Cognition 202 (2020) 104328

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing

Matthew Lehet[a,b], Lori L. Holt[a,b,c,*]

[a] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15232, USA
[b] Center for the Neural Basis of Cognition, Pittsburgh, PA 15232, USA
[c] Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15232, USA

ARTICLE INFO

Keywords:
Speech perception
Normalization
Dimension-based statistical learning
Perceptual learning

ABSTRACT

Speech is notoriously variable, with no simple mapping from acoustics to linguistically-meaningful units like words and phonemes. Empirical research on this theoretically central issue establishes at least two classes of perceptual phenomena that accommodate acoustic variability: normalization and perceptual learning. Intriguingly, perceptual learning is supported by learning across acoustic variability, but normalization is thought to counteract acoustic variability leaving open questions about how these two phenomena might interact. Here, we examine the joint impact of normalization and perceptual learning on how acoustic dimensions map to vowel categories. As listeners categorized nonwords as *setch* or *satch*, they experienced a shift in short-term distributional regularities across the vowels' acoustic dimensions. Introduction of this 'artificial accent' resulted in a shift in the contribution of vowel duration in categorization. Although this dimension-based statistical learning impacted the influence of vowel duration on vowel categorization, the duration of these very same vowels nonetheless maintained a consistent influence on categorization of a subsequent consonant via duration contrast, a form of normalization. Thus, vowel duration had a duplex role consistent with normalization and perceptual learning operating on distinct levels in the processing hierarchy. We posit that whereas normalization operates across auditory dimensions, dimension-based statistical learning impacts the connection weights among auditory dimensions and phonetic categories.

1. Introduction

The ease of perceiving speech masks the perceptual challenges inherent in mapping from highly variable acoustic input to an inventory of linguistically-relevant phonetic category representations. This variability arises from many sources. For example, speaker-based variability such as accent or dialect (Babel, 2010; Clarke & Garrett, 2004; Floccia, Goslin, Girard, & Konopczynski, 2006; Kraljic, Brennan, & Samuel, 2008; Labov, Ash, & Boberg, 2005; Ladefoged, 1989), age (Lee, Potamianos, & Narayanan, 1999), gender (Perry, Ohde, & Ashmead, 2001), style of speech (Lindblom, 1990) and differences in vocal tract size and shape (Fitch & Giedd, 1999; Peterson & Barney, 1952) all affect the detailed acoustic realization of speech. As a result, the acoustic information associated with a particular speech category can vary widely. For example, under some circumstances the acoustics of a talker's /æ/ vowel (as in *sand*) may overlap considerably with another talker's /ɛ/ (as in *send*). Understanding the nature of the complex mapping between speech acoustics and speech categories has been a theoretically-central issue in understanding perception of speech.

A rich body of research has identified multiple classes of perceptual phenomena that may aid listeners in coping with the acoustic variability in speech. Although these phenomena have been studied deeply in isolation, few studies have attempted to examine how they may interact in speech perception. Yet, examination of interactions may offer new opportunities to advance theoretical models of spoken language processing. We do not yet know, for example, whether the approaches the perceptual system brings to dealing with acoustic variability in speech involve operations at distinct levels in a processing hierarchy and, if so, how these processes cooperate. Here, we examine the joint influence of two classes of phenomena – normalization and perceptual learning – on how acoustic dimensions are mapped to speech categories. We specifically seek to understand how a single segment of acoustic speech information might have multiple roles in perceptual processing that are differently affected by normalization versus perceptual learning.

[*] Corresponding author.
E-mail address: loriholt@cmu.edu (L.L. Holt).

https://doi.org/10.1016/j.cognition.2020.104328
Received 1 June 2018; Received in revised form 4 May 2020; Accepted 13 May 2020

1.1. Normalization

To the extent that acoustic variability is a challenge in mapping sound to speech categories, perceptual processes capable of transforming the input in a manner that compensates for acoustic input variability may be an effective solution. Here, we refer to processes thought to explicitly transform speech by stripping away acoustic variability under the umbrella term 'normalization'. Indeed, there are diverse instantiations of normalization that rely upon a range of proposed mechanisms to transform speech acoustics. These mechanisms include interactions in general auditory processing, references to speech motor plans, and accommodation via talker-specific internal models (e.g., Bourguignon, Baum, & Shiller, 2016; Broadbent, Ladefoged, & Lawrence, 1956; Dechovitz, 1977; Garvin & Ladefoged, 1963; Johnson, 1990; Joos, 1948; Ladefoged, 1989; Ladefoged & Broadbent, 1957; Liberman & Mattingly, 1985; Lotto & Kluender, 1998; Mann, 1980; Miller & Liberman, 1979; Miller & Volaitis, 1989; Sjerps, Fox, Johnson, & Chang, 2019; Zhang & Chen, 2016). Across normalization accounts, the common conjecture is that perceptual processing counteracts acoustic input variability by transforming input with the support of relational or contextual information available to the listener.

Some forms of normalization may arise even in general auditory processing that plays out prior to speech categorization (Holt, 2006; Holt, Lotto, & Kluender, 2000; Huang & Holt, 2009, 2012; Hufnagle, Holt, & Thiessen, 2013; Lotto & Holt, 2006; Lotto & Kluender, 1998; Stephens & Holt, 2003; Wade & Holt, 2005a, 2005b). To illustrate, it is useful to begin with an example. Many speech categories are differentiated by duration. For example, the vowel /ε/ (as in *send*) typically has a shorter duration than the vowel /æ/ (as in *sand*; Hillenbrand, Clark, & Houde, 2000; Hillenbrand, Getty, Clark, & Wheeler, 1995; Liu & Holt, 2015). Likewise, the duration of voice onset time (VOT) provides information distinguishing /b/ from /p/. But, in natural speech, rate can vary quite a lot, generating substantial variability in the acoustic realization of vowel duration and VOT (J. L. Miller, Grosjean, & Lomanto, 1984; Pellegrino, Coupé, & Marsico, 2011; Quené, 2008, 2013). A fast-paced conversation may lead to /æ/ productions that are shorter than /ε/ vowels produced in a more relaxed conversation.

Listeners appear to use rate-of-speech information to perceive duration relationally. When duration-dependent speech categories are heard in the context of faster adjacent speech, speech categorization contrastively shifts toward longer-duration percepts (e.g., /æ/ or /p/) in a manner that appears to normalize, or compensate for, variable acoustics according to the rate of surrounding speech (Bosker & Reinisch, 2015; Liberman, Delattre, Gerstman, & Cooper, 1956; J. L. Miller & Liberman, 1979; Reinisch, 2016). Perception of the very same speech signal shifts in an opposing, contrastive direction (e.g., toward shorter-duration percepts such as /ε/ or /b/) in the context of slower adjacent speech. Thus, there appears to be perceptual normalization of variable input acoustics as a function of the rate or duration of adjacent context.

This contrastive adjustment for speech rate persists even when the context is a single phoneme (J. L. Miller & Baer, 1998; J. L. Miller & Liberman, 1979; J. L. Miller & Wayland, 1993; Newman & Sawusch, 1996; Sawusch & Newman, 2000). For example, the duration of a preceding vowel can influence categorization of subsequent consonants as /b/ versus /p/. A shorter preceding vowel, like /ε/, leads listeners to categorize consonants with perceptually-ambiguous VOT more often as longer-VOT, /p/, whereas a longer preceding vowel, like /æ/, leads the same consonants to be categorized as /b/ (Denes, 1955; Kluender, Diehl, & Wright, 1988; Port & Dalby, 1982; Raphael, 1972). The manner by which the perceptual representation of VOT in these situations may be affected by context – and representational changes associated with normalization, in general – are the subject of a longstanding debate (Kingston et al., 2014; Lotto & Holt, 2006; Viswanathan, Fowler, & Magnuson, 2009; C. Zhang & Chen, 2016). However, it is well-established that speech categorization shifts in a context-dependent manner.

This form of normalization may have its roots in general auditory processing playing out at a pre-categorical level prior to speech categorization because *nonspeech* signals varying in duration also elicit rate and duration-dependent speech categorization. Sequences of nonspeech tones (Wade & Holt, 2005b), or single tones differing in duration (Diehl & Walsh, 1989), shift speech categorization in a contrastive manner. Just as a shorter vowel shifts /b/ − /p/ categorization toward the longer-VOT /p/, a short nonspeech tone shifts speech categorization toward /p/ and a long tone shifts categorization toward /b/.

In all, temporal information in adjacent acoustic input, whether speech or nonspeech, appears to 'normalize' speech categorization by contrastively shifting it relative to the acoustic context. Like other forms of normalization, duration contrast appears to adjust perception of acoustic input as a function of context, thereby counteracting acoustic variability across situational differences such as faster or slower talkers. Critical to the approach we will take in the present study, the existence of an influence of simple tones on speech categorization implicates rather low-level, general auditory perceptual processes operating precategorically in duration contrast (Diehl & Walsh, 1989).

1.2. Perceptual learning

Perceptual learning refers here to another class of phenomena that appears to aid listeners in coping with acoustic variability in speech. When listeners rely on contextual information to resolve ambiguous speech acoustic input, they learn in a manner that influences how they later categorize acoustically-ambiguous speech even when the disambiguating context is no longer present (Bertelson, Vroomen, & de Gelder, 2003; Guediche, Fiez, & Holt, 2016; Idemaru & Holt, 2011; Kraljic & Samuel, 2005; Maye, Aslin, & Tanenhaus, 2008; Norris, McQueen, & Cutler, 2003). For example, listeners exposed to distorted (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Guediche, Holt, Laurent, Lim, & Fiez, 2014; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Norris et al., 2003; Samuel, 1997) or accented (Bradlow & Bent, 2008; Maye et al., 2008) speech exhibit rapid perceptual learning to details of speech acoustics such that later encounters with similarly distorted or accented speech exhibits accommodation of the experienced variability (see Guediche et al., 2014; Samuel & Kraljic, 2009), observed as increases in intelligibility or a shift in speech categorization.

Although many examples of perceptual learning are driven by lexical knowledge (Kraljic & Samuel, 2005, 2006; Kraljic, Samuel, & Brennan, 2008; Maye et al., 2008; Norris et al., 2003; Samuel & Kraljic, 2009) or explicit, often orthographic, feedback that directs listeners to map distorted acoustics to familiar words (Davis et al., 2005; Greenspan, Nusbaum, & Pisoni, 1988; Guediche et al., 2016; Schwab, Nusbaum, & Pisoni, 1985), the locus of learning is pre-lexical, as evidenced by the fact that the learning generalizes to novel words not experienced in training (Eisner & McQueen, 2005; Greenspan et al., 1988; Schwab et al., 1985) and exerts its influence on speech categorization in nonword contexts (Kraljic & Samuel, 2006, 2007; Norris et al., 2003).

Across various instances of perceptual learning, the common conjecture is that *learning* about short-term input regularities in recently heard speech neutralizes some of the challenges presented by acoustic variability in speech. This perceptual learning is possible because supportive information conveyed by lexical (Kraljic, Brennan, & Samuel, 2008; Kraljic & Samuel, 2005, 2006; Norris et al., 2003), visual (Vroomen & Baart, 2009; Vroomen, van Linden, de Gelder, & Bertelson, 2007; Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004), or acoustic (Idemaru & Holt, 2011, 2014; Liu & Holt, 2015; Schertz, Cho, Lotto, & Warner, 2015) context helps to resolve ambiguous speech acoustics. When context repeatedly disambiguates acoustically ambiguous speech input, perceptual learning leads to shifts in speech categorization that are apparent even in the absence of disambiguating context. In this way, perceptual learning may provide a means by which to adjust to the short-term input regularities that differ as a function of accent, talker, and other sources.

For example, in English, vowel duration covaries with the spectral quality of /ɛ/ and /æ/ vowels such that /æ/ is typically longer than /ɛ/. Listeners are sensitive to this regularity and use duration to disambiguate vowels when spectral quality is ambiguous (Liu & Holt, 2015). In *dimension-based statistical learning*, distributional regularities across acoustic input dimensions serve as the context that drives perceptual learning. When the relationship of vowel duration departs from its typical mapping whereby /ɛ/ is shorter and /æ/ is longer in an 'artificial accent,' on the influence of duration in signaling /ɛ/ − /æ/ vowel categories rapidly changes (Liu & Holt, 2015).

1.3. Interactions of normalization and perceptual learning

There is considerable evidence that both normalization and perceptual learning are available to listeners to meet the challenge of acoustic variability in speech input. However, studies have not yet examined how they might interact. This is an important gap in that it leaves open crucial theoretical issues. If listeners learn to re-weight an acoustic dimension based on short-term regularities experienced in the input, does that learning modify how that acoustic dimension drives normalization? Do these classes of phenomena operate at distinct levels in the processing hierarchy? Empirical results indicate that each process impacts speech categorization, but their interaction is not understood. Here, we examine whether a specific type of perceptual learning, *dimension-based statistical learning* (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2016; Liu & Holt, 2015; Schertz et al., 2015; Zhang & Holt, 2018), interacts with a specific type of normalization, *duration contrast* (Diehl & Walsh, 1989; Kluender et al., 1988; Raphael, 1972; Wade & Holt, 2005b; Kluender et al., 1988). We specifically test whether vowels varying across an acoustic dimension rendered less effective in signaling vowel category identity by perceptual learning continue to evoke normalization. Experiment 1 demonstrates a baseline effect size for a duration contrast effect and then Experiment 2 examines how this duration contrast effect is impacted by dimension-based statistical learning that influences the effectiveness of duration in signaling vowel categories.

2. Experiment 1

In Experiment 1, we demonstrate a form of duration contrast by preceding a /ba − /pa/ voicing judgment with long and short vowels. We predict, in line with prior research (Denes, 1955; Kluender et al., 1988; Port & Dalby, 1982; Raphael, 1972), that a shorter preceding vowel will shift /b/ − /p/ categorization toward the longer-VOT /p/ whereas a longer preceding vowel will shift categorization toward /b/. Here, the preceding vowels are spectrally ambiguous between /ɛ/ and /æ/, with either long or short duration. The overarching goal of this first experiment is to identify the stimulus along the VOT series between /ba/ and /pa/ for which the contrastive normalization impacts vowel categorization most robustly.

2.1. Methods

2.1.1. Participants

Fifteen monolingual English Carnegie Mellon University undergraduates with self reported normal hearing participated in Experiment 1 (mean age 22.2 years, 9 female).

2.1.2. Stimuli

We constructed a 10-step series varying acoustically in VOT in 5-ms increments and perceptually from /ba/ to /pa/ using a cross-splicing method in which voiceless segments of the /pa/ (including the burst and acoustic evidence of voicing onset) were spliced onto the /ba/ token (Idemaru & Holt, 2011; Lehet & Holt, 2016; Liu & Holt, 2015; McMurray & Aslin, 2005). These tokens were natural /ba/ and /pa/ productions of the same adult female native-English speaker who recorded the *set* and *sat* (and the *setch,* and *satch)* utterances described below.

Syllables from this /ba/ − /pa/ series were presented alone (syllable condition) or preceded by /sɛ/ and /sæ/ syllables that had either a long vowel duration or a short vowel duration and acoustically ambiguous spectral quality. These syllables were constructed from the *setch-satch* stimuli described for Experiment 2 (below, see Fig. 2, Step 4 of Spectral Quality, with 225 and 425 ms vowel duration).

When combined, the /sɛ/ − /sæ/ fricative-vowel segments and the /ba/ or /pa/ tokens took the nonword forms *seppa, sebba, sappa, and sabba*. The final /a/ vowel of each of the /ba/ and /pa/ syllables was manipulated using PSOLA (Praat; Boersma & Weenik, 2013) to be 325 ms, the average duration of the initial vowels forming the *seppa, sebba, sappa, and sabba* nonwords. The fundamental frequency (F0) of the final /a/ vowel was adjusted to align with the F0 of the proceeding vowel (170 Hz).

This allowed three series to be constructed. One varied from /ba/ to /pa/ across VOT, one varied from *sebba* to *seppa* for which a short initial vowel duration was the only distinguishing cue to initial vowel identity, and one varied from *sabba* to *sappa* with long vowel duration distinguishing the initial vowel identity. The /ba/ and /pa/ tokens were identical across all three series.

2.1.3. Procedure

Each participant heard each of these three stimulus series (/ba/ or /pa/ syllable only, short vowel plus syllable, long vowel plus syllable) at a comfortable listening level, and randomized in order of presentation across 10 repetitions. On each trial, a 500-ms blank screen was followed by acoustic stimulus presentation diotically over headphones along with the appearance of orthographic response options (/ba/ or /pa/) that corresponded spatially to response buttons on a standard keyboard. Participants were instructed to report whether they heard a /ba/ or a /pa/ on each trial with a keypress.

2.1.4. Statistical analysis and power

Using G-power (logistic regression: compute achieved power using probabilities; Faul, Erdfelder, Lang, & Buchner, 2007) we estimated the power for the duration contrast effect based on a 21% difference in /pa/ judgments between long (47%) and short vowel duration (68%) contexts (in line with the data reported below), 150 observations of each stimulus with long and short vowel duration contexts, and an alpha of 0.05. Therefore, given our effect size, we achieve 83.0% power with our sample size. All models were run in R (R Core Team, 2013) and logistic mixed effects models used the lme4 package (Bates, Sarkar, Bates, & Matrix, 2007).

2.2. Results

2.2.1. Identifying the ambiguous VOT

We examined the full range of the VOT series to identify the stimulus step for which the duration of a preceding vowel most influenced categorization of the stimulus as /ba/ versus /pa/ (Fig. 1). Based on prior research, we expected the most perceptually ambiguous stimulus along the VOT series in the syllable-only condition to be most influenced by the duration of the preceding vowel context. Therefore, we used a linear mixed effects model with a binomial linking function to predict the categorization across the syllable-only series as /pa/ (see Fig. 1a). Participant was included as a random intercept to account for random variance across participants (see Table s1).

The most /ba/ − like stimulus (the first VOT Step) was used as the baseline for which /ba/ responses were significantly more likely than /pa/ responses (ß = −2.62, SE = 0.36, $p < .001$, odds of /p/ responses = 0.07:1). VOT Steps 2 (5 ms VOT) through 4 (15 ms VOT)

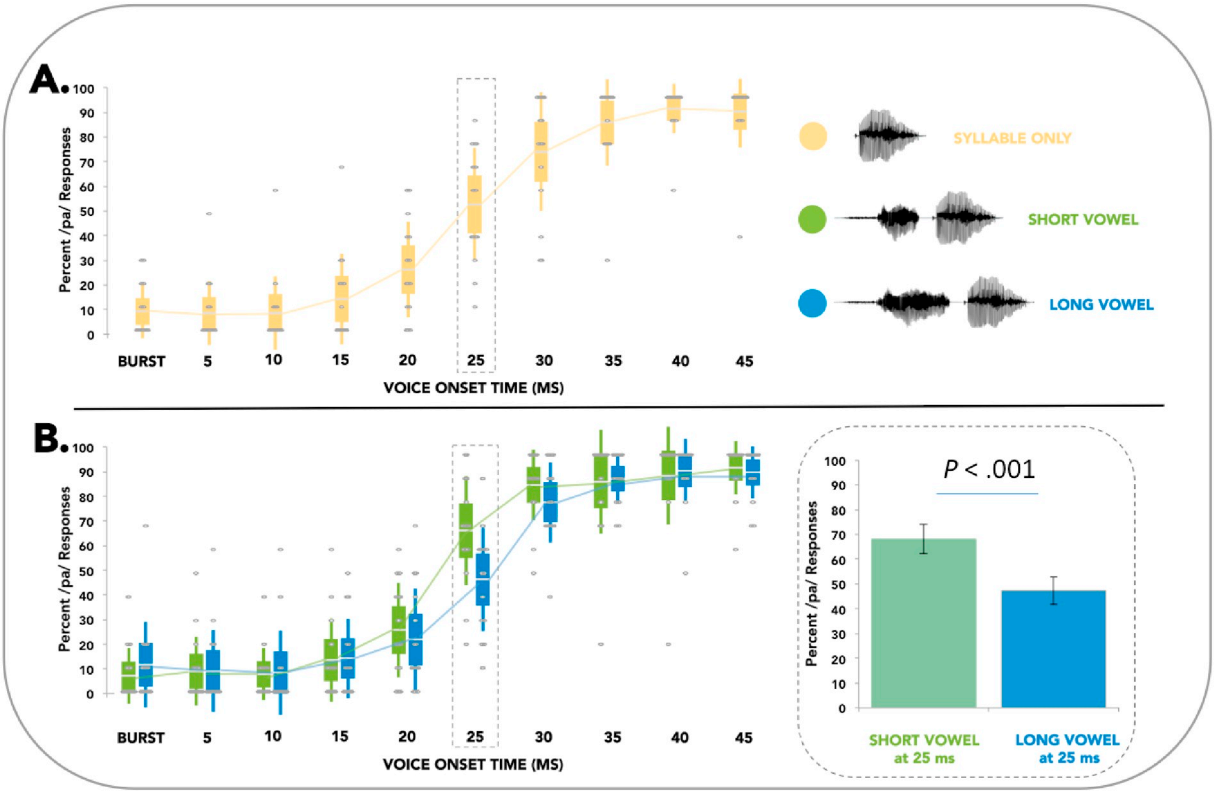

Fig. 1. Results of Experiment 1. (A) Mean categorization of /ba/ − /pa/ syllables with no preceding vowel revealed the most perceptually ambiguous stimulus (25 ms VOT, identified by the dashed box). (B) Short versus long preceding vowels in /sV/ contexts (see inset in (A)) shifts ba/ − /pa/ categorization, consistent with duration contrast. The shorter preceding vowel results in more /pa/ categorization responses compared to the longer preceding vowel. The inset at the right highlights the average percent /pa/ responses for the most perceptually ambiguous syllable identified in (A). In each panel, individual dots show participant means, vertical lines represent the standard deviation from the mean, boxes depict standard error, and horizontal lines in each box depict the mean. Error bars in the bar plot represent standard error of the mean.

were not significantly different from this baseline (all $p > .1$), but there were significantly more /pa/ responses at Steps 5 through 10 compared to baseline (all $p < .001$). Step 6 (25 ms VOT) was the most perceptually ambiguous (M = 54.0%, SE = 6.31%), with close to even odds of /ba/ and /pa/ responses (ß = 2.79, SE = 0.35, $p < .001$, odds of /pa/ responses = 1.19:1).

2.2.2. Duration contrast effect

We compared the influence of the preceding short versus long vowel contexts on perception of /pa/ (Fig. 1b, right panel) at the most perceptually ambiguous stimulus along the VOT stimulus series (Step 6, 25 ms VOT) using a logistic mixed effects model with participant as a random intercept (see Table s2). The short vowel duration was used as a reference category. In the short vowel duration context /pa/ responses were significantly more likely (short vowel: ß = 0.86, SE = 0.28, $p = .002$, odds of /pa/ response = 2.35:1). However, in the long vowel duration context /pa/ responses were significantly less likely (ß = −0.99, SE = 0.26, $p < .001$, odds of /pa/ response = 0.88:1). When the short vowel preceded the perceptually ambiguous consonant-vowel syllable, the odds of /pa/ categorization were 2.69 times greater than when the long vowel preceded it. This durationally-contrastive shift in consonant categorization presented as a 20.7% mean difference in categorization responses at Step 6 (SE = 4.5%, $M_{\text{Short Vowel}}$ = 68.0%, SE = 5.95%, $M_{\text{Long Vowel}}$ = 47.33%, SE = 5.65%). In sum, preceding vowel contexts significantly shifted speech categorization of the most perceptually ambiguous stimulus along the VOT series in a contrastive manner. The same syllables were more often categorized as /pa/ (with longer VOT) following a shorter vowel than when they were preceded by a longer vowel. This establishes a duration contrast effect size that we can use to predict normalization effects in Experiment 2.

3. Experiment 2

In Experiment 2 we examined the interaction of the Experiment 1 duration contrast effects with dimension-based statistical learning, a form of perceptual learning that presents the opportunity to examine detailed interactions between auditory input dimensions and phonetic category representations and that allows for precise manipulation of short-term acoustic input regularities implicitly available to listeners (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2016; Liu & Holt, 2015; Schertz et al., 2015; Zhang & Holt, 2018). In dimension-based statistical learning exposure to an 'artificial accent' that manipulates short-term acoustic regularities across acoustic dimensions affects the perceptual weight, or influence, of specific acoustic dimensions in signaling phonetic categories (e.g. the influence of vowel duration on vowel identity as in Liu & Holt, 2015). In this context, Experiment 2 investigates whether the duration contrast effect of a preceding vowel on /ba/ − /pa/ categorization observed in Experiment 1 is consistent across conditions in which dimension-based statistical learning leads the duration of the preceding vowel to have changing influences on categorization of that vowel. We examine whether the impact of dimension-based statistical learning on the perceptual weight of vowel duration impacts how that duration evokes duration contrast effects on categorization of a subsequent /ba/ − /pa/ syllable.

Recall from above that duration contrast is thought to operate at a pre-categorical level of speech processing (Diehl & Walsh, 1989;

Kluender et al., 1988; Raphael, 1972). If dimension-based statistical learning affects early, pre-categorical perceptual processing of acoustic dimensions like duration, then re-weighting vowel duration through perceptual learning may render vowel duration unavailable to elicit duration contrast on subsequent /ba/ − /pa/ categorization. Alternatively, normalization may play out at a different level of the speech processing hierarchy than perceptual learning. In this case, the very same acoustic duration information that is re-weighted in its influence on vowel categorization may nevertheless persist in its availability to elicit duration contrast on subsequent consonant categorization.

3.1. Methods

3.1.1. Participants

Thirty-three monolingual English Carnegie Mellon University undergraduates with self-reported normal hearing participated (mean age 22.1 years, 26 female) were split into two groups (n = 17, n = 16) that received complimentary designs, as described below.

3.1.2. Stimuli

A set of forty-nine speech stimuli was constructed in the manner described by Liu and Holt (2015). Stimuli varied in vowel *spectral quality* (7 Steps) and *vowel duration* (7 Steps) in nonword *setch* and *satch* contexts. Stimulus construction began with natural recordings of *set, sat, setch,* and *satch* utterances by an adult female native-English speaker, with slightly exaggerated vowel duration lengths (see Liu & Holt, 2015). Two exemplars, one *set* and one *sat* were selected based on vowel quality and roughly equivalent vowel durations to serve as the starting point for stimulus construction.

The relatively steady-state portions of the vowels /ɛ/ and /æ/ were digitally spliced from *set* and *sat,* respectively, at zero crossings of the waveform. The first four formant frequency trajectories of these isolated vowels were extracted using Burg's formant extraction algorithm (maximum 5 formants; maximum formant value = 5500 Hz; 0.025 s time window; pre-emphasis from 50 Hz) using Praat (Boersma & Weenik, 2013). The values of each of the formant trajectories were linearly interpolated at equal steps between /ɛ/ and /æ/ using R (R Development Core Team, 2008), and these values were entered into Praat to generate a 7-step series varying in spectral quality, a measure used to index spectral changes in the full-spectrum profile of formant frequencies across the vowels.

Next, the duration of each of the vowels along this 7-step spectral series was reduced using Praat's PSOLA function to yield identical spectral series varying in vowel duration from 175 to 475 ms in 50-ms steps. The intermediate duration along this series, 325 ms, straddles the boundary between average adult native English-speaking female /ɛ/ and /æ/ vowel acoustics reported by Hillenbrand et al. (1995). This approach yielded a 7 × 7 grid of 49 vowels.

Returning to the original natural speech recordings, the frication corresponding to /s/ was digitally spliced from a clearly-articulated instance of *setch,* as was the final consonant affricate /tʃ /. The /s/ was appended to the onset of each of the 49 vowels varying in the spectral quality x duration acoustic space and the /tʃ / was appended to each vowel's offset. This created nonword *setch-satch* contexts for which the initial and final consonants were acoustically identical across the 49 stimuli, with vowels varying in spectral quality and duration.

A second stimulus, derived from the Experiment 1 /ba/ − /pa/ series, was also used. Including these syllables advanced two goals: 1) to measure the impact of dimension-based statistical learning on subsequent duration contrast effects; and 2) to examine generalization of dimension-based statistical learning to a nonword frame across which the artificial accent had not been experienced. The stimuli with long and short vowel duration and 25 ms VOT that showed the largest duration contrast effect in Experiment 1 (Step 6, Fig. 1b) were chosen for inclusion in this portion of the experiment.[1] These Test stimuli could be perceived as *seppa, sebba, sappa, or sabba* and contained a perceptually ambiguous /ba/ − /pa/ syllable preceded by either a long or short duration vowel with ambiguous spectral quality. To achieve this, we spliced fricative-vowel /sɛ/ − /sæ/ segments from the two-dimensional stimulus grid described above for vowels with perceptually ambiguous spectral quality (Step 4) and either long (425 ms) or short (225 ms) duration. Thus, these stimuli were created such that categorization as *seppa, sebba, sappa, or sabba* is only possible using duration to identify the initial vowel, and duration contrast from that vowel to identify the stop consonant as /ba/ versus /pa/ (since VOT is perceptually ambiguous).

3.1.3. Procedure

There were four blocks of stimuli, each separated by brief, self-timed breaks. Fig. 2 illustrates the block structure. Open symbols indicate the 49-stimulus grid varying acoustically in vowel spectral quality and duration and perceptually from /ɛ/ to /æ/. The filled symbols indicate the stimuli presented within each of the block types (Baseline, Reverse, Canonical). Within a block, participants heard 10 repetitions of each *setch-satch* Exposure stimulus (black squares, Fig. 2), 10 repetitions of each of the two *setch-satch* Test stimuli (orange, purple diamonds, Fig. 2), and 10 repetitions of each of the *seppa, sebba, sappa, sabba* Test stimuli (orange, purple diamonds, Fig. 2). The order of these stimuli was randomized within each block.

3.1.4. Exposure stimuli

The filled black squares in Fig. 2 indicate *Exposure* stimuli. In each block, Exposure stimuli comprised the majority of stimuli (75%) to convey a short-term regularity between the vowel spectral quality and duration. In the Baseline Block, the 25 central stimuli within the 49-stimulus grid comprised the Exposure stimuli**.** This resulted in a neutral sampling such that listeners experienced no short-term correlation between vowel spectral quality and duration. This block provides a measure of listeners' baseline reliance on spectral quality and duration in vowel categorization, without the potential for short-term regularities to influence reliance upon the acoustic dimensions for vowel categorization. Each of the 25 exemplars was presented 10 times for a total of 250 trials.

In Canonical blocks, Exposure stimuli sampled vowel exemplars for which the relationship between vowel spectral quality and duration modeled the relationship typical of English speech productions (Hillenbrand et al., 1995; 2000); vowels with more /æ/ − like spectral quality possessed longer durations and vowels with more /ɛ/ − like spectral quality possessed shorter durations. The sampling included nine stimuli with /ɛ/ − like spectra and shorter vowel durations and another nine stimuli with /æ/ − like spectra and longer vowel durations. Each of the 18 Exposure stimulus exemplars was presented 10 times for a total of 180 Exposure trials in Canonical blocks.

In Reverse blocks, Exposure stimuli sampled vowel exemplars that created an 'artificial accent' whereby the relationship between spectral quality and duration was reversed relative to native-listeners' long-term experience with English speech productions. In these blocks, vowels with more /æ/ − like spectra possessed *shorter* durations and vowels with more /ɛ/ − like spectra possessed *longer* durations. Prior research has demonstrated that native-English listeners tend to perceptually weight spectral quality, relying upon it most to signal the /ɛ/ − /æ/ vowel categories (Liu & Holt, 2015). In this way, the Exposure stimuli in Reverse blocks are perceptually *unambiguous*; prior research supports

[1] An additional pair of /ba/ − /pa/ stimuli (Step 5, 20 ms VOT; see Figure 1) was originally included, but are not reported here. Experiment 2 data in group 1 was collected concurrently with Experiment 1. Since a robust duration contrast effect was observed only for Step 6 along the VOT series, we did not further analyze or report data for Step 5. Data are available upon request.

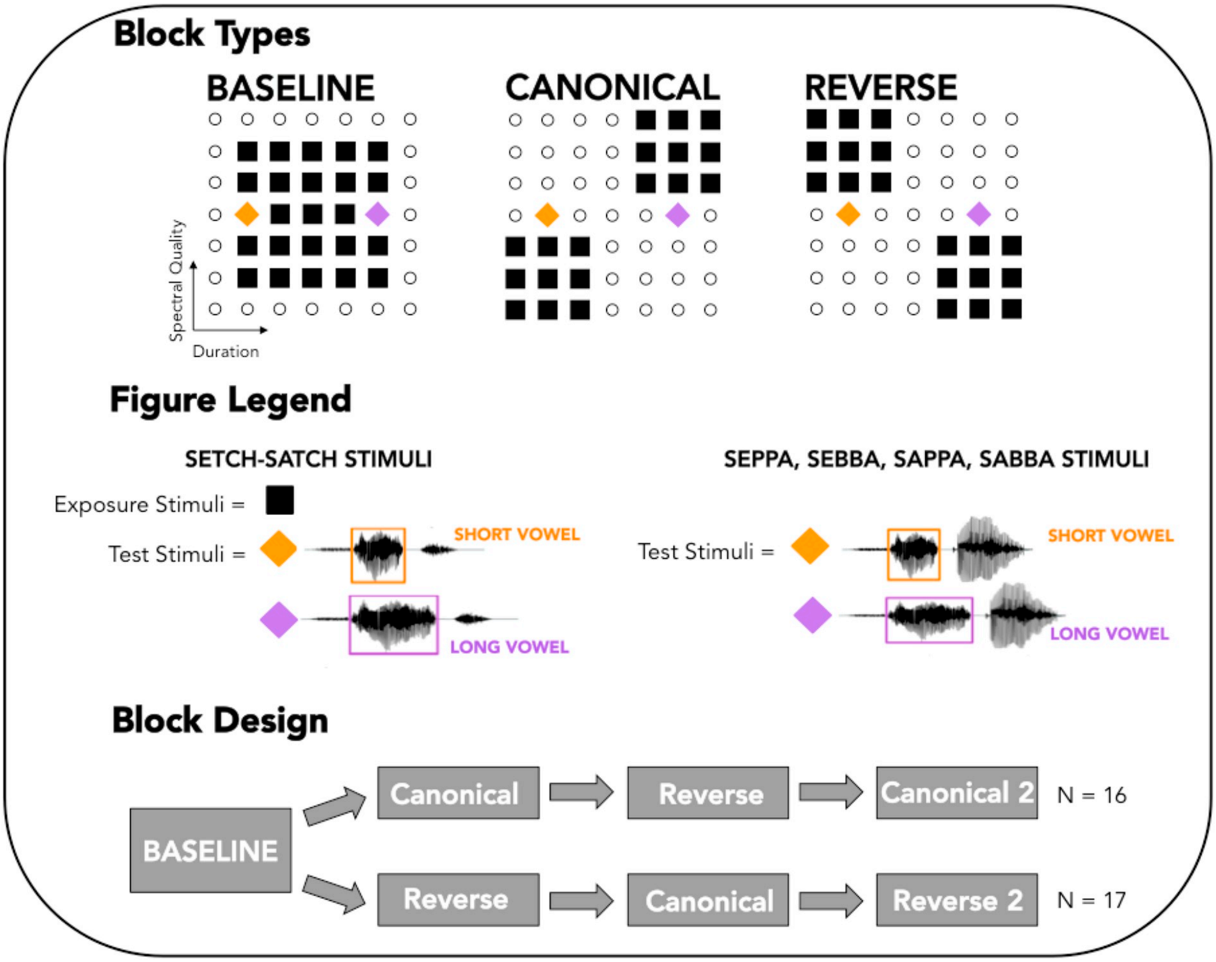

Fig. 2. Experiment 2 Design. There were three types of blocks in Experiment 2: Baseline, Canonical, and Reverse, defined by the sampling of stimuli across a two-dimensional acoustic space varying in vowel spectral quality and duration across an /ɛ/ to /æ/ perceptual space. In the top row, the black filled squares indicate *setch-satch* Exposure trials that sample the spectral quality x duration space neutrally in the Baseline block, with a correlation between dimensions consistent with English in the Canonical blocks, and with an 'artificial accent' that reverses the typical English dimension covariation in the Reverse blocks. Orange and purple diamonds indicate Test trials of two forms, *setch-satch* and *seppa, sebba, sappa, sabba*. The spectral quality x duration vowel acoustics for Test stimuli was identical across forms, but nonword context varied. As depicted in the bottom panel, all participants experienced the same Baseline block. One group then experienced blocks of stimuli corresponding to blocks of Canonical, Reverse, and Canonical stimuli. The other group experienced blocks of Reverse, Canonical, and Reverse stimuli. The initial fricative was constant across all stimuli, the /tʃ / was consistent across all *setch-satch* stimuli, and the ambiguous /ba/ − /pa/ token was identical for the *seppa, sebba, sappa, sabba* stimuli (from the 25 ms VOT Step, validated in Experiment 1). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the prediction that stimuli with /æ/ − like spectral quality will be categorized as /æ/ and stimuli with /ɛ/ − like spectral quality will be categorized as /ɛ/, despite the reversal of the typical correlation with duration (Liu & Holt, 2015). As in the Canonical blocks, each of the 18 Exposure stimulus exemplars was presented 10 times for a total of 180 Exposure trials.

3.1.5. *Test stimuli*

Within each block, a minority of trials (25%) involved presentation of Test stimuli (orange and purple diamonds, Fig. 2). There were two classes of Test stimuli with identical vowels and distinct nonword frames. The first class had the same nonword frame as Exposure trials. These *setch-satch* Test stimuli possessed a perceptually ambiguous spectral quality between /æ/ and /ɛ/, and either a short (orange diamond, 225 ms) or a long (purple diamond, 425 ms) vowel duration. Holding spectral quality constant neutralized native-English listeners' dominant cue to vowel identity and provided a test of the impact of the secondary duration dimension on vowel categorization. Note that these two Test stimuli were identical across blocks, allowing a test of the influence of the short-term regularities conveyed by the Canonical versus Reverse Exposure stimuli on reliance upon the duration dimension to categorize Test stimuli. The *setch-satch* Test stimuli were presented 10 times per block (20 total stimuli), randomly interspersed among Exposure stimuli. (In the Baseline block, the two Test stimuli were included in the 5 × 5 grid of stimuli to maintain a neutral sampling of the acoustic space).

A second set of Test stimuli of the form *seppa, sebba, sappa,* and *sabba* measured generalization of dimension-based perceptual learning to a new nonword form. This allowed for a test of whether short-term regularities conveyed by the *setch-satch* Exposure stimuli across blocks generalized to influence /ɛ/ − /æ/ vowel categorization in a nonword context not experienced in the 'artificial accent.' These same Test stimuli also provided a test of the primary research question: whether dimensional re-weighting of vowel duration as a result of short-term regularities experienced across the Exposure stimuli influences the vowels' ability to exert a duration contrast effect on /ba/ − /pa/ categorization. The *seppa, sebba, sappa, sabba* Test stimuli (the long and short vowel segments preceding the 6th VOT step from Experiment 1) were presented 10 times each, for 10 total Test stimuli that were randomly interspersed within each block (an additional 10 stimuli based on the

5th VOT step were also included, but are not reported here; see Footnote 1).

Participants were not made aware of the changing relationship between vowel spectral quality and duration across blocks, nor of a distinction between Exposure and Test stimuli. As Fig. 2 illustrates, the overall range of acoustic variation along each dimension was the same across blocks; the distributional shifts were subtle.

On each trial, a 500-ms blank screen was followed by acoustic stimulus presentation diotically over headphones along with the appearance of orthographic response options on the screen that corresponded spatially to response buttons on a standard keyboard. For *setch-satch* Exposure and Test trials, participants responded *setch* or *satch*. For the other Test trials, participants responded with *seppa, sebba, sappa,* or *sabba;* in this way, responses for these trials indicated participants' categorization of both the initial vowel and the final consonant. Participants made their responses on a keyboard, seated comfortably in a sound-treated booth.

3.1.6. Statistical power

Using G-power (logistic regression: compute achieved power using probabilities; Faul et al., 2007) we estimated power for the duration contrast effect based on the 20.7% effect size (short preceding vowel led to 68.0% /pa/ response versus long preceding vowel led to 47.3% /pa/ response) observed in Experiment 1, 490 observations of each vowel duration stimulus in Canonical and 500 observations in Reverse conditions when combining both groups, and an alpha of 0.05, we achieve 99.9% power of observing a duration contrast effect with our sample size in each block type. Similarly, given a 56% difference in /æ/ perception between short and long vowel duration Test stimuli, as reported by Liu and Holt (2015), and 490 observations in Canonical blocks and 500 observations in Reverse blocks we expect over 99.9% power of seeing a dimension-based statistical learning effect.

3.2. Results

3.2.1. Baseline perception: vowel categorization in the Baseline block

Fig. 3a plots the average percent *satch* responses to the 5 × 5 grid of *setch-satch* stimuli presented in the Baseline block. We first examined each acoustic dimensions' perceptual weight, the extent to which listeners relied upon vowel spectral quality versus duration in /ɛ/ − /æ/ perception. Following the approach of Holt and colleagues (Holt & Lotto, 2006; Idemaru, Holt, & Seltman, 2012; Liu & Holt, 2015), we submitted each listener's percent /æ/ response to a simple linear regression (lm function in R) to characterize the contribution of each dimension in predicting vowel perception across all stimuli (see Table s3). At the group level both spectral quality (ß = 0.73, SE = 0.02, $p < .001$) and duration (ß = 0.21, SE = 0.02, p < .001) significantly contributed to vowel perception. Next, we iteratively examined each individual's data using the same model (see Table s3) and normalized the absolute values of the individual coefficients to sum to one to express the *relative* perceptual weight of each dimension for each listener. This allowed us to calculate coefficients for duration and spectral quality for each individual and normalize those coefficients by their sum for each participant. Across listeners, spectral quality (M = 0.77, SE = 0.03) contributed more robustly to vowel categorization than duration (M = 0.23, SE = 0.03). Fig. 3b plots individual participant's normalized perceptual weights, illustrating that just two of the 33 participants tended to perceptually weight vowel duration more than spectral quality, a pattern closely replicating the observations of Liu and Holt (2015).

Duration was especially influential when vowel spectral quality was ambiguous. This can be seen clearly in categorization of the *setch-satch* Test stimuli (diamonds, Fig. 2; highlighted with a dark square in Fig. 3a) when spectral quality was acoustically ambiguous. A linear mixed effects model with random slopes and intercepts per vowel duration per participant revealed a significant effect of Test stimulus vowel duration on /æ/ responses (see Table s4). Short vowel duration was treated as the reference condition and showed marginally fewer /æ/ responses than /ɛ/ (M_{Short} = 43.94%, SE = 3.74%; β = −0.27, SE = 0.16, p = .10, odds of /æ/ responses = 0.76:1). The long vowel duration test stimulus was categorized as /æ/ significantly more frequently than the short vowel duration (M_{Long} = 72.73%, SE = 3.46%; β = 1.39, SE = 0.23, $p < .001$, odds of /æ/ responses = 3.06:1). This indicates that, at baseline, participants used vowel duration to categorize vowels with perceptually ambiguous spectral quality.

3.2.2. Artificial accent exposure: vowel categorization of setch-satch exposure stimuli across Canonical and Reverse blocks

The results from the Baseline block confirm that participants tended to rely most upon vowel spectral quality in categorization. Thus, across Canonical and Reverse blocks, spectral quality is likely to unambiguously signal vowel category identity for Exposure stimuli such

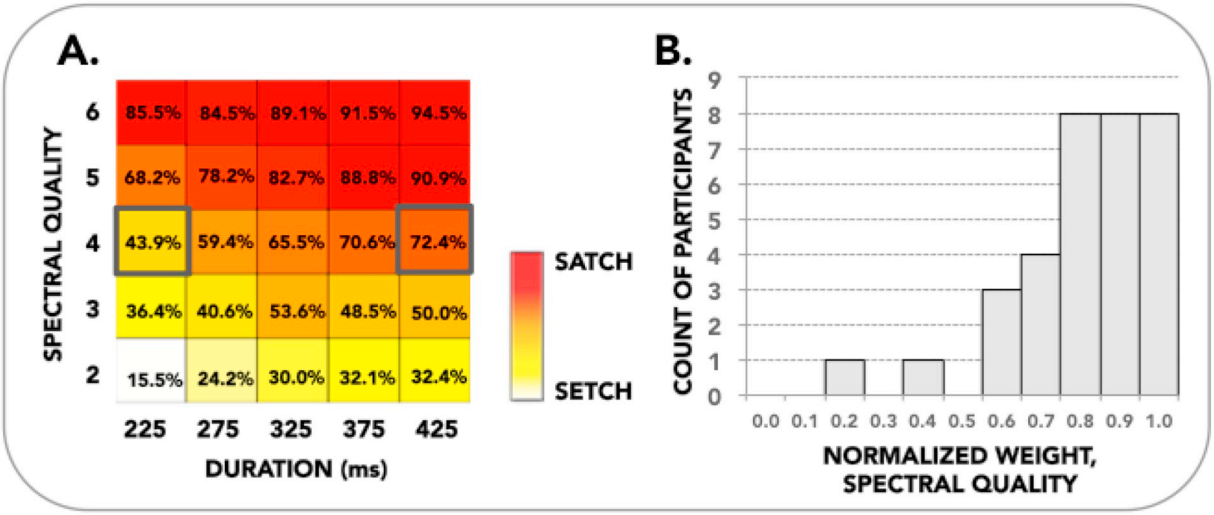

Fig. 3. Experiment 2: Vowel Categorization at Baseline. (A) The heat map shows vowel categorization across the 5 × 5 grid of Baseline block stimuli with percent /æ/ responses presented in each cell. White corresponds to no *satch* responses (all *setch* responses) whereas dark red corresponds to all *satch* responses (no *setch* responses). The dark boxes highlight the Test stimuli presented also in Canonical and Reverse blocks (diamonds, Fig. 2). (B) The histogram indicates that most native-English participants perceptually weighted vowel spectral quality more than duration, as has been reported previously (Liu & Holt, 2015). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that categorization responses are expected to align with the vowel spectral quality dimension. Indeed, across the three experimental blocks, participants exhibited categorization consistent with the perceptually dominant vowel spectral quality dimension (Group 1: $M_{Canonical1}$ = 91.11%, SE = 2.67%; $M_{Reverse}$ = 90.63%, SE = 1.94%; $M_{Canonical2}$ = 91.18%, SE = 2.30%; Group 2: $M_{Reverse1}$ = 86.01%, SE = 3.06%; $M_{Canonical}$ = 91.76%, SE = 3.31%; $M_{Reverse2}$ = 90.69%, SE = 3.23%). These patterns (demonstrated in heat maps in Fig. 1a) confirm that Exposure stimuli resulted in robust and consistent vowel category activation, hypothesized in previous work to be the driver of dimension-based statistical learning (Idemaru & Holt, 2011; Liu & Holt, 2015). Further, they demonstrate consistent listener task engagement across the experiment.

3.2.3. Dimensional-based perceptual learning: vowel categorization of setch-satch Test stimuli

We next examined vowel categorization of the *setch-satch* Test stimuli across Canonical and Reverse blocks to evaluate dimension-based statistical learning. Recall that these Test stimuli had equivalent, and perceptually-ambiguous, vowel spectral quality and differed only in vowel duration. Thus, following the logic of prior research (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2016; Liu & Holt, 2015; Schertz et al., 2015), categorization of these vowels serves to index listeners' reliance on the vowel duration dimension in categorization as a function of the differences in short-term acoustic regularities conveyed across the Reverse and Canonical blocks. Fig. 4a plots the results across experimental blocks for each group.

Setch-satch Test data from both groups was combined in a linear mixed effects model with a binomial linking function where the fixed factors of block type (Canonical, Reverse), vowel duration (Short, Long) and their interaction were used to predict the /æ/ responses on each trial. Random intercepts and random slopes by vowel duration*exposure type were included per participant to account for variance based on baseline differences across participants and differences in how responses changed over repeated experience. This model is summarized in Table s5.

The Test stimuli with shorter vowel duration in the Canonical block were treated as the reference category whereby /æ/ responses were significantly less likely (ß = −0.91, SE = 0.18, $p < .001$, odds of /æ/ responses = 0.40:1). In the context of the Canonical block, there were

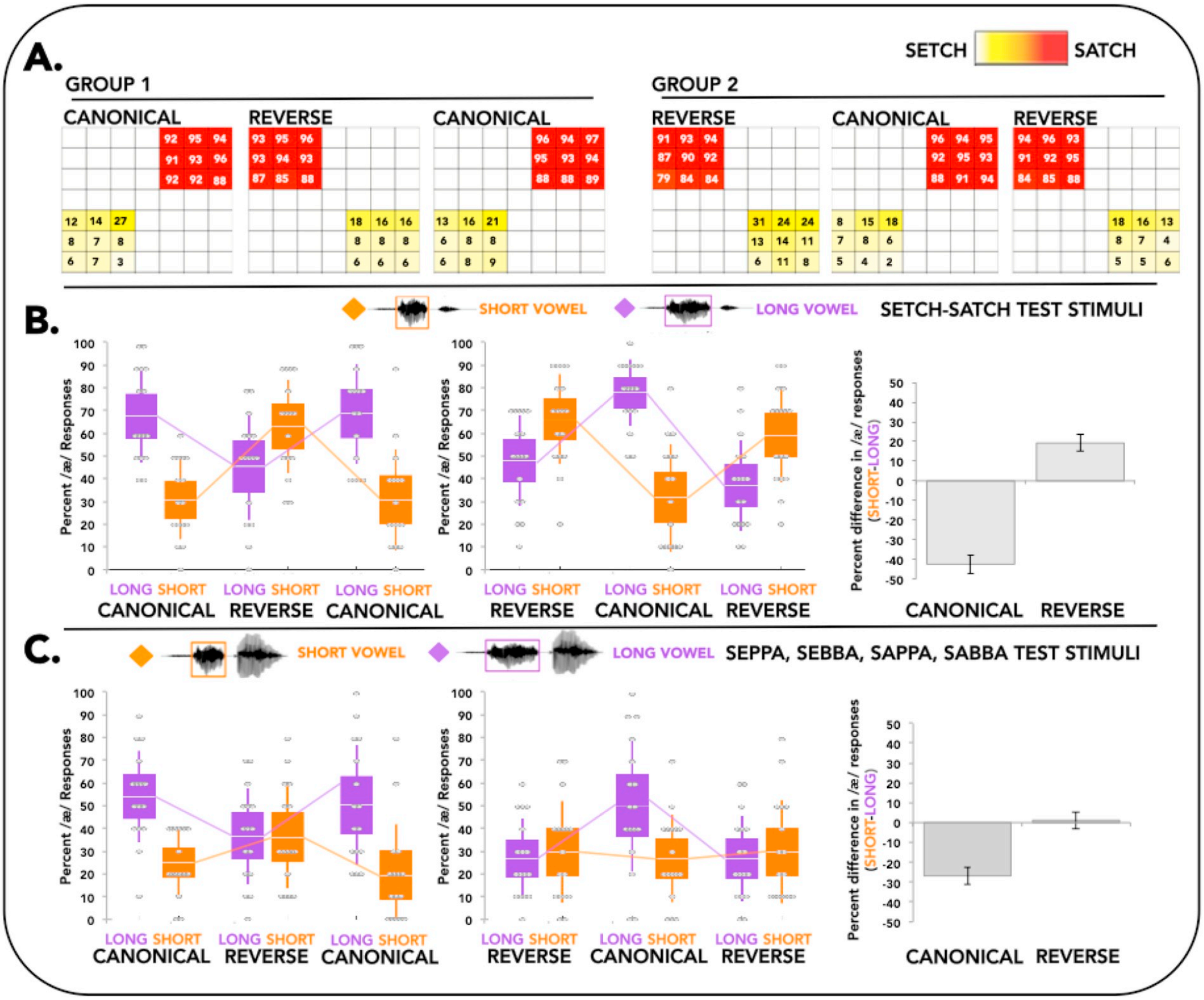

Fig. 4. Experiment 2: Dimension-based statistical learning. (A) Categorization of *setch-satch* Exposure stimuli revealed a strong use of spectral quality to identify Exposure stimuli across blocks. Numbers in each cell reflect the percent categorization of that stimulus as /æ/. (B) Categorization of *setch-satch* Test stimuli reveals a significant interaction of Test stimulus vowel duration with the short-term regularities conveyed by Exposure stimuli across blocks, indicative of perceptual learning via dimension-based statistical learning in both groups. (C) Learning as a consequence of short-term regularities experienced across *setch-satch* Exposure stimuli carried over to influence vowel categorization in *seppa, sebba, sappa, sabba* context. Individual dots in plots represent participant means, vertical lines represent standard deviation from the mean, boxes depict standard error, and horizontal lines in each box depict the group mean. Differences in categorization between short and long vowel duration Test stimuli are summarized across groups in the right panel for all Canonical and Reverse blocks.

significantly more /æ/ responses to the long duration vowel than to the short duration vowel (β = 2.01, SE = 0.25, p < .001, odds of /æ/ responses = 3.00:1), consistent with the correlation between vowel spectral quality and duration typical of English; longer vowels were more often categorized as *satch* (M_{Long} = 73.94%, SE = 2.88%) than shorter vowels (M_{Short} = 31.52%, SE = 3.62%).

However, consistent with the dynamic perceptual re-weighting of vowel duration observed in prior research (Liu & Holt, 2015), exposure to the 'artificial accent' sampling vowels with a reversed correlation between vowel spectral quality and duration resulted in a different pattern categorization across Test stimuli. In the Reverse blocks (see Table s5), the odds of /æ/ responses to short-vowel-duration setch-satch Test stimulus increased 4.05 times the rate observed in Canonical blocks (ß = 1.50, SE = 0.19, *p* < .001, odds of /æ/ responses = 1.81:1). The opposite pattern of response was apparent for the long-vowel-duration setch-satch Test stimulus (the interaction term of block type and vowel duration from Table s5), with significantly fewer /æ/ responses in the Reverse blocks than in Canonical blocks (ß = −2.87, SE = 0.30, p < .001, odds of /æ/ responses = 0.77:1). This pattern of responses demonstrates a significant difference in vowel categorization across Test stimuli in the Reverse blocks. Interestingly, the directionality of this difference in Reverse blocks was *opposite* that observed in the Baseline and Canonical blocks: the long-vowel-duration Test stimulus was less often heard as /æ/ (M_{Long} = 44.39%, SE = 3.52%) than the short-vowel-duration Test stimulus (M_{Short} = 63.64%, SE = 3.33%).

This pattern of results is consistent with perceptual learning via dimension-based statistical learning, and indicates dynamic re-weighting of the effectiveness of the duration dimension in signaling vowel category as a function of short-term input regularities (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2016; Liu & Holt, 2015; Schertz et al., 2015). Here, the influence of encountering an 'artificial accent' in the Reverse blocks was sufficient to invert the typical relationship of vowel duration to spectral quality. We return to this point in the General Discussion, as it is notable in relation to prior research (Idemaru & Holt, 2011; Liu & Holt, 2015).

3.2.4. Dimensional-based perceptual learning: vowel categorization of *seppa, sebba, sappa, sabba* Test stimuli

Since the artificial accent was experienced only in *setch-satch* context, listeners' responses to the *seppa, sebba, sappa, sabba* Test stimuli provide a measure of the extent to which dimension-based statistical learning generalizes across contexts.

Fig. 4b plots the average percent of /æ/ responses to the initial vowel of the *seppa, sebba, sappa, sabba* Test stimuli as a function of block for both groups. A linear mixed effects model with a binomial linking function was used to predict /æ/ responses based on the fixed effects of vowel duration and block type and their interaction. Participant and block number were included as random intercepts to account for random variance across participants and change across blocks (see Table s6). Broadly, we see the same pattern of responses for categorization of the initial vowel in the *seppa, sebba, sappa, sabba* Test stimuli as we observed for the setch-satch test stimuli.

The Canonical short vowel duration Test stimulus was used as the reference category, for which /æ/ responses were significantly less likely (ß = −1.30, SE = 0.19, *p* < .001, odds of /æ/ responses = 0.27:1). The Canonical long vowel duration led to significantly more /æ/ responses than the short vowel duration (ß = 1.37, SE = 0.22, p < .001, odds of /æ/ responses = 1.07:1). These effects indicate a significant difference in percent /æ/ responses between long and short vowel durations in the Canonical blocks (M_{Long} = 51.52%, SE = 4.35%; M_{Short} = 24.70%, SE = 3.04%). In the Reverse blocks, the short vowel duration led to significantly more /æ/ responses than in Canonical blocks (ß = 0.43, SE = 0.20, *p* = .034, odds of /æ/ responses = 0.42:1). The long vowel duration in the Reverse blocks (the interaction of vowel duration and block type) resulted in significantly fewer /æ/ responses than in the Canonical block (ß = −1.37, SE = 0.30, p < .001, odds of /æ/ responses = 0.42:1) reflecting a difference in categorization across long and short vowel duration Test stimuli in the Reverse blocks (M_{Long} = 31.82%, SE = 3.24%; M_{Short} = 33.03%, SE = 3.84%). This pattern of results indicates that the influence of vowel duration on categorization of the initial /ɛ/−/æ/ vowel was impacted by the short-term distributional regularities experienced across the *setch-satch* Exposure stimuli. In short, experiencing the artificial accent in *setch-satch* context generalized to influence initial-vowel categorization in *seppa, sebba, sappa, sabba* context in which the artificial accent was never heard.

To compare the size of the effect between the *setch-satch* stimuli and the *seppa, sebba, sappa, sabba* stimuli, linear mixed effects models with random intercepts and slopes for vowel duration and lexical frame per subject with binomial linking functions were used to predict /æ/ responses separately for Canonical and Reverse blocks comparing categorization across the *setch-satch* versus *seppa, sebba, sappa, sabba* Test stimulus contextual frames.

In the Canonical blocks (see Table s7), the reference category (short vowel duration for *setch-satch* nonword frame) exhibited a significant bias toward /ɛ/ responses (β = −0.87, SE = 0.17, *p* < .001, odds of /æ/ responses = 0.42:1). The long vowel duration in *setch-satch* Test stimuli instead exhibited a significantly greater likelihood of /æ/ responses (β = 1.97, SE = 0.24, *p* < .001, odds of /æ/ responses = 3.02:1). The *seppa, sebba, sappa, sabba* Test stimuli showed significantly lower likelihood of /æ/ responses than the *setch-satch* frame for the short vowel duration (β = −0.42, SE = 0.15, *p* = .006, odds of /æ/ responses = 0.28:1). There was a significant interaction between nonword frame and vowel duration (β = −0.58, SE = 0.21, p = .006, odds of /æ/ responses = 1.12:1) that indicates the effect of the nonword frame was also different in long vowel duration *seppa, sebba, sappa, sabba* stimuli. In other words, in the Canonical blocks the effect size of the influence vowel duration on categorization was significantly reduced in *seppa, sebba, sappa, sabba* Test stimuli relative to *setch-satch* Test stimuli.

In the Reverse blocks (see Table s8), the reference category (short vowel duration for *setch-satch* nonword frame) had a significant bias toward /æ/ responses (ß = 0.61, SE = 0.15, p < .001, odds of /æ/ responses = 1.84:1). The long vowel duration in *setch-satch* Test stimuli exhibited a significantly lower likelihood of /æ/ responses (β = −0.87, SE = 0.18, p < .001, odds of /æ/ responses = 0.77:1). The *seppa, sebba, sappa, sabba* Test stimuli showed a significantly lower likelihood of /æ/ responses for the short vowel duration (β = −1.42, SE = 0.20, p < .001, odds of /æ/ responses = 0.44:1). There was a significant interaction between nonword frame and vowel duration (β = 0.79, SE = 0.20, p < .001, odds of /æ/ responses = 0.41:1) that indicates that the effect of nonword frame was different in long vowel duration stimuli. Overall, across both Canonical and Reverse blocks, the influence of the short-term input regularities on listeners' reliance on vowel duration for categorization was significantly reduced for the *seppa, sebba, sappa, sabba* Test stimuli relative to *setch-satch* stimuli. Thus, although there is significant generalization to nonwords not experienced in the 'artificial accent', the impact of the short-term regularity is diminished. Liu and Holt (2015) report a similar pattern of 'incomplete' generalization to stimuli not experienced in the artificial accent.

3.2.5. Duration contrast: consonant categorization of *seppa, sebba, sappa, sabba* Test stimuli

Recall that the response options for the generalization and duration contrast Test trials required a four-alternative forced choice between the nonwords *seppa*, *sebba*, *sappa*, and *sabba*. Therefore, each trial measured both categorization of the first vowel (as /ɛ/−/æ/, Fig. 4b) and the subsequent consonant (/b/−/p/). The results plotted in Fig. 4b indicate that the short-term input regularities across Canonical and Reverse blocks impacted listeners' reliance on vowel duration in /ɛ/−/æ/ categorization across the *seppa, sebba, sappa, and sabba* Test

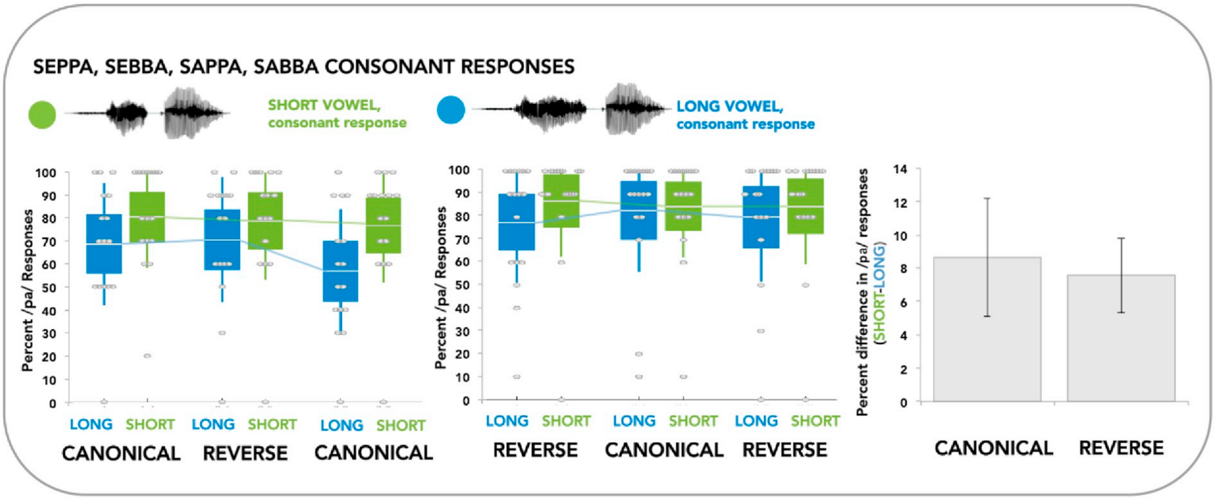

Fig. 5. Experiment 2: Normalization of consonant categorization by duration contrast from the preceding vowel. Percent /pa/ responses are plotted as a function of the preceding vowel duration (Short, Long) and the short-term regularities experienced within Canonical and Reverse blocks for both groups. Duration contrast is evident as greater /pa/ responses following a short vowel compared to a long vowel. Individual dots in plots represent participant means, vertical lines represent standard deviation from the mean, boxes depict standard error, and horizontal lines in each box depict the group mean. The right panel summarizes the data, demonstrating that duration contrast is present, and does not differ in magnitude, across Canonical and Reverse blocks.

stimuli. We next addressed the question of whether this dimension-based perceptual learning impacted the influence of vowel duration in eliciting normalization through duration contrast effects on categorization of the following /b/ − /p/ consonant.

Fig. 5 plots the consonant categorization results. With VOT and vowel duration of the /ba/ − /pa/ segment held constant across these stimuli, and spectral quality held constant across the preceding /ɛ/ − /æ/ vowel, only the duration of the preceding /ɛ/ − /æ/ is available as a means of differentially categorizing the /ba/ − /pa/ consonant. Duration contrast (Diehl & Walsh, 1989; Kluender et al., 1988; Lotto & Kluender, 1998; Raphael, 1972) predicts that preceding /ɛ/ − /æ/ vowels with shorter durations will result in more /pa/ categorization responses (longer perceived VOT) whereas longer preceding vowels will result in more /ba/ responses (shorter perceived VOT). Crucially, if the influence of dimension-based statistical learning on listeners' reliance on vowel duration in signaling /ɛ/ − /æ/ categories (as in Fig. 4b) impacts these vowels' ability to exert duration contrast on subsequent consonant categorization, then we expect reduced duration contrast (evidenced by reduced /ba/ − /pa/ categorization differences) in Reverse blocks. We tested this with a linear mixed effects model with a binomial linking function that predicted /pa/ responses based on the fixed effects of vowel duration, block type, and their interaction. Random slopes and intercepts for vowel duration and exposure condition per participant were included to account for variance across participants (see Table s9).

Categorization of consonants preceded by short vowels in the Canonical block was used as the reference category, for which /pa/ responses were significantly more likely (ß = 2.09, SE = 0.34, $p < .001$, odds of /pa/ responses = 8.12:1). In Canonical blocks categorization of consonants preceded by a long vowel exhibited significantly fewer /pa/ responses (ß = −0.87, SE = 0.21, p < .001, odds of /pa/ responses = 3.38:1) typical of duration contrast. Overall, in Canonical blocks VOT-ambiguous /ba/ − /pa/ stimuli following short vowels were more often categorized as /pa/ (M_{Short} = 81.82%, SE = 3.95%) than long vowels (M_{Long} = 73.18%, SE = 4.84%), consistent with predictions from normalization via duration contrast.

It is possible that perceptual learning renders vowel duration unavailable to elicit duration contrast. Thus, the crucial test comes in examining the magnitude of the Canonical-block duration contrast effect relative effects of duration contrast in the Reverse blocks. We examine this across the *seppa, sebba, sappa, sabba* Test stimuli, for which listeners rely on vowel duration for vowel categorization in the Canonical, but not the Reverse, blocks (see Fig. 4b). We observe that the magnitude of the duration contrast effect is consistent across Canonical and Reverse blocks. The logistic mixed effect model (from Table s9) revealed a non-significant effect of block, indicating the likelihood of /pa/ responses when the consonant was preceded by a short vowel did not differ across Canonical and Reverse blocks (ß = 0.17, SE = 0.25, p = .680, odds of /pa/ responses = 9.65:1). Similarly, there was no significant difference in the interaction of block and vowel duration indicating that across Canonical and Reverse blocks the short preceding vowel and long preceding vowel differed in /pa/ responses by a similar amount (ß = 0.13, SE = 0.28, p = .472, odds of /pa/ responses = 4.59:1). In sum, the duration of the preceding vowels impacted categorization of the Test stimuli as /ba/ versus /pa/ similarly across Canonical and Reverse blocks.

Although the duration of *seppa, sebba, sappa, sabba* Test stimulus vowels did not contribute to *vowel categorization* in the Reverse blocks (see the bar plot in Fig. 4b), it *did* exert a duration contrast effect on subsequent consonant categorization. Consonant categorization was significantly different for long versus short preceding vowel durations in Reverse blocks (M_{Short} = 82.42%, SE = 4.34%; M_{Long} = 74.85%, SE = 4.65%) and this difference was the same magnitude as in Canonical blocks (reported above), as evidenced by the linear mixed effects model in Table s9. Thus, the duration contrast effect elicited by the preceding vowel is stable across blocks even as listeners' reliance on those vowels duration for categorization is modulated by short-term speech input regularities.

Examination of the duration contrast effects plotted in Figs. 1 and 5 illustrates a relatively greater shift in consonant categorization as a function of preceding vowel in Experiment 1 (20.7%) compared to Experiment 2 (8.64% Canonical; 7.57% Reverse). We examined this using linear mixed effects models with random intercepts and slopes by experiment per participant and binomial linking functions. We predicted /pa/ responses for Experiment 2 Canonical blocks versus Experiment 1 (see Table s10) and, separately, Experiment 2 Reverse blocks versus Experiment 1 (see Table s11) across the *seppa, sebba, sappa, sabba* Test stimuli. In the Experiment 2 Canonical blocks, there was a significant shift toward a greater overall proportion of /pa/ responses compared to Experiment 1 that affected categorization no matter the duration of the preceding vowel (see Table s10). The reference category (short vowel duration for 25 ms VOT in Experiment 1) showed a significant bias toward /pa/ responses (ß = 0.86, SE = 0.28, p = .002, odds of /pa/ responses = 2.35:1). The long vowel duration context in Experiment 1 resulted in significantly less likelihood of /pa/ responses (ß = −0.99, SE = 0.26, p < .001, odds of /pa/ responses = 0.88:1) than the short vowel duration context. In contrast, the Canonical blocks in Experiment 2 showed significantly higher likelihood of /pa/ categorization for the short vowel duration (ß = 1.26, SE = 0.42, p = .002, odds of /pa/ responses = 8.26:1) and an equivalent shift toward /pa/ in the long vowel duration (as shown by the non-significant interaction term: ß = 0.17, SE = 0.31, p = .588, odds of /pa/ responses = 3.65:1). A very similar pattern was apparent for the Experiment 2 Reverse blocks compared to Experiment 1 (see Table s11). In Experiment 1 we see a bias toward /pa/ for short vowel duration context (ß = 0.86, SE = 0.28, p = .002, odds of /pa/ responses = 2.35:1) and significantly less likelihood of responding /pa/ in the long vowel duration context (ß = −0.99, SE = 0.26, p < .001, odds of /pa/ responses = 0.88:1). The Reverse blocks in Experiment 2 exhibited consistent shifts toward /pa/ relative to Experiment 1 for both vowel durations, as indicated by the significant effect of experiment (ß = 1.43, SE = 0.47, p = .002, odds of /pa/ responses in the short vowel duration = 9.88:1) and a non-significant interaction between experiment and vowel duration (ß = 0.27, SE = 0.33, p = .415, odds of /pa/ responses long vowel duration = 4.80:1). Overall, this demonstrates that whereas the magnitude of the duration contrast effect was consistent across Canonical and Reverse blocks in Experiment 2, this magnitude was diminished in comparison to Experiment 1 due to an overall shift toward /pa/ responses. The reason for this is unclear, but may be due to the heterogeneity of stimulus types in Experiment 2 compared to Experiment 1.

To return to the larger question of the interplay between perceptual learning and normalization, the 'duplex' nature of the acoustic information conveying vowel duration in Experiment 2 is notable. The duration of the preceding /ɛ/ − /æ/ vowels was re-weighted in its influence on vowel categorization as a function of short-term exposure to an artificial accent among the *seppa, sebba, sappa, sabba* test stimuli (Fig. 4b). Nevertheless, it persists. Vowel duration had an effect on subsequent consonant categorization in the same blocks in which it was re-weighted for vowel categorization (Fig. 5).

4. General discussion

The acoustic variability inherent in speech input presents a challenge to the perceptual system in mapping from acoustics to meaning. Rich literatures demonstrate that multiple processes contribute to resolving speech despite acoustic variability, but less is known about how these different processes may interact. The interaction between normalization and perceptual learning is especially intriguing. Normalization appears to adjust perception to account for acoustic variability (Diehl & Walsh, 1989; Joos, 1948; J. D. Miller, 1989; Raphael, 1972; Wade & Holt, 2005b; C. Zhang & Chen, 2016); whereas perceptual learning leverages this variability to tune the perceptual system (Idemaru & Holt, 2011; Kraljic & Samuel, 2005, 2006; Kraljic, Samuel, & Brennan, 2008; Liu & Holt, 2015; Maye et al., 2008; Norris et al., 2003; Reinisch & Holt, 2014; Samuel & Kraljic, 2009; Zhang & Samuel, 2014). However, perceptual learning – particularly dimension-based statistical learning – theoretically could change whether acoustic dimensions are available to be utilized in normalization.

Here, seeking to begin to understand interactions of different phenomena that accommodate variability in speech perception, we examined the mutual influences of one normalization process, duration contrast (Kluender et al., 1988), and a perceptual learning phenomenon, dimension-based statistical learning (Liu & Holt, 2015). We created stimuli for which vowel duration potentially had a dual role: it could inform /ɛ/ − /æ/ vowel identity and also serve to produce duration contrast on a following /ba/ − /pa/ consonant. As listeners encountered an artificial accent in which the typical correlation of acoustic dimensions contributing to vowel categorization was reversed, we observed that reliance on vowel duration was re-weighted in these stimuli such that it no longer robustly informed vowel categorization (Liu & Holt, 2015). We sought to examine whether this perceptual learning impacts how vowel duration impacts subsequent consonant categorization via a specific type of normalization, duration contrast.

Replicating the results of Liu and Holt (2015), we observed that introduction of an artificial accent that reversed the relationship between spectral quality and vowel duration impacted the effectiveness of duration in signaling vowel categories. The present results elaborate upon prior research by demonstrating that this dimension-based statistical learning generalizes to vowels in contexts not experienced in the artificial accent (from *setch-satch* to *seppa, sebba, sappa, sabba*).

Most intriguing, perception across these *seppa, sebba, sappa, sabba* stimuli highlights a curious duplex quality to the initial vowel's duration. The short-term regularities of the artificial accent result in reweighting of vowel duration in vowel categorization in the *seppa, sebba, sappa, sabba* stimuli. Yet, whereas vowel duration no longer contributes to vowel categorization, it continues to exert an influence on subsequent consonant categorization in the manner of duration contrast. This may suggest parallel representations of vowel duration. Or, as we hypothesize below this may suggest that the two processes are operating at different levels of representation or processing, and that these levels are hierarchical such that normalization occurs at lower levels of the hierarchy.

Prior research indicates that the locus of dimension-based statistical learning is pre-lexical (Liu & Holt, 2015), and phonetic-category-specific at least for stop consonants (Idemaru & Holt, 2014). But, the specific locus of this perceptual learning has not yet been determined. Duration contrast appears to arise in the encoding of acoustic dimensions prior to activation of phonetic category representations because even nonspeech sounds elicit duration contrast effects on speech categorization (Diehl & Walsh, 1989; Wade & Holt, 2005b). Since duration contrast is likely to occur pre-categorically, the present results suggest that dimension-based statistical learning does not impact the representation of acoustic dimensions per se. Instead, we argue, it may affect the effectiveness of integrating these dimensions into vowel judgments. Specifically, we posit that dimension-based statistical learning impacts the effectiveness of auditory dimensions in activating phonetic categories by modulating connection weights that link auditory representations to phonetic category representations.

Liu and Holt (2015) proposed that dimension-based statistical learning may be accounted for by a multilevel network of representations with assumptions similar to interactive activation models like TRACE (McClelland & Elman, 1986a; Mirman, McClelland, & Holt, 2006b), whereby feature (or auditory; McClelland, Mirman, & Holt, 2006) level representations of acoustic input dimensions activate phonetic categories through weighted connections. These phonetic category representations inhibit competitors and share mutually excitatory, interactive connections among lexical and feature representations that are congruent with the phonetic category (also see Joanisse & McClelland, 2015; Magnuson, Mirman, & Harris, 2011; McClelland et al., 2006; McClelland & Elman, 1986a, 1986b; Mirman, McClelland, & Holt, 2006a). In line with previous proposals, contrast effects could be instantiated in such models by allowing influence of earlier auditory information on the encoding of later auditory information at a pre-categorical, auditory level of representation (McClelland et al., 2006).

In the context of such a model, the initial connections among auditory dimensions and phonetic categories are related to perceptual weights, and are established by experience with long-term regularities in the perceptual environment. In the model, these connection weights are approximated by the perceptual weights observed at baseline (Fig. 3) across balanced, orthogonally-varying acoustic dimensions. By this view, the perceptual weights of vowel spectral quality in signaling /ɛ/ − /æ/ reflect strong connection weights that are highly effective in activating /ɛ/ versus /æ/ vowel category representations. The relatively weaker perceptual weights for vowel duration would reflect weaker connection weights less effective at /ɛ/ – /æ/ vowel activation. When listening to non-accented speech as in the Canonical block, vowel spectral quality and duration are aligned with long-term representations such that they collaborate to robustly activate the vowel category with which they are associated.

Liu and Holt (2015) proposed that these connection weights are modifiable, accounting for the rapid perceptual re-weighting of dimensions upon introduction of the artificial accent. Category activation based on spectral quality – the dominant cue – may allow the system to derive an internal prediction of the expected duration from the connection weights established by long-term experience. Mismatches between these predictions and the actual sensory input, such as may occur upon introduction of the artificial accent, may result in an internally-generated error signal that can drive adaptive adjustments to improve alignment between future predictions and input. Thus, according to this conceptual model, dimension-based statistical learning is hypothesized to exert its influence on the connection weights between auditory representations of input dimensions like vowel spectral quality or duration, and phonetic categories through a process of self-supervision via activation of internal category representations.

The present results are consistent with this model, and the model provides a means of reconciling the duplex nature of vowel duration we observe in the present results. Even as vowel duration was re-weighted in its influence on vowel categorization, its impact on subsequent consonant categorization through duration contrast remained consistent across blocks. This provides compelling evidence that dimensional re-weighting in dimension-based statistical learning, arising from the mechanism sketched above or via other possible mechanisms, does not directly modify pre-categorical auditory representations. Here, the results are clear that the utilization of duration information affected by dimension-based statistical learning is distinct from that involved in duration contrast. This may point to the weighted connections among auditory dimensions and phonetic categories as the locus of dimension-based statistical learning, leaving pre-categorical auditory representation of duration intact and able to elicit duration contrast effects. In this manner, it would be possible to observe both perceptual learning and normalization in perception of the same nonword stimulus because they impact different levels of processing.

Liu and Holt (2015) speculated that two specific mechanisms could account for how error-driven learning, arising from self-supervision by activation of internal category representations and a mismatch of the input with predictions derived from this activation, could produce dimensional re-weighting in dimension-based statistical learning. An error signal might broadly destabilize the connection weights uniformly across a dimension like vowel duration, producing a dampening of the effectiveness of the dimension in signaling phonetic categories. Alternatively, connection weights may adjust to better match the short-term input. For example, the formerly weak connection weights between short vowel durations and /æ/ may be strengthened upon experience with the artificial accent resulting in a different balance competitive inhibitory dynamics at the level of phonetic category processing.

The present results align with the latter alternative, in that the influence of vowel duration on vowel categorization fully reverses for the *setch-satch* stimuli in which the artificial accent is experienced. This is notable, as prior studies of dimension-based statistical learning across consonants observed re-weighting, but no reversal in influence, even after five days of exposure to the artificial accent with reversed dimension correlations (Idemaru & Holt, 2011). The ultimate explanation must be more complex because the generalization evident from experience with the artificial accent in *setch-satch* to vowel categorization in *seppa, sebba, sappa, sabba* exhibits a reduced impact whereby the influence of duration on vowel categorization is greatly diminished, but not reversed. This is consistent with the generalization effects reported by Liu and Holt (2015) in that re-weighting generalized, albeit more weakly, to vowels produced by a voice whose speech was not experienced in the artificial accent. A context-sensitive system with very rapidly adapting connection weights could learn the mappings from acoustic information to phonetic categories specific to the current environment and provide a degree of specificity to the learning that might not be accounted for in previously described approaches. Further research examining the conditions under which the system comes to mirror the statistics of the artificial accent with a reversal versus when shifts priority of input dimensions via re-weighting will be needed to resolve this issue, and to extend mechanistic conceptual models.

The current results also inform theoretical models of perceptual learning, normalization, and speech perception, more broadly. Bayesian approaches using distributional regularities reflecting the probability that acoustic evidence signals a given phoneme (Kleinschmidt & Jaeger, 2015; Kuperberg & Jaeger, 2015) have gained traction. But, like other models situated at Marr's computational level of analysis (Marr & Poggio, 1976), the focus of explanation is on specifying the challenges involved in adaptive plasticity in speech perception in a computationally rigorous manner. Such models do not specify *how* the challenges are solved by the cognitive system, leaving open questions of which representations contribute, which processes may be employed to manipulate representations, and whether there are interactions among multiple processes. These questions lie at the intersection of the algorithms underlying computation and the implementation of mechanisms that give rise to such algorithmic explanation.

In this regard, the present results indicate that an early, pre-categorical auditory representation of acoustic dimension input is unaffected by dimension-based perceptual learning, consistent with models positing that such learning operates not upon pre-categorical auditory representations but upon the connections weights that establish the effectiveness of auditory representations in activating phonetic category representations. The results further suggest that pre-lexical representations involved in contrastive normalization are not affected by dimension-based statistical learning, establishing that normalization and perceptual learning phenomena may be available to account for variability occurring at different stages in the processing hierarchy. Most broadly, these findings emphasize the utility in beginning to wed phenomena of speech perception that have been investigated largely in isolation.

CRediT authorship contribution statement

Matthew Lehet: Data curation, Formal analysis, Investigation, Software, Writing - original draft. **Lori L. Holt:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Visualization, Writing - review & editing.

Acknowledgements

Research was supported by the National Institutes of Health (R01DC004674, T32GM081760). Thanks to Christi Gomez for support in testing human participants and Dr. Ran Liu for help generating stimuli and thoughtful discussions in motivating this work. Correspondence regarding this article should be addressed to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, loriholt@cmu.edu.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104328.

References

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*(4), 437–456. https://doi.org/10.1017/S0047404510000400.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R Package*

Version, 2(1), 74.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x.

Boersma, P., & Weenik, D. (2013). Praat: Doing phonetics by computer (version 5.3. 39) computer program. Retrieved January 29, 2013.

Bosker, H. R., & Reinisch, E. (2015). *Normalization for speechrate in native and nonnative speech*.

Bourguignon, N. J., Baum, S. R., & Shiller, D. M. (2016). *Please say what this word is—Vowel-extrinsic normalization in the sensorimotor control of speech.* Human Perception and Performance: Journal of Experimental Psychology1–10. https://doi.org/10.1037/xhp0000209.

Bradlow, A. R., & Bent, T. (2008). *Perceptual adaptation to non-native speech, 106(2), 707–729.* https://doi.org/10.1016/j.cognition.2007.04.005.

Broadbent, D. E., Ladefoged, P., & Lawrence, W. (1956). Vowel sounds and perceptual constancy. *Nature, 178*(4537), 815–816. https://doi.org/10.1038/178815b0.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658. https://doi.org/10.1121/1.1815131.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222.

Dechovitz, D. (1977). *Information conveyed by vowels: A confirmation.* (Haskins Laboratory Status Report).

Denes, P. (1955). Effect of duration on the perception of voicing. *The Journal of the Acoustical Society of America, 27*(4), 761–764. https://doi.org/10.1121/1.1908020.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America, 85*(5), 2154–2164. https://doi.org/10.1121/1.397864.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224–238. https://doi.org/10.3758/BF03206487.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America, 106*(3), 1511–1522. https://doi.org/10.1121/1.427148.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1276–1293. https://doi.org/10.1037/0096-1523.32.5.1276.

Garvin, P. L., & Ladefoged, P. (1963). Speaker identification and message identification in speech recognition. *Phonetica, 9*(4), 193–199. https://doi.org/10.1159/000258404.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 421–433.

Guediche, S., Fiez, J. A., & Holt, L. L. (2016). *Adaptive plasticity in speech perception: Effects of external information and internal predictions.* Journal of Experimental Psychology: Human Perception and Performancehttps://doi.org/10.1037/xhp0000196.

Guediche, S., Holt, L. L., Laurent, P., Lim, S.-J., & Fiez, J. (2014). Evidence for cerebellar contributions to adaptive plasticity in speech perception. Cerebral Cortex, (bht428).

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance, 34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America, 108*(6), 3013–3022. https://doi.org/10.1121/1.1323463.

Holt, L. L. (2006). Speech categorization in context: Joint effects of nonspeech and speech precursors. *The Journal of the Acoustical Society of America, 119*(6), 4016–4026. https://doi.org/10.1121/1.2195119.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America, 119*(5), 3059–3071. https://doi.org/10.1121/1.2188377.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America, 108*(2), 710–722. https://doi.org/10.1121/1.429604.

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *The Journal of the Acoustical Society of America, 125*(6), 3983–3994. https://doi.org/10.1121/1.3125342.

Huang, J., & Holt, L. L. (2012). Listening for the norm: Adaptive coding in speech categorization. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00010.

Hufnagle, D. G., Holt, L. L., & Thiessen, E. D. (2013). Spectral information in nonspeech contexts influences children's categorization of ambiguous speech sounds. *Journal of Experimental Child Psychology, 116*(3), 728–737. https://doi.org/10.1016/j.jecp.2013.05.008.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37*(6), 1939–1956. https://doi.org/10.1037/a0025641.

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 40*(3), 1009–1021. https://doi.org/10.1037/a0035269.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America, 132*(6), 3950. https://doi.org/10.1121/1.4765076.

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews. Cognitive Science, 6*(3), 235–247. https://doi.org/10.1002/wcs.1340.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America, 88*(2), 642–654. https://doi.org/10.1121/1.399767.

Joos, M. (1948). Acoustic phonetics. *Language, 24*(2), 5. https://doi.org/10.2307/522229.

Kingston, J., Kawahara, S., Chambless, D., Key, M., Mash, D., & Watsky, S. (2014). Context effects as auditory contrast. *Attention, Perception & Psychophysics, 76*(5), 1437–1464. https://doi.org/10.3758/s13414-013-0593-z.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203. https://doi.org/10.1037/a0038695.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics, 16*, 153–169.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation. *Dialects, idiolects, and speech processing, 107*(1), 54–81. https://doi.org/10.1016/j.cognition.2007.07.013.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262–268. https://doi.org/10.3758/BF03193841.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19*(4), 332–338. https://doi.org/10.1111/j.1467-9280.2008.02090.x.

Kuperberg, G. R., & Jaeger, T. F. (2015). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*.

Ladefoged, P. (1989). A note on "Information conveyed by vowels". *The Journal of the Acoustical Society of America, 85*(5), 2223–2224. https://doi.org/10.1121/1.397821.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104. https://doi.org/10.1121/1.1908694.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455–1468. https://doi.org/10.1121/1.426686.

Lehet, M., & Holt, L. L. (2016). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science.*. https://doi.org/10.1111/cogs.12413.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., & Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*(2), 127–137. https://doi.org/10.1037/h0041240.

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

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H Theory. In W. J. Hardcastle, & A. Marchal (Vol. Eds.), *Speech production and speech modelling (NATO ASI series. Series D: behavioural and social sciences)*. *Vol. 55*. *Speech production and speech modelling (NATO ASI series. Series D: behavioural and social sciences)* (pp. 403–439). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-2037-8_16.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1783–1798. https://doi.org/10.1037/xhp0000092.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics, 68*(2), 178–183. https://doi.org/10.3758/BF03193667.

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60*(4), 602–619. https://doi.org/10.3758/BF03206049.

Magnuson, J. S., Mirman, D., & Harris, H. D. (2011). *Computational models of spoken word recognition, 1–31*.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*(5), 407–412. https://doi.org/10.3758/BF03204884.

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry*.

Maye, J., Aslin, R., & Tanenhaus, M. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science: a Multidisciplinary Journal, 32*(3), 543–562. https://doi.org/10.1080/03640210802035357.

McClelland, J. L., & Elman, J. L. (1986a). *Interactive processes in speech perception: The TRACE model.* (Parallel Distributed Processing).

McClelland, J. L., & Elman, J. L. (1986b). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10*(8), 363–369. https://doi.org/10.1016/j.tics.2006.06.007.

McMurray, B., & Aslin, R. N. (2005). *Infants are sensitive to within-category variation in* *speech perception, 95(2), B15–B26*. https://doi.org/10.1016/j.cognition.2004.07.005.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *The Journal of the Acoustical Society of America, 85*(5), 2114–2134. https://doi.org/10.1121/1.397862.

Miller, J. L., & Baer, T. (1998). Some effects of speaking rate on the production of /b/ and /w/. *The Journal of the Acoustical Society of America, 73*(5), 1751–1755. https://doi.org/10.1121/1.389399.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica, 41*(4), 215–225. https://doi.org/10.1159/000261728.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*(6), 457–465. https://doi.org/10.3758/BF03213823.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*(6), 505–512. https://doi.org/10.3758/BF03208147.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics, 54*(2), 205–210. https://doi.org/10.3758/BF03211757.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006a). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958–965. https://doi.org/10.3758/BF03213909.

Mirman, D., McClelland, J. L., & Holt, L. L. (2006b). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*(6), 958–965.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics, 58*(4), 540–560.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language, 87*(3), 539–558. https://doi.org/10.1353/lan.2011.0057.

Perry, T. L., Ohde, R. N., & Ashmead, D. H. (2001). The acoustic bases for gender identification from children's voices. *The Journal of the Acoustical Society of America, 109*(6), 2988–2998. https://doi.org/10.1121/1.1370525.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24*(2), 175–184.

Port, R. F., & Dalby, J. (1982). *Consonant/vowel ratio as a cue for voicing in English.* (Attention).

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America, 123*(2), 1104–1113. https://doi.org/10.1121/1.2821762.

Quené, H. (2013). Longitudinal trends in speech tempo: The case of Queen Beatrix. *The Journal of the Acoustical Society of America, 133*(6), EL452–EL457. https://doi.org/10.1121/1.4802892.

R Core Team (2013). *R: A language and environment for statistical computing*.

Raphael, L. J. (1972). Preceding vowel duration as a Cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America, 51*(4B), 1296–1303. https://doi.org/10.1121/1.1912974.

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied PsychoLinguistics, 37*(6), 1397–1415. https://doi.org/10.1017/S0142716415000612.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance, 40*(2), 539–555. https://doi.org/10.1037/a0034409.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32*(2), 97–127. https://doi.org/10.1006/cogp.1997.0646.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics, 71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate. II: Effects of signal discontinuities. *Perception & Psychophysics, 62*(2), 285–300.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics, 52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27*(4), 395–408.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications, 10*(1), 2465.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech (L). *The Journal of the Acoustical Society of America, 114*(6), 3036–3039. https://doi.org/10.1121/1.1627837.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review, 16*(1), 74–79. https://doi.org/10.3758/PBR.16.1.74.

Vroomen, J., & Baart, M. (2009). Recalibration of phonetic categories by Lipread speech: Measuring aftereffects after a 24-hour delay. *Language and Speech, 52*(2–3), 341–350. https://doi.org/10.1177/0023830909103178.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*(3), 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031.

Vroomen, J., van Linden, S., Keetels, M., de Gelder, B., & Bertelson, P. (2004). *Selective adaptation and recalibration of auditory speech by lipread information: Dissipation, 44(1–4), 55–61.* https://doi.org/10.1016/j.specom.2004.03.009.

Wade, T., & Holt, L. L. (2005a). Effects of later-occurring nonlinguistic sounds on speech categorization. *The Journal of the Acoustical Society of America, 118*(3), 1701–1710. https://doi.org/10.1121/1.1984839.

Wade, T., & Holt, L. L. (2005b). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics, 67*(6), 939–950. https://doi.org/10.3758/BF03193621.

Zhang, C., & Chen, S. (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1252–1268. https://doi.org/10.1037/xhp0000216.

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance, 44*(11), 1760.

Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 200–217. https://doi.org/10.1037/a0033182.