Psychonomic Bulletin and Review
https://doi.org/10.3758/s13423-020-01748-1

THEORETICAL REVIEW

A graph-theoretic approach to identifying acoustic cues for speech sound categorization

Anne Marie Crinnion[1] · **Beth Malmskog**[2,3] · **Joseph C. Toscano**[4]

Abstract

Human speech contains a wide variety of acoustic cues that listeners must map onto distinct phoneme categories. The large amount of information contained in these cues contributes to listeners' remarkable ability to accurately recognize speech across a variety of contexts. However, these cues vary across talkers, both in terms of how specific cue values map onto different phonemes and in terms of which cues individual talkers use most consistently to signal specific phonological contrasts. This creates a challenge for models that aim to characterize the information used to recognize speech. How do we balance the need to account for variability in speech sounds across a wide range of talkers with the need to avoid overspecifying which acoustic cues describe the mapping from speech sounds onto phonological distinctions? We present an approach using tools from graph theory that addresses this issue by creating networks describing connections between individual talkers and acoustic cues and by identifying subgraphs within these networks. This allows us to reduce the space of possible acoustic cues that signal a given phoneme to a subset that still accounts for variability across talkers, simplifying the model and providing insights into which cues are most relevant for specific phonemes. Classifiers trained on the subset of cue dimensions identified in the subgraphs provide fits to listeners' categorization that are similar to those obtained for classifiers trained on all cue dimensions, demonstrating that the subgraphs capture the cues necessary to categorize speech sounds.

Keywords Speech perception · Language processing · Computational modeling · Graph methods

Human speech contains a wide variety of acoustic cues. A classic problem in speech perception concerns the fact that these cues—measurable spectral, temporal, and amplitude differences in the speech signal—do not map onto phoneme categories in a one-to-one way. Despite an extensive search for invariant features, researchers have been unable to identify individual cues that are used by all talkers in all contexts to signal specific phonemes. Rather, speech sounds are highly context dependent (Liberman et al., 1967), an issue that is compounded by variability between talkers and differences between individual utterances produced by the same talker (Peterson & Barney, 1952; Hillenbrand et al., 1995; Kleinschmidt & Jaeger, 2015; Toscano & Allen, 2014).

Despite the inherent complexity of this problem, many human listeners have a relatively easy time identifying speech sounds. One proposed mechanism that allows for this is the fact that there are multiple, redundant cues for specific phonological distinctions (Lisker, 1986; Jongman et al., 2000; Hillenbrand et al., 1995). Thus, listeners may combine information from multiple cues to achieve accurate speech recognition. Indeed, perceptual experiments have demonstrated that listeners do this (Miller & Liberman, 1979; Repp, 1982). Moreover, a number of models use *cue-integration* as a key principle for describing how

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13423-020-01748-1) contains supplementary material, which is available to authorized users.

✉ Anne Marie Crinnion
anne.crinnion@uconn.edu

[1] Department of Psychology, Harvard University, 945 Memorial Drive, Cambridge, MA 02138, USA

[2] Department of Mathematics and Computer Science, Colorado College, Colorado Springs, CO, USA

[3] Department of Mathematics & Statistics, Villanova University, Villanova, PA, USA

[4] Department of Psychological & Brain Sciences, Villanova University, Villanova, PA, USA

Published online: 15 July 2020

listeners combine sets of acoustic cues in order to classify speech sounds (Bejjanki et al., 2011; Oden & Massaro, 1978; Nearey, 1997; Smits, 2001; Toscano & McMurray, 2010), including recent approaches that take into account known sources of contextual variability (Cole et al., 2010; McMurray & Jongman, 2011; McMurray et al., 2011; Kleinschmidt, 2019).

However, the complexity of cue-integration approaches makes the problem of describing the relevant cue dimensions for specific speech sounds that much more difficult, as any statistically informative cue might be included in these models. There may be ten (for vowels Hillenbrand et al., 1995) or 20 (for fricatives McMurray & Jongman, 2011) different acoustic dimensions that are informative for a given class of speech sounds. However, it is unclear whether all statistically informative acoustic dimensions are necessary for categorization or whether we can achieve similar performance with a subset of cues that still capture the relevant distinctions between specific phonemes.

Previous work has largely relied on phonetic analyses of single cues (or very small sets of cues), perceptual experiments (in which sounds are systematically varied along specific cue dimensions), or predictions from phonological theory to determine how acoustic differences are mapped onto phoneme categories. These studies have primarily approached this problem by seeking out acoustic cue dimensions that are informative for distinctive features (e.g., voice onset time [VOT] as a cue for the feature *voicing*). While this is useful, it does not address the questions posed above. In addition, because speech contains multiple cues, large sets of cues are often found to be statistically significant for one or more feature dimensions (McMurray & Jongman, 2011), making it difficult to identify a subset of cues that can still provide enough information to account for differences across talkers.

Here, we address the problem of identifying specific acoustic dimensions that are informative for specific phonemes, and assess whether specific subsets of cues contain information for classifying speech sounds across different talkers. To accomplish this, we introduce a new computational technique from graph theory (Scott et al., 2005; Ideker et al., 2002; Bailly-Bechet et al., 2011) that aims to reduce a large set of possible cue dimensions (with potentially large variability across talkers) to a subset that is still informative, providing a simpler model and reduced set of acoustic cues. Current models of speech perception have addressed similar questions; however, they typically focus on only individual cues or small sets of them (Kleinschmidt & Jaeger, 2015; Toscano & McMurray, 2010), or they use many cues but measure average performance across talkers (McMurray & Jongman, 2011). Although these models provide key insights into the information available to listeners for speech categorization, the relationship between acoustic cues and specific phonemes has not been fully evaluated.

To study these issues, we create networks (graphs) that connect cue dimensions and individual talkers based on phonetic measurements of their speech (in this case, based on production of fricative sounds; Jongman et al., 2000; McMurray & Jongman, 2011). We then identify subgraphs of these fully connected graphs that include a reduced set of cues for that specific phoneme, while still connecting all talkers in the graph (i.e., ensuring that sufficient information is captured to account for variability across the talkers). We then evaluate classifiers trained on the set of cues identified across all phonemes, comparing performance to classifiers trained on all cue dimensions.

In the next section, we briefly review the acoustic characteristics of our target set of speech sounds—English fricatives (Jongman et al., 2000)—as well as previous work that has developed models of speech sound categorization that incorporate multiple cues and address the problem of talker variability. We then present the methods used to create the talker-cue graphs, and the results of this procedure for a large dataset of speech sounds, specifically the dataset of fricative sounds presented by McMurray and Jongman (2011). We focus on this well-studied dataset with well-defined cues in order to provide a useful starting point for evaluating this approach.

Acoustic characteristics of fricatives

The current study focuses on fricatives, a class of consonants produced via partial constrictions in the vocal tract. In English, there are eight fricatives (/f,v,θ,ð,s,z,ʃ,ʒ/) that vary in place of articulation (i.e., where in the vocal tract the constriction is made; e.g., /f/ vs. /θ/ vs. /s/ vs. /ʃ/ and voicing (whether or not the sound is produced coincident with periodic vibration of the glottis; e.g., /ʃ/ vs. /ʒ/). We can further divide place of articulation into sibilants (/s,z,ʃ,ʒ/) and non-sibilants (/f,v,θ,ð/). Overall, fricatives represent one of the largest classes of speech sounds in English.

Previous work has identified 24 acoustic feature *dimensions* that signal differences between these sounds. Fourteen of these cue dimensions were studied by Jongman et al. (2000) and an additional ten were studied by McMurray and Jongman (2011). These cues include temporal differences (e.g., duration of frication is lower for /v/ than for /θ/), static spectral differences (e.g., spectral mean during frication is lower for /ʃ/ than for /s/, dynamic spectral changes (e.g., formant transitions; differences in F2 onset for /ʃ/ vs. /s/), and amplitude differences (e.g., amplitude of frication is lower for nonsibilants than for sibilants; see Fig. 6 for examples of some of these cues). Many of these cues are also affected by context, including coarticulatory effects and

variability between talkers (Jongman et al., 2000). Indeed, compensating for this contextual variability yields improvements in fricative classification (McMurray & Jongman, 2011); thus, effects of contextual variability are considered in the current study as well.

Details of the acoustic characteristics of fricatives can be found in Jongman et al. (2000) and McMurray and Jongman (2011), and a summary of the cues is provided in Table 1.[1] The results of these studies suggest that there are a large number of cues available to listeners for identifying fricative voicing, place, sibilance, or some combination of these features. In addition, none of these cues provides sufficient information on its own and many are context dependent (i.e., varying across talkers and coarticulatory contexts). Thus, in order to accurately recognize fricatives, listeners must overcome the variability across talkers that is present in individual cues. This provides an ideal test case for the current study.

Table 1 Description of cue dimensions

Cue	Description
DUR_F	Duration of frication
DUR_V	Duration of vowel
F0	Fundamental frequency at vowel onset
F1	First formant frequency at vowel onset
F2	Second formant frequency at vowel onset
F3	Third formant frequency at vowel onset
F4	Fourth formant frequency at vowel onset
F5	Fifth formant frequency at vowel onset
M1	Spectral center of gravity during frication
M2	Spectral variance during frication
M3	Spectral skew during frication
M4	Spectral kurtosis during frication
M1TRANS	Spectral center of gravity at vowel onset
M2TRANS	Spectral variance at vowel onset
M3TRANS	Spectral skew at vowel onset
M4TRANS	Spectral kurtosis at vowel onset
MAXPF	Max peak frequency during frication
$F3AMP_F$	Third formant amplitude during frication
$F3AMP_V$	Third formant amplitude at vowel onset
$F5AMP_F$	Fifth formant amplitude during frication
$F5AMP_V$	Fifth formant amplitude at vowel onset
LOW_F	Amplitude for frequencies below 500 Hz during frication
RMS_F	Root-mean-square amplitude during frication
RMS_V	Root-mean-square amplitude at vowel onset

Models of speech categorization

As noted previously, one way to mitigate talker-level variability in speech is by combining information from multiple cues via a process of *cue-integration*. A number of models of speech categorization have used this approach, and it has been applied to other questions in cognitive science as well, for example, in ideal observer models of depth perception (Ernst & Banks, 2002; Jacobs, 2002). One of the earliest and most influential cue-integration models is Oden & Massaro's (1978) fuzzy logical model of perception (FLMP). FLMP proposes that listeners first extract individual acoustic cues from the speech signal, match cue values to prototype representations of phonemes, and finally compute the probability that a given speech sound is a member of a phoneme category. Oden & Massaro demonstrate that by combining information from two cues and by using the same cue to identify different phonological features, the model can more accurately match listeners' identification responses. Several other models have proposed similar cue-integration principles (Nearey, 1990; Smits, 2001; Toscano & McMurray, 2010). In each of these models, information from multiple sources is combined to arrive at an overall percept (and often the same cue can be used for different types of phonological feature distinctions).

Generally, these approaches have focused on only a small set of cues (two or three), and they may or may not account for contextual variability in a principled way (e.g., they may simply pool information from multiple cues rather than trying to systematically account for contextual differences). McMurray and Jongman (2011) addressed these issues using a dataset of fricatives with a large number of measured cues, based on the initial set reported by Jongman et al. (2000). Listeners heard CVC syllables in which the initial consonant was one of the eight English fricatives (/f,v,θ,ð,s,z,ʃ,ʒ/) and the final consonant was a /p/. The sounds were produced in six different vowel contexts by 20 different talkers. The results demonstrated that listeners' performance depended on contextual factors—talker and vowel identity. Listeners also varied in their accuracy at identifying the fricatives: accuracy was 91% when they heard the complete syllable, but 76% when they only heard the frication portion of the syllable (chance was 12.5%). Listener accuracy also varied depending on the phoneme, with the highest accuracy for /ʒ/ (99.6%) and the lowest for /ð/ (74.4%).

Listeners' performance on this task was compared with predictions from three types of models, evaluated using multinomial regression classifiers: (1) a naïve invariance model, which used a small number of cues expected to be informative from previous studies; (2) a cue-integration model, which used all 24 cue dimensions from the dataset;

[1] Note that the labels for the cue dimensions used in the current study are the same as those used by McMurray and Jongman (2011).

and (3) the Computing Cues Relative to Expectations (C-CuRE) model. In C-CuRE, cues are encoded relative to the talker and vowel context in order to eliminate variation caused by these factors. Specifically, the model uses linear regression to predict the cue value relative to the context, meaning talker and vowel effects are first partialed out and the residual cue values are used to classify speech sounds into phoneme categories. Note that this requires the listener to know the acoustic properties of the context (e.g., the talker's mean cue values); how listeners learn this is a question that remains open and is being examined in work on perceptual adaptation (Eisner & McQueen, 2005; Kleinschmidt & Jaeger, 2015; Kraljic & Samuel, 2005).

The classifiers were trained using the 24 acoustic cue dimensions measured from the fricatives. Overall, they found that no single cue was completely robust for just one feature (e.g., there was no cue that uniquely provided information for place of articulation). In addition, C-CuRE matched listeners' responses more closely than the other models, and the cue-integration model performed better than naïve invariance. These results fit with previous work showing that listeners shift their perceptual responses as a function of both coarticulatory context (Mann & Repp, 1980) and indexical differences between talkers (Niedzielski, 1999; Strand & Johnson, 1996), suggesting that they compensate for contextual variability, as C-CuRE does.

Other types of models, such as exemplar models (Goldinger, 1998; Johnson, 1997) have also demonstrated the importance of accounting for talker-level variability. Recently, Kleinschmidt and Jaeger (2015) presented a computational approach, the ideal adapter framework, in which speech categorization is viewed as a problem of making inferences under uncertainty (e.g., uncertainty about the acoustic cue distributions of a novel talker). They propose three general principles that guide a listener in categorizing speech sounds: (1) recognize the familiar (e.g., identify a specific talker who was encountered previously), (2) generalize to the similar (e.g., another talker with similar acoustic characteristics to those encountered previously), and (3) adapt to the novel (e.g., a novel accent, with acoustic cue distributions shifted in a systematic way). This framework can explain how listeners adapt to shifts along specific acoustic cue dimensions and how they can learn talker-specific mappings between acoustic cues and categories. Moreover, the ideal adapter model can be used to predict indexical information about talkers from their acoustic cues (e.g., age, sex, and dialect; Kleinschmidt et al., 2018) and to quantify the extent to which talker-level factors influence cues to speech perception (Kleinschmidt, 2019).

In sum, models of speech categorization demonstrate a clear need to integrate multiple acoustic cues and handle talker-level variability in the mapping between cues and categories. Thus, it is likely that, for any given set of speech sounds, there will be dozens of acoustic cues (or possibly more) that provide statistically meaningful differences between speech sounds. In addition, a number of models demonstrate the importance of accounting for talker-level variability, including exemplar models (Goldinger, 1998), context compensation models (McMurray & Jongman, 2011), and the ideal adapter framework (Kleinschmidt, 2019). However, it remains unclear whether all statistically informative cue dimensions are needed to account for the acoustic variability across talkers. A reduced set of cues, defined in a principled way, may allow us to simplify the information needed in models of speech categorization, while still accounting for talker variability.

An approach from graph theory

Techniques from graph theory may offer a way to answer these questions and provide a tool for researchers to identify subsets of relevant acoustic cues. The use of graphs, in the form of connectionist or neural networks, is familiar to many cognitive scientists, and network-based models have played a significant role in speech perception and language processing (McClelland & Elman, 1986; Elman, 1990; McClelland & Rumelhart, 1981; Magnuson et al., 2007; Dell, 1988). In connectionist models, neurons (referred to as *nodes* or *vertices* in graph theory) are connected via weights (*edges*) of varying strengths.[2] The types of connectionist networks typically used to model cognitive phenomena, however, represent only a subset of possible graphs (e.g., a perceptron is a bipartite directed graph). Moreover, there is evidence that broader classes of graphs may characterize neural information processing. Work from Bullmore and Sporns (2009), for example, emphasizes that graph-theoretic analyses provide a basis for describing networks in the brain. Graphs also form the basis of approaches to modeling other psycholinguistic processes, such as the relationships between concepts in semantic maps (Haspelmath, 2003).

Many other cognitive science applications of graphs make use of their ability to elegantly capture information about connections and relationships, illuminating essential structures within data. Algorithms devised to identify important structures within large graphs can be used to find strong connections and structures within the underlying phenomenon, as in the case of identifying social networks using data from disease outbreaks (Angluin et al., 2010) and

[2]Note that in connectionist models, higher weights indicate stronger connections between nodes. In other work using graph theory, however, edges are often characterized as representing a *cost* of connecting two nodes. The edge weights in the current study use the latter approach (i.e., higher edge weights represent higher costs).

inferring semantic maps from cross-linguistic data (Regier et al., 2013).

In the current study, we use algorithms to identify *Steiner Trees*, subgraphs within larger networks, to illuminate important connections between talkers and acoustic cue dimensions signaling specific phonemes. In order to develop this approach further, we first define the general form of a graph and provide some examples.

Let $G = (N, E)$ be an edge-weighted graph, with node set N and weighted edge set E. Elements of E are of the form $e = (\{n_1, n_2\}, w_e)$, where n_1 and n_2 are nodes connected by the edge e and w_e is the weight of the edge. We say two edges are *incident* if they share a node. Thus, graphs are topological constructions that contain information about how some set of entities (the nodes) are connected, while not specifying geometric relationships between the entities. Edge-weighted graphs carry the additional information of a numeric value associated to each edge, which can be used to model the importance or cost of the connection (similar to weights in a connectionist network).

One well-studied problem—and one directly related to the goals of the present study—is the *(graph) Steiner tree problem*. If G is a graph, a *subgraph* of G is a graph H with the properties that the node set of H is a subset of the node set of G, and the edge set of H is a subset of the edge set of G. Given an edge-weighted graph G and a subset S of the node set N, a *Steiner tree* T for S in G is a minimal-weight subgraph of G such that all nodes in S are connected by some sequence of edges in T (i.e., a subgraph whose total edge weight is minimal while meeting certain constraints; see Fig. 1). The nodes in S are called the *distinguished nodes*. Nodes included in T but not in S

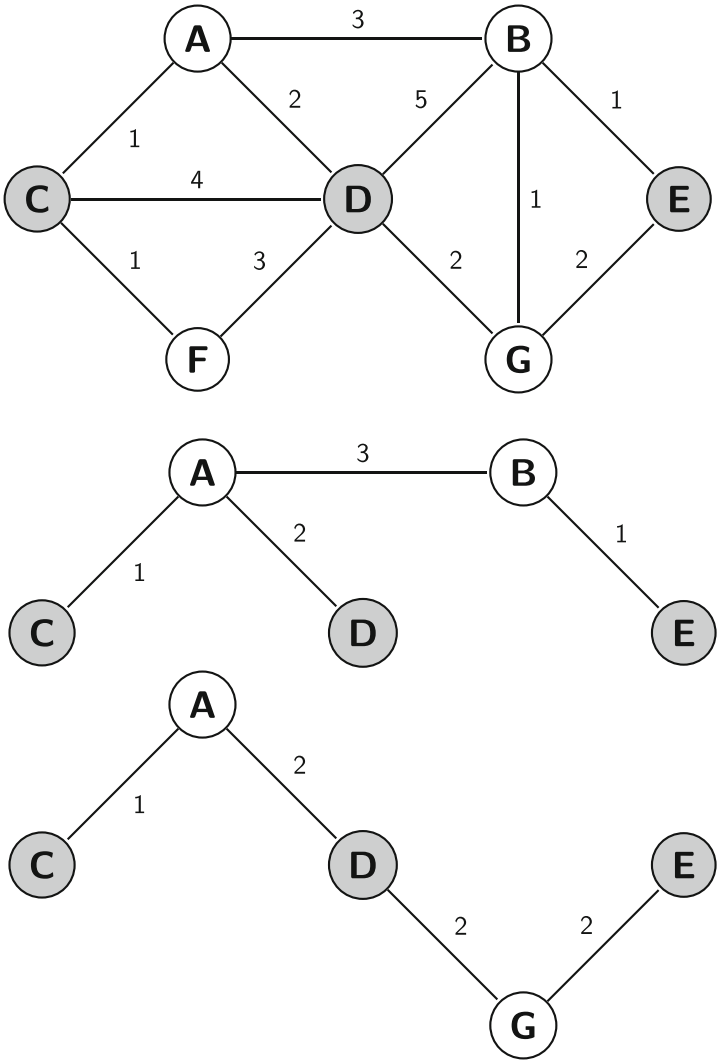

Fig. 1 At the *top*, an example edge-weighted graph. In the *middle*, a Steiner tree for the graph and distinguished nodes $S = \{C, D, E\}$. Steiner nodes are A and B. At the *bottom*, another Steiner tree for the graph and the same distinguished set, with Steiner nodes A and G. Both Steiner trees have a total edge weight of 7. There exists at least one additional Steiner tree for this edge-weighted graph and distinguished set

are called *Steiner nodes*. In general, a *tree* is a graph in which any two nodes are connected by some sequence of edges, and which contains no cycles (i.e., there is a unique sequence of edges connecting a node to any other node). Therefore, a graph with k nodes is a tree if and only if it has $k - 1$ edges and no cycles.

The Steiner tree problem and its variants have been very well studied in mathematics and computer science. The problem is generally computationally expensive to solve. The graph version is NP-hard (Garey et al., 1977), though some special cases of the problem can be solved exactly in polynomial time. Approximation algorithms exist, but finding a provably extremely close approximation is also NP-hard (Chlebík & Chlebíková, 2008).

Despite the general complexity of the problem, advances in algorithms and computational power have enabled researchers to solve or approximate the Steiner tree problem in increasingly large networks. The last 15–20 years have seen an explosion of Steiner trees and other graphical methods in the biological sciences. For example, Steiner tree methods have been employed in analyzing biochemical interactions including protein interactions, gene-protein interactions, and gene expression pathways (Scott et al., 2005; Ideker et al., 2002; Bailly-Bechet et al., 2011). Often, in these graphs, the nodes correspond to proteins or other biochemical substances, genes, observed symptoms/gene expressions, or a combination of these categories. Weighted edges correspond to strength of statistical evidence for interactions or known biochemical reactions relating substances. A variant of the problem also involves weighted nodes, with node weights based on importance of corresponding genes, molecules, or expressions. Researchers have been successful in identifying potential gene-expression pathways and other biologically relevant subnetworks by finding Steiner trees, and in some cases, the involvement of previously unknown links have been experimentally verified (Scott et al., 2005; Bailly-Bechet et al., 2011).

Inspired by these successful applications in systems biology, we seek to use Steiner tree methods to identify networks of cues employed in human speech. We employ nodes corresponding to acoustic cue dimensions and talkers (e.g., Fig. 2a), and weighted edges are derived from the probability that a specific talker uses a given cue to produce a specific phoneme. Weighted edges also exist between nodes corresponding to cues that co-occur for a given phoneme. There are no edges between the talkers in the model, and separate graphs are constructed for each phoneme (i.e., the eight fricative categories).

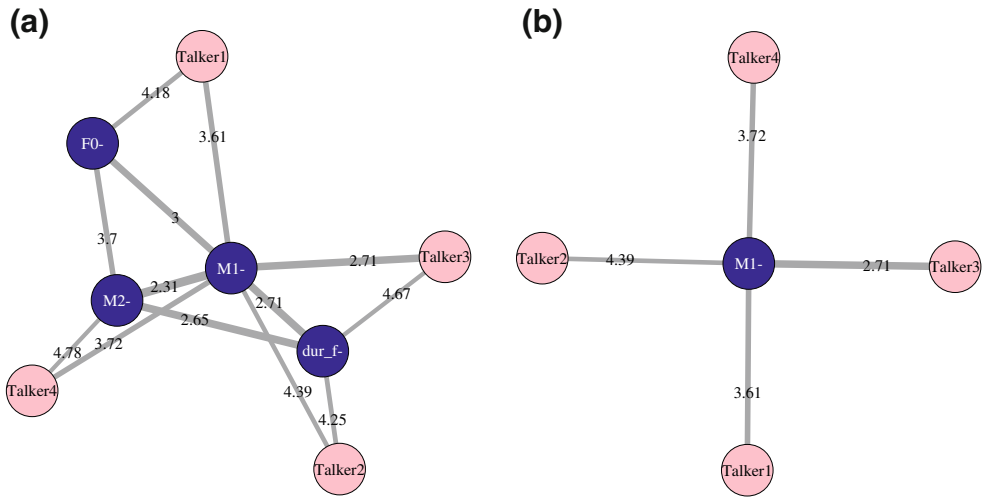

Fig. 2 **a** Example network for four cues and four talkers, derived from a subset of the acoustic data for /ʒ/ (details about how the graphs are constructed, how edge weights are determined, etc., are provided below in the main text). *Blue nodes* represent cues and *pink nodes* represent talkers. Each talker is connected to each cue unless the probability of a talker indicating a /ʒ/ using that cue was very low (<20%), in which case the connection was removed. Note that in these and other graphs, length of the edges is not meaningful; nodes are placed in locations that help to visualize the connections in the network. In this example, it is not costly to go from Talker 3 to Cue M1– because the probability of Talker 3 producing /ʒ/ tokens that contained that cue, relative to other phonemes, is high. Edges between cues indicate the probability that those cues co-occur for /ʒ/ tokens. **b** In this simplified network, the optimal Steiner tree connecting all four talkers is easy to identify: Each talker is connected to M1–, and this is the only cue that they are all connected to. Additionally, the edge weights are lowest for this solution

Thus, the graph represents the way that individual talkers communicate specific phonemes, via an array of acoustic cues.

Such graphs can be easily created from acoustic measurements for datasets with large numbers of phonetic cues. This allows us to visualize, for instance, the degree to which an individual talker uses differences along an acoustic dimension to signal a specific phoneme, with the edge weight inversely proportional to the probability that they produced sounds containing certain cue values as instances of that phoneme (Fig. 2a). The edge weights in these networks represent the cost of connecting nodes, allowing us to study this as an edge-weight minimization problem. Thus, edges that are highly informative for a given phoneme (i.e., a talker that uses a cue only for that specific phoneme) will have low weights. On the other hand, connections between talkers and cues that are used less consistently will be penalized (and hence, weighted) more.

The critical problem is to identify subgraphs where talkers are fully connected using the minimum total edge weight. This allows us to maintain the information needed to categorize the sounds (i.e., capturing information about the mapping from cues to categories across the entire set of talkers) while potentially reducing the number of cues needed to classify a given phoneme. In the language of Steiner trees, if G is the graph created from the data of talkers producing a specific phoneme, and S is the set of nodes corresponding to the talkers, we wish to find a Steiner tree for S in G. We predict that the cues in the Steiner tree will be those that are most useful for identifying the phoneme across the set of talkers, simplifying the model while preserving the information needed to achieve good categorization performance.

Goals and predictions

Despite our knowledge of the phonetic properties of fricatives and the cues used by listeners to identify them, the original problem remains: do we need all statistically-informative cues to accurately describe and classify specific phonemes, or can we achieve similar results with a simpler model that uses a subset of cues? In answering this question, talker variability must be accounted for in some way. Previous approaches provide principled ways of determining which acoustic cues are most influenced by differences between talkers (Kleinschmidt, 2019) and methods for factoring out talker-level variability in specific cue values (McMurray & Jongman, 2011). However, it is unclear whether talkers all use the same acoustic dimensions to signal specific phonemes, or whether they simply vary in the specific cue values they use. If talkers all use the same set of cues, we may be able to achieve accurate categorization with a very small set of cues for each phoneme. On the other hand, if talkers vary idiosyncratically in which cue dimensions they rely on to signal specific phonemes, it may be that all statistically meaningful cue dimensions must be included in a model. The Steiner tree approach described above offers a way to find a middle ground between these two extremes, balancing model complexity with the need to account for variability across talkers.

In order to develop a graph-theoretic approach to solving this problem, we focus on the 24 cue dimensions identified by McMurray and Jongman (2011). For each acoustic dimension, we look at the presence or absence of low versus high values along that dimension in a given speech sound token (e.g., spectral mean is coded as high or low, based on z-scores obtained from the dataset of acoustic measurements). This conceptualization allows us to translate the problem into a network-based representation and offers the possibility of identifying the relevant cue dimensions for each phoneme. Throughout the remainder of the text, we will use the term "cue" to refer to specific high or low values along each acoustic dimension (e.g., a short duration of frication, $\text{DUR}_{\text{F}}-$) that are either present or absent in individual tokens.

Our approach is as follows. First, we create cue-talker graphs for each phoneme (the eight fricatives) that include connections based on all possible cues in the (McMurray & Jongman, 2011) dataset, with edges weighted based the probability of a talker using a particular cue to signal a specific phoneme (talker-cue edges) or the probability of two cues co-occurring for that phoneme (cue-cue edges). We then identify Steiner trees in each of these graphs to obtain cues that connect all individual talkers with minimal edge weights. We predict that this will reveal cues that are used distinctively to indicate specific phonemes. We also measure how the cues are distributed across speech sound tokens, allowing us to determine whether multiple cues are present in individual sounds.

The set of cues identified in these trees (i.e., the *Steiner cues*) are evaluated using multinomial regression classifiers, similar to previous models of speech categorization. In particular, we look at (1) classification accuracy as a function of phoneme, (2) model fit to the training dataset of speech sounds, and (3) model fit to listeners' perceptual responses for novel sounds (the same metrics used by McMurray & Jongman, 2011, in their classifiers). These classifiers serve to verify that the subgraphs identified in our networks capture relevant cue dimensions for recognizing the phonemes. Thus, while the goal of the current study is not to create a model of speech perception, we can evaluate whether the cues identified in the Steiner trees provide information needed to categorize speech sounds similarly to human listeners.

Method

Acoustic data

Acoustic measurements used in the model come from the dataset reported in McMurray & Jongman (2011, see Supplementary Material for that paper). The dataset consists of 2880 utterances produced by 20 talkers (10 female) in six vowel contexts. From these tokens, values along 24 acoustic cue dimensions were measured (e.g., F1 onset, frication duration). To convert these continuous cue values into nodes that could be represented in a graph, we divided each cue dimension into high and low values, creating 48 possible cue-nodes. This was done by first converting numeric cue values along each dimension to z-scores across all talkers. Each token was then classified as high along a dimension if its z-score was positive and low if its z-score was negative. For example, the frication duration dimension was divided into two cues: DUR_F+ and DUR_F−, corresponding to a positive z-score and negative z-score, respectively. If a token had a long duration of frication, DUR_F+ would be present for that token, and DUR_F− would be absent. Thus, there were 48 possible cues, half of which were present in each token (i.e., if one cue was present in a token [e.g., DUR_F+], its inverse [DUR_F−] was not).[3]

We evaluated Steiner trees using two versions of the acoustic data. First, we examined the effectiveness of raw cue values, converting them into high and low cues as described above. Second, we applied the C-CuRE approach (McMurray & Jongman, 2011) prior to calculating z-scores, whereby contextual variability caused by talker and vowel differences is first factored out by running a linear regression on each cue dimension using talker and vowel as predictors. The residual values were then used as the cue values when computing z-scores. This produces cues that are less context dependent and more closely reflect listener performance, as demonstrated by McMurray and Jongman (2011).

Talker-cue graphs

For each fricative F, we construct an edge-weighted graph G_F. The node set N_F consists of 68 nodes, corresponding to 20 talkers and 48 cues. For a given talker and cue, an edge exists between the two corresponding nodes if the talker ever produced tokens of F that included the given cue. The edge set E_F contains some talker-cue and some cue-cue edges. Edges between talkers and cues are weighted by the inverse probability that a talker produced a phoneme F given that the cue c was used. Edges between two cues are weighted by the inverse probability that the phoneme F was produced given that the two cues co-occurred.[4]

Given a talker t and a cue c, if t ever uses c when producing F, let the conditional probability that t was producing the phoneme F given that t used the cue c be denoted by $P(F|(c,t))$. Based on this, we define a talker-cue edge weight as

$$w_F(c,t) = \frac{1}{P(F|(c,t))} = \frac{\#\text{ instances in which } t \text{ uses } c \text{ for any phoneme}}{\#\text{ instances in which } t \text{ uses } c \text{ when producing } F}. \quad (1)$$

In addition, because high edge weights indicate a low probability that a talker used a cue to signal a given phoneme, we initially prune the graph of edges with weights >5 (corresponding to a probability of 0.2). This decision was made to optimize performance while maintaining computational tractability, and variations in this parameter did not adversely affect the results (see Sensitivity Analysis below for further discussion). Thus, if

$$0 < w_F(c,t) \leq 5 \quad (2)$$

then the edge $e_{(c,t)}$ connecting nodes for c and t exists in E_F, and has weight $w_F(c,t)$. A similar approach is used to compute weights for cue-cue edges (see Supplemental Material for equations used to calculate edge weights).

As a specific example, suppose that 20 of the tokens produced by Talker 2 had a high spectral mean, and of these, ten were /s/ tokens (i.e., M1+ was present in those sounds). This results in an edge weight of 2 (1) between the node for Talker 2 and the M1+ node. This approach thus reveals which cues are uniquely informative for specific phonemes and individual talkers. If Talker 2 always had a high spectral mean (M1+ was present in all of their tokens, perhaps because the talker is a woman, not because she uses this cue to distinguish the phonemes), the edge weight would be 8 (all tokens produced by that talker, divided by the eight fricatives, which were equally likely). In this case, the cue would not be informative. Indeed, the edge would be >5 and thus would be pruned from the graph entirely. Conversely, if the talker had only ever used that cue when producing /s/, it would be *extremely* informative (edge weight of 1, which is the minimum possible edge weight that we would observe in a graph; an edge weight of zero would signify that there is no edge in the graph.).

[3] Note that this way of characterizing the acoustic cue dimensions does not allow us to make inferences about the specific fine-grained cue values that are used to signal different phonemes, which is important for perception (McMurray et al., 2002; Toscano et al., 2010). While our goal here is to evaluate the efficacy of the Steiner tree method and provide an initial estimate for the consistency with which cue dimensions are used across talkers, future work should investigate the fine-grained information in more detail. See Discussion section for additional details.

[4] Other monotonic functions would work as well, varying in how much they over/under weight edges with values near the extremes (e.g., close to 1).

Steiner trees

After creating each fricative graph, we identified low-weight subgraphs connecting all talker nodes. In many cases the identified subgraphs were provably optimal, and hence Steiner trees. To keep computation time manageable, we also accepted nearly optimal subgraphs, as described below. For simplicity, all identified subgraphs are referred to as Steiner trees.

Integer linear programming (Williams, 2009), which is a common method of finding Steiner trees, was used to identify the subgraphs. In general, a linear programming problem requires maximizing or minimizing a linear function of several variables, subject to a system of linear constraints, where the constraints take the form of equations or inequalities. Integer linear programming refers to the situation in which the variables are further constrained to be integers. One way to translate the Steiner tree problem into this framework is outlined in Joyner et al. (Section 11.4, 2010): define a variable x_e for each edge e in the graph $G = (N, E)$. A subgraph H of G may be defined by assigning either $x_e = 0$ or $x_e = 1$ for each edge e in E. If $x_e = 0$, then the edge e is not present in the subgraph; if $x_e = 1$, then e is present in the subgraph. We then define a variable y_n for each node n in N. Define $y_n = 1$ if there is some edge in the subgraph that contains n, and $y_n = 0$ otherwise. Our subgraph contains all nodes n such that $y_n = 1$.

The Steiner tree problem for S in G can be thought of as minimizing the sum of the weights of the edges in a subgraph, subject to the constraint that the subgraph must connect all distinguished nodes. The function to be minimized is then

$$\sum_{e \in E} x_e w_e, \tag{3}$$

subject to the constraints that (1) the subgraph is a tree and (2) the subgraph contains all of the nodes in S. The first constraint may be tested by checking for acyclicity (a condition which may be stated as a linear inequality using maximum average degree; Joyner et al., 2010, Section 11.1) and requiring that the number of edges is one less than the number of nodes in the subgraph. The second constraint may be stated as $y_n = 1$ for all $n \in S$.

Once the Steiner tree problem has been translated into the linear programming setting, existing linear programming software can be used. Many authors have developed ways to improve the efficiency of linear programming approaches to Steiner tree problems (e.g., Stanojevic & Vujoševic, 2006; Althaus et al., 2003). However, the simple approach described above has been implemented in SageMath (The Sage Developers, 2017) and allows users to employ their choice of linear programming solvers. The SageMath function `Steiner_Tree()` with LP solver GLPK (GNU Linear Programming Kit) was used to find all Steiner trees described here.

To reduce computing time, the relative mixed integer programming gap tolerance was increased, meaning the subgraphs returned by the `Steiner_Tree()` function were sometimes approximate solutions to the Steiner tree problem, as mentioned above. We note that in this context, nearly optimal approximate solutions also provide potentially meaningful information about useful cue networks (we verify this using the multinomial regression classifiers described below). There also may be multiple optimal (or nearly optimal) Steiner trees for a given graph and distinguished set. Therefore, the `Steiner_tree()` function was run 100 times on each phoneme graph to identify the cues that were most consistently included in the Steiner trees. Note that the distinguished nodes in all cases were the set of all nodes corresponding to talkers, and the additional (Steiner) nodes included in the returned trees corresponded to cues. We hypothesized that cues appearing in many of the trees would be those that are most useful in signaling specific phonemes across the population of talkers.

Multinomial regression classifiers

In order to determine whether the cues identified in the Steiner trees contain sufficient information for fricative classification, multinomial regression classifiers were trained on both the entire set of cues and the Steiner cues that were included in 5% or more of the solutions for each fricative (see below for a justification of this criterion). Separate classifiers were trained for raw cues and context-compensated (C-CuRE) cues, with cues coded as either present or absent in each token. This led to four sets of classifiers: (1) a classifier based on raw cue values with all possible cues; (2) a classifier based on raw cue values and those identified in the Steiner trees; (3) a classifier based on context-compensated cues (computed using C-CuRE) for all possible cues; and (4) a classifier based on context-compensated Steiner cues.[5] This approach follows from several previous models that have used logistic and multinomial regression to model speech categorization (Nearey, 1997; McMurray & Jongman, 2011, see Supplemental Material for additional details on the multinomial regression classifiers used here).

Classifiers were trained and tested using the same approach as McMurray and Jongman (2011), who presented three measures of model performance: (1) classifier accuracy, (2) measures of model fit to the training data, and

[5] Classifier performance alone does not tell us which cues are informative for which phonemes, since a cue from the Steiner trees for one phoneme could be used to classify other phonemes as well. Rather, the classifiers provide a way to determine how the reduced set of cues identified in the Steiner trees performs overall.

(3) similarity of model responses to listeners' perceptual data. Training data for the classifiers came from tokens with a complete set of acoustic measurements (26 tokens had missing or undefined cue values and were not included); a separate set of tokens from the McMurray & Jongman dataset with perceptual data from 40 listeners was used for testing. This yielded 2617 tokens in the training dataset and 237 tokens in the test set.

To determine how well each classifier performed, we first looked at overall classifier accuracy as a function of phoneme. Next, we computed Bayesian Information Criterion (BIC) scores, a measure of goodness-of-fit that takes into account the number of free parameters in a model (Schwarz et al. 1978). This allows us to compare models with different numbers of cues, such as the all-cue and Steiner-cue classifiers. Together, these two measures indicate how well the model can identify the intended fricatives and how well it accounts for the training data.

Although these measures are useful for evaluating the models, they do not provide an indication of how well they perform relative to human listeners' perceptual judgments (i.e., did the model correctly classify the same tokens as listeners, and misidentify the same ones?). To measure how well the classifiers accounted for listeners' perceptual data, we computed the log-likelihood of each model given the listener data, following the procedure outlined in McMurray and Jongman (2011). Briefly, these likelihoods describe the probability of the listeners' distribution of responses given the model's predictions (i.e., how likely was a given model to generate the data observed from human listeners?). We take the log of these probabilities across the individual tokens and then compute the sum to determine the overall log-likelihood. This method allows us to compare how well different models match listener performance on individual tokens. Equations describing how to calculate the likelihoods are given in McMurray and Jongman (2011).

Results

Steiner trees

We first examined the output of the Steiner tree solutions. In general, 2–6 cue dimensions per phoneme were necessary to connect all of the talkers. Figure 3a shows the complete graph for /z/, and Fig. 3b shows a representative Steiner tree solution for /z/ using context-compensated (C-CuRE) cues. In each graph, pink nodes represent talkers (i.e., the distinguished nodes) and blue nodes represent possible cues. As the figures illustrate, the solution for /z/ required three cues to connect each talker, with two large clusters of talkers using M1TRANS+ and M4+ distinctively to signal /z/, and one smaller cluster of three talkers using LOW_F+ to indicate this phoneme. Thus, the space of 24 possible cues has been reduced to three.

Steiner tree solutions for other phoneme categories took different forms. Figure 4 shows examples for each phoneme. In these trees, /f,v,θ,z/ required a small number of cues to connect all talkers, whereas /ð,s,ʃ,ʒ/ required many more. Different patterns emerge within these graphs as well: /f/ has two equal-sized clusters of talkers, whereas /v/ has a large number of talkers using one cue (M2+) with only a small number connected to the other cue (DUR_F–).

The graphs reveal that, in minimizing the total edge weight, talkers were not always connected to the same cue node, suggesting that cue-use varies across talkers. Alternatively, if all talkers were consistently using one cue for a specific phoneme, the graphs would include a single cue node. In addition, clusters of talkers are generally

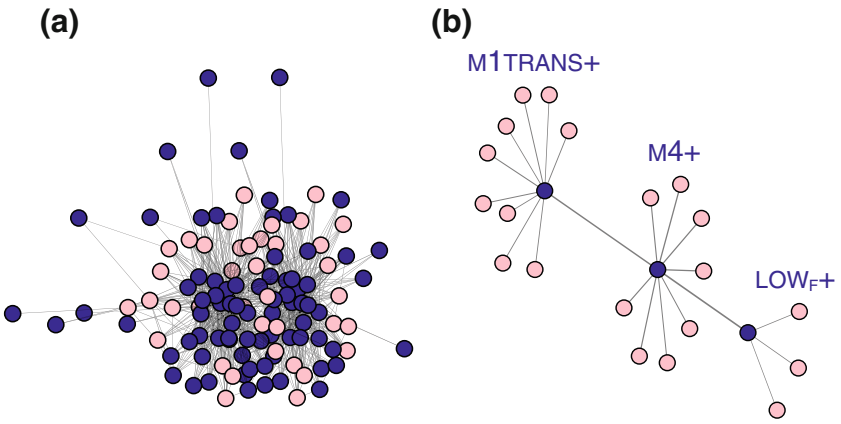

Fig. 3 **a** Complete graph for /z/ generated using context-compensated (C-CuRE) cue values. *Pink nodes* indicate talkers (which are the distinguished nodes in the Steiner tree) and *blue nodes* indicate possible cues. **b** Representative Steiner tree solution for the complete graph in the left panel. The solution reduces the number of cues to three while still connecting all talkers (in this case, via the three cues, M1TRANS+, M4, and LOW_F+)

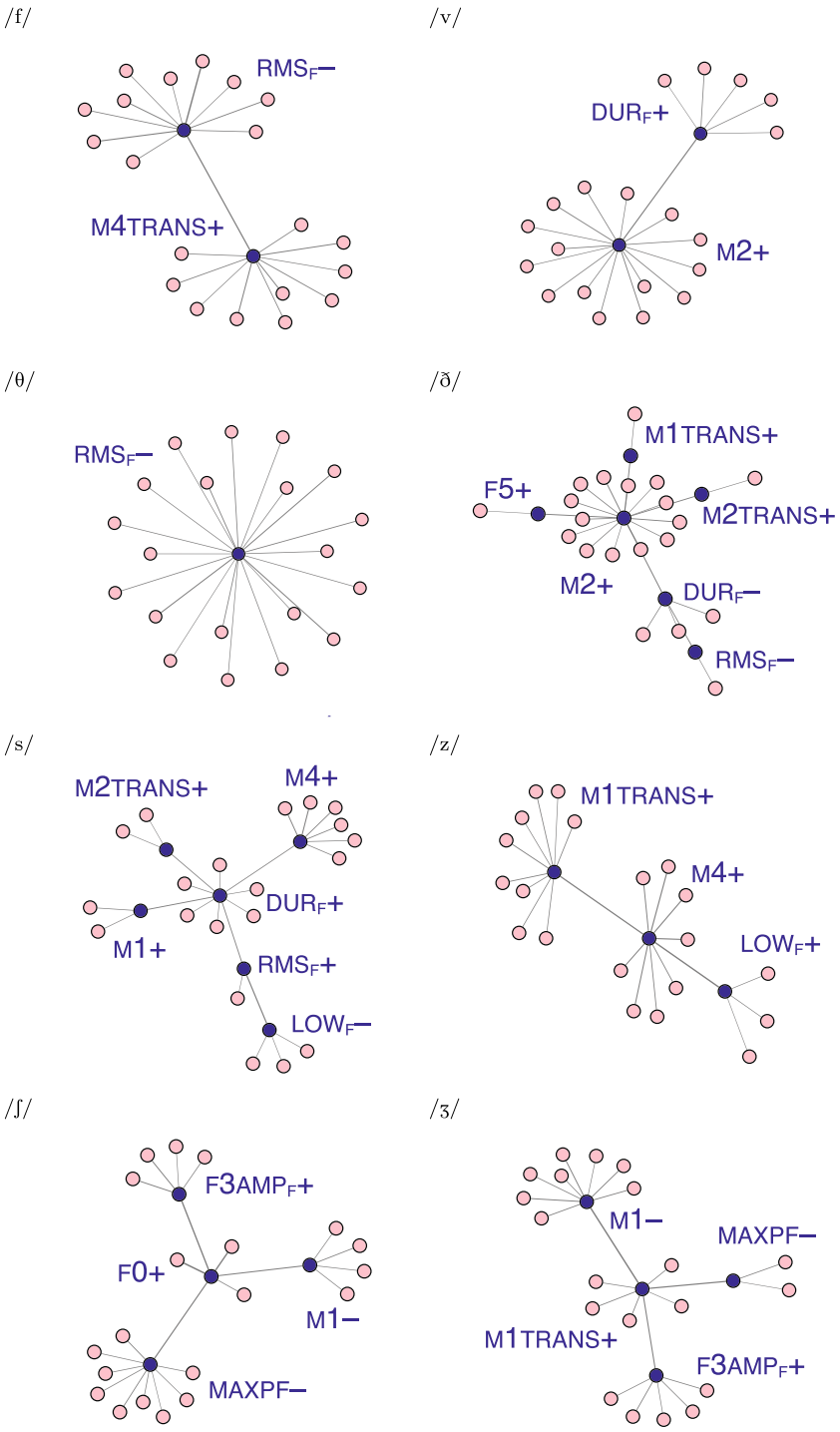

Fig. 4 Representative Steiner Trees for each fricative using context-compensated cues. *Pink nodes* represent individual talkers, and *blue nodes* represent cues. Each fricative has a distinct pattern of cue clustering, with some graphs showing a single cluster for all talkers (e.g., /θ/) indicating that all talkers tend to use this cue; and others showing a large number of cues (/ð/), indicating that individual talkers tend to use different cues

connected to each other via connections between correlated cues, rather than between individual talkers connecting clusters (i.e., by having an edge between two cues). Note that the lack of an edge between a talker and cue does not necessarily mean that the talker doesn't *also* use that cue. That is, because there are no cycles in the Steiner trees; only the edges needed to connect the talkers with minimal edge weight (i.e., those that produce the lowest total weight in the network) are shown. The results also suggest that when more than one cue is needed to connect all talkers, the

additional cues co-occur with each other more consistently than talker-cue connections that could be used to bridge clusters of talkers. This leads to lower cue-cue edge weights compared to talker-cue edge weights. Indeed, the mean cue-cue weight across phonemes in the complete graphs for context-compensated cues is lower (3.81) than the mean talker-cue weight (4.18). This demonstrates a key principle of the Steiner trees, which balance the need to include the most robust cues from each talker with the need to have a compact representation and to avoid including cues in the network that account for just one or two individual talkers (which could result in a much higher total edge weight for the graph).

We next looked at how often each cue was identified over the 100 simulation runs for each fricative graph. Figure 5 summarizes the cues identified in the C-CuRE Steiner trees as a function of phoneme. As the figure shows, many cues are rarely (<5%) included in the Steiner trees for certain phonemes (e.g., none of the formants were used for /f/), suggesting that these cues are generally not robust in signaling those phonemes. Other cues were found in almost all Steiner trees for a given fricative, suggesting that these are highly reliable (e.g., M1TRANS+ for /z,ʒ/).

A number of other patterns emerge from this analysis, several of which are consistent with previous phonetic analyses of these sounds (Jongman et al., 2000; Klatt, 1976; Baum & Blumstein, 1987; Haley et al., 2010; McMurray & Jongman, 2011; Stevens et al., 1992; Shadle & Mair, 1996; Hughes & Halle, 1956). For example, a high spectral mean (M1+) signals /s/, while a low spectral mean (M1–) signals /ʃ/. Several cues also reflect previously described differences along phonological feature dimensions. Steiner trees for the voiceless sounds (/f,θ,s,ʃ/) had high f0 values (F0+) and generally had longer frication durations (DUR_F+, except for /θ/). Non-sibilants (/f,v,θ,ð/) all included low frication amplitude as a cue (RMS_F–), and conversely, two of the sibilants had high frication amplitudes (RMS_F+). $F5AMP_F$+ and M3+ also tend to be present in sibilants but not in non-sibilants. The specific place of articulation for some phonemes was also distinguished, with M3TRANS+

Cue Type	Cue	f	v	θ	ð	s	z	ʃ	ʒ
Temporal	DUR_F	+	−		−	+	−	+	
	DUR_V		+		+				
Spectral (dynamic)	F0	+		+	−	+		+	
	F1								
	F2		−						
	F3								+
	F4				+			−	
	F5				+				
Spectral (static)	M1			+		+		−	−
	M2	+	+		+	−			
	M3			−		+		+	+
	M4		+	+	+	+	+		
	M1TRANS				+		+		+
	M2TRANS		+		+	+			
	M3TRANS	+	+		−		−		
	M4TRANS	+	+						
	MAXPF							−	−
Amplitude	$F3AMP_F$							+	+
	$F3AMP_V$								
	$F5AMP_F$		−		−	+	+		+
	$F5AMP_V$								
	LOW_F				+	−	+	−	+
	RMS_F	−	−	−	−	+	+		
	RMS_V					−			

Fig. 5 Chart of Steiner cues as a function of fricative derived from graphs using context-compensated cues. *Positive signs* indicate that Steiner tree solutions identified that cue dimension as high for that phoneme; *negative signs* indicate Steiner tree solutions with low cue values for that dimension. *Shading* indicates the proportion of simulation runs for which a cue was included in the Steiner tree (black=100%; white=0%). *Empty cells* indicate cue dimensions that were included <5% of the time in the Steiner trees for a given phoneme. Cue labels are the same as those used in McMurray and Jongman (2011), which also gives the definition of each cue. See Supplemental Material for the proportion of Steiner trees that included each cue

present for labiodental fricatives (/f,v/) but not other phonemes, and several cues (F3AMP$_F$+, M1–, M3+, and MAXPF–) distinguishing /ʃ,ʒ/ from other phonemes.

Interestingly, the patterns for several cues are more idiosyncratic and do not reflect articulatory feature dimensions or previously described phonetic patterns. For example, /v,θ,ð,s,z/ all include a high degree of kurtosis during frication as a cue (M4+), and /v,ð,s/ use M2TRANS+, yet these sounds differ in place, voicing, and sibilance. Similarly, phonemes that share similar features may be signaled by different sets of cues. This suggests that not all acoustic cues used by talkers are linked to phonological feature dimensions. Moreover, there are several cues that appear to be uniquely used to signal specific phonemes (e.g., F3+ for /ʒ/, F2– for /v/, and RMS$_V$– for /s/), though they are not invariant (i.e., they do not connect all the talkers with edge weights of 1).

Overall, these results demonstrate that multiple cue dimensions are necessary to describe how talkers indicate specific phonemes. In particular, they suggest that talkers may preferentially use different acoustic dimensions for specific phonemes, with networks revealing that several talkers often cluster around a specific cue (as opposed to all talkers clustering around a single cue, though this happens for some phonemes). It is worth noting that the complexity of the tree structure does not predict ease of recognition by listeners, but rather uniformity of production among talkers. In some cases, there may be a large group of talkers who use a cue, while in others there may be only a few; critically, both cues are needed to account for variability in the production of the sound across the population of talkers. This result is reminiscent of models suggesting talker-specific phonological and lexical representations in speech perception (Goldinger, 1998; Johnson, 1997; Kleinschmidt & Jaeger, 2015, though note that this does not necessarily mean that listeners are sensitive to these talker-specific cue dimensions in perception). These results also mirror the conclusions of McMurray and Jongman (2011) that multiple cues are necessary. At the same time, they provide a simpler description of the cues necessary for categorizing individual phonemes.

The results also suggest specific cues that researchers might measure from spectrograms in individual tokens for each phoneme. Figure 6 summarizes this information graphically: it shows example waveforms and spectrograms for each fricative with the most relevant cues (the four cues most frequently found in the Steiner tree solutions for that phoneme) overlaid on top, illustrating how each cue can be identified in individual tokens. Given token-level variability between speech sounds and its impact on speech perception

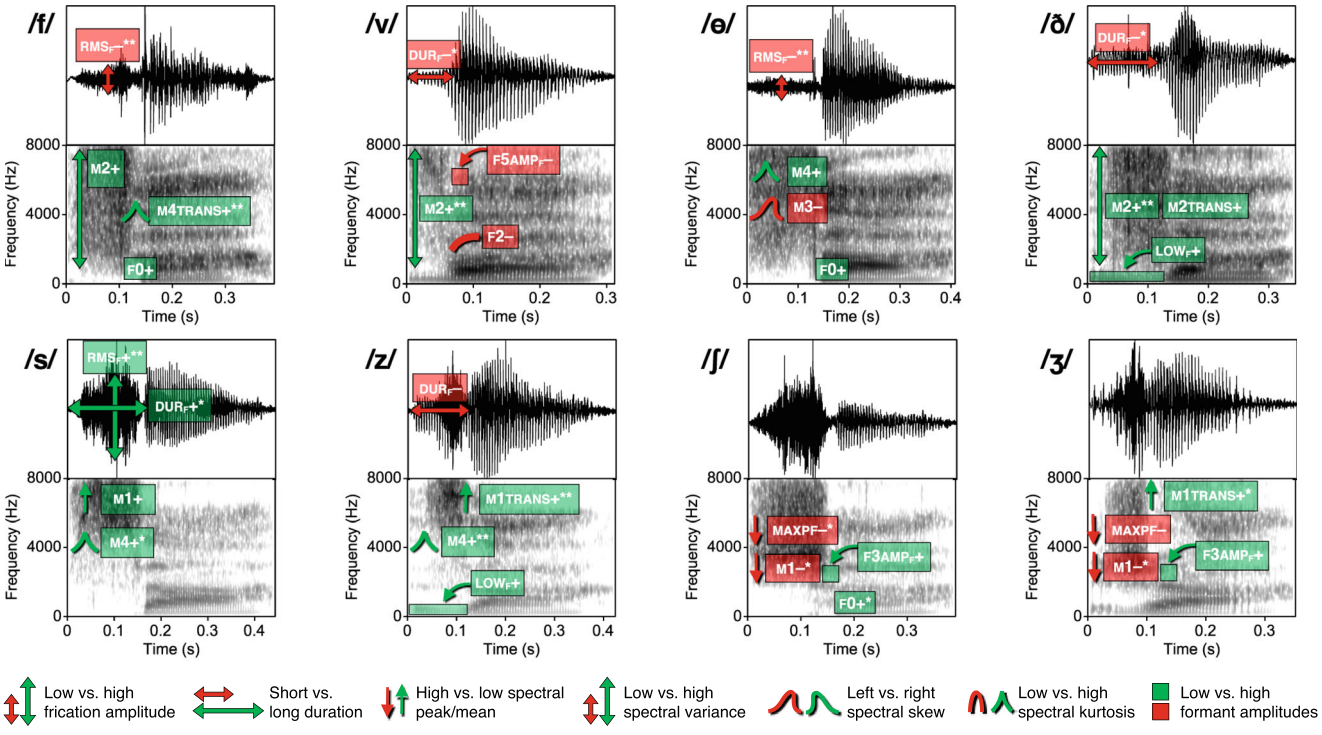

Fig. 6 Example waveforms and spectrograms for each fricative with the most frequent Steiner cues overlaid on top. ** indicates that the cue was identified in 100% of the solutions, and * indicates that the cue was identified in at least 90% of the solutions. Cues labeled *in red* correspond to low values along that dimension; *green* indicates high cue values along that dimension. None of the fricatives has the exact same set of the top four relevant cues. Example spectrograms are from tokens in the Schatz et al. (2015) corpus

(Li et al., 2012; Toscano & Allen, 2014), this provides a useful measure for determining whether a given sound should yield robust recognition (i.e., whether it contains informative cues).

Lastly, we examined how many cues a given token contains. Given the Steiner tree solutions we found, it may be that each token only contains a single informative cue (e.g., a cue that is characteristic of each talker, who were only linked to individual cues in the Steiner trees). Thus, although multiple cues are needed to account for variability across all the talkers, it is not clear whether listeners must integrate cues on a token-by-token basis. To investigate this, we calculated the proportion of tokens that contained different numbers of Steiner cues (though not necessarily the same Steiner cues) as a function of phoneme. Figure 7 shows the results of this analysis. For each phoneme, all tokens in the dataset contained at least one Steiner cue, and the majority (¿96.9%) contained at least three cues. Beyond this number, the proportion of tokens that contain multiple cues begins to fall off, though this varies as a function of phoneme. This is partly due to the fact that different phonemes use different numbers of cues (e.g., /ð/ has 13 Steiner cues, so there is a higher likelihood that a greater number of these cues will be present in the tokens). Because almost all tokens contain at least three cues, we would expect listeners to integrate these cues during perception of natural speech, similar to processes observed with experimentally manipulated speech sounds (Miller & Liberman, 1979; Repp, 1982; Toscano & McMurray, 2012; 2015). This result also demonstrates that although talkers vary in which cues connect them to other talkers in the Steiner trees, it is likely that many of them are using similar sets of acoustic cues to signal specific phonemes (though they may vary in which cues they use most consistently).

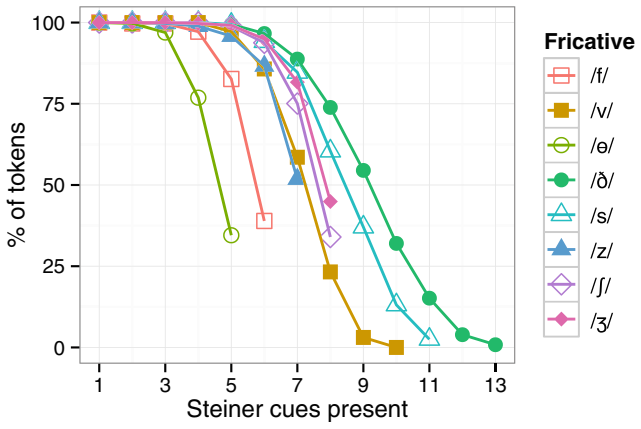

Fig. 7 Proportion of tokens by phoneme that have a given number of Steiner cues present. Tokens generally have at least three Steiner cues, and most have four (except for /θ/). For example, approximately 80% of all /f/ tokens in the dataset contain five out of the six total Steiner cues for that phoneme, where the six Steiner cues are those that were found in at least 5% of the Steiner trees for /f/

Classifier performance

In order to determine whether the cues identified in the Steiner trees do indeed provide useful information for speech sound categorization, we trained multinomial regression classifiers to categorize the eight phonemes based on different subsets of cues. Classifiers were tested on tokens with perceptual data that were withheld from training, as in McMurray and Jongman (2011). We compared the performance of the Steiner-cue model to classifiers that had all cue dimensions, examining both raw cues and context-compensated cues in each case.[6] In all analyses described in this section, we used a 5% threshold for including Steiner cues (i.e., cues were present in at least 5% of Steiner tree solutions; this is also reflected in the analyses above; Fig. 5). Justification for this inclusion threshold is provided below in the Sensitivity Analysis.

For the classifiers trained on raw cues, performance was similar for both the all-cue (74.7%) and Steiner-cue (75.9%) models, with slightly better performance for the Steiner-cue model (Table 2). Performance was lower than the raw-cue classifiers in McMurray and Jongman (2011), which had a mean accuracy of 85.0%. The decreased performance may be due to the use of discrete cues rather than continuous cue values (see C-CuRE results below for an analysis of this). The Steiner-cue model yielded a better fit to the training dataset (all-cue BIC: 4503; Steiner-cue BIC: 4434).[7] Thus, overall classifier performance is not impacted by reducing the cues to just those included in the Steiner trees and the classifier provides a better fit to the data. These results suggest that the Steiner cues provide a useful set of cues for categorization. Overall, however, both classifiers performed low relative to human listeners. Figure 8a shows classifier performance as a function of phoneme, along with the perceptual data from listeners presented in McMurray and Jongman (2011). This replicates the findings of McMurray & Jongman for classifiers based on raw cues. Despite this, the Steiner-cue model still shows a better fit to listeners' data (all-cue log-likelihood: -5069; Steiner-cue log-likelihood: -5047).

Next, we examined classifiers trained using context-compensated (C-CuRE) cues. Recall that these cues were created by first factoring out talker and vowel context and then z-scoring the residual values to determine whether a token had a positive or negative value along a cue

[6] Separate Steiner trees were computed for raw cues and context-compensated cues.

[7] Smaller BIC values indicate better fits, and BIC penalizes more complex models (in this case, models with more cues). Thus, the fit is better for the Steiner-cue model.

Table 2 Classifier performance

Cues used	Context compensation	Accuracy	BIC	Log-likelihood
All	Raw cues	74.7%	4503	–5069
Steiner	Raw cues	75.9%	4434	–5047
All	C-CuRE	87.3%	3748	–4050
Steiner	C-CuRE	85.7%	3672	–3781

BIC measures model fit for the talker's intended phoneme category, and log-likelihood measures fit to listeners' perceptual data

dimension. Overall, these classifiers performed much better, with a mean accuracy of 87.3% for the all-cue model, and a mean accuracy of 85.7% for the Steiner-cue model (Fig. 8b). Note that performance for the all-cue model is lower than the C-CuRE classifier in McMurray and Jongman (2011), which was 92.9%. This represents the cost of converting from continuous cue values to discrete positive and negative cue dimensions that are present or absent in each token. This is an inherent limitation of the graph-based approach: information must be represented as discrete nodes in the graphs. However, if one wanted to obtain finer-grained estimates of the cue values (e.g., for improving classifier performance or for deriving stimuli to be used in a perceptual experiment), this could be accomplished by identifying Steiner trees that divide the cue dimensions into smaller bins.

Restricting the set of cues to the Steiner cues had an impact on overall accuracy (1.6%), but the Steiner-cue model had a better fit to the training dataset (all-cue BIC: 3748; Steiner-cue BIC: 3672). In addition, both models provided a close match to listeners' responses (Fig. 8b), with the Steiner-cue model providing the closest match overall (all-cue log-likelihood: -4050; Steiner-cue log-likelihood: -3781). Again, these results demonstrate that the Steiner-cue classifier shows similar performance to the all-cue classifier, despite having three fewer cue dimensions.[8]

To see whether comparable performance is achieved when continuous cue values are used, we trained an additional C-CuRE classifier using continuous cues for acoustic dimensions identified in the Steiner trees, providing a more direct comparison to the classifiers used by McMurray and Jongman (2011). This resulted in an accuracy of 89.5% (BIC: 3086; log-likelihood: -3156), illustrating the slight cost of including fewer cues in the classifier (i.e., a classifier using only 21 acoustic cue dimensions, as opposed to the full set of 24 cue dimensions).

[8]Because of the way cues are defined in the Steiner trees, as long as one cue was identified in the tree, the cue dimension was included in the classifier. Thus, while the Steiner-cue classifiers included 21 of the 24 cue dimensions, this corresponded to only 31 out of 48 possible Steiner cue nodes.

Lastly, a follow-up analysis revealed that the Steiner cues are less reliable in tokens that produce classification errors. The average absolute value of the z-scores for the Steiner cues for each phoneme was 0.95 for correct tokens and 0.75 for incorrect tokens. Thus, as expected, the Steiner cues are more distinct in tokens that are classified correctly.

Sensitivity analysis

Creating and analyzing the talker-cue graphs for each phoneme required a number of choices to be made. For example, creating the graphs necessitated choosing some discrete categories to classify the cue values along each dimension. While many systems would have been reasonable, we used high- and low-values for each cue dimension as an initial approximation. Additional cue value categories along each dimension would yield finer-grained measurements of the cues used in each Steiner tree graph (see Discussion for additional details).

Similarly, though other choices might have been reasonable, edge weights were determined by inverse conditional probabilities, and thresholds were specified for including edges in graphs and including cues in later analyses. These thresholds were generally chosen to make computations feasible, while still making meaningful distinctions and preserving as much information as possible. A full exploration of the entire space of possible parameter choices and methods used to construct the graphs is beyond the scope of the current study. However, in the spirit of sensitivity analysis and parameter space exploration (e.g., Pitt et al., 2006), we assessed whether small changes to some of these parameters would drastically change the results. Here, we give data to indicate that our results are robust in the sense that, for the examined parameters, small changes in parameter values are unlikely to significantly affect the results.

Two parameter values for which sensitivity analysis is appropriate are the thresholds mentioned above. The edge inclusion threshold is the maximum value such that edges of that weight are included in the phoneme graphs. Our analysis uses a threshold of 5, meaning that the edge must correspond to at least a 20% conditional probability of the phoneme, given the information of the two incident nodes. For context, since eight phonemes are considered, randomly assigning phonemes to the tokens would result in average weights of 8 for all edges such that the talker-cue pair or two cues corresponding to the incident nodes were ever linked. Thus, edges with weights of 8 or greater are unlikely to be informative. We initially considered using a threshold of 8, but this was not practical due to computational feasibility.[9]

[9]Simulations were run on a desktop computer; a high-performance computing system may overcome this limitation.

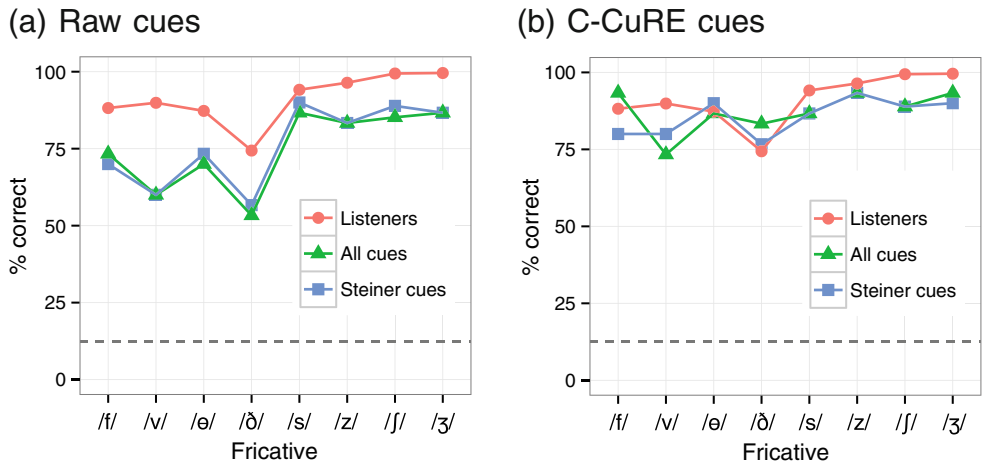

Fig. 8 **a** Listener and classifier performance for raw cues. Both the all-cue and Steiner-cue classifiers perform more poorly than listeners, though the Steiner-cue model provides a better fit to listeners' data. **b** Performance for context-compensated (C-CuRE) cues. Here, both classifiers closely match listener performance across the set of phonemes, with a better match for the Steiner-cue classifier. Note also that the overall pattern of classifier accuracy follows that of the listeners (e.g., poorer performance on non-sibilants than on sibilants)

To determine whether a threshold of 5 was reasonable, we ran the Steiner tree algorithm ten times for each phoneme on graphs created using other edge weight thresholds. Edge weights were reduced until computational run time was lowest while still being able to find solutions for all phonemes. Edge weights started at 8 (corresponding to chance) and were reduced to 7, 6, 5, and 4. However, in graphs with maximum edge weights of 4, a fully connected subgraph could not be found for all phonemes. With the threshold set at 5, 100 iterations for each phoneme were run (those reported above). Additional analyses were run to identify ten solutions including edge weights of 6, 7, and 8. The most common cues ($\geq$ 20%) identified in our Steiner tree solutions were also identified in the 10 runs for each phoneme for edge weight thresholds of 6, 7, and 8 with just three exceptions: for a cutoff of 6, F5AMPF- (26%) was not found for /v/; for a cutoff of 8, M2+ (20%) was not found for /f/ and F5AMPF+ (22%) was not found for /ʒ/. The total number of cues found with probabilities $\geq$ 20% was 38 (including repeats for different phonemes). Our sample ten runs finds 37/38 cues for a cutoff of 6, 38/38 cues for a cutoff of 7, and 36/38 cues for a cutoff of 8. Thus, while some variation in less common cues occurred, the most frequent cues for all eight phonemes were essentially the same as those reported above for graphs with edge weights of 6, 7, and 8. Thus, the threshold of 5 is robust, while also minimizing run time and producing fully connected subgraphs for each phoneme.

A second factor we can examine is the cue inclusion threshold. This is the percentage of Steiner tree solutions that must include a cue for that cue to be included as a Steiner cue in the classifier analyses. The result of changing this threshold for the C-CuRE Steiner tree classifier is shown in Fig. 9. A threshold of 5% was chosen to balance classifier performance with having a compact set of cues. Decreasing this threshold to 0% (i.e., including all cues found in any Steiner tree), results in a small improvement in classifier performance, but at the cost of having to include nine additional cues. Conversely, increasing the threshold from 5% to 10% results in a larger performance drop than going from 0% to 5%, while only reducing the set of cues by

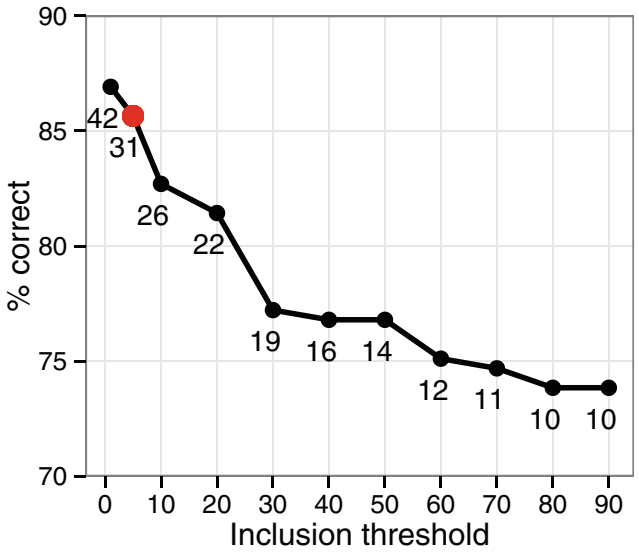

Fig. 9 Classifier performance as a function of the proportion of Steiner tree solutions for context-compensated cues that included a given cue. *Numbers next to data points* indicate the specific number of cues included in the classifier. *Data point highlighted in red* corresponds to the 5% inclusion threshold (i.e., cues that appeared in at least 5% of trees for a specific phoneme). This set of Steiner cues was used in all other analyses

five. It is worth noting that even using an inclusion threshold of 90%, which leads to only ten Steiner cues total, the classifier still achieves around 74% accuracy. Overall, these results indicate that the model is robust to small changes in the cue inclusion threshold. Furthermore, one could alter these parameters in future work depending on the goal (e.g., including every possible relevant cue).

Discussion

By finding subgraphs known as Steiner trees, we found that we can identify specific cues for speech sound classification that account for differences between talkers. Classifiers trained on the Steiner cues show similar accuracy, better fits to the training data, and closer matches to listeners' perceptual responses than classifiers trained on all possible cues. Thus, the Steiner tree approach provides a viable method for identifying informative acoustic cues in speech. Given that this approach can be used with cues defined by any spectro-temporal feature in the speech signal, this provides a useful method of reducing the number of cues needed to recognize specific phonemes while still accounting for variability in the relevant cue dimensions across talkers. In the following sections, we discuss how these results fit with previous work on phonetic cues to fricatives, existing models of speech sound categorization, and possibilities for future work using these techniques.

Phonetic cues to fricatives

The Steiner tree solutions reinforce several findings from previous phonetic analyses of fricatives. Spectral mean (M1), for instance, was found to be informative for the distinction between /ʃ/ and /s/, consistent with previous acoustic analyses (Haley et al., 2010). Spectral peak (MAXPF) distinguished place of articulation, specifically separating /ʃ,ʒ/ from the other phonemes (Jongman et al., 2000). Similarly, frication duration (DUR_F) and low-frequency frication amplitude (LOW_F) distinguished voiced and voiceless sounds, consistent with previous phonetic analyses (Klatt, 1976; Jongman et al., 2000; McMurray & Jongman, 2011; Stevens et al., 1992). This provides additional confirmation that the Steiner trees included phonetically relevant cues. Not all previously identified cues were used in the Steiner tree solutions, however. F2 onset, in particular, has been argued to be a cue to place of articulation (Jongman et al., 2000), but we found that it only occurred as in at least 5% of the Steiner trees for one phoneme (F2– in /v/).

We can also compare these results to those of McMurray and Jongman (2011), who used the same dataset in their models, but created classifiers for continuously valued cues. There are a number of cues identified in the Steiner trees that are consistent with the patterns they report (compare Fig. 5 in the current paper and Table 3 from McMurray & Jongman, 2011). For instance, they found that RMS_F and DUR_F are cues to both sibilance and voicing, which we find as well. There are also some cues that they reported as indicating multiple phonological features (e.g., LOW_F as a cue to both sibilance and voicing) that are only used for one feature in the Steiner trees (voicing). Thus, the Steiner tree solutions reduce the cues to a subset needed to connect all the talkers, even though those cues may distinguish additional phonological features statistically.

There were also cues included in the Steiner trees that McMurray & Jongman do not report as distinguishing any features. This occurs, in particular, for cases where the cue was primarily used for a single phoneme, making its function more idiosyncratic than cues that serve to provide information about all phonemes that share a particular phonological feature (e.g., F0+ was found in the Steiner trees for the four voiceless fricatives, but primarily it was used in /ʃ/; McMurray & Jongman did not report this as a cue for voicing).

The Steiner tree solutions also revealed patterns not previously observed. Of particular interest are cases where a cue dimension provides information for specific fricatives that vary along several phonological feature dimensions (e.g., M4+ as a cue for /v,θ,ð,s,z/). This result may inform the debate over whether the mental representations used to recognize speech are based on acoustic features or articulatory gestures (Fowler, 1984; Liberman & Mattingly, 1985; Nearey, 1997; Ohala, 1996; Lotto & Kluender, 1998; Viswanathan et al., 2010). Consistent with models arguing for auditory-based representations (Nearey, 1997; Ohala, 1996; Lotto & Kluender, 1998), the results presented here reveal cues that are not tied specifically to feature dimensions and suggest that relying solely on feature-based approaches may not provide a complete description of the information available to a listener.

The results also confirm that cues to fricative identity are talker-specific. In general, multiple cues were needed to connect talkers. Combined with the approach of identifying cues specific to a phoneme category—instead of a phonological feature dimension—this suggests a change in the way we think about phonetic cues. Rather than high spectral variance (M2+) providing a cue to sibilance, the Steiner trees reveal that this cue dimension is used, in particular, for signaling /v/ and /ð/, and even then, this cue by itself does not provide enough information to account for variability in production across all talkers (Figs. 4 and 5). Similarly, differences along an acoustic dimension do not always indicate contrasting phonological features (e.g., a high f0 indicates a voiceless phoneme, but a low f0 does not necessarily indicate a voiced phoneme). Instead, the current

results suggest that we might describe cues to specific phonemes as being present or absent in individual speech sounds. Such a description suggests a critical rethinking of how we define phonetic cues.

Classifying cues as high/low values

The implementation of the Steiner trees presented here groups cues into high and low values along each dimension. Grouping cue values in some way is necessary to create nodes in the graph, and there is a cost to dichotomizing the cue dimensions in this way. Indeed, by converting from continuously valued cue dimensions to discrete cues that are present or absent in each sound, we lose information. This can be seen by directly comparing the accuracy of the all-cue C-CuRE classifier in the current study (87.3%) with the mean accuracy from McMurray & Jongman (92.9%). Moreover, given that listeners are highly sensitive to fine-grained acoustic differences in speech (Andruski et al., 1994; McMurray et al., 2002; Toscano et al., 2010), careful consideration should be made about this choice.

First, note that the use of high/low values along each cue dimension is not a fundamental limitation of the technique, but rather a simplification we made for the current study in order to evaluate the approach and to derive the simplest description of the relevant cue dimensions. Alternatively, we could have divided cue values into quartiles, into ten bins along each dimension, or into 100 bins, and so on. Indeed, we could divide a spectrogram into any number of spectro-temporal features, allowing us to reduce a very large set of potential cues to a smaller subset. This approach would result in a greater number of cue nodes in the network, and would likely lead to improved performance for the classifiers. Moreover, this may allow us to draw more precise conclusions about the informativeness of specific cues in the classifiers, since each cue would represent a smaller range of cue values along that acoustic dimension, which may be informative for a specific subset of phonemes. A more precise parsing of the cue values, however, would have also resulted in a much more complex description of which cue dimensions are most informative for a given phoneme. Thus, depending on the goal of the study, different choices about how to discretize the cue values can be made (including not discretizing them at all beyond the level of precision offered in the original measurements). Alternatively, other approaches use data generated from the distributional statistics of acoustic cues (Toscano et al., 2010; Kleinschmidt, 2019), which allows cues to be defined based on continuous dimensions and may be advantageous for addressing certain questions.

Nonetheless, the results presented here demonstrate that, even with the most sparse cue specification (i.e., only two possible cue values for each dimension), the cue dimensions identified in the Steiner trees are informative, and the classifier provides a close match between model and listener performance, even though overall accuracy for both sets of classifiers is lower than with continuously-valued cues. Because we are looking at high vs. low values (denoted as either having a positive or negative z-score), it is all the more revealing that this approach pulls out relevant information. Z-values close to zero are not very informative but are necessarily discretized as either high or low (e.g., a z-score of 1 and a z-score of 0.01 are both coded as positive); as a result, less informative cue values could potentially lead to poorer model performance. The fact that the classifier still performs well despite this suggests that this isn't particularly problematic for the classification problem studied here. This may be partly attributable to the use of C-CuRE: because contextual variability was factored out of the cue values, the residual distributions would be expected to show more distinct clusters around the phoneme categories (cf. Cole et al., 2010). Additionally, no model found both a negative and a positive cue to be informative for a given fricative, indicating that, while individual talkers vary in which cue dimensions they use, they do not seem to contradict each other in their use of specific cues. This might have occurred, for instance, if cue values were closely distributed around 0.

Future work aimed at examining more complex graphs with multiple cue values along a dimension may yield more specific cues. The current results provide a baseline for such work; with more bins along each cue dimension, cue values would be represented a more continuous way, providing more detailed information about the range of cue values that are informative for each phoneme and group of talkers.

Comparison with models of speech categorization

The approach used here shares several principles with current models of speech perception, such as the need to account for differences between talkers (Kleinschmidt & Jaeger, 2015) and the use of C-CuRE to factor out contextual variability (McMurray & Jongman, 2011). The current results also argue for the use of multiple cues for distinguishing different phonemes, a key principle in models of speech categorization and perceptual studies (Cole et al., 2010; Escudero & Boersma, 2004; Holt & Lotto, 2006; Kim et al., 2017; McMurray & Jongman, 2011; Oden & Massaro, 1978; Nearey, 1997; Smits, 2001; Toscano & McMurray, 2010). By finding multiple solutions to the Steiner tree problem, we discovered that there can be multiple low-weight solutions for a given phoneme's graph, suggesting that a variety of cues provide useful information.

Would other approaches yield similar results? For instance, we could use regression to identify subsets of informative cues, as in Nearey and colleagues pattern

recognition models (Nearey & Assmann, 1986; Nearey, 1997; Hillenbrand & Nearey, 1999). We could also apply a simple threshold to the edge weights or calculate the amount of information in specific talker-cue connections. It is likely that these approaches would lead to similar sets of cues, and it could be useful to compare subsets of cues identified with these methods with those identified in the Steiner trees. Some key differences might emerge however, given differences between these approaches. For example, consider a case where a talker uses a somewhat atypical set of cue values to signal a given phoneme (relative to the other talkers in the population) but does so reliably. The best cue for this talker (as determined, for example, by their lowest absolute edge weight) would not necessarily be included in the Steiner tree, unless incorporating it resulted in the lowest overall edge weight for the graph (i.e., adding an additional cue node into the network could increase the total network weight more than if that talker were connected via a cue that is already present in the network). Because the Steiner tree method seeks to minimize the total edge weight for the entire network while still connecting individual talkers, it could arrive at a different—and possibly reduced—set of cues.

We can also consider this approach in comparison to standard connectionist networks. One of the main advantages of the Steiner tree method is that it allows us to identify sub-graphs, which is useful for reducing the set of cues needed to connect all talkers in the graph. Translating this problem into a connectionist framework would require us to build a network in which cues are connected to each other and talkers are connected to possible cues. This differs from a typical input-hidden-output layer structure, but it could still be implemented in a neural network. Weights could be set based on the strength of connections between cues and talkers, similar to the way they were set in the graphs used here, but it is unclear how to prune away cue nodes to arrive at a subset of cues. The Steiner tree approach, on the other hand, provides precisely this information. Thus, while connectionist approaches are valuable for studying problems in speech perception, as demonstrated by the success of models like TRACE (McClelland & Elman, 1986), the Steiner tree method is better suited to the specific questions addressed in the current study.

Lastly, we can consider these results in light of models that address the problem of talker variability, Here, we adopt the C-CuRE approach to factor out talker-level variability, but even after doing so, we still find variation in cue use across talkers. The ideal adapter framework (Kleinschmidt & Jaeger, 2015) provides another approach for handling talker-level variability in speech. It can be used, for example, to quantify the information available in phonetic cues for distinguishing indexical characteristics of talkers (Kleinschmidt et al., 2018) and to determine which of those factors should be accounted for to improve speech perception (Kleinschmidt, 2019). This provides a method for measuring talker-level variability in a way that could yield insights into why certain groups of talkers cluster together in the Steiner trees. It also provides a way to quantify how informative talker differences are for specific cues (e.g., talker identity provides more information about formant frequencies than VOT; Kleinschmidt, 2019). This could be used to determine which phonemes we would expect to show simpler Steiner trees (e.g., phonemes for which talker differences do not affect the primary cue, resulting in all talkers connected to a single cue node) versus more complex Steiner trees.

The take home message is that different models provide complementary tools for addressing related questions about talker variability in speech: C-CuRE factors out talker differences to achieve accurate speech recognition, the ideal adapter framework quantifies how talker-level differences affect phonetic cues and identifies which factors a listener should take into account when perceiving speech, and the Steiner tree approach provides a way to identify the relevant cue dimensions while accounting for variability across a population of talkers.

Future work and additional tests

Could the Steiner trees themselves serve as models of listeners' perceptual representations? Given that the Steiner cues provide the closest match to responses from human listeners (relative to the all-cue classifiers we evaluated), these trees may capture the same information used by listeners to perceive speech, including the connections between cues and individual talkers. Such a model is reminiscent of exemplar models of speech perception (Goldinger, 1998; Johnson, 1997), whereby talker-specific representations are retained. It is also consistent with proposals suggesting that listeners track talker-specific distributions of acoustic cues to deal with between-talker variability (Idemaru & Holt, 2011; Munson, 2011; Xie & Myers, 2017). However, more work must be done to establish whether listeners' representations directly correspond to the Steiner tree solutions (e.g., if a listener knows that Talker *X* uses Cue *Y* to signal /s/, do they primarily use that cue from that talker, rather than other cues?). The techniques introduced here could also be applied to listener data (i.e., to identify listener–cue relationships), and it would be interesting to address whether or not the cues identified for listeners are the same as for talkers.

Furthermore, these techniques could be used to classify other types of speech sounds, such as vowels, and capture variability across other indexical variables, such as

dialectical or sociophonetic differences. Vowels are highly talker-dependent (Cole et al., 2010; Hillenbrand et al., 1995), and thus represent an interesting comparison to the fricatives studied here. It may be that more cues are needed per vowel, but the Steiner tree approach could reveal some small subset of cues that connects all talkers for each vowel. Similarly, these techniques could be used to address coarticulatory effects. Rather than nodes representing talkers, graphs could be created in which the distinguished nodes represent a specific context, such as a particular consonant context for a vowel. The goal would then be to connect all coarticulatory contexts to capture the cues needed across all instances of the vowel.

These methods also offer promise for generating hypotheses in other areas of language research. For example, they may serve as a useful tool for describing the necessary connections between phonological and lexical representations in spoken word recognition. In general, in any space where there are multiple possible connections between entities, Steiner trees can allow us to parse out a subset of information necessary to represent the relationships between them.

Conclusions

In conclusion, the results demonstrate that Steiner trees reveal networks of phonetic cues connecting multiple talkers, and that this approach provides a useful tool for reducing a large set of cues to a smaller, informative subset for a specific phoneme. Overall, the techniques presented here offer a new approach for studying speech sound classification, and they shed light on the information that allows listeners to recognize speech from different talkers on the basis of multiple phonetic cues. Beyond questions about speech perception, these techniques can be used in other domains for studying how a large number of feature dimensions can be reduced while still producing informative descriptions of complex cognitive processes.

Acknowledgements This work was supported by a Herschel Smith Undergraduate Research Fellowship from Harvard University awarded to AMC. We would like to thank Florian Jaeger, Terry Nearey, and several anonymous reviewers for helpful feedback on earlier versions of this work.

Open Practices Statement The code for the models is available in the Supplemental Material.

References

Althaus, E., Polzin, T., & Daneshmand, S. V. (2003). Improving linear programming approaches for the Steiner tree problem. In Jansen, K., Margraf, M., Mastrolilli, M., & Rolim, J. (Eds.) *Experimental and efficient algorithms. WEA 2003. Lecture notes in computer science*, Vol. 2647: Springer.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–187.

Angluin, D., Aspnes, J., & Reyzin, L. (2010). Inferring social networks from outbreaks. In Hutter, M., Stephan, F., Vovk, V., & Zeugmann, T. (Eds.) *Algorithmic learning theory*, (pp. 104–118). Berlin: Springer.

Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., François, J.-M., & Zecchina, R. (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, *108*, 882–887.

Baum, S. R., & Blumstein, S. E. (1987). Preliminary observations on the use of duration as a cue to syllable-initial fricative consonant voicing in English. *Journal of the Acoustical Society of America*, *82*(3), 1073–1077.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PloS One*, *6*(5), e19812.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186.

Chlebík, M., & Chlebíková, J. (2008). The Steiner tree problem on graphs: Inapproximability results. *Theoretical Computer Science*, *406*(3), 207–214.

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167–184.

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429.

Escudero, P., & Boersma, P. (2004). Bridging the gap between l2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, *36*, 359–368.

Garey, M. R., Graham, R. L., & Johnson, D. S. (1977). The complexity of computing Steiner minimal trees. *SIAM Journal on Applied Mathematics*, *32*(4), 835–859.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Haley, K. L., Seelinger, E., Mandulak, K. C., & Zajac, D. J. (2010). Evaluating the spectral distinction between sibilant fricatives through a speaker-centered approach. *Journal of Phonetics*, *38*(4), 548–554.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In Tomasello, M. (Ed.) *The new psychology of language*, (Vol. 2, pp. 211–242): Lawrence Erlbaum.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized/hvd/utterances: Effects of formant contour. *Journal of the Acoustical Society of America*, *105*(6), 3509–3523.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

Hughes, G. W., & Halle, M. (1956). Spectral properties of fricative consonants. *Journal of the Acoustical Society of America*, *28*, 303–310.

Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, *18*, S233–S240.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939.

Jacobs, R. A. (2002). What determines visual cue reliability? *Trends in Cognitive Sciences*, *6*(8), 345–350.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson, K., & Mullenix, J. W. (Eds.) *Talker variability in speech processing*, (pp. 145–165). London: Academic Press.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *108*, 1252–63.

Joyner, D., Van Nguyen, M., & Cohen, N. (2010). Algorithmic graph theory. Google Code.

Kim, D., Clayards, M., & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. *Linguistics Vanguard*, *3*, 1.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition, and Neuroscience*, *34*(1), 43–68.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kleinschmidt, D. F., Weatherholtz, K., & Florian Jaeger, T. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science*, *10*(4), 818–834.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Li, F., Trevino, A., Menon, A., & Allen, J. B. (2012). A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise. *Journal of the Acoustical Society of America*, *132*(4), 2663–2675.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language & Speech*, *29*, 3–11.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception and Psychophysics*, *60*, 602–619.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *30*, 133–156.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [f]-[s] distinction. *Perception and Psychophysics*, *28*, 213–228.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375.

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

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–42.

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In Clements, G. N., & Ridouane, R. (Eds.) *Where do features come from?* (pp. 197–236).

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, *25*, 457–65.

Munson, C. M. (2011). Perceptual learning in speech reveals pathways of processing. PhD thesis, University of Iowa.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241–3254.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347–373.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*(5), 1297–1308.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*, 62–85.

Oden, G., & Massaro, D. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*(3), 1718–1725.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*(1), 57.

Regier, T., Khetarpal, N., & Majid, A. (2013). Inferring semantic maps. *Linguistic Typology*, 17.

Repp, B. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*, 81–110.

Schatz, T. et al. (2015). Articulation Index LSCP LDC2015S12. Linguistic Data Consortium.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scott, M. S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D. Y., & Hallett, M. (2005). Identifying regulatory subnetworks for a set of genes. *Molecular & Cellular Proteomics*, *4*(5), 683–692.

Shadle, C. H., & Mair, S. J. (1996). Quantifying spectral characteristics of fricatives. In *Proceedings of the fourth international conference on spoken language processing*, (pp. 1521–1524).

Smits, R. (2001). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception and Psychophysics*, *63*, 1109–1139.

Stanojevic, M., & Vujoševic, M. (2006). An exact algorithm for Steiner tree problem on graphs. *International Journal of Computers Communications & Control*, *1*(1), 41–46.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics

of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America*, *91*, 2979–3000.

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In Gibbon, D. (Ed.) *Natural language processing and speech technology*, (pp. 14–26).

The Sage Developers (2017). SageMath, the Sage Mathematics Software System (Version 7.4). http://www.sagemath.org.

Toscano, J. C., & Allen, J. B. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research*, *57*, 2293–2307.

Toscano, J., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition, and Neuroscience*, *30*(5), 529–543.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*, 1532–1540.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology. Human Perception and Performance*, *36*, 1005–1015.

Williams, H. P. (2009). *Logic and integer programming*, 1st edn. Springer Publishing Company, Incorporated.

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language*, *97*, 30–46.