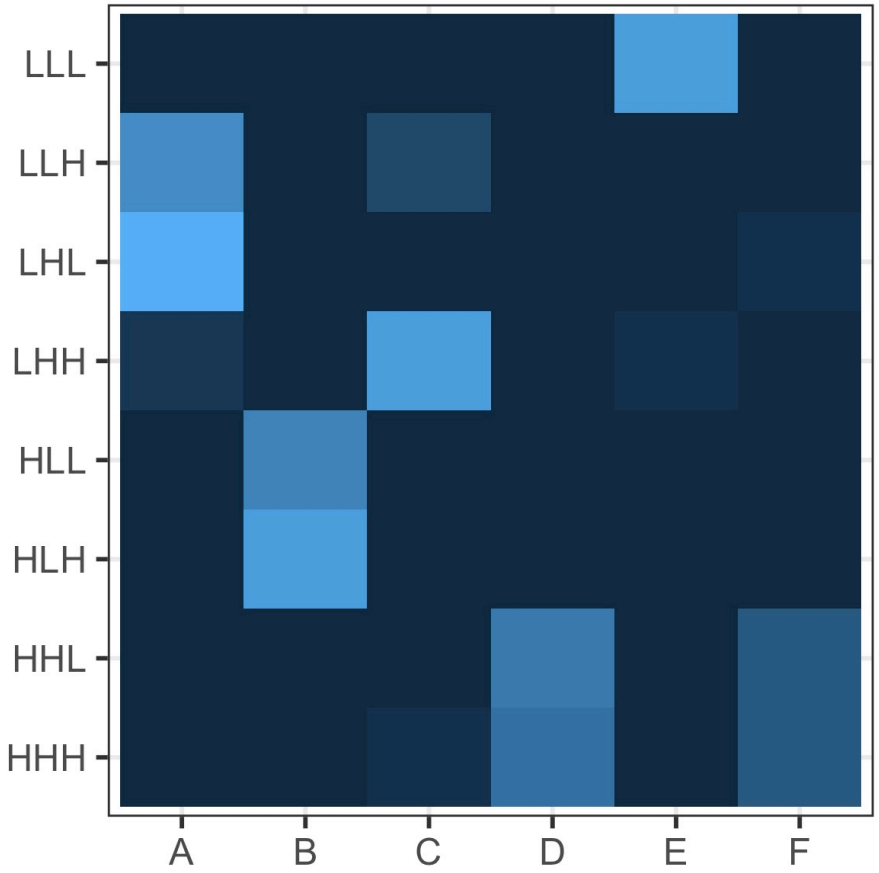
Veilleux, N. Shattuck-Hufnagel, S., and Brugos, A. (2006) *6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI.* January IAP. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* [PhD]. MIT.

Shue, Y.-L. (2010), The voice source in speech production: Data, analysis and models. UCLA dissertation.

Appendix: The 6 and 7 cluster speakers





Appendix: Calinski Harabatz Criterion

- Calinski-Harabasz criterion: the ratio of between cluster dispersion to within cluster dispersion (higher values = better)

$$\frac{dispersion_{between}}{dispersion_{within}} \times \frac{N-k}{k-1} = CH$$

- Where
  - N is the number of vectors (trajectories)
  - k is the number of clusters

Appendix: Calinski Harabasz Criterion

- dispersion is computed with the time series vectors via matrices:

$$\mathrm{B} = \sum_{m=1}^{k} n_m \, (\overline{y_m} - \bar{y})(\overline{y_m} - \bar{y})^T$$

$$\mathrm{W} = \sum_{m=1}^{k} \sum_{l=1}^{n_m} (\, y_{ml} - \overline{y_m})(y_{ml} - \overline{y_m})^T$$

- Where
  - n_m the number of trajectories in cluster m
  - $\overline{y_m}$ is the mean of the trajectories in cluster m
  - $\bar{y}$ is the mean of all the trajectories
  - v^T is the transposition of vector v

Appendix: Calinski Harabasz Criterion

- Finally, the trace of covariance matrix B and W is taken (summing the diagonal coefficients in each matrix)

$$\frac{trace(B)}{trace(W)} \times \frac{N-k}{k-1} = CH$$