Research Article

Psychological Science
1–10

DOI: 10.1177/0956797615609578
pss.sagepub.com

What Comes After /f/? Prediction in Speech Derives From Data-Explanatory Processes

Bob McMurray[1] and Allard Jongman[2]
[1]Department of Psychological and Brain Sciences, University of Iowa, and [2]Department of Linguistics, University of Kansas

Abstract

Acoustic cues are short-lived and highly variable, which makes speech perception a difficult problem. However, most listeners solve this problem effortlessly. In the present experiment, we demonstrated that part of the solution lies in predicting upcoming speech sounds and that predictions are modulated by high-level expectations about the current sound. Participants heard isolated fricatives (e.g., "s," "sh") and predicted the upcoming vowel. Accuracy was above chance, which suggests that fine-grained detail in the signal can be used for prediction. A second group performed the same task but also saw a still face and a letter corresponding to the fricative. This group performed markedly better, which suggests that high-level knowledge modulates prediction by helping listeners form expectations about what the fricative should have sounded like. This suggests a form of data explanation operating in speech perception: Listeners account for variance due to their knowledge of the talker and current phoneme, and they use what is left over to make more accurate predictions about the next sound.

Keywords

speech perception, anticipation, predictive coding, generative models, social expectations, auditory processing, open data

Received 12/13/14; Revision accepted 9/10/15

Speech perception poses two challenges. First, acoustic cues are highly variable because of talker differences, overlap between phonemic segments (coarticulation), and differences in speaking rate. Second, speech arrives over time—cues are short-lived and occur asynchronously. While most listeners solve these problems effortlessly, how they do so remains elusive.

With respect to the variability of acoustic cues, the starting assumption is often a bottom-up approach in which cues are mapped to units such as phonemes via boundaries or templates (Nearey, 1990; Oden & Massaro, 1978). However, such approaches are not consistently successful because of the substantial variability in speech (Blumstein & Stevens, 1979; McMurray & Jongman, 2011; Smits, 2001). A compelling alternative is that the perceptual system engages in something akin to *data explanation*. In this view, the bottom-up sensory signal is not the sole basis of perception. Rather, listeners use what they know to account for or explain what they have heard thus far. As these properties are tagged, the remainder has less variance and is used for further inference. The degree to which an explanation (or prediction) fails to fully account for the input suggests that other factors may have shaped it. Rather than perception proceeding from input to higher-level representations, perception is a constant comparison between the input and expectations.

Data-explanatory accounts have been largely developed in research on vision and motor control (Rao & Ballard, 1999; Rhodes & Leopold, 2011; Wolpert & Flanagan, 2001). In speech, the fact that every aspect of the signal is simultaneously the product of multiple factors (talker, neighboring phonemes) makes data explanation more compelling. Indeed, similar principles appear

Corresponding Author:
Bob McMurray, University of Iowa, Department of Psychological and Brain Sciences, 11 Seashore Hall East, Iowa City, IA 52242
E-mail: bob-mcmurray@uiowa.edu

in various theories (Fowler & Smith, 1986; Gow, 2003; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011; Smits, 2001). They suggest that if listeners know some of the factors underlying the form of a segment (e.g., the talker was female), they can account for this variability to extract additional information. For example, while the absolute pitch of a segment is only moderately useful for identifying consonant voicing, if listeners know a man is speaking, they can identify the current pitch as high for a man, which makes pitch a more informative cue for voicing.

There is partial evidence that listeners use data explanation to solve the problem of acoustic variability. Computational models, such as *computing cues relative to expectations* (C-CuRE), demonstrate how data-explanatory principles improve categorization of highly variable corpora of phonetic measurements (Cole, Linebaugh, Munson, & McMurray, 2010), and such models yield listener-like accuracy levels and patterns of errors (McMurray & Jongman, 2011). Empirical studies also suggest that expectations about talkers bias listeners' categorization of ambiguous phonemes (Hay & Drager, 2010; Johnson, Strand, & D'Imperio, 1999); however, it is not clear that these expectations improve accuracy—as opposed to simply shifting categorization boundaries—in the context of variable speech. Thus, there is only partial evidence that data-explanatory processes are how listeners attain accuracy when faced with acoustic variability.

The problem of the temporally unfolding signal seems to demand different solutions. Classic accounts suggest that listeners solve this problem by gradually accumulating partial evidence for multiple candidates in parallel (McClelland & Elman, 1986). Listeners may also cope with the problem of time by predicting upcoming material. Phonetic analyses suggest that substantial coarticulatory information precedes any phoneme (Beddor, Harnsberger, & Lindemann, 2002; Daniloff & Moll, 1974), and listeners use these fine-grained details to anticipate the next phoneme (Gow, 2001; Martin & Bunnell, 1981; Salverda, Kleinschmidt, & Tanenhaus, 2014; Yeni-Komshian & Soli, 1981).

In the experiment reported here, we asked what mechanisms underlie prediction. A bottom-up approach to prediction is simple and compelling. Such mechanisms could use coarticulatory detail to activate upcoming phonemes earlier or more efficiently, and the time demands on prediction might argue for a rapid, autonomous process. Alternatively, prediction may derive from data-explanatory processes. In these schemes, as aspects of the signal (the current phoneme) are identified, variance that is not accounted for may signal upcoming phonemes. Consequently, prediction is simultaneously an expectation about future material and the residual of data explanation. If predictions derive from (and participate in) data explanation, this mechanism would offer a unifying solution for the problems of both acoustic variability and time.

A critical test of such an account is whether high-level knowledge about a segment enhances the accuracy of predictions. Such knowledge should not directly indicate the upcoming sound (as in classic top-down effects). Rather, these cues should inform listeners about sources of variability only in the current phoneme, enabling them to extract more predictive information from it for further judgments.

In our experiment, participants heard isolated fricatives such as /s/ and /ʃ/ and predicted the upcoming vowel. While fricatives contain substantial coarticulatory detail (Daniloff & Moll, 1974; Jongman, Wayland, & Wong, 2000), these cues are variable and inconsistent. However, if participants have expectations about what a specific fricative should sound like (e.g., knowing it was an /s/ spoken by a man), the value of the information in the signal for the vowel should increase. We manipulated expectations by showing or not showing participants a static picture of the talker and the orthographic label of the fricative. This visual information may allow participants to form more precise expectations about what the fricative should sound like, improving detection of subtle deviations from these expectations that signal the upcoming vowel.

Method

Participants

Participants were 84 undergraduates at the University of Kansas. Forty-one were assigned to the no-expectations group, which did not receive any visual stimuli; 43 were assigned to the face+letter group, which did. An additional 4 participants were run in the no-expectations group but excluded from analysis for being native speakers of more than one language. We attempted to screen participants in advance for knowledge of languages other than English. However, we did not discover that these participants were multilingual until they completed the language-background questionnaire after the experiment. Participants received course credit for participation and gave informed consent in accordance with the University of Kansas Institutional Review Board protocols. A sample size of 40 in each group was targeted at the onset of the study on the basis of our experience conducting similar work. Several extra participants were run in case any needed to be excluded. The data were not examined until all data were collected, and all participants (other than the multilinguals) were included in the analysis.

Design and stimuli

Participants heard short segments of frication noise (e.g., /s/ by itself) that had been excised from complete recordings of fricative-vowel utterances. Their task was to predict which of four vowels would come next. The primary factor of interest was whether participants' expectations affected their accuracy. Expectation was manipulated between participants either by presenting the auditory stimulus (the frication noise) in isolation (the no-expectations condition) or by preceding it with a visual stimulus consisting of the still face of the talker and the letter corresponding to the fricative (the face+letter condition). In the face+letter condition, the visual stimulus did not provide any direct information about the upcoming vowel.

Three stimulus factors were manipulated within participants: the fricative itself (we tested all eight fricatives of English: /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, /ʒ/), the talker (10 talkers; 5 female, 5 male), and the vowel from which the fricative originally was excised (four vowels: /i/, /u/, /æ/, /ɑ/). These factors maximized the generality of our findings by forcing participants to cope with substantial variability to accurately anticipate the vowels.

Auditory stimuli were drawn from a corpus of fricatives recorded and phonetically analyzed by Jongman et al. (2000), which we have used in two prior perceptual studies (Apfelbaum, Bullock-Rest, Rhone, Jongman, & McMurray, 2014; McMurray & Jongman, 2011). The original recordings consisted of fricative-vowel pairs comprising all eight fricatives of English, spoken by 20 talkers, followed by six different vowels. For the present experiment, we randomly chose 10 talkers and used only the four corner vowels. Jongman et al. recorded three repetitions of each fricative; we chose the second repetition for the majority of the stimuli and the first or third if that token was unclear.

After selecting the 320 tokens (10 talkers × 4 vowels × 8 fricatives), we removed the vocoid to create the final stimuli—fricatives in isolation. For voiceless fricatives, the vocoid was cut at the first evidence of any periodicity in the waveform; for voiced fricatives (which show periodicity during the frication), the vocoid was cut at the first point where high-frequency frication energy decreased substantially. Phonetic analysis did not reveal any vocalic formants present in the final stimuli.

Faces were drawn from an Internet library. They were converted to gray scale, and each was randomly assigned to one of the original talkers of the study. Faces were selected on the basis of a pilot norming study to ensure that each was representative of its gender (i.e., that it looked extremely masculine or extremely feminine). The assignment of faces to auditory stimuli preserved gender (e.g., female faces were assigned only to female voices). Faces contained no articulatory information, nor did they provide any information concerning vocal tract size. Rather, faces primarily provided participants with knowledge of the talker's gender. Secondarily, since faces consistently appeared with fricatives from the same talker, they provided a cue as to which fricatives should sound similar. In the face+letter condition, participants saw the common orthographic representations of fricatives (e.g., "S," "SH," "TH"), with the exception that "DH" was used to indicate /ð/, and "ZH" was used to represent /ʒ/.

Procedure

Stimuli were presented via headphones. On each trial in the no-expectations condition, participants heard a single fricative and indicated whether they thought the missing vowel was /i/ as in *bead*, /æ/ as in *bad*, /u/ as in *booed*, or /ɑ/ as in *bod* by pressing the appropriate button on a response box. No feedback was given after this response, and the next trial began 500 ms afterward. Button order was counterbalanced across participants. After 20 practice trials, three repetitions of the 320 tokens were presented in random order (960 total trials). The entire experiment took approximately 1 hr.

The face+letter condition was identical to the no-expectations condition, except that each trial started with a fixation point (500 ms) followed by a picture of the face of the talker as well as the letter (or letters) representing the fricative on a computer screen (participants were instructed on these letter strings prior to the experiment). One second after the presentation of face and letters, the auditory stimulus was presented. Visual information remained on the screen until participants responded. (See Videos S1 and S2 in the Supplemental Material available online for sample trial sequences in the no-expectations and face+letter conditions, respectively.)

Statistical analysis

Data were analyzed with logistic mixed-effects models predicting the accuracy of the response (correct vs. incorrect). The fixed effects were as follows: expectancy condition (face+letter = +.5, no expectations = −.5; then centered), talker gender (+.5 = male, −.5 = female), and vowel, which was contrast coded as two variables: height (−.5 = high: /i/, /u/; +.5 = low: /æ/, /ɑ/) and frontness (+.5 = front: /i/, /æ/; −.5 = back: /u/, /ɑ/). Fricative was the last fixed effect. It had eight levels and was coded with three contrast codes. The first two reflected voicing (+.5 = voiced: /ð/, /v/, /z/, /ʒ/; −.5 = voiceless: /θ/, /f/, /s/, /ʃ/) and sibilance (+.5 = sibilant: /s/, /z/, /ʃ/, /ʒ/; −.5 = nonsibilant: /f/, /v/, /θ/, /ð/). The third contrast reflected the relative place of articulation within a sibilant

class (+.5 = labiodentals and coronals: /f/, /v/, /s/, /z/; −.5 = interdentals and post-alveolars: /θ/, /ð/, /ʃ/, /ʒ/).

Because of difficulties in fitting the models with all of the fixed effects simultaneously, we created two mixed models testing different sets of fixed effects. These simplified models, while not ideal, are justified because the effects that were split across models were all within participants, orthogonal to each other (there was no shared variance), and not the primary experimental factor. The first model (the fricative model) examined the effect of information in the stimulus (the three fricative contrast codes as well as the gender of the talker). The second model (the vowel model) examined properties of the response (the two vowel codes). In both models, each fixed effect was added to the model along with its interaction with expectancy condition. Higher-order interactions were not included because these models did not converge. Participant was the only random effect with enough levels to be estimated (there were only 10 talkers, and these were split among the fixed effect, gender). We used the maximal random-effects structure with random slopes of gender, fricative, and vowel on participant (in the relevant models).

The significance of the intercept was used to determine whether prediction accuracy within a condition was greater than chance. Typical statistical tests on the intercept for a logistic model compare the coefficient with 0, which assumes a chance level of .5. However, chance here was .25. Thus, to evaluate the model predictions against .25, we added ln(3)[1] to the original intercept (this will be reported as the adjusted intercept). We then computed a Wald *Z* statistic by dividing the adjusted intercept by the original standard error estimated from the model. Thus, this tested whether the mean performance exceeded 25% correct responses. Models were implemented with the LME4 package (Version 1.1-7; Bates & Sarkar, 2011) in the R programming environment (Version 3.2.0; R Development Core Team, 2008).

Results

Prediction with no context

Before examining the effect of context, we first focused on whether participants were able to anticipate vowels at greater-than-chance level in the no-expectations condition alone. The results of the statistical models are shown in Table 1, and data are shown in Figure 1. For both models, the adjusted intercept was highly statistically significant ($p < .0001$), which indicates above-chance prediction accuracy. While the overall magnitude of the prediction was small ($M = 32.28\%$ correct; $SD = 6.3\%$), 37 of 41 participants predicted the vowel with a level of accuracy that was numerically higher than chance (and 23 participants exceeded 30%). This finding suggests a small—though widespread—ability to predict the vowel from the preceding fricative alone.

Table 1. Results of the Models Examining Prediction Accuracy in the No-Expectations Condition

Model and predictor	*b*	*SE*	*Z*	*p*
Fricative				
Intercept (original)	–0.760	0.046	—	—
Intercept (adjusted)	0.339	0.046	7.40	< .00001
Gender	0.117	0.030	3.86	.00011
Fricative voicing	–0.171	0.032	–5.41	< .00001
Sibilance	0.036	0.026	1.39	.17
Place within sibilance	0.148	0.035	4.18	.00003
Vowel				
Intercept (original)	–0.788	0.046	—	—
Intercept (adjusted)	0.311	0.046	6.79	< .00001
Vowel height	–0.285	0.097	–2.93	.0034
Vowel frontness	0.287	0.094	3.04	.0024

We also found a number of effects of properties of the stimulus. The fricative model revealed that vowels were predicted more accurately following voiceless fricatives than voiced fricatives ($p < .00001$) and that fricatives from male talkers yielded more accurate predictions than fricatives from female talkers ($p = .00011$). Further, while there was no overall difference between accuracy following sibilants and nonsibilants, there was a main effect of place within fricatives ($p = .00003$), which suggests that particular fricatives appear to be better carriers of coarticulatory information than others (overall, the labiodentals and coronals, /f/, /v/, /s/, /z/, were superior to the interdentals and post-alveolars, /θ/, /ð/, /ʃ/, /ʒ/; Fig. 1a). The vowel model found a significant effect of vowel height ($p = .0034$) and vowel frontness ($p = .0024$), which suggests that more information was conveyed by the fricative for certain vowel targets (high vowels and front vowels) than for others. The fricative preceding the high front vowel /i/ in particular led to quite good prediction (~40% correct; Fig. 1b). Overall, there is clearly substantial information in the signal that participants can use to anticipate the vowel, even as different vowels appear to influence the fricative more, and different fricatives appear to carry that information more clearly.

Effect of expectancy condition

Our primary analysis examined the effect of expectancy condition on prediction accuracy. The results of these analyses are shown in Table 2 and Figure 2. Again, the adjusted intercepts were highly statistically significant in both the fricative and vowel models ($p < .00001$), which confirms that participants performed above chance when predicting the vowel. A similar pattern of main effects as in the no-expectations condition alone was also observed.

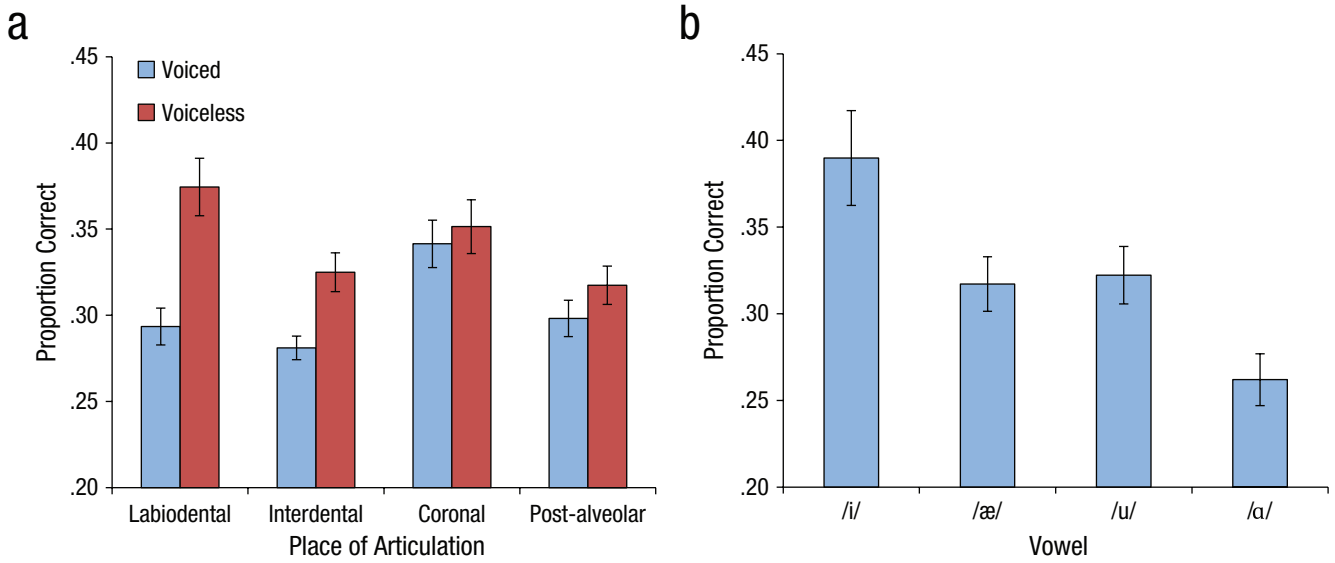

Fig. 1. Mean proportion of correct responses in the no-expectations condition (n = 41) as a function of (a) place of articulation and voicing of the fricative and (b) vowel sound being predicted. Error bars indicate ±1 *SEM*.

There was a significant effect of talker gender ($p < .00001$, with better prediction for male than for female talkers), voicing ($p < .00001$, with better prediction following voiceless than voiced fricatives), and place of articulation within sibilance class ($p < .00001$, with better prediction following labiodentals and coronals than interdentals and post-alveolars). This time, the main effect of sibilance was also significant ($p = .0015$), with better prediction for the sibilants than the nonsibilants. As before, some vowels were easier to anticipate than others, with main effects of vowel height ($p < .00001$; high vowels were more accurately predicted than low vowels) and frontness ($p < .00001$; front vowels were more accurately predicted than back vowels). Thus, the same pattern of signal-driven effects observed in the no-expectations condition held in the full data set.

Table 2. Results of the Primary Analyses Examining the Effect of Expectancy Condition on Prediction Accuracy

Model and predictor	*b*	*SE*	*Z*	*p*
Fricative				
Intercept (original)	–0.678	0.035	–19.51	< .001
Intercept (adjusted)	0.420	0.035	12.09	< .00001
Expectancy condition	0.167	0.069	2.41	.0161
Gender	0.173	0.022	7.88	< .00001
Fricative voicing	–0.225	0.023	–9.84	< .00001
Sibilance	0.054	0.017	3.18	.0015
Place within sibilance	0.201	0.024	8.29	< .00001
Expectancy Condition × Gender	0.117	0.044	2.67	.0075
Expectancy Condition × Voicing	–0.114	0.046	–2.50	.0123
Expectancy Condition × Sibilance	0.035	0.034	1.03	.3
Expectancy Condition × Place Within Sibilance	0.107	0.048	2.20	.0279
Vowel				
Intercept (original)	–0.702	0.036	–19.68	< .001
Intercept (adjusted)	0.397	0.036	11.13	< .00001
Expectancy condition	0.176	0.071	2.48	.0132
Vowel height	–0.345	0.062	–5.59	< .00001
Vowel frontness	0.327	0.060	5.49	< .00001
Expectancy Condition × Vowel Height	–0.123	0.122	–1.00	.3
Expectancy Condition × Vowel Frontness	0.079	0.118	0.67	.5

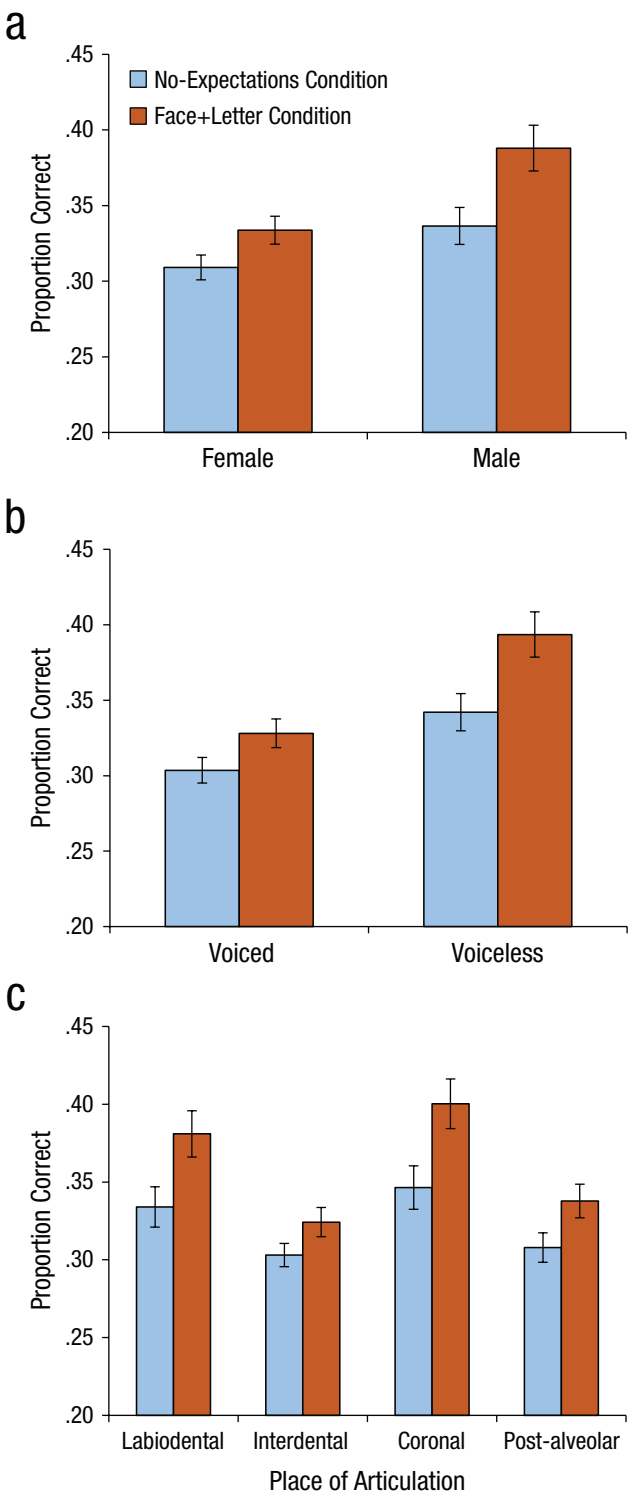

Fig. 2. Mean proportion of correct responses as a function of (a) talker's gender, (b) voicing of the fricative, and (c) place of articulation of the fricative, separately for participants in the no-expectations (n = 41) and face+letter (n = 43) conditions. Error bars indicate ±1 *SEM*.

Most important, the effect of expectancy condition was significant in both the fricative model (p = .0161) and the vowel model (p = .0132). Participants in the face+letter condition anticipated the vowel more accurately than those in the no-expectations condition. As Figure 2 shows, with context to guide expectations, performance rose from 32.3% correct to 36.1% correct (*SD* = 7.8%). By and large, this effect of expectancy condition was not strongly moderated by the other factors we manipulated. There were no interactions of expectancy condition with properties of the vowel, and only gender, voicing, and fricative frontness interacted with expectancy condition.

These interactions were examined in separate follow-up analyses. Since there were no interactions with vowel properties, these analyses were run using a reduced version of the fricative model. While all of the fixed effects from the fricative model were included in each model, we report only the main effect of expectancy condition.

To understand the interaction of expectancy condition and gender, we ran separate models for male and female talkers. These found the main effect of expectancy condition to be significant for both genders (male: b = 0.226, SE = 0.086, Z = 2.623, p = .0087; female: b = 0.110, SE = 0.056, Z = 1.97, p = .049). This suggests that the Gender × Expectancy Condition interaction was driven by the stronger effect of expectancy for male voices, even though both voices showed the effect (Fig. 2a). The analysis of the Expectancy Condition × Voicing interaction showed a similar pattern: The effect of expectancy was significant for voiceless fricatives (b = 0.224, SE = 0.084, Z = 2.64, p = .0083) and marginally significant for voiced ones (b = 0.111, SE = 0.059, Z = 1.88, p = .060). Finally, separate analyses at each place of articulation (Fig. 2c) showed significant effects of expectancy condition for labiodentals (b = 0.206, SE = 0.097, Z = 2.36, p = .018), coronals (b = 0.236, SE = 0.094, Z = 2.52, p = .012), and post-alveolars (b = 0.138, SE = 0.065, Z = 2.12, p = .034), as well as a marginal effect for interdentals (b = 0.094, SE = 0.056, Z = 1.69, p = .092). Thus, these interactions were not driven by a reversal or absence of the expectancy effect, but rather by differences in its magnitude.

Across all of these interactions, a clear pattern was observed: The conditions in which the baseline level of prediction was the lowest (female talkers, voiced fricatives, and interdentals) showed the smallest effect of expectancy condition. This suggests that when the raw acoustics of the fricative contain the most bottom-up information, participants are even better able to harness it if they have greater expectations about the talker and fricative. Thus, in accord with data-explanatory accounts, expectations enhanced participants' use of information in the signal, rather than providing information that was not present (as is usually observed in studies of top-down effects). However, this effect was limited to properties of the fricative (which are likely to mask the coarticulatory information); variation in the degree to which different

vowels could be predicted did not interact with expectancy condition. This is also in accord with data-explanatory accounts. The expectations we gave participants help to account for variation in the fricative's usefulness in uncovering the vowel; they do not offer much information that would differentially affect different vowels.

Reaction times (RTs)

Finally, we analyzed RTs to determine whether the availability of additional information makes prediction more efficient or whether this added information result in slower processing. RTs were slow overall, reflecting the difficulty of the task, but they were more than 100 ms faster in the face+letter condition (M = 1,494 ms) than in the no-expectations condition (M = 1,636 ms). However, RTs were variable both across participants (SD = 456 ms), and within participants (mean of the SD = 855 ms). To better understand this variability, we conducted an ex-Gaussian analysis, which models the distribution of RTs as the product of a Gaussian distribution (with parameters μ and σ) and an exponential function (to model the long tails, with parameter τ). We fit this distribution to each participants' RTs (for trials with correct responses only) using a gradient descent algorithm that maximized the log-likelihood of the data. (Fits could not be obtained for 4 participants in the no-expectations condition).

This analysis showed much lower values of μ (M = 710 ms) than the raw means, which suggests that the high average RTs may have been driven by long tails of the distribution. More important, participants in the face+letter condition showed significantly faster values of μ (M = 644.7 ms) than those in the no-expectations condition (M = 787.6 ms), $t(78) = 2.05$, $p = .044$; the two groups did not differ for either σ or τ ($ts < 1$). While there were insufficient data for an ex-Gaussian analysis within various subconditions, a comparison of RTs suggests that for most of the fricatives and vowels, RTs were faster when expectations were available than when they were not (Fig. 3).

Discussion

There were three key findings of this experiment. First, we showed that listeners can use expectations about what a phoneme should sound like to make faster and more accurate predictions about it. Second, these expectations can derive from sources that are arbitrarily related to speech categories and not directly related to what is being predicted. Finally, this expectation effect is greater when the stimulus contains more coarticulatory information to be uncovered. This supports the idea that prediction (and speech perception more generally) derives from processes attempting to explain the input, not simply categorize it. Our results hint at a chain of predictions: Listeners use context to develop expectations for the specific sound of the fricative; when the fricative is heard, they compare these expectations with the input to make further predictions about the vowel to be heard.

Several concerns remain. First, our experiment does not offer a clear picture of the time course of these effects. While eye tracking studies suggest that predictions based on other types of anticipatory coarticulation occur before the predicted phoneme (Gow & McMurray, 2007; Salverda et al., 2014), it is unknown whether contextually mediated prediction operates similarly.

Second, we cannot separate the contributions of the social and orthographic cues. However, we suspect the former are more important. Since the fricatives were unambiguous recordings that can be identified accurately (McMurray & Jongman, 2011), the orthography may not have offered participants new information. In contrast, the talker cannot be easily identified from frication alone. Moreover, single letters do not strongly bias speech perception (Fowler & Dekle, 1991), while talker effects are robust (Strand, 1999). Either way, neither information source directly predicts the vowel, and both are arbitrarily related to the fricative. This suggests that abstract, learned information augments prediction.

Third, we cannot make strong claims about information flow. Contextual expectations could exert a true feedback effect, altering perceptual encoding of the fricative. This would be consistent with predictive-coding accounts from neuroscience (Rao & Ballard, 1999) and with relative-cue-encoding accounts of speech (McMurray & Jongman, 2011). In these accounts, expectations take the form of perceptual-level predictions about what a stimulus should sound like to enable more rapid comparison with the actual input. Conversely, context could participate at later decision stages, biasing the system to select an interpretation (fricative, talker, and vowel identities) that best explains the data, consistent with Bayesian (Kleinschmidt & Jaeger, 2015) and information-integration (Smits, 2001) accounts.

Our study echoes the long-running debate on feedback in speech perception (McClelland, Mirman, & Holt, 2006; Norris, McQueen, & Cutler, 2000) but with an important distinction. Traditional feedback accounts emphasize how high-level information fills in or restores missing or ambiguous sounds, aligning the percept with expectations. Our account instead stresses the lack of alignment or contrast between signal and expectations. For example, an /s/ with a lower-than-expected frequency indicates that a /u/ is next. We showed that context information—which (unlike lexical information) does not directly cue the relevant phoneme—leads to more precise expectations and better detection of the violations.

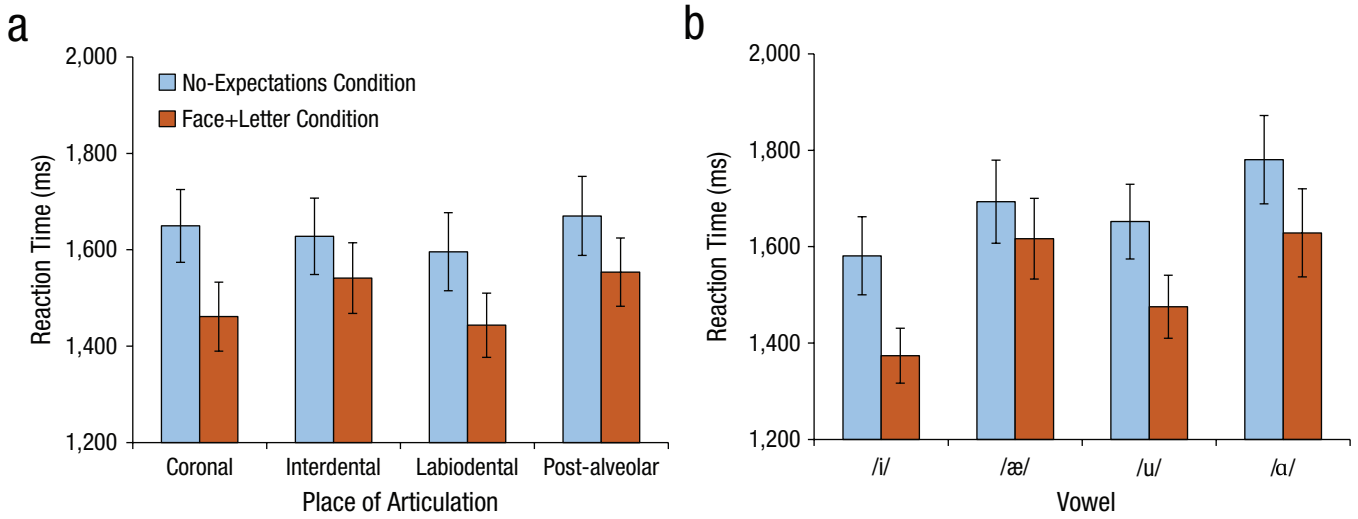

Fig. 3. Mean reaction time as a function of (a) place of articulation of the fricative and (b) vowel sound being predicted, separately for participants in the no-expectations (n = 41) and face+letter (n = 43) conditions. Error bars indicate ±1 *SEM*.

Our findings have implications for the problem of acoustic variability. The idea that listeners are actively forming (and evaluating) hypotheses about talkers and phonemes is a clear prediction of computational models, such as C-CuRE (McMurray & Jongman, 2011), which have been shown to account for much of the acoustic variability in speech, and our results confirm a prediction made by Cole et al. (2010) in this framework. More broadly, our results are consistent with data-explanatory principles in other frameworks (Kleinschmidt & Jaeger, 2015; Smits, 2001) and with neuroscience suggesting independent pathways for processing a talker's voice and for phonological information (von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010). Data explanation also offers an alternative account of findings that knowledge of talker voice influences lexical processing (Goldinger, 1998; Nygaard, Sommers, & Pisoni, 1994); rather than requiring detailed exemplar memory, similar effects could be accomplished by interactions across pathways. This may explain why failure of these interactions is involved in phonological impairments such as dyslexia (Perrachione, Del Tufo, & Gabrieli, 2011).

Our results speak to the problem of time and suggest that the same data-explanatory mechanisms that help listeners account for variability also underlie prediction: Residual variance that cannot be attributed to current expectations must be due to upcoming material. However, expectations may not only function prospectively. In speech (and other areas of perception), contextual factors are not always time-locked to stimuli: Sometimes the talker is known before speech begins; other times, only once they have spoken. A companion study using fricatives from the same corpus shows that contextual expectations can also function retrospectively. Apfelbaum et al. (2014) manipulated the gender and vowel in the vocoid following the fricative—misleading listeners about the source of the fricative after they heard it. This reduced fricative-identification accuracy, which suggests that listeners use later information to form expectations about what a previous phoneme should have sounded like and revise previous (partial) interpretations.

These findings also dovetail with neuroscience on predictive coding that reports reduced neural activity in auditory cortex when inputs match expectations (suggesting a reduced error signal; see Gagnepain, Henson, & Davis, 2012; Houde, Nagarajan, Sekihara, & Merzenich, 2002; and see Blank & von Kriegstein, 2013, for analogues in visual speech). This hints that our effect may be situated in early auditory areas. Our work complements these studies by showing that context-driven expectations do not just make processing more efficient—they make it more accurate. Further, while neuroscience has examined expectations from nearby representations that directly predict specific sounds (lexical processes, speech production, and visual speech), our study demonstrates that expectations can also be based on distal representations that provide only context.

Data-explanatory or generative processes are part of many theories of speech perception. These results help refine understanding of such mechanisms. Recognition by synthesis, a key part of motor theory (Liberman & Mattingly, 1985), evaluates hypotheses about the underlying cause of acoustic input (see also Pickering & Garrod, 2013, for broader applications in language). Our study suggests that contextual information is not solely articulatory, broadening the factors involved in such analysis. Likewise, Bayesian frameworks offer a similar data-explanatory

approach as a form of rational inference (Kleinschmidt & Jaeger, 2015). However, data explanation extends beyond any particular meta-theoretical framing. It can be implemented by something as simple as linear regression (in speech perception; Cole et al., 2010; McMurray & Jongman, 2011) or as distance between normative (prototype) representations and the input (e.g., in face perception; Rhodes & Leopold, 2011).

Data explanation and predictive coding are important theories in visual perception (Rao & Ballard, 1999) and motor control (Miall & Wolpert, 1996; Wolpert & Flanagan, 2001) and have been posited as unifying theories for cognitive science (Clark, 2013). Our work suggests a striking analogue in speech. Listeners evaluate speech relative to perceptual expectations—expectations that can be shaped by high-level knowledge. This comparison helps anticipate future events and enhance perceptual analysis by accounting for, rather than simply categorizing, the input.

Author Contributions

The study was conceived and designed by A. Jongman and B. McMurray. A. Jongman developed the stimuli, implemented the study, and collected the data. B. McMurray analyzed the data. B. McMurray and A. Jongman wrote the manuscript. Both authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by National Institutes of Health Grant No. DC0008089 awarded to B. McMurray.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data have been made publicly available via Open Science Framework and can be accessed at https://osf.io/hjfw9/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received a badge for Open Data. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Note

1. We used ln(3) because it is the log odds ratio when chance is .25: ln(.25/[1 − .25]) = ln(3).

References

Apfelbaum, K. S., Bullock-Rest, N., Rhone, A., Jongman, A., & McMurray, B. (2014). Contingent categorization in speech perception. *Language, Cognition and Neuroscience*, *29*, 1070–1082.

Bates, D., & Sarkar, D. (2011). lme4: Linear mixed-effects models using S4 classes (Version 1.1-7) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*, 591–627.

Blank, H., & von Kriegstein, K. (2013). Mechanisms of enhancing visual–speech recognition by prior auditory information. *NeuroImage*, *65*, 109–118. doi:10.1016/j.neuroimage.2012.09.047

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*, 1001–1017.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36*, 181–204.

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

Daniloff, R., & Moll, K. (1974). Coarticulation of lip rounding. In N. J. Lass (Ed.), *Experimental phonetics* (pp. 100–114). New York, NY: MSS Information Corp.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 816–828.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. S. Perkell & D. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 123–136). Hillsdale, NJ: Erlbaum.

Gagnepain, P., Henson, R. N., & Davis, M. H. (2012). Temporal predictive codes for spoken words in auditory cortex. *Current Biology*, *22*, 615–621.

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

Gow, D. W., Jr. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language*, *45*, 133–159.

Gow, D. W., Jr. (2003). Feature parsing: Feature cue mapping in spoken word recognition. *Perception & Psychophysics*, *65*, 575–590.

Gow, D. W., Jr., & McMurray, B. (2007). Word recognition and phonology: The case of English coronal place assimilation. In J. Cole & J. I. Hualde (Eds.), *Laboratory phonology 9* (pp. 173–200). New York, NY: Mouton de Gruyter.

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics*, *48*, 865–892.

Houde, J. F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the auditory cortex during speech:

An MEG study. *Journal of Cognitive Neuroscience*, *14*, 1125–1138.

Johnson, K. C., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *24*, 359–384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *106*, 1252–1263.

Kleinschmidt, D., & Jaeger, F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*, 148–203.

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

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, *69*, 559–567.

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

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*, 363–369.

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*, 219–246.

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*, 1265–1279. doi:10.1016/S0893-6080(96)00035-4

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

Norris, D., McQueen, J., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral & Brain Sciences*, *23*, 299–370.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*, 42–46.

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

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science*, *333*, 595. doi:10.1126/science.1207327

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral & Brain Sciences*, *36*, 329–347. doi:10.1017/S0140525X12001495

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*, 79–87.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Retrieved from http://www.R-project.org

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 263–286). Oxford, England: Oxford University Press.

Salverda, A. P., Kleinschmidt, D., & Tanenhaus, M. K. (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *Journal of Memory and Language*, *71*, 145–163.

Smits, R. (2001). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1145–1162.

Strand, E. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*, 86–100.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *The Journal of Neuroscience*, *30*, 629–638. doi:10.1523/jneurosci.2742-09.2010

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology*, *11*, R729–R732.

Yeni-Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: Perceptual evidence of fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, *70*, 966–975.