**A graph theoretic approach for generating hypotheses about phonetic cues in speech**

Human speech is highly variable, and as a result, the speech signal does not contain invariant cues for identifying phonemes. Instead, models of speech perception suggest that combining multiple cues and factoring out contextual variability may be the best way to optimize classifier performance (McMurray and Jongman, 2011). How do we identify what the relevant cues are? Previous work has generally relied on phonetic analyses and perceptual data to identify candidate cues. Here, we employ tools from graph theory, providing a novel method for building perceptual models.

We investigate perception of fricatives (/f,v,θ,ð,s,z,ʃ,ʒ/), drawing on previous work (Jongman et al., 2000; McMurray and Jongman, 2011) that has identified 24 cue dimensions for classifying these sounds. Using the dataset from McMurray and Jongman (2011), we create graphs (networks) that depict the likelihood of individual talkers producing each phoneme using each of these cues, along with the likelihoods that cues co-occur in individual tokens. Our goal is to identify subgraphs that connect all of the talkers using minimal weights/maximal likelihoods, creating a model that depicts the essential cues needed to correctly classify fricatives produced by the entire set of talkers.

**Method.** Graphs consist of nodes and weighted edges, with separate graphs for each phoneme. Nodes represent individual talkers (N=20) and cues. An edge weight between a talker and cue is defined as the inverse probability that the talker produced a given phoneme with that cue; edge weights between cues are the inverse probability that tokens contain both cues. To create discrete nodes for cues, we converted the 24 cue dimensions in the McMurray and Jongman dataset into 48 possible cues by first factoring out contextual differences (using the C-CuRE approach; McMurray and Jongman, 2011) and then z-scoring the resulting cue-values. Each token was then classified high or low along each cue dimension (e.g., an F1 value was coded as either F1LOW or F1HIGH).

From these graphs, we identified subgraphs that fully connected each of the 20 talkers with minimal edge weights (Fig. a); edges with weights $\geq$5 (corresponding to probabilities of $\leq$20%) were not included. The resulting subgraphs are known as Steiner trees, and this method has been used recently in other fields for similar types of problems (e.g., identifying gene-protein networks in biology; Bailly-Bechet et al., 2011). The search algorithm was run 100 times for each fricative; a cue was included in analyses if $\geq$5% of the simulation runs included that cue. We then trained multinomial regression classifiers, evaluating classifier accuracy using tokens held out from training and comparing performance to listeners' responses. We also compared the Steiner tree classifiers to those trained on the entire set of cues to determine if the algorithm successfully identified the subset of cues needed without sacrificing accuracy or goodness-of-fit.

**Results.** We first evaluate the overall classifier performance when trained on all 48 cues, providing an upper bound on model performance (Fig. b). Overall, this model had a mean accuracy of 87.3%. Note that this is slightly lower than the 92.9% accuracy reported by McMurray and Jongman (2011) for classifiers trained on continuous cue-values—the difference reflects the cost of coding the data in terms of high-low values only. We also evaluated this model's goodness-of-fit to the training dataset using the Bayesian Information Criterion (BIC) and compared the pattern of model responses to those of human listeners by measuring the likelihood of the model given the listener data. The overall BIC for the model was 3747.7 and the log-likelihood was -4050.4.

We then performed the same analyses for classifiers trained only using the Steiner tree cues. The number of cues identified for each fricative varied: /f/ has 6 relevant cues, whereas /ð/ has 13. This reflects the variability in which cues are used by individual talkers, and the overall reliability of cues for identifying specific phonemes. Overall, there were 31 cues (out of 48 possible) identified across the set of fricatives. The classifier trained on these cues had a mean accuracy of

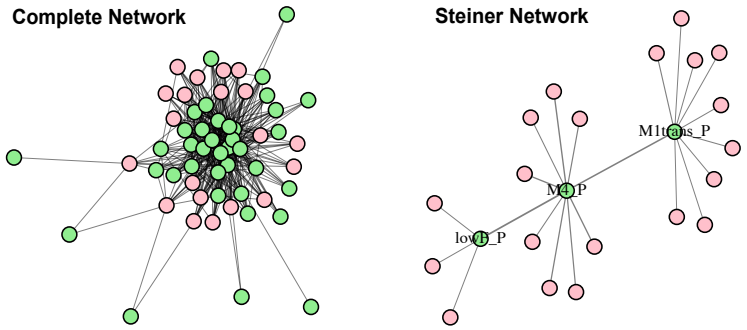


*Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 211-212.
Salt Lake City, Utah, January 4-7, 2018

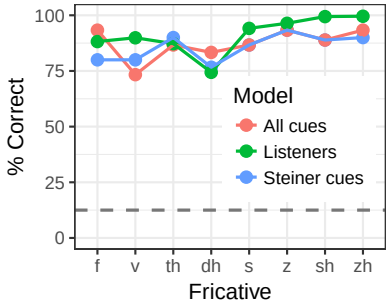
85.7%, slightly lower, but very close to the performance for the entire set of cues. The Steiner tree model, however, had a lower BIC (3671.7) and higher log-likelihood (-3780.5), both indicating better fits to the data. Thus, the Steiner tree model provided a better match to listeners' responses.

Examining the Steiner trees reveals which cues are useful for identifying specific phonemes. In the figure below, the Steiner tree for /z/ reveals three cues with clusters of talkers surrounding them; several talkers use a high spectral mean at the vowel transition to indicate this sound (M1TRANS P), and a small group use voicing during frication (LOWF P). This information allows us to make predictions about how /z/ tokens will be classified by listeners. Indeed, the Steiner tree could represent listeners' knowledge of the way this group of talkers produces /z/ sounds. Combined with the subgraphs for the other phonemes, this allows us to generate hypotheses about the cues listeners use to categorize speech sounds.

**Conclusions.** Graph algorithms that find Steiner trees provide a novel method of identifying phonetic cues used by talkers to reliably indicate specific phonemes. This, in turn, allows us to build computational models and classifiers that use these specific cues to recognize speech. Reducing the space of potential cues to just those identified in Steiner trees has only a small impact on performance and produces results that are closer to listeners' performance. The method also enables easy identification of cues in new speech sound tokens because we do not need to know specific cue-values; by examining a spectrogram and waveform, we can tell whether values are high or low along each cue dimension, which in turn allows us to classify cues as either present or absent in a given speech sound. In sum, this technique provides a useful tool that draws out highly informative data from acoustic measurements of speech sounds that can be used to understand the cues used by listeners to accurately recognize speech.




(a) Example graphs for /z/. Pink dots are talker nodes; green dots are cues. Left: complete graph. Right: Subgraph identifying three cues needed to account for all talkers.




(b) Model and listener accuracy as a function of phoneme. Dashed line is chance performance.

## References.


Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., and Zecchina, R. (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proc Nat Acad Sci*, 108:882–887.

Jongman, A., Wayland, R., and Wong, S. (2000). Acoustic characteristics of English fricatives. *J Acoust Soc Am*, 108:1252–63.

McMurray, B. and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychol Rev*, 118:219–46.