Speech Perception is Speech Learning

Lori L. Holt

The University of Texas at Austin,

Department of Psychology and Center for Perceptual Systems

ABSTRACT

Speech conveys both linguistic messages and a wealth of social and identity information about a talker. This information arrives as complex variation across many acoustic dimensions. Ultimately, speech communication depends upon experience within a language community to develop shared long-term knowledge of the mapping from acoustic patterns to the category distinctions that support word recognition, emotion evaluation, and talker identification. A great deal of research has focused on the learning involved in acquiring long-term knowledge to support speech categorization. Inadvertently, this focus may give the impression of a mature learning endpoint. Instead, there seems to be no firm line between perception and learning in speech. The contributions of acoustic dimensions are malleably reweighted continuously as a function of regularities evolving in short term input. In this way, continuous learning across speech impacts the very nature of the mapping from sensory input to perceived category. Broadly, this presents a case study in understanding how incoming sensory input – and the learning that takes place across it -- interacts with existing knowledge to drive predictions that tune the system to support future behavior.

KEYWORDS

Speech Perception, Perceptual Weights, Statistical Learning, Categorization

Perception allows us to interpret the present in relation to what we have experienced in the past. To take a simple example from the world of sound, consider when a new acquaintance shares her name. The utterance reaches our ears in the present moment and interacts with long-term representations we have built from past speech experience. The complex speech acoustics reveal her name, and likewise communicate information about her gender, race, socioeconomic status, education, native language, and emotional state (see Kraus et al., 2019; Kutlu et al., 2022). Across a lifetime of listening experience, we have built representations that support the effortless mapping of complex speech acoustics to these category distinctions. Understanding how this knowledge develops across long-term experience, such as the learning necessary to acquire native- or second-language speech sounds (see Gervain, 2022; Baese-Berk et al.,2022), has been a major focus of research.

More recently, multiple literatures have converged to highlight that the learning does not conclude upon having built these representations. Instead, listeners continuously forage acoustic input, discovering patterns of statistical regularities that influence the very mapping of sensory input to speech representations. This review provides examples to make the case that speech perception illustrates a lesson general to cognitive science: there is no firm line between perception and learning. Rather, (speech) perception *is* (speech) learning.

INPUT DIMENSIONS CARRY DIFFERENT PERCEPTUAL WEIGHT

A simple utterance like *beer* is distinguished from its near-neighbor *pier* by as many as 16 acoustic dimensions (Lisker, 1986). Expression of these dimensions varies with whether *beer* is part of a story told by Matteo or Sofia, whether the talker speaks Scottish or American English, and whether the storytelling venue is quiet or noisy. The traditional focus of research has been to wrestle with

how complex and variable acoustic speech input maps to native-language representations like phonemes, the units of sound like [b] and [p] in *beer* versus *pier*. The immense variability would make speech comprehension difficult if the mapping between acoustic input and speech representations were fixed and unchanging, as traditional theoretical accounts had once presumed (Blumstein & Stevens, 1981; Liberman & Mattingly, 1985). But, as the next sections highlight, contemporary research demonstrates that the mapping from acoustics to speech is inherently labile, not fixed. The relationship of acoustic input to speech representations is dynamic, and closely related to learning across both long- and short-term speech input.

To observe this malleability, it helps to examine a 'baseline' perceptual state. Characterizing the perceptual space for even simple utterances like *beer* and *pier* is complicated because acoustic dimensions do not necessarily contribute equivalently. Shape and color can each inform whether a fruit is a lemon or a lime, for example, but color tends to be more diagnostic. So, too, for the acoustic dimensions of speech. Figure 1 illustrates this across two (of the at least 16) acoustic dimensions that signal *beer* versus *pier:* voice onset time (VOT, measured as the time between acoustic markers of the release of lips to the start of vocal fold vibration) and fundamental frequency (F0, the rate of that vibration). When American English listeners categorize the utterances defined across the grid shown at the top of Figure 1 in quiet listening conditions both VOT and F0 inform category decisions, but VOT carries more weight. This can be quantified by the normalized coefficients of regression indicating how well each dimension predicts category identity (Figure 1, middle). On average, young adult listeners rely predominantly on VOT in quiet listening contexts, with F0 playing a secondary role. Within this aggregate data, there are individual differences (Wu & Holt, 2022) that are stable across time (Idemaru, Holt, & Seltman, 2012; Schertz et al., 2015), suggesting underlying processing differences rather than measurement fluctuation (Figure 1, bottom).

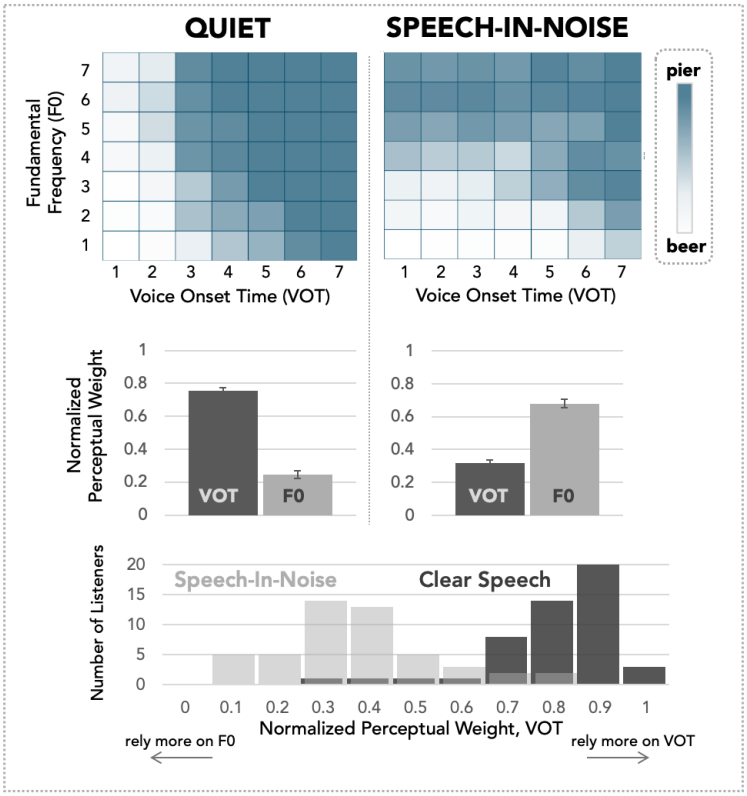

Figure 1. Each square illustrates an utterance varying in F0 and VOT, with average perceptual categorization responses painted along a spectrum from blue (*pier*) to white (*beer*) in the top row. Notice the strong reliance on VOT in quiet, with secondary contribution from F0. This is quantified in the middle row as normalized perceptual weights. Perception of the same speech sounds shifts in quiet versus noisy listening contexts, with F0 more informative in perceptual categorization decisions in noise. The same listeners rely on different acoustic dimensions to categorize speech across different listening contexts. Stable individual differences in these perceptual weights exist, as shown by plotting the normalized perceptual weight of VOT according to the number of listeners exhibiting that perceptual weight (bottom row).

This simple example extends broadly to other consonants (Idemaru & Holt, 2014) and vowels (Liu & Holt, 2015), as well as prosodic focus (exemplified in *Breaking Bad*: “I don’t need a criminal lawyer. I need a *criminal* lawyer”; Jasmin et al., 2023). Moreover, perceptual weights appear to emerge and develop over a rather long timeline. In category decisions like the one described above, four- and six-year-old listeners rely very little on F0, in contrast to young adults (Bernstein,

1983). Older adults, instead, rely *more* on F0 than younger adults (Toscano & Lansing, 2019), perhaps due to age-related changes in hearing. Overall, the perceptual weight of acoustic dimensions is language- and dialect-specific (Escudero & Boersma, 2004), with aggregate patterns of perception reflecting global distributional patterns of speech typical of a language community and smaller, individual differences potentially reflecting idiosyncratic distinctions in experience.

PERCEPTUAL WEIGHTS SHIFT WITH LISTENING CONTEXTS

Consider what happens upon a slight alteration of listening context. The right column of Figure 1 shows the same listeners' perceptual categorization of the same *beer-pier* utterances with a modest change in context: the speech is embedded in noise. This small change in listening context has a substantial impact on how the acoustic dimensions map to speech categories. Whereas VOT carries greater perceptual weight in quiet, F0 more effectively signals category identity in modest noise (Winn et al., 2013; Wu & Holt, 2022).

This means that everyday events like a noisy air conditioner turning on can completely upend the information that listeners rely upon to guide speech perception. The perceptual system adapts rapidly, but the changes do not overwrite the influence of long-term learning that established baseline perceptual weights in line with language community norms (Wu & Holt, 2022). When the air conditioner turns off and a quiet listening context is restored, the baseline perceptual weights are immediately reestablished. The mapping from acoustic input to speech is flexible, not fixed. As a result, understanding the representation of VOT (or any acoustic dimension) in brain or behavior only takes us part of the way in understanding how the complex acoustics of speech meet communicative demands of everyday listening.

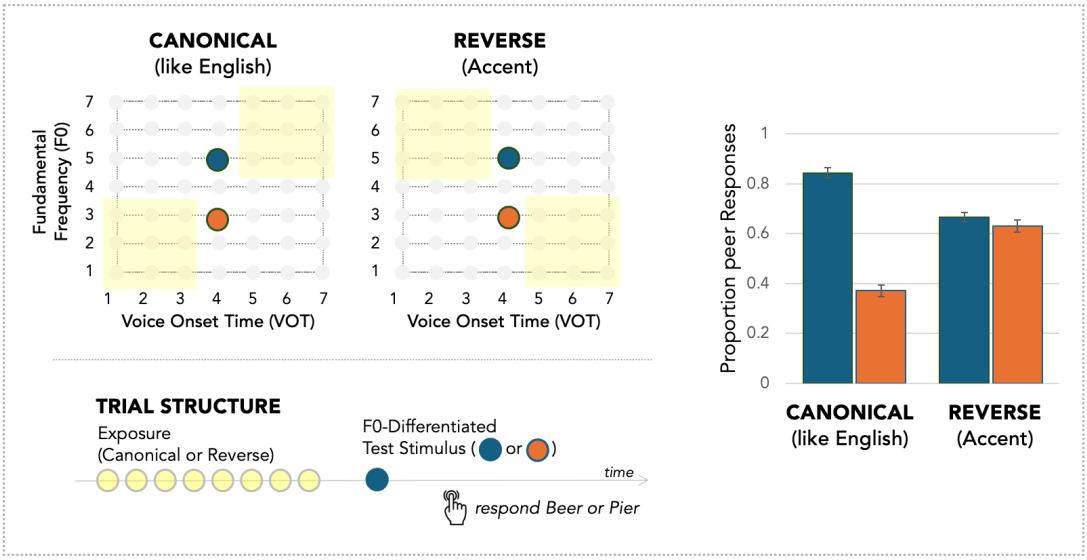

Figure 2. Dimension-based statistical learning. The top left shows the same VOTxF0 stimulus space plotted in Figure 1. The yellow highlighted regions indicate selective sampling of the space to create short-term speech regularities that match American English norms (Canonical) or violate them to create an accent (Reverse). Each trial of passive listening across 8 of these exposure stimuli (yellow) followed by one of two F0-differentiated test stimuli (blue, orange). Participants categorize the final, test stimulus as *beer* or *pier.* The data at the right illustrate the influence of statistical learning across passive exposure on the effectiveness of F0 in signaling *beer* vs. *pier.* In the context of the accent, F0 is not a reliable cue to category identity (see Idemaru & Holt, 2011; Hodson et al., 2023).

LEARNING RAPIDLY CHANGES PERCEPTUAL WEIGHTS

The rapid shift in perceptual weights arising from a noisy air conditioner might arise from lower-level sensory interactions that advantage one dimension over another. For example, F0 may be more robust to noise than VOT. Yet, there is an active role for learning in adjusting the mapping from acoustics to speech. Statistical learning across *patterns* of speech experienced over time dynamically shifts perceptual weights.

As reviewed above, baseline perceptual weights are sculpted by long-term experience in a language community. But community norms are just that – norms, not inviolate laws. We encounter talkers with nonnative accents, different dialects, and head colds. Each of these everyday contexts shifts how speech is realized within multidimensional acoustic space. As a concrete example, the acoustics of the vowel in the word *meet* uttered by an Australian English talker are very close to how an American English talker utters *mate* (Wells, 1982; opening potential for comedic misunderstanding when an Australian asks an American friend "When did you first meet?"). Speech comprehension suffers when speech departs from language community norms, but the perceptual system rapidly adapts, and comprehension improves (Bradlow & Bent, 2008). The mapping from acoustics to speech representations flexibly accommodates foreign accents, signal distortions, and even audio-visual mismatch (see Ullas et al., 2022). In a broad sense, the very acoustic dimensions that signal speech representations are dynamically – and rapidly – adjusted in online speech processing to accommodate regularities in the ambient speech environment.

A phenomenon called dimension-based statistical learning illustrates. Recall the VOTxF0 acoustic space of Figure 1. Baseline perceptual weights are evident when this space is sampled equiprobably. In quiet, this results a pattern of perceptual categorization that aligns with American English patterns of speech: longer VOTs and higher F0s tend to signal *pier* whereas shorter VOTs and lower F0s signal *beer*. This pattern of perception mirrors typical patterns of American English speech, reflecting an influence of long-term learning on perception.

An influence of learning is also revealed at shorter timescales. In a now well-replicated study, Idemaru and Holt (2011) selectively sampled stimuli from the VOTxF0 space to create American-

English-like short-term speech regularities or a reversed pattern that created a subtle accent (Figure 2, top). In a passive exposure version of the dimension-based statistical learning paradigm (Hodson et al., 2023; Murphy et al., 2023), listeners hear a sequence of 8 *beer* and *pier* utterances sampled to convey one of the Figure 2 distributional regularities (in yellow). The final stimulus is always one of two F0-differentiated test stimuli. With only F0 available to convey category identity, test stimuli categorization indexes listeners' reliance on F0 in speech categorization, its perceptual weight.

When passive exposure conveys a short-term regularity that conforms to their language community, American English listeners use F0 to inform category decisions when VOT is ambiguous: the higher-F0 test stimulus is more often labeled as *pier* and the lower-F0 test stimulus is more often *beer,* consistent with long-term language community norms. But these same listeners' reliance on F0 in speech categorization rapidly shifts when passive listening conveys a subtle accent. Now, listeners rely very little on F0 as a signal of category identity (Figure 2). Rapid learning across the reversal in VOTxF0 correlation conveyed by the accent has an immediate influence on how F0 contributes to speech categorization. The very same test stimuli are perceived differently as a function of the input regularities that came before them. Learning across a brief period of passive listening is sufficient to shift how acoustics map to speech.

Though this accent is subtle, occurs across a single voice, and is largely unbeknownst to listeners, its influence is fast (emerging in a few trials), and is evident for vowels and prosodic contrasts and for both words and nonwords (Jasmin et al., 2023; Lehet & Holt, 2020; Liu & Holt, 2015). Listeners can even juggle competing statistics evolving across two voices, adjusting the mapping from acoustics to speech for each according to the regularities of the specific voice (Zhang & Holt,

2021). These findings demonstrate reliable perceptual changes in how speech maps to categories, as a function of brief exposure to subtle shifts in the statistical properties of acoustic speech input.

There would seem to be an obvious benefit to this flexibility in accommodating accented speech, but at what cost? As in all cognitive systems, there is inherent tension in flexibility and stability: it would be undesirable to have a lifetime of speech learning overwritten by a conversation. Indeed, the fingerprints of long-term learning persist, even as the system flexibly adjusts to the accent. Five consecutive days of experience with the accent produces persistent F0 down-weighting, but not a complete remapping to mirror the statistics of the accent (Idemaru & Holt, 2011).

Liu and Holt (2015) propose that the basis of this arises from interaction of existing speech representations with input – such as an accent – that mismatches the predictions they generate. Imagine basic perceptual representations of VOT and F0 capable of activating a speech category representation for /b/ or /p/ that will inform *beer-pier* perception. The connections between these representations are 'weighted' in the sense that they more – or less – efficient. This provides a proxy for the baseline perceptual weights illustrated in Figure 1. Among most English listeners hearing *beer-pier* in quiet, VOT will carry greater weight. This bottom-up sensory information will be highly effective in selectively activating /b/ versus /p/.

The proposal is that this activation drives corresponding predictions about the nature of weights that other acoustic dimensions typically carry in the native language community. In the case of an accent, bottom-up VOT information is as unambiguous as it is in the case of English-consistent speech. But, for the accent, F0 is discrepant. Liu and Holt propose that this discrepancy between

actual and expected input generates an error signal that drives adaptive adjustments to the F0 connection weight, effectively rendering it less efficient in activating /b/ versus /p/ categories. Adjustment of the effectiveness of input in activating existing category representations would allow the system to balance flexibility (through connection weight change) even as it maintains stability in speech representations. In this way, short-term flexibility may be re-*weighting* rather than re-*mapping* perceptual space.

Several lines of evidence support this possibility. The magnitude of down-weighting is closely related to successful category activation by stimuli that convey the accent, as measured by overt categorization decisions (Wu and Holt, 2022). Further, context manipulations that shift baseline perceptual weights – like the noise of an air conditioner – change the direction of perceptual down-weighting. In noise, F0 is the more robust signal of category identity providing a route to category activation and driving VOT to be down-weighted with a reversal in VOTxF0 statistics. In quiet, as in the examples above, VOT drives activation and F0 is down-weighted (Wu & Holt, 2022). In more natural contexts, this is observed in Korean, where modern-era language change – driven largely by young women – is shifting baseline Seoul Korean perceptual weights away from VOT and toward F0 (Kang & Guion, 2008). Correspondingly, patterns of adjustment to shifting speech statistics are predictable based on listeners' baseline perceptual cue weights (Schertz et al., 2015).

Tellingly, bottom-up from sensory input need not be the driver of category activation. Zhang, Wu, & Holt (2021) neutralized the dominant VOT dimension and capitalized on well-established top-down effects of word knowledge to internally activate /b/ versus /p/ (Ganong, 1980). They created a 'phantom' accent with constant and ambiguous VOT stimuli. Lower F0 stimuli were presented

in the context of __*eace* (*peace* is a word but *beace* is not, resulting in /p/ activation from top-down word knowledge) and high-F0 stimuli conveyed by __*eef* (*beef-peef;* greater /b/ activation). This phantom accent was not present in the sensory input; VOTxF0 distributions were identical and only the word frame varied to create English-consistent or accented 'phantom' distributions from word knowledge. This top-down activation of speech categories results in F0 down-weighting for the phantom accent.

Thus, the detailed nature of existing, long-term speech representations is crucial how learning across short-term input regularities plays out, with category activation a driving force in the re-weighting of acoustic input dimensions in speech categorization. This may allow short-term changes to co-exist as connection weight changes, even as long-lasting representations persist thus contributing to the balance of flexibility with stability.

SUMMARY

In summary, results like these demonstrate that the mapping of acoustic input to speech representations is dynamically altered by online statistical learning about the patterns of speech that have recently preceded the current input. How we interpret the sounds arriving at our ears right now is a compromise among the input, long-term experience, such as with our native language norms, and short-term experience evolving across a much more rapid timescale (see Gwilliams & Davis, 2022). Listeners actively forage the acoustic environment for regularities that support online adjustments in the mapping of acoustics to speech, presenting an inherent interdependency of perception and learning that blurs the artificial lines we traditionally draw in cognitive science.

Along with other research examining recalibration, perceptual learning, and adaptation effects in speech perception (see Ullas et al., 2022 for review), this makes the case that there is no single mapping from acoustics to speech to be discovered, and no firm line between perception and learning. In speech perception (and likely other perceptual domains), the "learning is always on." Simply learning a probabilistic mapping from acoustics to a phoneme like 'b' or a name like 'Sofia' is not enough to account for the richness of speech perception. The relative contributions of acoustic dimensions are continuously and dynamically reweighted.

This has important implications. If the mapping of acoustic dimensions to speech representations is not rigidly fixed by long-term experience, as these results suggest, the hunt for the neural code for speech features may end empty-handed, or at least incomplete: we can expect the code to shift and change according to context. Research has traditionally focused on how learning establishes long-term speech representations, but results like these make clear that ongoing learning continuously tunes how information interacts with existing representations, comingling speech perception and learning and requiring more dynamic models. Further, although this short-term learning can develop across passive exposure it depends upon interaction with existing speech representations and, presumably, the predictions they can generate. This likely requires learning mechanisms distinct from both those that have been traditionally considered as 'statistical' learning over passive listening, and those proposed to drive acquisition of categories over long-term learning. Learning across speech will likely drive multiple learning mechanisms.

Ultimately, a better understanding of the dynamic, flexible nature of speech processing will advance understanding of communication in listening conditions that are more typical of natural environments. This will be especially important for understanding listeners with hearing loss and communication disorders, for whom baseline perceptual weights often differ from healthy

listeners. Even more broadly, speech provides a rich and ecologically significant testbed for understanding how incoming sensory input – and the learning that takes place across it -- interacts with existing knowledge to drive predictions that tune the system to support future behavior.

RECOMMENDED READING

These recommended readings relate to the relate to the current review and are written to engage a broad audience of readers.

Gwilliams, L., Davis, M.H. (2022). Extracting Language Content from Speech Sounds: The Information Theoretic Approach. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.) *Speech Perception. Springer Handbook of Auditory Research, vol 74*. Springer, Cham.

The chapter summarizes a growing literature demonstrating that speech processing is influenced by statistical properties of the information conveyed and introduces a modeling approach to understand these influences.

Ullas, S., Bonte, M., Formisano, E., Vroomen, J. (2022). Adaptive Plasticity in Perceiving Speech Sounds. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.) *Speech Perception. Springer Handbook of Auditory Research, vol 74*. Springer, Cham.

The chapter reviews how word knowledge and audio-visual correspondences can influence speech categorization, connecting with theories and neurobiological evidence to advance understanding of this perceptual plasticity.

Van Hedger, S.C., Johnsrude, I.S. (2022). Speech Perception Under Adverse Listening Conditions. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.) *Speech Perception. Springer Handbook of Auditory Research, vol 74*. Springer, Cham.

This chapter reviews how challenging listening contexts impact speech processing, and the cognitive and neurobiological processes that help listeners cope adverse listening environments.

NOTES

This work was supported by National Science Foundation Grants 2420979 and 2346989 and National Institute of Deafness and Communication Disorders Grants R01DC017734 and R21DC019217.

REFERENCES

Baese-Berk, M. M., Chandrasekaran, B., & Roark, C. L. (2022). The nature of non-native speech sound representations. *The Journal of the Acoustical Society of America, 152*(5), 3025-3034.

Bernstein, L. E. (1983). Perceptual development for labeling words varying in voice onset time and fundamental frequency. *Journal of Phonetics, 11,* 383-393.

Blumstein, S.E. & Stevens, K.N. (1981). Phonetic features and acoustic invariance in speech. *Cognition*, 10, 25-32.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729.

Escudero P., & Boersma P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*, 551–585.

Gervain, J. (2022). Development of Speech Perception. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (eds) Speech Perception. Springer Handbook of Auditory Research, vol 74. Springer, Cham

Gwilliams, L., Davis, M.H. (2022). Extracting Language Content from Speech Sounds: The Information Theoretic Approach. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.) *Speech Perception. Springer Handbook of Auditory Research, Vol 74*. Springer, Cham.

Hodson, A. J., Shinn-Cunningham, B.G., & Holt, L. L. (2023). Statistical learning across passive listening adjusts perceptual weights of speech input dimensions. *Cognition*, 238:105473.

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*, 3059–3071.

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.

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.

Idemaru, K., Holt, L.L., Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *Journal of the Acoustical Society of America*, *132*, 3950–3964.

Jasmin, K., Tierney, A., Obasih, C., & Holt, L. L. (2023). Short-term perceptual reweighting in suprasegmental categorization. *Psychonomic Bulletin & Review*, *30*(1), 373–382.

Kraus, M. W., Torrez, B., Park, J. W., & Ghayebi, F. (2019). Evidence for the reproduction of social class in brief speech. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(46), 22998–23003.

Kutlu, E., Tiv, M., Wulff, S., & Titone, D. (2022). Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cognitive Research: Principles and Implications*, *7*(1), 7.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328.

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

Lisker, L. (1986). “Voicing” in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, *29*(1), 3-11.

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

Murphy, T. K., Nozari, N., & Holt, L. L. (2023). Transfer of statistical learning from passive speech perception to speech production. *Psychonomic Bulletin & Review*, 10.3758/s13423-023-02399-8. Advance online publication. https://doi.org/10.3758/s13423-023-02399-8.

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

Toscano, J. C., & Lansing, C. R. (2019). Age-related changes in temporal and spectral cue weights in speech. *Language and Speech*, 62(1), 61-79.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464.

Ullas, S., Bonte, M., Formisano, E., Vroomen, J. (2022). Adaptive Plasticity in Perceiving Speech Sounds. In: Holt, L.L., Peelle, J.E., Coffin, A.B., Popper, A.N., Fay, R.R. (Eds.) *Speech Perception*. *Springer Handbook of Auditory Research, Vol 74*. Springer, Cham.

Wells, J.C. (1982). *Accents of English.* Cambridge University Press.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013). Roles of voice onset time and F0 in stop consonant voicing perception: effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research: JSLHR*, *56*(4), 1097–1107.

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology. Human Perception and Performance, 48*(9), 913–925.

Zhang, X. & Holt, L. L. (2018). Simultaneous tracking of co-evolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception & Psychophysics*, *44*, 1760-1779.

Zhang, X., Wu, X., & Holt, L. L. (2021). The learning signal in perceptual tuning of speech: Bottom-up vs. Top-down information. *Cognitive Science, 45*, e12947.

FIGURE CAPTIONS

Figure 1. Each square illustrates an utterance varying in F0 and VOT, with average perceptual categorization responses painted along a spectrum from blue (*pier*) to white (*beer*) in the top row. Notice the strong reliance on VOT in quiet, with secondary contribution from F0. This is quantified in the middle row as normalized perceptual weights. Perception of the same speech sounds shifts in quiet versus noisy listening contexts, with F0 more informative in perceptual categorization decisions in noise. The same listeners rely on different acoustic dimensions to categorize speech across different listening contexts. Stable individual differences in these perceptual weights exist, as shown by plotting the normalized perceptual weight of VOT according to the number of listeners exhibiting that perceptual weight (bottom row).

Figure 2. Dimension-based statistical learning. The top left shows the same VOTxF0 stimulus space plotted in Figure 1. The yellow highlighted regions indicate selective sampling of the space to create short-term speech regularities that match American English norms (Canonical) or violate them to create an accent (Reverse). Each trial passive listening across 8 of these exposure stimuli (yellow) followed by one of two F0-differentiated test stimuli (blue, orange). Participants categorize the final, test stimulus as *beer* or *pier.* The data at the right illustrate the influence of statistical learning across passive exposure on the effectiveness of F0 in signaling *beer* vs. *pier.* In the

context of the accent, F0 is not a reliable cue to category identity (see Idemaru & Holt, 2011; Hodson et al., 2023).