Learning acoustic features for English stops with graph-based dimensionality reduction

Patrick F. Reidy,[a] Mary E. Beckman,[b] Jan Edwards,[c] Benjamin Munson,[d] Allison A. Johnson[c]

[a]*Callier Center for Communication Disorders, University of Texas at Dallas* [b]*Dept. Linguistics, The Ohio State University* [c]*Dept. Hearing and Speech Sciences, University of Maryland* [d]*Dept. Speech-Language-Hearing Sciences, University of Minnesota*

Overview

- ▷ The computation of spectral features that cue segmental contrasts is a process of dimensionality reduction. Traditional approaches accomplish this reduction by mapping a high-dimensional observation (e.g., a spectrum) to a small number of pre-determined features (e.g., spectral moments; Forrest et al., 1988). Such approaches fail to exploit the distributional structure of the observations in the high-dimensional space and typically ignore superposing relationships among the observations, such as the word in which the segment occurs.
- ▷ This study adapts the Laplacian Eigenmaps algorithm (Belkin & Niyogi, 2003; Bengio et al., 2003) to learn acoustic features for /t/ versus /k/, consonants that contrast in terms of spectral shape and that differentially exhibit vowel-contextual variation in their spectral shape (see Fig. 1). The algorithm constructs a graph that simultaneously represents the high-dimensional structure of excitation patterns computed from a talker's productions, and aligns lexical correspondences between talkers. A function that embeds the excitation patterns into a two-dimensional feature-space is learned by computing the eigenvectors of the constructed graph.

Speech Production Data

- ▷ 21 adults (10 women, 11 men) completed a picture-prompted word repetition task.
- ▷ Two lists of words were used to elicit a variety of target consonants. Each list contained 32 words in which a target /t/ or /k/ occurred word-initially before a vowel (see footer at bottom for the stop-initial words in the two lists).
- ▷ Participants A50–A65 completed Lists A and B; participants A66–A70, only List A.
- ▷ Training set: List A productions by participants A50–A65 ($N = 493$).
- ▷ Test sets: List B from A50–A65 ($N = 447$); List A from A66–A70 ($N = 156$).
- ▷ Multitaper spectra were estimated from 25-ms windows around stop bursts, and then passed through an auditory (gammatone) filter bank, yielding excitation patterns.

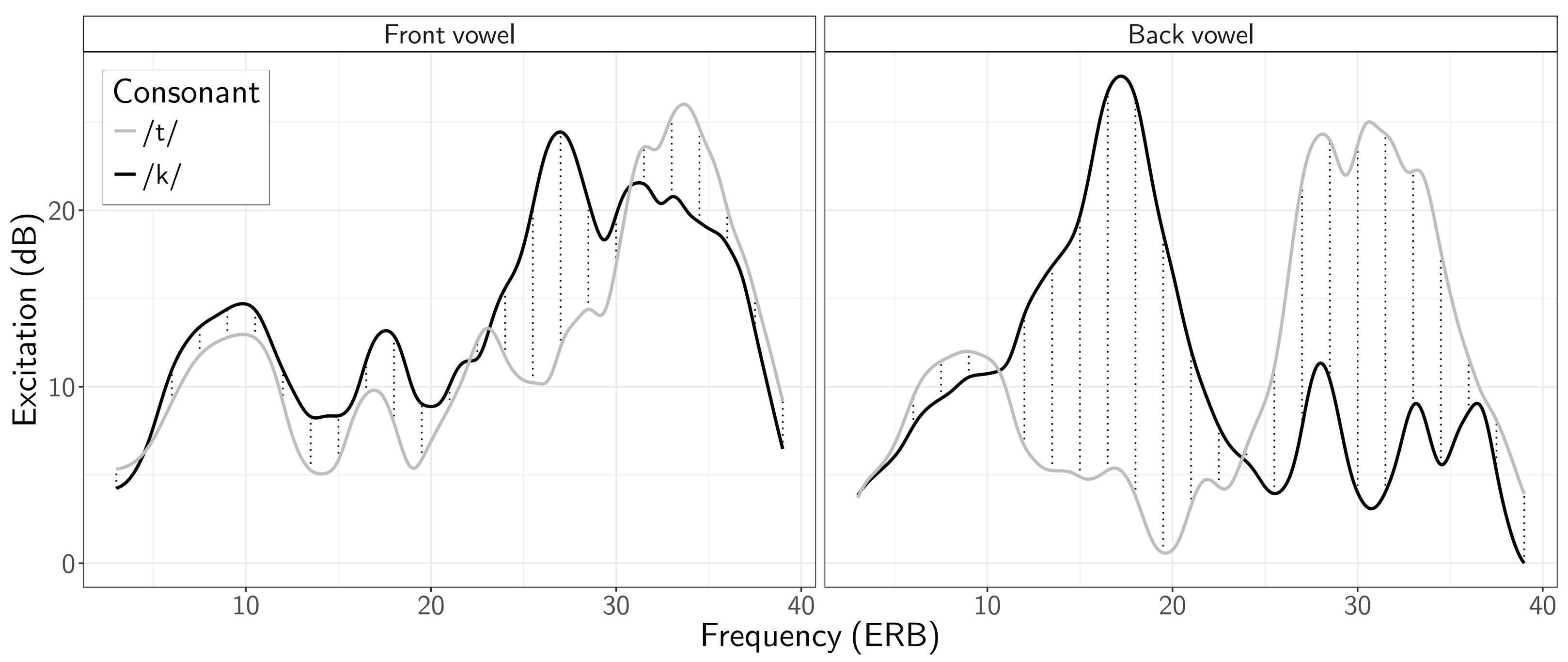


Figure 1: Excitation patterns computed from participant A54's productions of /t/ versus /k/ before the vowels /ɪ/ (left panel) and /oʊ/ (right panel). The dotted lines indicate a subset of the values used to compute the Kullback-Leibler divergence.

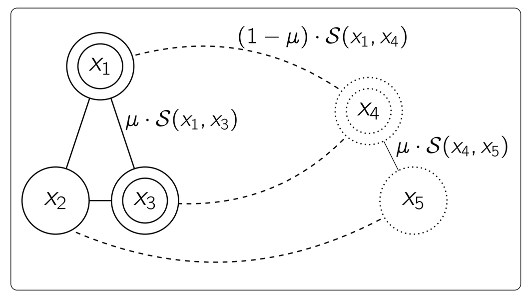
Laplacian Eigenmaps Algorithm

1. Let $X = \{x_1, \ldots, x_n\}$ be the training set of 493 excitation patterns (361-dimensional vectors). Each x_i is pre-processed to sum to 1, so that it may be treated as a probability mass function.
2. Define a similarity function $\mathcal{S}$ on X in terms of Kullback-Leibler divergence $\mathrm{D_{KL}}$ (see Fig. 1).

$$\mathrm{D_{KL}}(x_i \,||\, x_j) = \sum_{f=1}^{361} x_i[f] \ln \frac{x_i[f]}{x_j[f]} \qquad \mathcal{S}(x_i, x_j) = e^{-(\mathrm{D_{KL}}(x_i \,||\, x_j) + \mathrm{D_{KL}}(x_j \,||\, x_i))}$$

3. Define a function $\tilde{\mathcal{W}}$ on X that induces a weighted graph. Nodes correspond to observations in X (see diagram below, where $\{x_1, x_2, x_3\}$ versus $\{x_4, x_5\}$ represent productions by different talkers). Edges connect nodes corresponding either to productions by the same talker (solid lines) or to productions of the same target word by different talkers (dashed lines). Edge weights encode similarity between excitation patterns. Parameter $\mu \in (0, 1)$ adjusts the balance between preserving the structure of each talker's production-space and aligning multiple talkers' production-spaces.

$$\mathcal{W}(x_i, x_j) = \begin{cases} (1-\mu) \cdot \mathcal{S}(x_i, x_j) & \text{if } (\text{word}(x_i) = \text{word}(x_j)) \text{ and } (\text{talker}(x_i) \neq \text{talker}(x_j)) \\ \mu \cdot \mathcal{S}(x_i, x_j) & \text{if talker}(x_i) = \text{talker}(x_j) \\ 0 & \text{otherwise} \end{cases}$$



$$\tilde{\mathcal{W}}(x_i, x_j) = \frac{1}{n} \cdot \frac{\mathcal{W}(x_i, x_j)}{\sqrt{\mathbb{E}_x\left(\mathcal{W}(x_i, x)\right) \cdot \mathbb{E}_x\left(\mathcal{W}(x_j, x)\right)}}$$

4. Construct the graph's weighted adjacency matrix A, degree matrix D, and Laplacian matrix L.

$$A_{i,j} = \tilde{\mathcal{W}}(x_i, x_j) \qquad D_{i,i} = \textstyle\sum_j A_{i,j} \qquad L = D - A$$

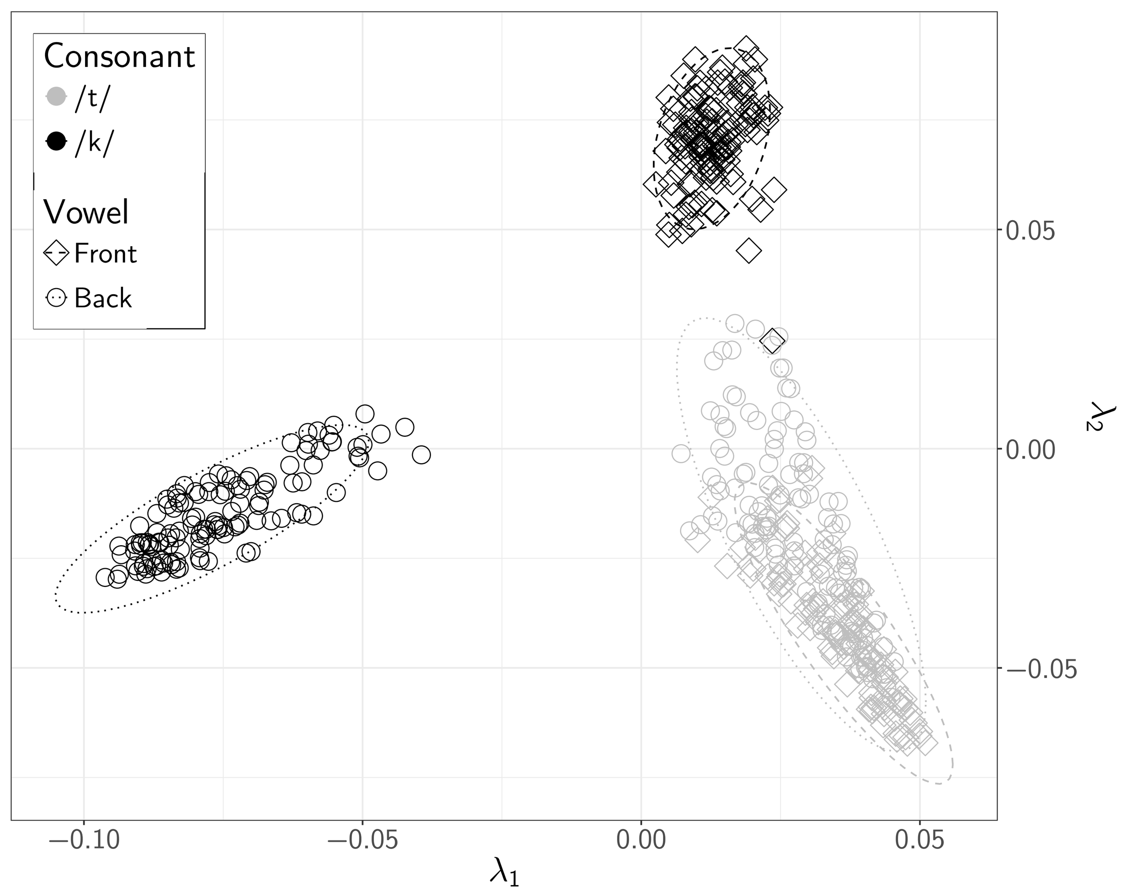
5. Solve the generalized eigenvalue problem $L\gamma = \lambda D\gamma$. The eigenvectors λ_1, λ_2 that correspond to the two least, non-zero eigenvalues embed X into 2-dimensional space: $x_i \mapsto \langle \lambda_1[i], \lambda_2[i] \rangle$.
6. Extend eigenvectors λ_1, λ_2 to eigenfunctions $\tilde{\lambda}_1$, $\tilde{\lambda}_2$. The projection $\tilde{\lambda}_k(x')$ of a test data point $x' \notin X$ onto dimension k of the embedding is a linear combination of the components of λ_k.

$$\tilde{\lambda}_k(x') = \sum_{i=1}^{n} \lambda_k[i] \cdot \tilde{\mathcal{W}}(x', x_i)$$

Eigenvector Embedding of Training Data (A50–A65, List A)

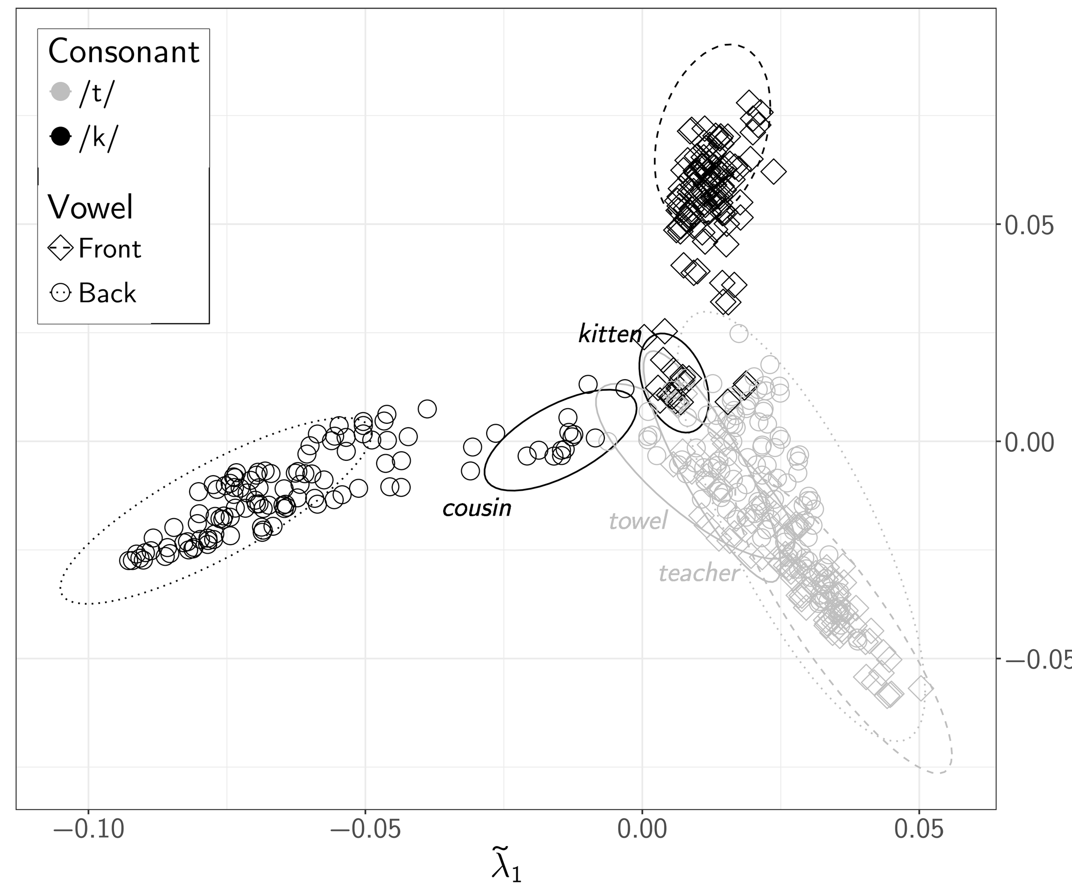
Figure 2: The image of the training data under the embedding given by Laplacian eigenvectors, λ_1 and λ_2, learned with $\mu = 1/6$. The ellipses denote 95% confidence regions, assuming a bivariate t-distribution.

(The front-vowel /k/ data point that overlaps with the /t/ productions was determined *post-hoc* to be a misarticulation.)



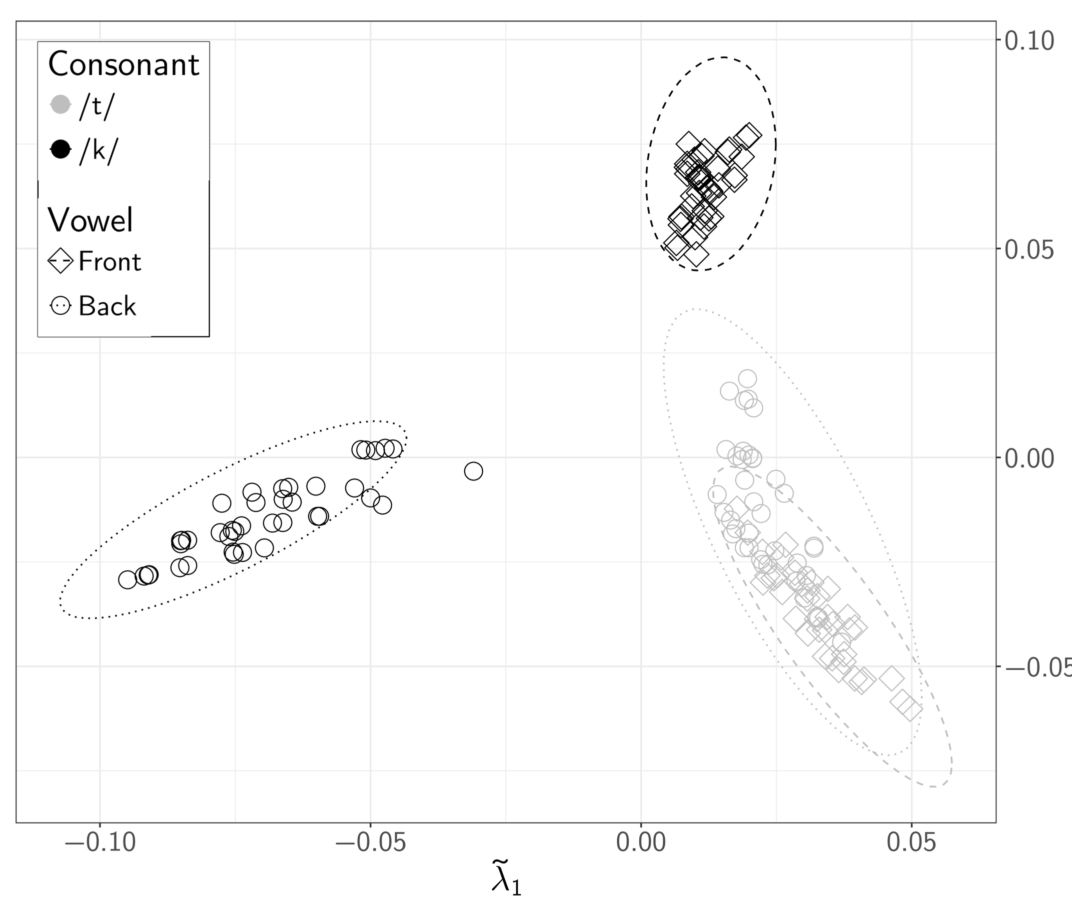
Eigenfunction Embedding of Test Set I (A50–A65, List B)

Figure 3: The image of the test data under the Laplacian eigenfunctions, $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$. The ellipses denote 95% confidence regions (bivariate t). The dashed and dotted ellipses were estimated from the training data. The solid ellipses were estimated from test productions of the 4 words that were not represented in the training data.



Eigenfunction Embedding of Test Set II (A66–A70, List A)

Figure 4: The image of the test data under the Laplacian eigenfunctions, $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$. The ellipses denote 95% confidence regions (bivariate t), estimated from the training data. Here, the embedding of the test data was guided solely by lexical information; comparison with Fig. 3 underscores the importance of lexical alignment across talkers.



Discussion and Future Directions

- ▷ The two-dimensional embedding that is learned by Laplacian Eigenmaps reflects well-established articulatory constriction features: λ_1 distinguishes /t/ versus back-vowel /k/, reflecting place of constriction (anterior versus posterior); λ_2 distinguishes /t/ versus front-vowel /k/, reflecting tongue shape (apical versus domed).
- ▷ We plan to extend this method to develop dynamic spectral features that model the transition from a stop burst to a vowel (see Nossair & Zahorian, 1991).

Word List A: tickle, tent (2x), teddy bear (2x), table, take, *tape, tooth,* toothbrush, toast, toaster, tongue (2x), tummy, tiger, *kitty, kitchen,* keys, cake, cat, candle, candy, *catch,* cookie (2x), comb, coat, coffee, *car,* cup, cutting. (Italicized words are unique to Word List A.)

Word List B: **teacher**, tickle, tent, teddy bear (2x), table, take, toothbrush, toast, toaster (2x), tongue, tummy (2x), tiger, **towel**, **kitten** (2x), keys, cake, cat, candle, candy (2x), cookie (2x), comb, coat, coffee, cup, cutting, **cousin**. (Boldface words are unique to Word List B.)

M. Belkin and P. Niyogi. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396.
Y. Bengio et al. (2003). Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering. *Advances in NIPS 16*, 177–184.
K. Forrest et al. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *JASA*, 84(1):115–123.
Z. B. Nossair and S. A. Zahorian. (1991). Dynamic spectral shape features as acoustic correlates for initial stop consonants. *JASA*, 89(6):2978–2991.

The research presented here was supported by NIDCD grant R01-02932 to M. E. Beckman, J. Edwards, and B. Munson and by a Callier Postdoctoral Fellowship to P. F. Reidy. We thank and acknowledge Alisha Blackman, for recruiting and testing subjects, and Mia Kim, for helping to annotate the speech production data.

Download poster at: https://patrickreidy.org/blog/asa-173
Correspondence: reidy@utdallas.edu