Published in final edited form as:
Lang Cogn Neurosci. 2014 ; 29(9): 1070–1082. doi:10.1080/01690965.2013.824995.

Contingent categorization in speech perception

Keith S. Apfelbaum,
Dept. of Psychology, University of Iowa, E11 SSH, Iowa City, IA 52242, (319)335-0692

Natasha Bullock-Rest[1],
Dept. of Psychology, University of Iowa, E11 SSH, Iowa City, IA 52242, (319)335-0692

Ariane E. Rhone,
Dept. of Neurosurgery, University of Iowa, 1825 JPP, Iowa City, IA 52242, (319)335-7049

Allard Jongman, and
Dept. of Linguistics, University of Kansas, 1541 Lilac Ln., Lawrence, KS 66044, (785)864-2384

Bob McMurray
Dept. of Psychology, Dept. of Communication Sciences and Disorders and Delta Center, University of Iowa, E11 SSH, Iowa City, IA 52242, (319)335-2408

Keith S. Apfelbaum: keith-apfelbaum@uiowa.edu; Natasha Bullock-Rest: natasha_bullock-rest@rush.edu; Ariane E. Rhone: ariane-rhone@uiowa.edu; Allard Jongman: jongman@ku.edu; Bob McMurray: bob-mcmurray@uiowa.edu

Abstract

The speech signal is notoriously variable, with the same phoneme realized differently depending on factors like talker and phonetic context. Variance in the speech signal has led to a proliferation of theories of how listeners recognize speech. A promising approach, supported by computational modeling studies, is contingent categorization, wherein incoming acoustic cues are computed relative to expectations. We tested contingent encoding empirically. Listeners were asked to categorize fricatives in CV syllables constructed by splicing the fricative from one CV syllable with the vowel from another CV syllable. The two spliced syllables always contained the same fricative, providing consistent bottom-up cues; however on some trials, the vowel and/or talker mismatched between these syllables, giving conflicting contextual information. Listeners were less accurate and slower at identifying the fricatives in mismatching splices. This suggests that listeners rely on context information beyond bottom-up acoustic cues during speech perception, providing support for contingent categorization.

Keywords

speech perception; contingent categorization; fricatives; expectation

Correspondence to: Keith S. Apfelbaum, keith-apfelbaum@uiowa.edu.

[1]Current affiliation: Dept. of Communication Sciences and Disorders, Rush University

[2]Here, and throughout, we use the term primary without any larger theoretical claims, simply as a way to describe the fact that a cue like VOT is one of the most important or most reliable cues to voicing.

Contingent categorization in speech perception

The most challenging problem of real-time speech perception is the fact that the acoustic form of any phoneme or word is multiply-determined. The acoustic cues that signal meaningful phonetic differences are influenced by many factors that are not directly related to phonological categories. These factors include speaking rate, talker identity, dialect, and neighboring phonetic material. Voice Onset Time (VOT), for example, is a primary[3] cue to stop consonant voicing (Lisker & Abramson, 1964), but is also affected by place of articulation (Lisker & Abramson, 1964; Nearey & Rochet, 1994), speaking rate (Kessinger & Blumstein, 1998; Miller, Green, & Reeves, 1986), the neighboring vowel (Nearey & Rochet, 1994), the talker (Allen, Miller, & DeSteno, 2003), stress (Smiljanic & Bradlow, 2009), and even whether it derives from a mispronunciation (Goldrick & Blumstein, 2006). Similarly, Jongman and colleagues (Jongman, Wayland, & Wong, 2000; McMurray & Jongman, 2011) analyzed over 20 cues to the eight English fricatives, and found that in addition to marking fricative identity, all 20 cues were also affected by the talker or neighboring vowel or both. Thus, a critical question in speech perception research is how listeners cope with these overlapping influences in the acoustic signal to accurately interpret speech (Fowler & Smith, 1986; McMurray, Cole, & Munson, 2011; McMurray & Jongman, 2011; Mermelstein, 1978; Nearey, 1997; Smits, 2001a, 2001b).

Several early approaches to this problem sought cues in the signal that were invariant across contexts (Blumstein & Stevens, 1979, 1980; Lahiri, Gewirth, & Blumstein, 1984). However, the invariant cues that were discovered did not always generalize well (e.g., cues for place of articulation do not hold up for different values of voicing or syllable position: Blumstein & Stevens, 1979), leading many researchers to abandon the search for invariant cues (Lindblom, 1996; McMurray & Jongman, 2011; Ohala, 1996).

In response, other studies examined *compound cues*, or cues constructed by combining two or more simple measurements. Sussman and colleagues' locus equations (Sussman, Fruchter, Hilbert, & Sirosh, 1998; Sussman & Shore, 1996) offer a cue to place of articulation that is more invariant to effects of neighboring vowel and talker by combining the formant frequency at the onset of the syllable with that at the vowel centroid. Similarly, a number of authors have proposed the consonant/vowel duration ratio as a cue to voicing that is relatively invariant to changes in speaking rate (Boucher, 2002; Port & Dalby, 1982).

While compound cue models are largely based on phonetic considerations, a parallel class of models suggests that auditory cues are processed in terms of their contrast from neighboring cue values as a general principle of the auditory system (Kluender, Coady, & Kiefte, 2003). For example, the third formant frequency (distinguishing /l/ and /r/) may be heard as higher after a low tone, and lower after a high tone (Holt, 2006; Lotto & Kluender, 1998; though see, Viswanathan, Fowler, & Magnuson, 2009). Similarly, the duration of various components of the signal may be treated in this contrast-driven manner (Diehl & Walsh, 1989). These general principles of auditory encoding may mimic effects of more specialized

[3]For vowel trials, we used these generic labels rather than the actual vowel identities because vowel pairing was a between-participants factor. Rather than change the labels before every participant, we identified the vowel-button pairings for each participant on the screen during each vowel identification trial.

compensation processes despite not being specialized for speech perception (Kluender et al., 2003; Lotto, Kluender, & Holt, 1997).

Both of these approaches have had some success in addressing the problem of lack of invariance. Compound cue approaches, when applied to corpora of phonetic measurements, separate categories fairly well. However, such cues are often constructed post-hoc on the basis of researchers' intuitions about what cues are of interest, and what other cues can be used as estimates of contextual factors. As a result, there is often no overarching criterion to determine which cues to combine for new phonological distinctions. Moreover, from a real-time processing perspective, such cues require listeners to delay phonetic decisions until a later point in the syllable. If the CV duration ratio is the cue to voicing, for example, the listener would have to wait until the end of the vowel to make a voicing decision on the preceding consonant. This is inconsistent with empirical work favoring immediacy (McMurray, Clayards, Tanenhaus, & Aslin, 2008; Toscano & McMurray, 2012), although some cues (like locus equations) can be computed earlier in processing than previously thought (Rhone & Jongman, 2012). In contrast, auditory accounts have received widespread, though not universal, empirical support in a number of studies (Holt, 2006; Kiefte & Kluender, 2008; Lotto & Kluender, 1998; though see, Viswanathan et al., 2009). However, such accounts have yet to be developed in a way that permits a test of their sufficiency to classify stimuli based on real phonetic measurements, and in many cases the speech signal doesn't present clear contrasting information, yet listeners still perform well in the face of variation. Additionally, such accounts show context effects only from neighboring acoustic information of similar types. That is, F3 judgments are affected by neighboring information in the F3 region. This neglects the potential for information from other portions of the acoustic signal (e.g., McMurray & Jongman, 2011) or from other levels of processing, like cognitive expectations, the topic of the present paper (e.g., Carden, Levitt, Jusczyk, & Walley, 1981; Drager, 2011; Hay & Drager, 2010; Johnson, Strand, & D'Imperio, 1999; Magnuson & Nusbaum, 2007; Niedzielski, 1999; Nusbaum & Magnuson, 1997; Strand, 1999) to affect perception, despite strong empirical evidence for such influences.

Contingent Categorization and Cue Sharing

The preceding approaches treat speech perception as a fundamentally bottom-up process, where acoustic information only flows forward, with decisions based on single acoustic cues or these cues combined with other, neighboring cues. In contrast, several accounts have argued for a more interactive approach in which listeners simultaneously extract multiple aspects of the acoustic signal, and decisions made about one feature influence decisions about others (Cole, Linebaugh, Munson, & McMurray, 2010; Fowler & Brown, 2000; Gow, 2003; Jongman et al., 2000; McMurray et al., 2011; Pardo & Fowler, 1997; Smits, 2001a, 2001b; Whalen, 1989). This suggests that speech categorization is a *contingent* process, involving both bottom-up and top-down information sources. That is, decisions made for one purpose affect decisions made for another. Conceptually, such models suggest that listeners try to account for all variance in the acoustic signal. While models vary in terms of how they conceptualize the categorization process, and in terms of which processes interact, they offer a roughly similar description. For example, once the listener can identify one acoustic property or cue-value as affected by a neighboring phoneme (for instance), they can

then account for the effect of that coarticulation (e.g. partial this out of the signal) before making other decisions, like identifying the current phoneme (e.g., on the basis of the residual; Cole et al., 2010; McMurray et al., 2011; McMurray & Jongman, 2011). Doing all of this simultaneously helps listeners achieve a correct "parse" of the signal.

Much prior empirical work testing predictions of contingent categorization models derives from work on *cue sharing*: many acoustic cues are affected by multiple phonological features. In stop consonants, for example, VOT is affected by voicing primarily, but secondarily by place of articulation. This raises the possibility that once the listener identifies the place of articulation (on the basis of other cues), they can better interpret VOT with respect to voicing.

Sawusch and Pisoni (1974) examined this by relating listeners' categorization of a two-dimensional continuum varying in both place of articulation and voicing (/b/→/t/) to categorization on each of the two dimensions independently (/b/→/p/ and /b/→/d/). They examined the ability of both additive and contingent models (in which place and voicing judgments were not independent) to fit the data as a whole; contingent models consistently outperformed the additive ones. Later work by Oden (1978), however, showed that the additive Fuzzy Logical Model of Perception (FLMP, Oden & Massaro, 1978) was also able to capture this pattern of data.

The same basic story has been repeated for several other phonetic contrasts. Mermelstein (1978) examined multi-dimensional continua spanning /æ/→/ɛ/ and /t/→/d/ (the words *bad/bed/bat/bet*). Here, duration is a primary cue for voicing, but can secondarily cue vowel differences, while first formant frequency is a primary cue for vowel identity, but also contributes to voicing. Mermelstein found little correlation between voicing and vowel judgments, suggesting additivity. However, Whalen (1989) replicated Mermelstein (1978) with more statistical power and extended it to new contrasts (fricative-vowel interactions, /s/→/ʃ/ and /i/→/u/, where the spectral mean cuing a fricative is affected by rounding in the vowel) and found evidence for non-independence of categorization judgments.

This series of studies launched an extensive debate over how to interpret evidence for contingent categorization. Much of this debate has been conducted in terms of increasingly sophisticated categorization models and large sets of speech sounds that vary in many dimensions. Nearey (1990; see also, Nearey, 1997) reexamined Whalen's (1989) results using the Normalized a Posteriori Probability model (NAPP) and accounted for listeners' categorizations without substantive interactions, as long as they were weakly biased to prefer particular pairs of phonemes (e.g., when they respond /s/ they should also be more likely to respond /u/, regardless of the stimulus; though see, Whalen, 1992). Under this diphone bias, listeners know that /u/ and /s/ are likely to co-occur, but they do not condition their interpretation of the cues for an /s/ on the fact that they chose /u/ for the vowel – they bias their /s/ decision on the vowel they chose (or vice versa). In contrast, Smits (2001a, 2001b) presented another analysis of Whalen's (1989) data, along with several new datasets, and showed that the best fitting model was one in which the interpretation of specific cues to the fricative was biased by the decision made on the vowel.

The final conclusion of this debate appears to favor a contingent account of speech categorization. This debate has benefited significantly from the use of a set of theoretical and computational modeling tools based on logistic regression, (Oden & Massaro, 1978; Oden, 1978; Nearey, 1990; Smits, 2001a, 2001b). However, there have been several limitations to this line of work that prevent drawing a firm conclusion.

First, the empirical results are somewhat opaque, and evidence for or against categorization-contingent processes are derived from complex model fits. There is no clear behavioral marker for contingent categorization. This is problematic, as the theoretical conclusions are dependent on the assumptions of the specific computational models. The similarity of the relevant models can help with this, as the models differ largely on the issue of contingency in categorization, and are quite similar in their other properties. However, more transparent evidence from behavioral paradigms would be valuable.

Second, it is unclear how to implement contingent categorization based on these models. HICAT (Smits, 2001a) implements contingent categorization by conditionalizing the interpretation of cue values on other decisions. For example, the F1 boundary for distinguishing /æ/ from /ɛ/ is dependent on the decision about fricative voicing. This solves the problem (for /æ-ɛ/), but it may not scale up, as how voicing affects F1 for one distinction (/æ-ɛ/) can't be immediately generalized to other dimensions (e.g., place of articulation). HICAT is optimized for a single phonetic contrast (though it can be optimized for any contrast), and neglects the broader problem of simultaneously identifying all of the factors that influence the signal. A model that considers how a factor like voicing influences various cues in general may be more valuable than one in which the influence of voicing on a cue is stored with respect to some other specific categorization. As we describe shortly, recent work (McMurray & Jongman, 2011) proposes such a model.

Finally, this debate has focused largely on interactions of multiple phonetic features in the signal, but has generally ignored other factors that cause variance in the signal, such as speaking rate or talker. Parallel work suggests that talker also takes part in contingent categorization, as listeners' identification of phonetic features is contingent on their identification of the talker. Nygaard, Sommers, and Pisoni (1994) trained listeners to identify a set of talkers by name, and then found subsequent improvements in word recognition when the stimuli were spoken by the same talkers (but less when the voice was novel), although these differences could be attributed to more general perceptual learning.

More impressively, Strand and colleagues (Johnson et al., 1999; Strand & Johnson, 1996; Strand, 1999) demonstrated that simply alerting listeners to the talker's gender (e.g., presenting videos of the speaker talking) affects fricative and vowel identification. Listeners' identification of the talker (or their gender) may also participate contingently in phonetic categorization, something that none of these models (or empirical paradigms) have considered (for similar effects in other domains, see Drager, 2011; Hay & Drager, 2010; Magnuson & Nusbaum, 2007; Niedzielski, 1999; Nusbaum & Magnuson, 1997). This evidence for expectation-driven changes in speech perception is consistent with contingent categorization. However, much of the evidence for this comes from cross-domain work, leaving it unclear whether contingent processing occurs without cross-domain inferences;

that is, can contingent processing arise from auditory information at different points in the signal?

Computing Cues Relative to Expectations (C-CuRE)

McMurray and colleagues (Cole et al., 2010; McMurray et al., 2011; McMurray & Jongman, 2011) proposed an account of contingent speech categorization that addresses several of the aforementioned issues. This account, Computing Cues Relative to Expectations (C-CuRE), suggests that speech is initially coded as multiple continuous cues, such as formant frequencies or durations. These can be mapped directly to categories like phonetic features or talker identity using a similar logistic model as FLMP (Oden & Massaro, 1978; Oden, 1978), NAPP (Nearey, 1990, 1997) or HICAT (Smits, 2001a). However, cues can also be computed relative to expectations about how they should behave in various contexts, making them more flexible. As categories like a neighboring phoneme or talker are identified, individual acoustic cues are recoded as the difference from the expected cue-values for these categories, and this difference serves as the input to the categorization model.

Consider F0, which is a good cue for talker gender, but a weak cue for voicing. Once the talker's gender is identified, F0 can be recoded relative to the expected F0 for that gender (e.g., unusually high for a man), making it more useful for voicing. The use and interpretation of F0 is contingent on other factors, but crucially the effect of such factors on F0 is stored as a general expectation about how F0 behaves, not as a component of specific voicing categories. The fact that male talkers generally have a low F0 can inform every phonetic decision that uses F0.

Conceptually, C-CuRE has much in common with parsing accounts of Fowler (Fowler & Brown, 2000; Fowler & Smith, 1986; Pardo & Fowler, 1997) and Gow (2003), as it attempts to simultaneously account for all sources of variance in the signal. Unlike these accounts, C-CuRE does not make strong claims about representation; it can use talker as a factor (treating it like other sources of expectations, like neighboring phonemes); and it has been formally implemented (using a combination of linear and logistic regression), offering precise and testable predictions.

In an initial test of this model, Cole et al. (2010) found that phonetic measurements processed with C-CuRE offered more power to classify vowels and predict upcoming vowels than did raw cues. More pertinent to the present study, McMurray and Jongman (2011) conducted an extensive test with a corpus of 2,880 fricatives (collected by Jongman et al., 2000). They measured 24 cues for each token in the corpus and computed the predicted category for each using either raw cues or relative cues processed with C-CuRE. These predictions were compared to listeners' categorizations of 240 tokens. The relative cue model performed quite similarly to listeners with accuracy of 87–92.9% (listeners: M=91.2%), and it showed the same pattern of errors across both fricatives and context vowels; the raw cue model was less accurate and showed a poor qualitative fit to error patterns. This model shows both listeners' level of accuracy and many of their errors for a large corpus of natural recordings.

The model also suggests a unique marker for contingent categorization. McMurray and Jongman found that listeners were substantially less accurate when categorizing fricatives in the absence of the vocalic portion ($M_{listeners}$ =76.3%) than with the complete syllable ($M_{listeners}$ =91.2%). Some of this difference likely arose because of secondary cues to fricative identity (like formant frequencies) in the vocalic portion of CV syllables. The model using raw cues offered a fairly close fit to the frication-alone stimuli (M_{model}=69.7–79.2%), but simply adding the secondary cues in the vocalic portion (M_{model}=79.2–85.0%) was not enough to match listeners' performance. Rather, the difference between the frication-only and the complete syllables appeared to be both the addition of the vocalic cues and the availability of information to support relative cue-encoding.

Without the vowel, listeners could not identify the talker or context to contingently categorize the cues in the frication, and were thus forced to rely on raw cues. This suggests that for fricatives, the vowel may uniquely contribute toward contingent categorization by allowing listeners to identify the vowel and talker, which they can then use to better interpret the fricative cues. Indeed, this makes sense as such factors are difficult to identify from the fricative alone (Lee, Dutton, & Ram, 2010).

This offers a compelling account of the behavioral difference that was observed, but this difference alone is not an unambiguous marker of contingent categorization. However, it does suggest that such effects may be visible in fricative categorization accuracy. It also suggests that fricatives may be a useful context for such an investigation because we can separate the portion of the acoustic signal useful for the phonetic categorization of interest (the fricative) from the portion of the signal used to identify the contextual factors (the vowel and talker). That is, while the cues in the frication are affected by both talker and vowel (Jongman et al., 2000; McMurray & Jongman, 2011), they do not contain sufficient information to unambiguously identify these factors; similarly, the vowel contains relatively weak cues for the fricative, but has sufficient information to identify the talker/vowel. Thus, it may be possible to manipulate the vocalic portion to mislead the listener about which talker produced the fricative, or which vowel context it was produced in.

An alternative approach used in many studies is to condition listener expectations on something outside the speech signal, such as a face showing the talker's gender (Johnson et al., 1999; Strand & Johnson, 1996; Strand, 1999) or expectations about the talker's dialect (Niedzielski, 1999). Although such studies show how listeners use knowledge of the talker to condition their categorization, this approach does not let us examine use of more proximal acoustic context, such as neighboring phonemes. Moreover, because the non-auditory expectations in studies using exogenous conditioning information are typically available before the trial, it is possible that expectations generated from information purely within the speech signal could not activate the appropriate representations quickly enough to play a meaningful role in perception, particularly when the conditioning information follows the segment to be categorized. When some version of this has been done, effects are often only seen for ambiguous stimuli, raising concerns about the presence of the effect in natural speech perception. Thus, it is not yet established that purely within-auditory expectations driven by *both* talker and neighboring phonemes can affect perception.

Empirical Paradigm

The present study assessed whether fricative categorization is contingent on the identification of the talker and the neighboring vowel. We constructed fricative-vowel stimuli in which the bottom-up cues to the fricative identity were consistent with the fricative, yet which would mislead listeners about the identity of the vowel and/or talker. In constructing these stimuli, we ensured that a purely bottom-up cue integration model would not predict a decrement in performance due to unfortuitously splicing tokens with weak cues to fricative identity. This was done by modeling performance of specific tokens, and choosing only combinations for which a bottom-up model predicts no performance decline. If listeners interpret fricative cues relative to expectations driven by the vowel and/or talker, this should cause them to err in their categorization on some small proportion of trials, even as the fricative cues were consistent with the right category. We also expect to see slower RTs on these trials as listeners must now resolve conflicting information in the signal. We thus use accuracy and RT in fricative identification given matching or mismatching vowel and talker information as a gauge for whether listeners use top-down expectations to adjust their speech perception.

Experiment

Methods

Participants—Forty-two undergraduates at the University of Iowa participated in this experiment. All were native English speakers with self-reported normal hearing. Participants received course credit or a small payment as compensation.

Design—Stimuli were constructed by cross-splicing the frication and vocalic portions from the CVC recordings in the corpus of (Jongman, et al., 2000; McMurray & Jongman, 2011). While this corpus contains recordings of all eight English fricatives (/f, v, Θ, ð, s, z, ʃ, ʒ/), we only used the voiceless fricatives (/f, Θ, s, ʃ/), as voicing during frication could give an indication of fundamental frequency or other cues to talker.

The primary factors of interest were 1) whether the talker that produced the fricative matched the talker that produced the vocalic portion; and 2) whether the vowel following the fricative matched the vowel from the context of the original fricative production. The secondary cues to the fricative identity (e.g., formant transitions from the onset of the vocoid) always matched the frication presented, so there were never any mismatching bottom-up cues to the fricative identity. Thus any performance decrement should derive from the listener being misled by the vocalic information, which signals a different talker or vowel from the context in which the fricative portion was produced.

We were secondarily interested in whether the effect of mismatching splices differed for sibilants and non-sibilants; sibilants are identified more accurately overall, and are identified quite well independent of vocalic information (McMurray & Jongman, 2011), suggesting that mismatching splices may result in smaller decrements for these tokens as contingent encoding may not be needed for these sounds.

The original corpus contained 20 talkers and six vowels, resulting in thousands of possible combinations (most of which would mismatch on both factors). To reduce the number of possible combinations, we selected four talkers (two male, two female) for our manipulations. Further, to maximize differences, we always crossed talkers across genders and used pairs of point vowels that contrasted along the diagonal (e.g., i/ɑ and æ/u). Crossing by gender and vowel led to four different splicing conditions: 1) *match-both*: the fricative and vowel were taken from different repetitions of the same token from the same talker (e.g. /s/ from M1, saying /su/, coded as <s_{mu}> and /u/ from M1, saying /su/, coded as <u_{ms}>); 2) *mismatch-vowel*: the fricative from one vowel context was spliced with a different vowel from the same fricative context spoken by the same talker (e.g. <s_{mu}+i_{ms}>); 3) *mismatch-talker*: the fricative from one talker was spliced with a matching vowel from the same fricative context, spoken by a different talker of the opposite gender (e.g. <s_{mu}+u_{fs}>); 4) *mismatch-both* the fricative was spliced to a different vowel from the same fricative context, spoken by a different talker of the opposite gender (e.g. <s_{mu}+i_{fs}>). In every case, the vocalic portion was spliced from a different token of the same fricative, so secondary cues in the vocalic portion supported the correct fricative.

We used two male and two female talkers, and always spliced across gender, resulting in four possible talker pairings. For each of these pairings, the vowel splices used either i/ɑ or æ/u. This resulted in eight different talker-vowel pairings. Five participants performed the study in each of these pairing conditions.

The Jongman et al. (2000) corpus has three recordings of each fricative/vowel pairing for each talker; we used all three used here to ensure that any findings were not artifacts of the particular tokens chosen. Our splicing always used different repetitions for the fricative and vowel in order to ensure that the *match-both* condition was also cross-spliced. There were six possible combinations of fricative-vowel splices just considering repetition (e.g. recording1/recording2, recording2/recording1, recording1/recording3, etc.). All six combinations were used for each experimental condition. This meant that for each talker/vowel pairing, there were 4 fricatives x 2 talkers x 2 vowels x 2 talker-match x 2 vowel-match x 6 splice directions, or 384 stimuli. Each participant was assigned to one talker-vowel pairing and was assessed on all 384 trials for that pairing without repetition.

Talker Selection—To select the four talkers used in this study, we wanted to ensure that differences between matching and mismatching conditions were solely the result of mismatches between the fricative and the vowel/talker leading listeners to misidentify the token. However, even though the frication and vowel matched on bottom-up cues to the fricative identity, some spliced tokens could still yield decrements in performance even assuming a bottom-up model, as there was likely variation in the overall quality of the cues to support fricative identity across talkers and tokens. That is, if cues in one portion of the syllable (e.g., the frication) are weak, a listener can rely on cues in the other portion (the vocoid). However, if a frication portion with weak cues is spliced onto a vocalic portion with weak cues, performance will decline. If a set of stimuli are chosen such that the cross-splices often result in such weak-weak stimuli, both a bottom-up model and a contingent encoding model would predict a performance decrement for mismatching splices.

We dealt with this concern by using acoustic measurements to balance the strength of cues to fricative identity in the frication and vowel portions of the stimuli across conditions. Because these tokens were derived from the Jongman et al. (2000) corpus, there were 24 measurements for each token, including 14 in the frication and 10 in the vowel or transition. We used multinomial logistic regression to find the optimal linear combination of the 24 cues to predict fricative identity and then transformed this into a probabilistic prediction about which category is most likely for each token. This combination of cues can be used to determine how likely a given token is to be classified as each of the eight English fricative categories (similar to the raw-cue model of McMurray and Jongman, 2011, as well as to the way FLMP and NAPP combine cues). This set of computations incorporates all available information for phonetic identity, giving the most optimal information for categorization for purely bottom-up cue combination.

To simulate splicing, we subdivided the measured cues in our stimuli (measurements from Jongman et al., 2000, McMurray and Jongman, 2011) between the frication and the vocalic portion (see Table 1). Cues that span the fricative/vocoid boundary (i.e. the spectral moments at the transition) were assigned to the vocoid. We “created” the cross-spliced stimuli by using the values of those cues present in the fricative as the fricative portion of the splice, and the values of those cues present in the vowel as the vocoid portion of the splice and submitted this hybrid stimulus to the model estimated from the unspliced corpus to determine how likely the output is to match the predicted fricative. By examining which talker pairs predict minimal performance changes for cross-splices using bottom-up encoding, we can ensure that our stimuli only predict decrements in the case of contingent processing.

For each possible pair of talkers (one female, one male), we computed every possible splice (given the three repetitions and thus six splice directions) and averaged across splices to determine the predicted output of the model for each vowel and talker pairing. There was a wide range of predictions. In some instances, the bottom-up model predicted a decrement of over 25% for mismatching conditions. In others, it predicted an *increase* in performance for mismatching splices as large as 6.5%. We examined each talker pairing to determine which predicted the most stable performance across conditions and used those talkers.

The selected set of talkers produced highly similar output predictions (Table 2). The *match-both* condition predicted performance of 76.5% correct; for the cross-spliced conditions, predicted performance was quite similar (*mismatch-talker*: 75.3%; *mismatch-vowel*: 77.3%; *mismatch-both*: 76.6%). In several conditions, performance actually slightly *improved* for mismatching splice conditions. A bottom-up cue integration approach thus does not predict decrements for these stimuli. As seen in the lower portion of Table 2, this pattern was quite consistent for all the pairwise combinations of the talkers.

Stimuli—Auditory stimuli were taken from the Jongman et al. (2000) fricative corpus. These recordings were made by native English speakers at Cornell University. Each fricative was embedded in a CVC frame (the final consonant was a /p/). Fricatives were recorded in the carrier phrase “Say ____ again” and isolated (see Jongman et al., 2000 for details).

From these CVp stimuli, fricative and vocalic portions were cut at the fricative/vowel juncture. This was defined as the point at which no more high-frequency frication could be observed in the waveform. To create the mismatching tokens heard in the experiment, a fricative was spliced onto a vowel that mismatched the fricative in either talker identity or vowel identity, or both. The vocalic portion always came from the same fricative context as the fricative it was spliced with (even in mismatching conditions) to ensure that bottom-up cues to fricative identity throughout the token were consistent with the correct fricative. Completely matching conditions (matching on both talker and vowel) were constructed by splicing tokens from two different recordings of the same fricative.

Procedure—On each trial, a screen displayed the arrangement of responses on the button box. The buttons were also labeled with orthographic representations of the four fricatives ("f", "th", "s", "sh"). Before the experiment, participants were given examples of the four fricatives in real-word contexts to ensure they could correctly apply the labels. During the experiment, syllables were played over high-quality headphones, and participants indicated which fricative they heard by pressing one of four buttons on a button box. The next trial began 500 ms after the response.

On a subset of trials ("catch-trials"), participants were asked to identify either the gender or the vowel presented in the token (36 trials of each; approximately 16% of total trials). These trials used different buttons labeled on the button box ("female" and "male" for gender trials; "V1 and V2" for vowel trials[3]), and the display screen changed to indicate the different response required for the trial. Additionally, the background color of the screen changed for these trials to cue participants that a different response was required.

These catch-trials were administered for two reasons: to ensure that listeners continued to attend to the information present in the vowel portion of the stimuli throughout the experiment; and to ensure that only participants who were responding appropriately were included in our analysis. The catch-trials were randomly selected from all splice types; because cues to talker and vowel are relatively weak in the frication portion of the stimulus, cross-splicing was not expected to affect identification performance on these catch-trials.

Results

We first examined the catch-trials to ensure that the participants were attending to the task. Two participants scored quite poorly (below 65% correct) and were excluded from further analysis. The remaining 40 participants scored at least 88% correct on the catch-trials (*M*=98.7%; *SD*=2.0%). For these participants, the cross-spliced catch-trials were identified as accurately as the same-splice items (*match-both: M*=98.7%; *mismatch-talker: M*=98.9%; *mismatch-vowel: M*=98.7%; *mismatch-both: M*=98.5%). Listeners were thus quite adept at identifying the gender of the talker and at identifying the vowel in the syllable.

To analyze the data of the 40 remaining participants, we used mixed effects models (Baayen, Davidson, & Bates, 2008; Jaeger, 2008) implemented in the LME4 package (Bates & Sarkar, 2011). We first conducted analyses examining the overall effect of talker- and vowel-match across all fricatives, looking at both accuracy and RT. We then conducted analyses that examine these effects in terms of sibilance and place of articulation of the

different fricatives used in the study. For every model, either accuracy or RT on each trial was selected as the DV, while *talker-match* condition (contrast coded: +.5 for match, −.5 for mismatch) and *vowel-match* condition (contrast coded: +.5 for match, −.5 for mismatch) were IVs. Further analyses included additional IVs, to determine if the primary results held across different conditions (e.g., fricative class). Models analyzing accuracy used a binomial linking function (a variant of logistic regression), whereas RT models used a linear linking function. Models analyzing RT considered only those trials where an accurate response was given. These data were trimmed to exclude trials with RTs faster than 400 ms or slower than 3000 ms (1.9% of trials); inspection of a histogram of RTs showed that these responses were well into the tails of the distribution. After trimming, RTs were log transformed.

As in any mixed effects model, it is typical to conduct a series of analyses in which models with the same fixed-effects but different random effects are tested to determine the appropriate random effects structure, prior to analyzing the fixed effects. Our models had a number of possible random effects: the participant, the block of trials (one block = 64 test trials), the vowel set used for that participant, the talker of the fricative in the stimulus, and the talker of the vowel in the stimulus. For each of these, random effects could be implemented as either the intercept of the model, or a random slope of the fixed factors. We included intercept random effects terms for each of the above factors, as well as random slopes of vowel- and talker-match by participant (each model offered a significantly better fit than the model with one fewer random effect using χ^2 tests, all $p<.05$). Across all analyses, the maximum correlation between fixed effects was .128. While for the (binomial) accuracy model, p-values could be computed directly from the Z statistic, computing p-values from linear models with random slopes is more difficult. Thus, for the RT analyses we computed p-values by comparing a full model that includes the factor of interest to one with only that factor removed, using the χ^2 test of model comparison.

Talker and vowel effects—Our primary questions were whether each type of mismatch affected accuracy and whether there was an additive or interactive effect of the two forms of mismatch. Figure 1A shows the effect of *vowel-condition* (match/mismatch) and *talker-condition* (match/mismatch) for accuracy, collapsed across all four fricatives, and across all talkers and vowels. There were small effects of both vowel- and talker-mismatch. The *mismatch-talker* condition showed a reduction in accuracy of about 1.6%; *mismatch-vowel* showed a similar-sized decrement of 1.7%; and *mismatch-both* showed a total decline of about 2.0%.

The overall effect of *vowel-match* on accuracy was significant (B=.13, SE=.053, Z=2.4, p=.017; Figure 1A), signaling better performance on trials with matching vowel information. The main effect of *talker-match* was also significant (B=.13, SE=.051, Z=2.5, p=.013), with better performance when the talker of the vocalic portion matched the talker of the fricative. The interaction of vowel- and talker-match was not significant (B=−.13, SE=.10, Z=1.2, p=.22), suggesting an additive effect of vowel and talker mismatches. As predicted by C-CuRE and other contingent categorization approaches, we found that performance was impaired when fricatives were heard in the context of mismatching vowels and talkers.

The RT results closely mirrored the accuracy results (Figure 1B). *Match-both* trials were completed the most quickly (*M*=1112 ms), while both types of mismatch slowed responses (*mismatch-vowel: M*=1145 ms, *B*=−.010, *SE*=.0021, χ^2=15.94, *p*<.0001; *mismatch-talker: M*=1136 ms, *B*=−.0056, *SE*=.0026, χ^2=4.38, *p*=.036). When both mismatched, the slowing was more pronounced (*M*=1153 ms). There was no interaction (*B*=−.0035, *SE*=.0044, χ^2=.62, *p*=.43), suggesting that slowing in the *mismatch-both* condition was approximately additive of the individual mismatch effects.

Effect of mismatch across fricatives—As sibilant fricatives are identified more accurately than non-sibilants, listeners may rely less on context when identifying these stimuli. We thus examined how well our effects held across the different voiceless fricatives used in this experiment. We briefly summarize the major findings here; full statistical analyses are available in Online Supplement S1.

As predicted, the sibilants were identified more accurately and more quickly than the non-sibilants. *Vowel-match* was significant in both accuracy and RT analyses and did not interact with sibilance type suggesting similar mismatch effects for both classes of fricatives. The main effect of talker-match was significant for RT, but not for accuracy. However we did see interactions of place of articulation, vowel-match and talker-match on accuracy. Follow-up analyses suggested that failure of talker-match to reach significance arose because of the pattern of responding for /ʃ/. All fricatives except /ʃ/ were identified most accurately in the *both-match* condition; for /ʃ/, this condition was least accurate (although all conditions were above 95% correct). However, for RT every fricative, including /ʃ/, was identified most quickly in the *match-both* condition. The unexpected accuracy performance for /ʃ/ may have thus been anomalous, likely owing to a ceiling effect. Overall, these analyses suggest that mismatching talker and vowel information exert a more pronounced effect on non-sibilants, where effects are apparent in both accuracy and reaction time. Sibilants are affected, but this is most notable for RT; contingent processing does not appear necessary to accurately identify sibilants, but it speeds the identification process.

Relationships among cues: Phonetic analyses—Our results suggest that mismatching talker or vowel information in the vocoid portion of our stimuli impairs identification of fricatives. We designed our cross-splicing manipulation to maintain bottom-up cues to the fricative even in mismatching conditions, such that detriments are best explained by misattribution from talker and vowel information. However, while this holds constant the relationship among cues and categories, it leaves unexamined the possibility that listeners' expectations about relationships between cues in the frication and vocoid may play a role independent of talker and vowel identity. For example, listeners may expect that spectral means in particular frequencies predict formant frequencies at particular locations. Disrupting these correlations may have disrupted performance. To examine whether such correlations could cause the effect, we conducted a mediation analysis (Baron & Kenny, 1986) on the Jongman et al corpus to determine whether correlations between frication and vocoid cues are mediated by talker and vowel context. The details of this analysis are available in Online Supplement S2. This analysis showed that the frication cues together accounted for about 10.1% of the variance in the vocoid cues (averaged across the 10 cues;

range: 3–24.7%). Talker and vowel identity accounted for over 57% of the variance in the vocoid cues. Moreover, much of the predictability of vowel cues from frication cues was mediated by the identity of the talker and/or vowel: when including these in the model, frication cues only accounted for an additional 1.9% of the variance across cues (range: . 1%–5.4%), or 81% less than when used alone. The large mediating effect of context suggests that much of the relationship between frication and vocoid cues can be accounted for by identifying the context. This makes it unlikely that predictability *between cues*, are powerful enough to drive the effects seen in this study, suggesting that contingent encoding using context is a stronger explanation.

Discussion

This experiment showed that listeners utilize vowel and talker information in the vocalic portion of CV stimuli to identify fricatives. This reliance is not simply based on bottom-up cues to fricative identity within the vowel; our cross-splicing manipulation coupled with careful measurements and a computational approach that allows us to combine measurements into an overall metric of quality ensured that cues in the vowel signaled the same fricative as cues in the frication. Moreover, this cannot be accounted for by disrupting correlations among cues between the frication and the vowel: our phonetic analysis suggests such effects are small at best. Instead, mismatching vowel and talker information changed the way bottom-up information was used by listeners, such that they misinterpreted the cues to frication. This resulted in a decrease in identification accuracy and an increase in reaction time. The effects of mismatching talker and vowel were of a similar magnitude, and mismatching both led to more pronounced decrements in performance.

These results suggest that listeners' phonetic processing is contingent on other judgments about the signal. Listeners are implicitly aware of how different factors affect the speech signal; for example, listeners recognize that females typically produce fricatives with higher spectral characteristics than do males. Upon identifying contextual information, listeners can better interpret the speech segment by attributing variability in the signal to known contextual factors. For example, if a listener is identifying a sound with spectral characteristics between /s/ and /ʃ/, the listener can use the gender of the talker to better interpret the intended segment; if the talker is female, the spectral characteristics appear lower than typical female productions of /s/, leading to a /ʃ/ judgment (Johnson et al., 1999; Mann & Repp, 1980; Strand & Johnson, 1996; Strand, 1999). Using such relative encoding mechanisms can help listeners overcome many of the sources of variability in the signal.

Such expectation-dependent encoding has been shown in cases where listeners are given overt information to establish expectations. For example, showing the listener the face of the talker generates gender-based expectations, which affects fricative identification (Johnson et al., 1999; Strand & Johnson, 1996; Strand, 1999). As listeners identify high-level information about the speech context, they generate expectations about productions. However, it has been less clear whether information from other portions of the speech signal can build these expectations. The rapid nature of speech would require listeners to form expectations extremely quickly, and in some instances use expectations that are generated

after receiving the bottom-up acoustic information to update prior decisions (or encodings of the signal).

Listeners in our study always heard the same pair of talkers and vowels, so they may have formed very precise expectations based on these contexts. In natural contexts, such precision may not be warranted; listeners hear countless talkers throughout their lives, and in many circumstances hear new talkers for whom they have no experience. Indeed, listeners entered this study with no experience hearing our talkers. Rather than contingency occurring on a talker-specific basis, listeners may generate expectations based on coarser information. For example, listeners can identify general production patterns among males and females, or among different accent groups, and then use these classifications to establish expectations. Alternatively, listeners may use a small number of "prototype talkers," and form expectations by comparing new talkers to these reference categories.

The catch trials in this study may have alerted participants to changes in talker/vowel information, increasing the likelihood that mismatching information would affect performance. Although such a concern is valid in the current study, previous studies without such catch trials show evidence of contingent encoding in fricative perception (McMurray & Jongman, 2011). In the present study, the catch trials may have increased the magnitude of the mismatch effect, but likely were not the sole provenance of this effect. Moreover, in natural speech perception listeners simultaneously make judgments about multiple factors (e.g., several phonemes plus the talker); our catch trials may have actually put the listeners in a more natural mode of processing multiple factors at the same time.

Our study shows expectations generated as a result of talker and vowel; however, other factors are also likely to drive expectations. Other factors that affect speech production, such as speaking rate (Summerfield, 1981), dialect or ambient acoustics, can allow listeners to form expectations. As a listener acquires increasingly detailed information about the context in which a sound occurs, she can form increasingly precise predictions about the form that a production will take. The listener can then use deviations from her predictions (i.e. residuals) to identify the current speech token and to tune later predictions for better precision. This form of prospective model places heavy emphasis on prediction in speech perception (see also, Clark, in press, for applications of this idea beyond speech). However, unlike many of these predictive models, our work suggests that such later-occurring information can also affect speech perception (and in the same general format) for perceptual information that has already been heard (through a re-analysis, or re-parsing). This "postdiction" argues against strictly forward-looking predictive accounts of processing.

Our data suggest that speech perception can accommodate rapid expectation generation. In our stimuli, listeners had no access to talker or vowel information until after hearing the frication. Nevertheless, they showed sensitivity to this manipulation, such that misleading contextual information in the vocalic portion impaired performance. Listeners incorporate expectations generated from information later in the speech stream before they make a categorization decision about fricative identity. This suggests quite rapid formation and use of expectations during speech perception. The retrograde direction of these effects seemingly contrasts with evidence of continuous lexical activation from acoustic

information (Allopenna, Magnuson, & Tanenhaus, 1998; McMurray et al., 2008; Zwitserlood, 1989). However, rather than suggesting that decisions are gated until all information is received, this use of later information can be accomplished through immediate, incremental decisions that are updated with later-occurring information. This is consistent with McMurray and Jongman's (2011) finding that in the absence of the vocoid, listeners can identify fricatives (albeit less accurately); listeners may rely on raw cues until information is available to generate a more contingent encoding. In that sense, it is quite consistent with the broader framework of continuous, parallel partial activation used to characterize spoken word recognition (e.g., Marslen-Wilson, 1987; McClelland and Elman, 1986). Although our approach is compatible with such continuous encoding theories, exactly how listeners use these asynchronous cues is still a matter of debate. Ongoing eye-tracking work using the paradigm developed by McMurray and colleagues (McMurray et al., 2008; Toscano & McMurray, 2010) addresses this.

In this sense, this approach instantiates a form of multiple constraint satisfaction (McClelland & Rumelhart, 1981), in which decisions are graded and continuously updated, and the product of multiple bottom-up and top-down constraints. Although activation begins immediately, these activations are malleable with later-occurring information and feedback. However, our approach differs from that of multiple constraint satisfaction and interactive activation in one key respect: whereas interactive activation typically uses feedback to reinforce partially-active representations, the contingent encoding approach advocated in this paper instead acts contrastively, changing decisions made rather than boosting them. For example, identifying the vowel as /u/ highlights the difference between the actual cues and the expectations of how those cues should behave in that context – and that difference drives perception. The typical form of feedback used in interactive activation models typically forces the percept to converge on the expected value (minimizing, not highlighting the deviation).

Our results support the kind of contingent encoding suggested by C-CuRE (Cole et al., 2010; McMurray et al., 2011; McMurray & Jongman, 2011). Under this account, listeners explicitly identify factors like vowel and talker, and then recode the cue to fricative identity relative to expectations about these factors. Rather than using raw acoustic information, listeners are constantly updating their perceptual judgments based on explicit decisions about other portions of the signal (or other contextual factors). Although listeners must ultimately rely on the acoustic signal for successful speech perception, they do not appear to use this information veridically for speech categorization. Instead, acoustic cues are computed relative to expectations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by National Institutes of Health Grant DC-008089 to Bob McMurray and the Ballard and Seashore Dissertation Year Fellowship to Keith Apfelbaum. We thank Yue Wang and Dan McEchron for help collecting and analyzing the acoustic data used in the simulations.

References

Allen JS, Miller JL, DeSteno D. Individual talker differences in voice-onset-time. The Journal of the Acoustical Society of America. 2003; 113(1):544.10.1121/1.1528172 [PubMed: 12558290]

Allopenna PD, Magnuson JS, Tanenhaus MK. Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. Journal of Memory and Language. 1998; 38(4):419–439.10.1006/jmla.1997.2558

Baayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language. 2008; 59(4):390–412.10.1016/j.jml.2007.12.005

Baron R, Kenny D. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986; 51(6):1173–1182. [PubMed: 3806354]

Bates, D.; Sarkar, D. R package version 099875-9. 2011. lme 4: Linear mixed-effects models using S4 classes.

Blumstein SE, Stevens KN. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America. 1979; 66(4):1001–1017. [PubMed: 512211]

Blumstein SE, Stevens KN. Perceptual invariance and onset spectra for stop consonants in different vowel environments. The Journal of the Acoustical Society of America. 1980; 67(2):648–62. [PubMed: 7358906]

Boucher VJ. Timing relations in speech and the identification of voice-onset times: a stable perceptual boundary for voicing categories across speaking rates. Perception & Psychophysics. 2002; 64(1): 121–30. [PubMed: 11916295]

Carden G, Levitt A, Jusczyk PW, Walley AC. Evidence for phonetic processing of cues to place of articulation: perceived manner affects perceived place. Perception & Psychophysics. 1981; 29(1): 26–36. 8. [PubMed: 7243528]

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. :1–86. (in press).

Cole J, Linebaugh G, Munson C, McMurray B. Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. Journal of phonetics. 2010; 38(2):167–184.10.1016/j.wocn.2009.08.004 [PubMed: 21173864]

Diehl RL, Walsh Ma. An auditory basis for the stimulus-length effect in the perception of stops and glides. The Journal of the Acoustical Society of America. 1989; 85(5):2154–64. [PubMed: 2732389]

Drager K. Speaker Age and Vowel Perception. Language and Speech. 2011; 54(1):99–121.10.1177/0023830910388017 [PubMed: 21524014]

Fowler CA, Brown JM. Perceptual parsing of acoustic consequences of velum lowering from information for vowels. Perception & Psychophysics. 2000; 62(1):21–32. [PubMed: 10703253]

Fowler, CA.; Smith, M. Speech perception as “vector analysis”: An approach to the problems of segmentation and invariance. In: Perkell, JS.; Klatt, D., editors. Invariance and Variability in Speech Processes. Hillsdale, NJ: Erlbaum; 1986. p. 126-136.

Goldrick M, Blumstein S. Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. Language and Cognitive Processes. 2006; 21(6):649–683.

Gow DW. Feature parsing: feature cue mapping in spoken word recognition. Perception & Psychophysics. 2003; 65(4):575–90. [PubMed: 12812280]

Hay J, Drager K. Stuffed toys and speech perception. Linguistics. 2010; 4(2010):865–892.10.1515/LING.2010.027

Holt LL. The mean matters: effects of statistically defined nonspeech spectral distributions on speech categorization. The Journal of the Acoustical Society of America. 2006; 120(5):2801–2817.10.1121/1.2354071 [PubMed: 17091133]

Jaeger TF. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of memory and language. 2008; 59(4):434–446.10.1016/j.jml.2007.11.007 [PubMed: 19884961]

Johnson K, Strand EA, D'Imperio M. Auditory–visual integration of talker gender in vowel perception. Journal of Phonetics. 1999; 27(4):359–384.10.1006/jpho.1999.0100

Jongman A, Wayland R, Wong S. Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America. 2000; 108(3 Pt 1):1252–63. [PubMed: 11008825]

Kessinger RH, Blumstein SE. Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. Journal of Phonetics. 1998; 26(2):117–128.10.1006/jpho.1997.0069

Kiefte M, Kluender KR. Absorption of reliable spectral characteristics in auditory perception. The Journal of the Acoustical Society of America. 2008; 123(1):366–376.10.1121/1.2804951 [PubMed: 18177166]

Kluender KR, Coady JA, Kiefte M. Sensitivity to change in perception of speech. Speech Commun. 2003; 41(1):59–69.10.1016/S0167-6393(02)00093-6

Lahiri A, Gewirth L, Blumstein SE. A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study. The Journal of the Acoustical Society of America. 1984; 76(2):391–404. [PubMed: 6480990]

Lee CY, Dutton L, Ram G. The role of speaker gender identification in relative fundamental frequency height estimation from multispeaker, brief speech segments. The Journal of the Acoustical Society of America. 2010; 128(1):384–8.10.1121/1.3397514 [PubMed: 20649232]

Lindblom B. Role of articulation in speech perception: Clues from production. The Journal of the Acoustical Society of America. 1996; 99(3):1683–1692.10.1121/1.414691 [PubMed: 8819859]

Lisker L, Abramson AS. A cross-langage study of voicing in initial stops: Acoustical measurements. Word. 1964; 20(3):384–422.

Lotto AJ, Kluender KR. General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Perception & Psychophysics. 1998; 60(4):602–619. [PubMed: 9628993]

Lotto AJ, Kluender KR, Holt LL. Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). The Journal of the Acoustical Society of America. 1997; 102(2):1134–1140.10.1121/1.419865 [PubMed: 9265760]

Magnuson JS, Nusbaum HC. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of experimental psychology. Human perception and performance. 2007; 33(2):391–409.10.1037/0096-1523.33.2.391 [PubMed: 17469975]

Mann VA, Repp BH. Influence of vocalic context on perception of the [ʃ]-[s] distinction. Perception & Psychophysics. 1980; 28(3):213–228. [PubMed: 7432999]

Marslen-Wilson WD. Functional parallelism in spoken word-recognition. Cognition. 1987; 25(1–2):71–102. [PubMed: 3581730]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognit Psychol. 1986; 18(1):1–86. [PubMed: 3753912]

McClelland J, Rumelhart D. An interactive activation model of context effects in letter perception: Part I. An account of basic findings. Psychological review. 1981; 88(5):375–407.

McMurray B, Clayards Ma, Tanenhaus MK, Aslin RN. Tracking the time course of phonetic cue integration during spoken word recognition. Psychonomic bulletin & review. 2008; 15(6):1064–71.10.3758/PBR.15.6.1064 [PubMed: 19001568]

McMurray, B.; Cole, JS.; Munson, C. Features as an emergent product of computing perceptual cues relative to expectations. In: Ridouane, R.; Clement, N., editors. Where Do Features Come From?. Amsterdam: John Benjamins Publishing; 2011. p. 197-236.

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review. 2011; 118(2):219–246.10.1037/a0022325 [PubMed: 21417542]

Mermelstein P. On the relationship between vowel and consonant identification when cued by the same acoustic information. Perception & Psychophysics. 1978; 23(4):331–336. [PubMed: 748856]

Miller JL, Green KP, Reeves A. Speaking rate and segments: A look a the relation between speech production and speech perception for the voicing contrast. Phonetica. 1986; 43:106–115.

Nearey TM. The segment as a unit of speech perception. Journal of Phonetics. 1990; 18:347–373.

Nearey TM. Speech perception as pattern recognition. The Journal of the Acoustical Society of America. 1997; 101(6):3241–54. [PubMed: 9193041]

Nearey TM, Rochet BL. Effects of place of articulation and vowel context on VOT production and perception for French and English stops. Journal of the International Phonetic Association. 1994; 24(1):1–18.

Niedzielski N. The Effect of Social Information on the Perception of Sociolinguistic Variables. Journal of Language and Social Psychology. 1999; 18(1):62–85.10.1177/0261927X99018001005

Nusbaum, HC.; Magnuson, JS. Talker normalization: Phonetic constancy as a cognitive process. In: Johnson, K.; Mullenix, JW., editors. Talker Variability in Speech Processing. Academic Press; 1997. p. 109-132.

Nygaard LC, Sommers MS, Pisoni DB. Speech perception as a talker-contingent process. Psychological Science. 1994; 5(1):42–46.10.1111/j.1467-9280.1994.tb00612.x [PubMed: 21526138]

Oden GC. Integration of place and voicing information in the identification of synthetic stop consonants. Journal of Phonetics. 1978; 6:83–93.

Oden GC, Massaro DW. Integration of featural information in speech perception. Psychological Review. 1978; 85(3):172–91. [PubMed: 663005]

Ohala JJ. Speech perception is hearing sounds, not tongues. The Journal of the Acoustical Society of America. 1996; 99(3):1718–25. [PubMed: 8819861]

Pardo JS, Fowler Ca. Perceiving the causes of coarticulatory acoustic variation: consonant voicing and vowel pitch. Perception & Psychophysics. 1997; 59(7):1141–52. [PubMed: 9360485]

Port RF, Dalby J. Consonant/vowel ratio as a cue for voicing in English. Perception & Psychophysics. 1982; 32(2):141–152. [PubMed: 7145584]

Rhone AE, Jongman A. Modified locus equations categorize stop place in a perceptually realistic time frame. The Journal of the Acoustical Society of America. 2012; 131(6):EL487–91.10.1121/1.4722169 [PubMed: 22713026]

Sawusch J, Pisoni D. On the identification of place and voicing features in synthetic stop consonants. Journal of Phonetics. 1974; (2):181–194.

Smiljanić R, Bradlow A. Speaking and hearing clearly: Talker and listener factors in speaking style changes. Language and linguistics compass. 2009; 3(1):236–264.10.1111/j.1749-818X.2008.00112.x.Speaking [PubMed: 20046964]

Smits R. Hierarchical categorization of coarticulated phonemes: a theoretical analysis. Perception & Psychophysics. 2001a; 63(7):1109–39. [PubMed: 11766939]

Smits R. Evidence for hierarchical categorization of coarticulated phonemes. Journal of experimental psychology. Human perception and performance. 2001b; 27(5):1145–62. [PubMed: 11642700]

Strand EA. Uncovering the Role of Gender Stereotypes in Speech Perception. Journal of Language and Social Psychology. 1999; 18(1):86–100.10.1177/0261927X99018001006

Strand, EA.; Johnson, K. Gradient and visual speaker normalization in the perception of fricatives. In: Gibbon, D., editor. Natural Language Processing and Speech Technology: Results of the 3rd KONVENS Conference. Berlin, Germany: Mouton de Gruyter; 1996. p. 14-26.

Summerfield Q. Articulatory rate and perceptual constancy in phonetic perception. Journal of experimental psychology. Human perception and performance. 1981; 7(5):1074–95. [PubMed: 6457109]

Sussman HM, Fruchter D, Hilbert J, Sirosh J. Linear correlates in the speech signal: The orderly output constraint. Behavioral and Brain Sciences. 1998; 21(02):241–299.10.1017/S0140525X98001174 [PubMed: 10097014]

Sussman HM, Shore J. Locus equations as phonetic descriptors of consonantal place of articulation. Perception & Psychophysics. 1996; 58(8):936–946. [PubMed: 8768188]

Toscano JC, McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive science. 2010; 34(3):434–464.10.1111/j.1551-6709.2009.01077.x [PubMed: 21339861]

Toscano JC, McMurray B. Cue-integration and context effects in speech: Evidence against speaking-rate normalization. Attention, perception & Psychophysics. 2012; 74(6):1284–301.10.3758/s13414-012-0306-z

Viswanathan N, Fowler Ca, Magnuson JS. A critical examination of the spectral contrast account of compensation for coarticulation. Psychonomic bulletin & review. 2009; 16(1):74–9.10.3758/PBR.16.1.74 [PubMed: 19145013]

Whalen DH. Vowel and consonant judgments are not independent when cues by the same information. Attention, Perception, & Psychophysics. 1989; 46(3):284–292.

Whalen DH. Perception of overlapping segments: Thoughts on Nearey's model. Journal of Phonetics. 1992; 20:493–496.

Zwitserlood P. The locus of the effects of sentential-semantic context in spoken-word processing. Cognition. 1989; 32(1):25–64. [PubMed: 2752705]

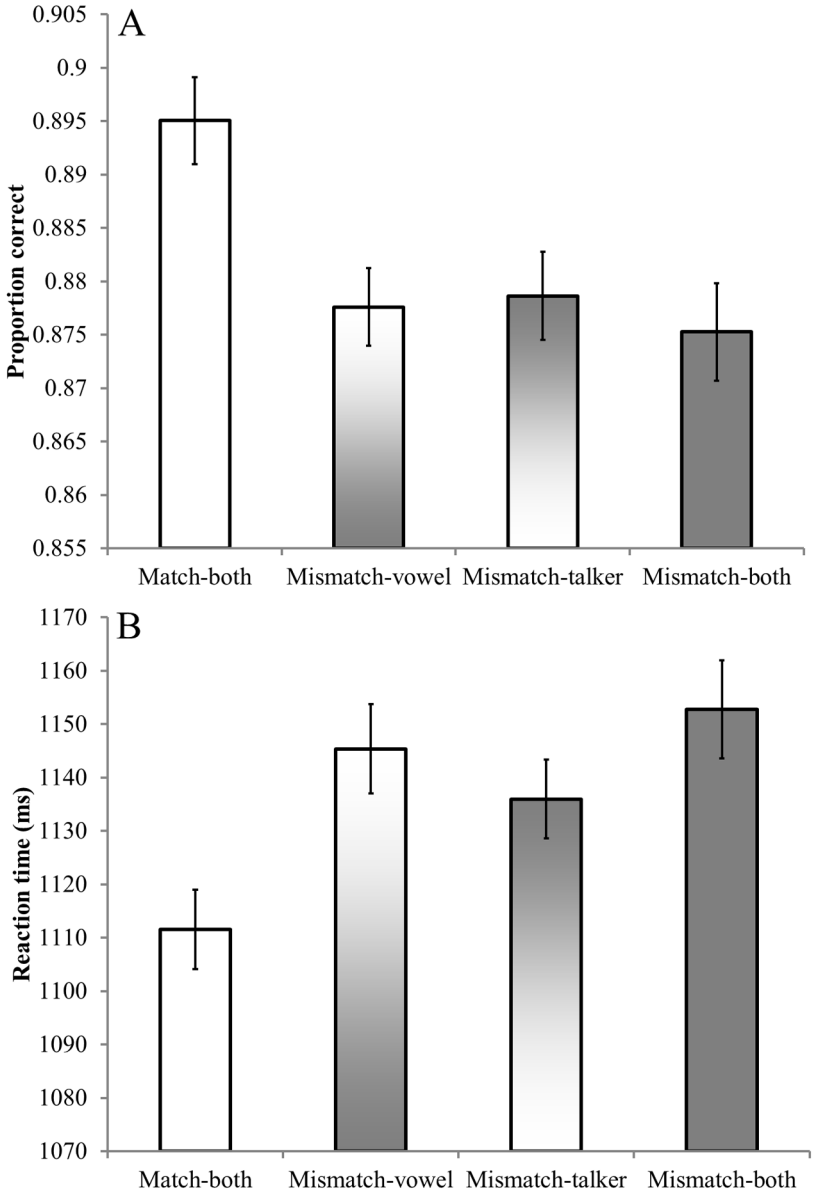

Figure 1.

Identification results from different splice conditions collapsed across all four fricatives. Error bars represent standard error for that condition. A) Accuracy; B) RT. Note that although raw RT values are displayed in the figures, log RT was used for analysis.

Table 1

Cues to fricative identity in the fricative and vowel portions of the stimuli, as used in simulations of splicing effects.

Cues within frication	Cues within vowel
Peak frequency	Pitch (F0) at vowel onset
RMS amplitude of frication	RMS amplitude of vowel
Duration of frication	Duration of vowel
Low-frequency energy (mean RMS below 500Hz)	F1, F2, F3, F4, F5 at vowel onset
Amplitude of frication at F3	Amplitude of vowel at F3
Amplitude of frication at F5	Amplitude of vowel at F5
Spectral mean (two windows in frication)	Spectral mean (at transition)
Spectral variance (two windows in frication)	Spectral variance (at transition)
Spectral skewness (two windows in frication)	Spectral skewness (at transition)
Spectral kurtosis (two windows in frication)	Spectral kurtosis (at transition)

Table 2

Percent correct fricative classification as estimated by a bottom-up cue integration model for the talkers chosen for this study. In McMurray and Jongman (2011), performance could be estimated using a probabilistic or a discrete linking rule. For convenience, we list only the results from the probabilistic rule; the discrete rule exhibited very similar performance.

Pairing	Match-both	Mismatch-talker	Mismatch-vowel	Mismatch-both
Overall	76.5	75.3	77.3	76.6
F1 and M1	74.8	73.1	75.3	74.0
F1 and M2	75.3	74.1	75.3	74.9
F2 and M1	77.7	76.2	79.3	78.3